

Karen E. Nelson
Editor

Encyclopedia of Metagenomics

Genes, Genomes and Metagenomes:
Basics, Methods, Databases and Tools

Encyclopedia of Metagenomics

Karen E. Nelson
Editor

Encyclopedia of Metagenomics

Genes, Genomes and Metagenomes:
Basics, Methods, Databases and
Tools

With 216 Figures and 64 Tables

 Springer Reference

Editor

Karen E. Nelson
J. Craig Venter Institute
Rockville, MD, USA

ISBN 978-1-4899-7477-8 ISBN 978-1-4899-7478-5 (eBook)
ISBN 978-1-4899-7479-2 (print and electronic bundle)
DOI 10.1007/978-1-4899-7478-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014954611

© Springer Science+Business Media New York 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Welcome to the Encyclopedia of Metagenomics. It is anticipated that the Encyclopedia will become a resource for tools, tool development and all things metagenomic. Volume 1 encompasses Genes, Genomes and Metagenomes. It covers a range of approaches to conduct metagenomics surveys including descriptions of analysis tools. Several of these approaches, including databases, have been under development from the beginning of the metagenome era and are enabling the analysis and interpretation of large microbial data sets from various environments.

“Genes, Genomes and Metagenomes” also covers DNA extraction, various cloning and sequencing approaches, quality control and experimental designs: all essential components of the microbiome and metagenomic sequencing process. These approaches have continued to evolve and be refined, and several improvements have been incorporated over the past few years. This has also been driven by a switch to next-generation sequencing (NGS) platforms including Ion Torrent, 454 and various Illumina technologies.

Post-sequencing genome assembly, alignment tools, gene prediction and annotation are also critical to successful data interpretation. Deeper dives in Vol. 1 discuss codon usage, clustering programs and functional gene characterization.

MD, USA
September 2014

Karen E. Nelson

About the Editor



Dr. Karen E. Nelson is the President of the J. Craig Venter Institute (JCVI). Prior to being appointed President, Dr. Nelson held a number of other positions at the Institute including Director of JCVI's Rockville Campus and Director of Human Microbiology and Metagenomics in the Department of Human Genomic Medicine at the JCVI. She is also a Professor at JCVI with an active research program in genomics and metagenomics.

Dr. Nelson has led several genomic and metagenomic efforts including those of several reference microbial genomes and the first human metagenomics study that was published in 2006. Additional ongoing studies in her group include metagenomic approaches to study the ecology of the gastrointestinal tract of humans and animals, studies on the relationship between the microbiome and various human and animal disease conditions, reference genome sequencing and analysis primarily for the human body, and other omics studies. Dr. Nelson also heads the microbiome group at Human Longevity Inc., which was recently formed in La Jolla, California.

Dr. Nelson received her undergraduate degree from the University of the West Indies and her Ph.D. from Cornell University. She has authored or coauthored over 100 peer-reviewed publications and edited three books and is currently Editor-in-Chief of the journal *Microbial Ecology*. She also serves on the Editorial Boards of *BMC Genomics*, *GigaScience*, and the *Central European Journal of Biology*. She is also a standing member of the NRC Committee on Biodefense, a member of the National Academy of Sciences Board of Life Sciences, and a Fellow of the American Academy of Microbiology. She was recently appointed an Honorary Professor at the University of the West Indies.

Contributors

Takashi Abe Graduate School of Science and Technology, Niigata University, Niigata, Japan

Yutaka Akiyama Department of Computer Science, Tokyo Institute of Technology, Meguro-ku Tokyo, Japan

Rudolf Amann Molecular Ecology Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Jaime Henrique Amorim Universidade Estadual de Santa Cruz, Laboratório de Biotecnologia Microbiana, Ilhéus, BA, Brazil

Luke D. Bainard Semiarid Prairie Agricultural Research Centre, Agriculture and Agri-Food Canada, Swift Current, SK, Canada

Annalisa Ballarini Laboratory of Microbial Genomics, Centre for Integrative Biology (CIBIO), University of Trento, Trento, Italy

Navneet Batra Department of Biotechnology, GGSDS College, Chandigarh, India

Arvind Behal Department of Biotechnology, GGSDS College, Chandigarh, India

Robert G. Beiko Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Terrence H. Bell Department of Natural Resource Sciences, McGill University, Sainte-Anne-de-Bellevue, QC, Canada

Johan Bengtsson-Palme Institute of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Göteborg, Sweden

Nicholas H. Bergman National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

Sonu Bhatia Department of Biotechnology, GGSDS College, Chandigarh, India

Kai Blin Interfakultäres Institut für Mikrobiologie und Infektionsmedizin Tübingen, Mikrobiologie/Biotechnologie, Eberhard-Karls Universität, Tübingen, Germany

Hervé M. Blottière INRA, AgroParisTech, Jouy en Josas, France

MetaGenoPolis, INRA, Jouy en Josas, France

Paul L. E. Bodelier Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, Netherlands

Germán Bonilla-Rosso Laboratorio de Evolución Molecular y Experimental, Instituto de Ecología UNAM, Universidad Nacional Autónoma de México, Mexico City, Mexico

Mark Borodovsky Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Center for Bioinformatics and Computational Genomics, Atlanta, GA, USA

Yan Boucher Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

Jean-Luc Bouchot Department of Mathematics, Drexel University, Philadelphia, PA, USA

Rainer Breitling Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

Florence Busato Laboratory for Epigenetics and Environment, Centre National de Génotypage, CEA- Institut de Génomique, Evry, France

Brandi Cantarel Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

Rebecca J. Case Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

Patrick Chain Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Chon-Kit Kenneth Chan Department of Mechanical Engineering, The University of Melbourne, Melbourne, VIC, Australia

Trevor C. Charles Department of Biology, University of Waterloo, Waterloo, ON, Canada

Chao Chen Dalian University of Technology, Dalian, China

Liangyu Chen Dalian University of Technology, Dalian, China

Tsute Chen Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA

Francis Y. L. Chin Department of Computer Science, The University of Hong Kong, Hong Kong, China

Marco Cosentino Lagomarsino Computational and Quantitative Biology, University Pierre et Marie Curie, Paris, France

CNRS, Paris, France

Paul Cotter Teagasc Food Research Centre, Moorepark, Fermoy, Co., Cork, Ireland

Alimentary Pharmabiotic Centre, University College, Cork, Ireland

Pedro Coutinho Centre National de la Recherche Scientifique & Aix-Marseille Université, Marseille, France

Don Cowan Centre for Microbial Ecology and Genomics (CMEG), Genome Research Institute (GRI), University of Pretoria, Hatfield, Pretoria, South Africa

David E. Crowley Environmental Sciences, University of California, Riverside, Riverside, CA, USA

Mulan Dai Semiarid Prairie Agricultural Research Centre, Agriculture and Agri-Food Canada, Swift Current, SK, Canada

Rolf Daniel Institute of Microbiology and Genetics, Georg-August-University Göttingen, Göttingen, Germany

Colin Davenport Hannover Medical School, Hannover, Germany

Tomas de Wouters INRA, AgroParisTech, Jouy en Josas, France

UMR Micalis, AgroParisTech, Jouy en Josas, France

Ye Deng Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

Chandrika Deshpande Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia

Floyd Dewhirst Department of Molecular Genetics, The Forsyth Institute, Cambridge, MA, USA

Greg Ditzler Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA

Joël Doré INRA, AgroParisTech, Jouy en Josas, France

US 1367 MetaGenoPolis, INRA, Jouy en Josas, France

UMR Micalis, AgroParisTech, Jouy en Josas, France

Inna Dubchak US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

Lisa Durso Agroecosystem Management Research Unit, US Department of Agriculture, University of Nebraska, Lincoln- East Campus, Lincoln, NE, USA

Chitra Dutta Structural Biology & Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, West Bengal, India

Akihito Endo Department of Food and Cosmetic Science, Faculty of Bioindustry, Tokyo University of Agriculture, Abashiri, Hokkaido, Japan

K. Martin Eriksson Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden

Jean Euzéby Society of Systematic Bacteriology and Veterinary (SBSV) & National Veterinary School de Toulouse (ENVT), Toulouse, France

James A. Foster Department of Biological Sciences, Institute for Bioinformatics & Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

Iddo Friedberg Department of Microbiology, Miami University, Oxford, OH, USA

Limin Fu Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, CA, USA

C. G. M. Gahan Department of Microbiology, School of Pharmacy & Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland

Xiang Geng Dalian University of Technology, Dalian, China

Jan Gerken Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Wolfgang Gerlach Institute for Genomics and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

Tarini Shankar Ghosh Biosciences R & D, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Limited, Pune, MH, India

Jack Gilbert Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA

Frank Oliver Glöckner Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Jacobs University Bremen gGmbH, Bremen, Germany

Johannes Goll Informatics, The J. Craig Venter Institute, Rockville, MD, USA

Juan M. Gonzalez Instituto de Recursos Naturales y Agrobiología, IRNAS-CSIC, Seville, Spain

Susumu Goto Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

Luigi Grassi Physics Department, Sapienza University of Rome, Rome, Italy

Stefan J. Green University of Illinois at Chicago, Chicago, IL, USA

Charles W. Greer National Research Council Canada, Montreal, QC, Canada

Igor V. Grigoriev US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

Jacopo Grilli Dipartimento di Fisica “G. Galilei”, CNISM and INFN, Università di Padova, Padova, Italy

Saman K. Halgamuge Department of Mechanical Engineering, The University of Melbourne, Melbourne, VIC, Australia

Chantal Hamel Semiarid Prairie Agricultural Research Centre, Agriculture and Agri-Food Canada, Swift Current, SK, Canada

Jun Hang Viral Diseases Branch, WRAIR, Silver Spring, MD, USA

Mohammed Monzoorul Haque Biosciences R & D, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Limited, Pune, MH, India

Stephen J. Harrop School of Physics, University of New South Wales, Sydney, NSW, Australia

Martin Hartmann Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Zurich, Switzerland

Zhili He Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

Bernard Henrissat Centre National de la Recherche Scientifique & Aix-Marseille Université, Marseille, France

Sarah Highlander Genomic Medicine, J. Craig Venter Institute, La Jolla, CA, USA

Colin Hill Alimentary Pharmabiotic Centre, Department of Microbiology, University College, Cork, Ireland

David Horn School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel

Arthur L. Hsu Department of Mechanical Engineering, The University of Melbourne, Melbourne, VIC, Australia

Gangqing Hu Systems Biology Center, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA

Shih-Ting Huang J. Craig Venter Institute, Rockville, MD, USA

Daniel H. Huson Center for Bioinformatics, Algorithms in Bioinformatics, University of Tübingen, Tübingen, Germany

Toshimichi Ikemura Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, Japan

Hachiro Inokuchi Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, Japan

Yuki Iwasaki Nagahama Institute of Bio-Science and Technology,
Nagahama, Shiga, Japan

Mukesh Jain Functional and Applied Genomics Laboratory, National
Institute of Plant Genome Research (NIPGR), New Delhi, India

Diego Javier Jiménez Department of Microbial Ecology, University of
Groningen, Center for Ecological and Evolutionary Studies (CEES),
Groningen, The Netherlands

Brian V. Jones Center for Biomedical and Health Science Research,
University of Brighton, School of Pharmacy and Biomolecular Sciences,
Brighton, East Sussex, UK

I. King Jordan School of Biology, Georgia Institute of Technology,
Atlanta, GA, USA

Amit Joshi Department of Biotechnology & Bioinformatics, SGS
College, Chandigarh, India

Olivier Jousson Laboratory of Microbial Genomics, Centre for Integrative
Biology (CIBIO), University of Trento, Trento, Italy

Minoru Kanehisa Bioinformatics Center, Institute for Chemical Research,
Kyoto University, Uji, Kyoto, Japan

Geun-Joong Kim Department of Biological Sciences, College of Natural
Sciences, Chonnam National University, Gwangju, Republic of Korea

Joel Kostka School of Biology and Earth & Atmospheric Sciences, Georgia
Institute of Technology, Atlanta, GA, USA

Masaaki Kotera Bioinformatics Center, Institute for Chemical Research,
Kyoto University, Uji, Kyoto, Japan

Renzo Kottmann Max Plank Institute for Marine Microbiology, Bremen,
Germany

Marcio R. Lambais Luiz de Queiroz College of Agriculture (ESALQ),
University of São Paulo (USP), Piracicaba, SP, Brazil

Ronald F. Lamont Department of Gynecology and Obstetrics, Clinical
Institute, University of Southern Denmark, Odense University Hospital,
Odense, Denmark

Division of Surgery, University College London, Northwick Park Institute of
Medical Research Campus, London, UK

Yemin Lan School of Biomedical Engineering, Science and Health, Drexel
University, Philadelphia, PA, USA

Nicolas Lapaque INRA, AgroParisTech, Jouy en Josas, France

Henry C. M. Leung Department of Computer Science, The University of Hong Kong, Hong Kong, China

Weizhong Li J. Craig Venter Institute, La Jolla, CA, USA

Mark Liles Department of Biological Sciences, Auburn University, Auburn, AL, USA

Ho-Dong Lim Department of Biological Sciences, College of Natural Sciences, Chonnam National University, Gwangju, Republic of Korea

Chien-Chi Lo Genome Science Group, Los Alamos National Laboratory, Los Alamos, NM, USA

Hernan Lorenzi Informatics, J. Craig Venter Institute, Rockville, MD, USA

Petra Louis Rowett Institute of Nutrition and Health, Microbiology Group, Gut Health Programme, University of Aberdeen, Aberdeen, UK

Connie Lovejoy Department of Biology, Laval University, Québec, QC, Canada

Vedran Lucić Molecular Biology Department, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

Wolfgang Ludwig Lehrstuhl Für Mikrobiologie, Technische Universität München, Freising, Germany

Haiwei Luo Department of Marine Sciences, University of Georgia, Athens, GA, USA

Bridget Mabbutt Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia

Norman J. MacDonald Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Emmanuelle Maguin INRA, AgroParisTech, Jouy en Josas, France

Sharmila Mande Biosciences R & D, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Limited, Pune, MH, India

Alan J. McCarthy Microbiology Research Group, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Liverpool, UK

Alice C. McHardy Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

David Mead Lucigen Corporation, Middleton, WI, USA

Marnix H. Medema Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Folker Meyer Institute of Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

Kentaro Miyazaki Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Sapporo, Japan

Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology, Sapporo, Japan

Yuki Moriya Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

Mark Morrison Diamantina Institute, The University of Queensland, Woolloongabba, Brisbane QLD, Australia

Michael J. Moser Lucigen Corporation, Middleton, WI, USA

Raul Munoz Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Illes Balears, Spain

Akira Muto Faculty of Agriculture and Life Science, Hirosaki University, Hirosaki, Aomori, Japan

Heiko Nacke Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Göttingen, Germany

Istvan Nagy Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary

Tania Nasreen Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

Shamima Nasrin Department of Biological Sciences, Auburn University, Auburn, AL, USA

Josh D. Neufeld Department of Biology, University of Waterloo, Waterloo, ON, Canada

R. Henrik Nilsson Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden

Beifang Niu Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, CA, USA

Brian D. Ondov National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

Orla O'Sullivan Teagasc Food Research Centre, Moorepark, Fermoy, Co., Cork, Ireland

Alimentary Pharmabiotic Centre, University College, Cork, Ireland

Asli Ismihan Ozen The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark

Stephan Pabinger Division of Bioinformatics, Biocenter, Innsbruck Medical University, Innsbruck, Austria

AIT – Austrian Institute of Technology, Health & Environment Department, Molecular Diagnostics, Vienna, Austria

Donovan H. Parks Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Australian Centre for Ecogenomics, University of Queensland, Brisbane QLD, Australia

Ravi K. Patel Functional and Applied Genomics Laboratory, National Institute of Plant Genome Research (NIPGR), New Delhi, India

Jörg Peplies Ribocon GmbH, Bremen, Germany

Adam M. Phillippy National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

Rob Phillips Departments of Applied Physics and Bioengineering California Institute of Technology, California Institute of Technology, Pasadena, CA, USA

Rembert Pieper J. Craig Venter Institute, Rockville, MD, USA

Om Prakash National Centre for Cell Science, Pune, Maharashtra, India

Elmar Pruesse Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Pei-Yuan Qian KAUST Global Collaborative Program, Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China

Christian Quast Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Jean-Baptiste Ramond Centre for Microbial Ecology and Genomics (CMEG), Genome Research Institute (GRI), University of Pretoria, Hatfield, Pretoria, South Africa

Rachel Rezende Universidade Estadual de Santa Cruz, Laboratório de Biotecnologia Microbiana, Ilhéus, BA, Brazil

Lavanya Rishishwar Bioinformatics, Georgia Institute of Technology, Atlanta, GA, USA

Francisco Rodriguez-Valera Microbiologia, Universidad Miguel Hernandez, Campus San Juan, San Juan, Alicante, Spain

Masa Roller Bioinformatics Group, Molecular Biology Department, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

Sandra Ronca Centre for Microbial Ecology and Genomics (CMEG), Genome Research Institute (GRI), University of Pretoria, Hatfield, Pretoria, South Africa

David J. Rooks Microbiology Research Group, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Liverpool, UK

Gail Rosen Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA

Paul Ross Teagasc Food Research Centre, Moorepark, Fermoy, Co., Cork, Ireland

Alimentary Pharmabiotic Centre, University College, Cork, Ireland

Ramon Rosselló-Móra Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Illes Balears, Spain

Isaam Saeed Optimisation and Pattern Recognition Group, Melbourne School of Engineering, The University of Melbourne, Parkville, Australia

Munmun Sarkar CSIR-Indian Institute of Chemical Biology, Kolkata, India

Tulasi Satyanarayana Department of Microbiology, University of Delhi, New Delhi, India

Karl-Heinz Schleifer Lehrstuhl Für Mikrobiologie, Technische Universität München, Freising, Germany

Thomas W. Schoenfeld Lucigen Corporation, Middleton, WI, USA

Matthew B. Scholz Genome Science Group, Los Alamos National Laboratory, Los Alamos, NM, USA

Timmy Schweer Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Vineet K. Sharma MetaInformatics Laboratory, Metagenomics and Systems Biology Group, Department of Biological Sciences, Indian Institute of Science Education and Research, Bhopal, India

Martin Sievers Zurich University of Applied Sciences, Institute of Biotechnology, Wädenswil, Switzerland

Jagtar Singh Department of Biotechnology, Panjab University, Chandigarh, India

Roy Sleator Department of Biological Sciences, Cork Institute of Technology, Cork, Co. Cork, Ireland

Jens Stoye Faculty of Technology, Bielefeld University, Bielefeld, Germany

Hikaru Suenaga Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology, Sapporo, Japan

Moo-Jin Suh J. Craig Venter Institute, Rockville, MD, USA

Fengzhu Sun Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Visaahini Sureshan Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia

Arbel D. Tadmor TRON – Translational Oncology at the University Medical Center of the Johannes Gutenberg University Mainz, Mainz Germany

Hideto Takami Microbial Genome Research Group, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

Eriko Takano Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

Sen-Lin Tang Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan

Shiyuyun Tang School of Biology, Biodiversity Research Center, Georgia Institute of Technology, Atlanta, GA, USA

Todd D. Taylor Laboratory for Integrated Bioinformatics, Core for Precise Measuring and Modeling, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

João Carlos Teixeira Dias Universidade Estadual de Santa Cruz, Laboratório de Biotecnologia Microbiana, Ilhéus, BA, Brazil

Torsten Thomas School of Biotechnology and Biomolecular Sciences & Centre for Marine Bio-Innovation, University of New South Wales, Sydney, NSW, Australia

Toshiaki Tokimatsu Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

Jörg Tost Laboratory for Epigenetics and Environment, Centre National de Génétique, CEA-Institut de Génomique, Evry, France

Zlatko Trajanoski Division of Bioinformatics, Biocenter, Innsbruck Medical University, Innsbruck, Austria

Susannah Tringe US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

Huai-Kuang Tsai Institute of Information Science, Academia Sinica, Taipei, Taiwan

Ching-Hung Tseng Bioinformatics Program, Taiwan International Graduate Program, Biodiversity Research Center, Institute of Information Science, Academia Sinica, Taipei, Taiwan

David Wayne Ussery Bioscience Division of Oak Ridge National Laboratory, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Joy D. Van Nostrand Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

Digvijay Verma Department of Microbiology, University of Delhi, New Delhi, India

Kristian Vlahoviček Bioinformatics Group, Molecular Biology Department, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

Jun Wang BGI Shenzhen, Shenzhen, China

Lingling Wang Department of Animal Sciences, The Ohio State University, Columbus, OH, USA

Tse-Yi Wang Department of Medical Research, Mackay Memorial Hospital, New Taipei City, Taiwan

Yi Wang Department of Computer Science, The University of Hong Kong, Hong Kong, China

Yong Wang Division of Deep Sea Science, Sanya Institute of Deep Sea Science and Engineering, San Ya, Hainan, China

Yumei Wang Dalian University of Technology, Dalian, China

Tandy Warnow Institute for Genomic Biology, University of Illinois, IL, USA

Tilmann Weber Interfakultäres Institut für Mikrobiologie und Infektionsmedizin Tübingen, Mikrobiologie/Biotechnologie, Eberhard-Karls Universität, Tübingen, Germany

Martin Wu Department of Biology, University of Virginia, Charlottesville, VA, USA

Sitao Wu Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, CA, USA

Li Charlie Xia Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Jianping Xu Department of Biology, McMaster University, Hamilton, ON, Canada

Yuko Yamada Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, Japan

Jian Yang MOH Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS&PUMC), Beijing, People's Republic of China

Pablo Yarza Ribocon GmbH., Bremen, Germany

Yuzhen Ye Indiana University, School of Informatics and Computing, Bloomington, IN, USA

Etienne Yergeau National Research Council Canada, Montreal, QC, Canada

Pelin Yilmaz Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

S. M. Yiu Department of Computer Science, The University of Hong Kong, Hong Kong, China

Zhongtang Yu Department of Animal Sciences, Environmental Science Graduate Program, The Ohio State University, Columbus, OH, USA

María Mercedes Zambrano Molecular Genetics and Microbial Ecology, Corporación CorpoGen, Bogotá, DC, Colombia

Xinqing Zhao School of Life Science and Biotechnology, Dalian University of Technology, Dalian, People's Republic of China

Jizhong (Joe) Zhou Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

Department of Environmental Science and Engineering, Tsinghua University, Beijing, China

Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Huaiqiu Zhu Department of Biomedical Engineering, and Center for Theoretical Biology, Peking University, Beijing, China

Zhengwei Zhu Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, CA, USA

A

A 123 of Metagenomics

Torsten Thomas¹, Jack Gilbert² and Folker Meyer³

¹School of Biotechnology and Biomolecular Sciences & Centre for Marine Bio-Innovation, University of New South Wales, Sydney, NSW, Australia

²Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA

³Institute of Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

Introduction

Microbial ecology aims to comprehensively describe the diversity and function of microorganisms in the environment. Culturing, microscopy, and chemical or biological assays were not too long ago the main tools in this field. Molecular methods, such as 16S rRNA gene sequencing, were applied to environmental systems in the 1990s and started to uncover a remarkable diversity of organisms (Barns et al. 1994). Soon, the thirst for describing microbial systems was no longer satisfied by the knowledge of the diversity of just one or a few genes. Thus, approaches were developed to describe the total genetic diversity of a given environment (Riesenfeld et al. 2004). One such approach is metagenomics, which involves sequencing the total DNA extracted

from environmental samples. Arguably, metagenomics has been the fastest growing field of microbiology in the last few years and has almost become a routine practice. The learning curve in the field has been steep, and many obstacles still need to be overcome to make metagenomics a reliable and standard process. It is timely to reflect on what has been learned over the past few years from metagenome projects and to predict future needs and developments.

This brief primer gives an overview for the current status and practices as well as limitations of metagenomics. We present an introduction to sampling design, DNA extraction, sequencing technology, assembly, annotation, data sharing, and storage.

Sampling Design and DNA Processing

Metagenomic studies of single habitats, for example, acid mine drainage (Tyson et al. 2004), termite hindgut (Warnecke et al. 2007), cow rumen (Hess et al. 2011), and the human gastrointestinal tract (Gill et al. 2006), have provided an insight into the basic diversity and ecology of these environments. Moreover, comparative studies have explored the ecological distribution of genes and the functional adaptations of different microbial communities to specific ecosystems (Tringe et al. 2005; Dinsdale et al. 2008; Delmont et al. 2011). These pioneering studies were predominately designed to develop

and prove the general metagenomic approach and were often limited by the high cost of sequencing. Hence, desirable scientific methodology, including biological replication, could not be adopted, a situation that precluded appropriate statistical analyses and comparison (Prosser 2010).

The significant reduction, and indeed continuing fall, in sequencing costs (see below) now means that the central tenants of scientific investigation can be adhered to. Rigorous experimental design will help researchers explore the complexity of microbial interactions and will lead to improved catalogs of proteins and genetic elements. Individual ecosystems can now be studied with appropriate cross-sectional and temporal approaches designed to identify the frequency and distribution of variance in community interaction and development (Knight et al. 2012). Such studies should also pay close attention to the collection of comprehensive physical, chemical, and biological data (see below). This will enable scientists to elucidate the emergent properties of even the most complex biological system. This capability will provide the potential to identify drivers at multiple spatial, temporal, taxonomic, phylogenetic, functional, and evolutionary levels and to define the feedback mechanisms that mediate equilibrium.

The frequency and distribution of variance within a microbial ecosystem are basic factors that must be ascertained by rigorous experimental design and analysis. For example, to analyze the microbial community structure from 1 l of seawater in a coastal pelagic ecosystem, one must also ideally define how representative this will be for the ecosystem as a whole and what the bounds of that ecosystem are. Numerous studies of marine systems have shown how community structure can vary between water masses and over time (e.g., Gilbert et al. 2012; Fuhrman 2009; Fuhrman et al. 2006, 2008; Martiny et al. 2006), and metagenomics currently helps further define how community structure varies in these environments (Ottesen et al. 2011; DeLong et al. 2006; Rusch et al. 2007; Gilbert et al. 2010a). In contrast, in soil systems variance in space appears to be far larger than in time

(Mackelprang et al. 2011; Barberan et al. 2012; Bergmann et al. 2011; Nemergut et al. 2011; Bates et al. 2011). Considerable work still is needed in order to determine spatial heterogeneity, for example, how representative a 0.1 mg sample of soil is with respect to the larger environment from which it was taken.

The design of a sampling strategy is implicit in the scientific questions asked and the hypotheses tested, and standard rules outside of replication and frequency of observation are hard to define. However, the question of “depth of observation” is prudent to address because researchers now can sequence microbiomes of individual environments with exceptional depth or breadth. By enabling either deep characterization of the taxonomic, phylogenetic, and functional potential of a given ecosystem or a shallow investigation of these elements across hundreds or thousands of samples, current sequencing technology (see below) is changing the way microbial surveys are being performed (Knight et al. 2012).

DNA handling and processing play a major role in exploring microbial communities through metagenomics (see also DNA extraction methods for human studies, “Extraction Methods, DNA” and “Extraction Methods, Variability Encountered in”). Specifically, it is well known that the type of DNA extraction used for a sample will affect the community profile obtained (e.g., Delmont et al. 2012). Therefore, with projects like the Earth Microbiome Project that aim to compare a large number of samples, efforts have been made to standardize DNA extraction protocols for every physical sample. Clearly, no single protocol will be suitable for every sample type (Gilbert 2011, 2010b). For example, a particular extraction protocol might yield only very low DNA concentrations for a particular sample type, making it necessary to explore other protocols in order to improve efficiency. However, differences among DNA extraction protocols may limit comparability of data. Therefore, researchers need to further define in qualitative and quantitative terms how different DNA extraction methodologies affect microbial community structure.

Sequencing Technology and Quality Control

The rapid development of sequencing technologies over the past few years has arguably been one of the driving forces in the field of metagenomics. While shotgun metagenomic studies initially relied on hardware-intensive and costly Sanger sequencing technology (Tyson et al. 2004; Venter et al. 2004) available only to large research institutes, the advent and continuous release of several next-generation sequencing (NGS) platforms has democratized the sequencing market and has given individual laboratories or research teams access to affordable sequencing data. Among the available NGS options, the Roche (Margulies et al. 2005), Illumina (Bentley et al. 2008), Ion Torrent (Rothberg et al. 2011), and SOLiD (Life Technologies) platforms have been applied to metagenomic samples, with the former two being more intensively used than the latter. The features of these sequencing technologies have been extensively reviewed – see, for example, Metzker (2010) and Quail et al. (2012) – and are therefore only briefly summarized here (Table 1).

Roche's platform utilizes pyrosequencing (also often referred to as 454 sequencing because of the name of the company that initially developed the platform) as its underlying molecular principle. Pyrosequencing involves the binding of a primer to a template and the sequential addition of all four nucleoside triphosphates in the presence of a DNA polymerase. If the offered

nucleoside triphosphate matches the next position after the primer, then its incorporation results in the release of diphosphate (pyrophosphate, or PPi). PPi production is coupled by an enzymatic reaction involving an ATP sulfurylase and a luciferase to the production of a light signal that is detected through a charge-coupled device. The Ion Torrent sequencing platform uses a related approach; however, here, protons that are released during nucleoside incorporation are detected through semiconductor technology. In both cases, the production of light or charge signals relates to the incorporation of the sequentially offered nucleoside, which can be used to deduce the sequence downstream of the primer. Homopolymer sequences create signals proportional to the number of positions; however, the linearity of this relationship is limited by enzymatic and engineering factors leading to well-investigated insertion and deletion (Indel) sequencing errors (Prabakaran et al. 2011; McElroy et al. 2012).

Illumina sequencing is based on the incorporation and detection of fluorescently labeled nucleoside triphosphates to extend a primer bound to a template. The key feature of the nucleoside triphosphates is a chemically modified 3' position that does not allow for further chain extension ("terminator"). Thus, the primer gets extended by only one position, whose identity is detected by different fluorescent colors for each of the four nucleosides. Through a chemical reaction, the fluorescent label is then removed, and the 3' position is converted into a hydroxyl group

A 123 of Metagenomics, Table 1 Next-generation sequencing technologies and their throughput, errors, and application to metagenomics

Machine (manufacturer)	Throughput (per machine run)	Reported errors	Error/metagenomic example references
GLX Titanium (454/Roche)	~1 M reads @ ~500 nt	0.56 % indels; up to 0.12 % substitution	(McElroy et al. 2012; Fan et al. 2012)
HiSeq 2000 (Illumina)	~3 G reads @ 100 nt	~0.001 % indels; up to 0.34 % substitution	(McElroy et al. 2012; Quail et al. 2012; Hess et al. 2011)
Ion Torrent PGM (Life Technologies)	~0.1–5 M reads @ ~200 nt	1.5 % indels	(Loman et al. 2012; Whiteley et al. 2012)
SOLiD (Life Technologies)	~120 M reads @ ~50 nt	Up to 3 %	(Salmela 2010; Zhou et al. 2011; Iverson et al. 2012)

allowing for another round of nucleoside incorporation. The use of a reversible terminator thus allows for a stepwise and detectable extension of the primer that results in the determination of the template sequence. In theory, this process could be repeated to generate very long sequences; in practice, however, misincorporation of nucleosides in the many clonal template strands results in the fluorescent signal getting out of phase, and thus reliable sequencing information is only obtained for about 200 positions (Quail et al. 2012).

SOLiD sequencing utilizes ligation, rather than polymerase-mediated chain extension, to determine the sequence of a template. Primers are extended through the ligation with fluorescently labeled oligonucleotides. The high specificity of the ligase ensures that only oligonucleotides matching the downstream sequence will be incorporated; and by encoding different oligonucleotides with different fluorophores, the sequence can be determined.

It is important to understand the features of the sequencing technology in terms of throughput, read length, and errors (see Table 1), because these will have a significant impact on downstream processing. For example, the relative high frequency of homopolymer errors for the pyrosequencing technology can impact ORF identification (Rho et al. 2010) but might still allow for reliable gene annotation, because of its comparatively long read length (Wommack et al. 2008). Conversely, the short read length of Illumina sequencing might reduce the rate of annotation of unassembled data, but the substantial throughput and data volume generated can facilitate assembly of entire draft genomes from metagenomic data (Hess et al. 2011). These considerations are also particularly relevant with new sequencing technologies coming online. These include single-molecule sequencing using zero-mode waveguide nanostructure arrays (Eid et al. 2009), which promises read lengths beyond 1,000 bp and has been shown to improve the hybrid assemblies of genomes (Koren et al. 2012), as well as nanopore sequencing (Schneider and Dekker 2012), which also promises long read lengths.

One important practical aspect to consider when analyzing raw sequencing data is the quality value assigned to reads. For a long time, the quality assessment provided by the technology vendor was the only available option for data consumers. Recently, however, a vendor-independent error detection and characterization has been described that relies on error estimate-based reads that are accidentally duplicated during the PCR stages (a fact described for Ion Torrent, 454, and Illumina sequencing technologies) (Trimble et al. 2012). Moreover, a significant number of publicly available metagenomic datasets contain sequence adaptors (apparently because quality control is often performed on the level of assembled sequences, not raw reads). Simple statistical analyses with tools such as FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) will rapidly detect most of these adapter contaminations. An important aspect of quality control is therefore that each individual dataset requires error profiling and that relying on general properties of the platform used is not sufficient.

Assembly

Assembly of shotgun sequencing data can in general follow two strategies: the overlap-layout-consensus (OLC) and the de Bruijn graph approach (see also “► [A De Novo Metagenomic Assembly Program for Shotgun DNA Reads](#)”). These two strategies are employed by a number of different genome assemblers, and this topic has been reviewed recently (Miller et al. 2010). Basically, the OLC assembly involves the pairwise comparison of sequence reads and the ordering of matching pairs into an overlap graph. These overlapping sequences are then merged into a consensus sequence. Assembly with the de Bruijn strategy involves representing each sequence’s reads in a graph of all possible *k*-mers. Two *k*-mers are connected when the sequence reads have them in sequential, overlapping positions. Thus, all reads of a dataset are represented by the connection within

the de Bruijn graph, and assembled contigs are generated by traversing these connections to yield a sequence of k-mers.

The OLC assembly has the advantage that pairwise comparison can be performed to allow for a defined degree of dissimilarity between reads. This can compensate for sequencing errors and allows for the assembly of reads from heterogeneous populations (Tyson et al. 2004). However, memory requirement for pairwise comparisons increases exponentially with the numbers of reads in the dataset; hence, the OLC assembler often cannot deal with large datasets (e.g., Illumina data). Nevertheless, several OLCs, including the Celera Assembler (Miller et al. 2008), Phrap (de la Bastide and McCombie 2007), and Newbler (Roche), have been used to assemble partial or complete draft genomes from metagenomic data; see, for example, Tyson et al. (2004), Liu et al. (2011), and Brown et al. (2012).

In contrast, memory requirements of de Bruijn assemblers are largely determined by the k-mer size chosen to define the graph. Thus, these assemblers have been used successfully with large numbers of short reads. Initially, de Bruijn assemblers designed for clonal genomes, such as Velvet (Zerbino and Birney 2008), SOAP (Li et al. 2008), and ABySS (Simpson et al. 2009), were used to assemble metagenomic data. Because of the heterogeneous nature of microbial populations, however, assemblies often ended up fragmented. One reason is that every positional difference between two reads from the same region of two closely related genomes will create a “bubble” in the graph. Another reason is that sequence errors in low-abundance reads cause terminating branches. Traversing such a highly branched graph leads to large number of contigs. These problems have been partially overcome by modification of existing de Bruijn assemblers such as MetaVelvet (Namiki et al. 2012) or by newly designed de Bruijn-based algorithms such as Meta-IDBA (Peng et al. 2011; see also “Meta-IDBA, overview”). Conceptually, these solutions often include the identification of subgraphs that

correspond to individual genomes or the abundance information of k-mers to find an optimal solution path through the graph.

These subdividing approaches are analogous to binning metagenomic reads or contigs, in order to identify groups of sequences that define a specific genome. These bins or even individual sequence reads can also be taxonomically classified by comparison with known reference sequences. Binning and classifying of sequences can be based on phylogeny, similarity, or composition (or combinations thereof), and a large number of algorithms and software is available. For recent comparisons and benchmarking of binning and classification software, please see Bazinet and Cummings (2012) and Droge and McHardy (2012). Obviously, care has to be taken with any automated process, since nonrelated sequences can be combined to produce genomic chimera bins or classes. It is thus advisable that any binning or classification strategy is thoroughly tested through appropriate *in vitro* and *in silico* simulations (Mavromatis et al. 2007; Morgan et al. 2010; McElroy et al. 2012). Also, manual curation of contigs and iterative assembly and mapping can produce improved genomes from metagenomic data (Dutilh et al. 2009). Through such carefully designed strategies and refined processes, nearly complete genomes can be assembled, even for low-abundance organisms from large numbers of short reads (Iverson et al. 2012).

Annotation

Initially, techniques developed for annotating clonal genomes were applied to metagenomic data, and several tools for metagenomic analysis, such as MG-RAST (Meyer et al. 2008) and IMG/M (Markowitz et al. 2008), were derived from existing software suites. For metagenomic projects, the principal challenges lie in the size of the dataset, the heterogeneity of the data, and the fact that sequences are frequently short, even if assembled prior to analysis.

The first step of the analysis (after extensive quality control; see above) involves identification

of genes from a DNA sequence. Fundamentally, two approaches exist: the extrinsic approach, which relies on similarity comparison of an unknown sequence to existing databases, and the intrinsic (or *de novo*) approach, which applies statistical analysis of sequence properties, such as frequently used codon usage, to define likely open reading frames (ORFs). For metagenomic data, the extrinsic approach (e.g., running a similarity search with BLASTX) comes at a significant computational cost (Wilkening et al. 2009), rendering it less attractive. *De novo* approaches based on codon or nucleotide k-mer usage are thus more promising for large datasets. *De novo* gene-calling software for microbial genomes are trained on long contigs and assume clonal genomes. For metagenomic datasets this approach is often however unsuitable, because training data is lacking and multiple different codon usage (or k-mer) profiles are present due to the multiple, different genomes present.

However, several software packages have been designed to predict genes for short fragments or even reads (see Trimble et al. 2012 for a review). The most important finding of that review is the effect of errors on gene prediction performance, reducing the reading frame accuracy of most tools to well below 20 % at 3 % sequencing error. Only the software FragGeneScan (Rho et al. 2010; see also FragGeneScan, overview) accounted for the possibility that metagenomic sequences may contain errors, thus allowing it to clearly outperform its competitors.

Once identified, protein-coding genes require functional assignment. Here again, numerous tools and databases exist. Many researchers have found that performing BLAST analysis against the NCBI nonredundant database adds little value to their metagenomic datasets. Preferable are databases that contain high-level groupings of functions, for example, into metabolic pathways as in KEGG (Kanehisa 2002) or into subsystems as in SEED (Overbeek et al. 2005). Using such higher-level groupings allows for the generation of

overviews and comparison between samples after statistical normalization.

The time and resources required to perform functional annotations are substantial, but approaches that project multiple results derived from a single sequence analysis into multiple namespaces can minimize these computational costs (Wilke et al. 2012). Numerous tools are also available to predict, for example, short RNAs and/or other genomic features, but these tools are frequently less useful for large metagenomic datasets that exhibit both low sequence quality and short reads.

Several integrations package annotation functionality into a single website. The CAMERA (Seshadri et al. 2007) website, for example, provides users with the ability to run a number of pipelines on metagenomic data. The Joint Genome Institute's IMG/M web service also provides an analysis for assembled metagenomic data, which has been used so far for over 300 metagenomic datasets. The European Bioinformatics Institute provides a service aimed at smaller, typically 454/pyrosequencing-derived metagenomes. The most popular service is the MG-RAST system (Meyer et al. 2008), used for over 50,000 metagenomes with over 140 billion base pairs of data. The system offers comprehensive quality control, tools for comparison of datasets, and data import and export tools to, for example, QIIME (Caporaso et al. 2010) using standard formats such as BIOM (McDonald et al. 2012).

Metadata, Standards, Sharing, and Storage

With over 50,000 metagenomes available, the scientific community has realized that standardized metadata ("data about data") and higher-level classification (e.g., a controlled vocabulary) will increase the usefulness of datasets for novel discoveries (see also ► [Metagenomics, Metadata, and Meta-analysis](#)). Through the efforts of the Genomic Standards Consortium (GSC) (Field et al. 2011), a set of minimal questionnaires has

been developed and accepted by the community (Yilmaz et al. 2010) that allows effective communication of metadata for metagenomic samples of diverse types. While the “required” GSC metadata is purposefully minimal and thus provides only a rough description, several domain-specific environmental packages exist that contain more detailed information.

As the standards evolve to match the needs of the scientific community, the groups developing software and analysis services have begun to rely on the presence of GSC-compliant metadata, effectively turning them into essential data for any metagenome project. Furthermore, comparative analysis of metagenomic datasets is becoming a routine practice, and acquiring metadata for these comparisons has become a requirement for publication in several scientific journals. Since reanalysis of raw sequence reads is often computationally too costly, the sharing of analysis results is also advisable. Currently only the IMG/M and MG-RAST platforms are designed to provide cross-sample comparisons without the need to recompute analysis results. In the MG-RAST system, moreover, users can share data (after providing metadata) with other users or make data publicly available.

Metagenomic datasets continue to grow in size. Indeed the first multi-hundred gigabase pair of metagenomes already exists. Therefore, storage and curation of metagenomic data have become a central theme. The on-disk representation of raw data and analyses has led to massive storage issues for groups attempting meta-analyses. Currently there is no solution for accessing relevant subsets of data (e.g., only reads and analyses pertaining to a specific phylum or a specific species) without downloading the entire dataset. Cloud technologies may in the future provide attractive computational solutions for storage and computing problems. However, specific and metadata-enabled solutions are required for cloud systems to power the community-wide (re-)analysis efforts of the first 50,000 metagenomes.

Conclusion

Metagenomics has truly proven a valuable tool for analyzing microbial communities. Technological advances will continue to drive down the sequencing cost for metagenomic projects and, in fact, the flood of current datasets indicates that funding to obtain sequences is not a major limitation. Major bottlenecks are encountered, however, in terms of storage and computational processing of sequencing data. With community-wide efforts and standardized tools, the impact of these current limitations might be managed in the short term. In the long term, however, large standardized databases will be required (e.g., a MetaGeneBank) to give information access to the entire scientific community. Every metagenomic dataset contains many new and unexpected discoveries, and the efforts of microbiologists worldwide will be needed to ensure that nothing is being missed. As for the data, whether raw or processed, it is just data. Only its biological and ecological interpretation will further our understanding of the complex and wonderful diversity of the microbial world around us.

Government License

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a US Department of Energy Office of Science Laboratory, is operated under Contract No. DE-AC02-06CH11357. The US Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

References

- Barberan A, Bates ST, et al. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 2012;6(2):343–51.
- Barns SM, Fundyga RE, et al. Remarkable archaeal diversity detected in a Yellowstone National Park hot

- spring environment. *Proc Natl Acad Sci U S A*. 1994;91(5):1609–13.
- Bates ST, Berg-Lyons D, et al. Examining the global distribution of dominant archaeal populations in soil. *ISME J*. 2011;5(5):908–17.
- Bazinnet AL, Cummings MP. A comparative evaluation of sequence classification programs. *BMC Bioinforma*. 2012;13(1):92.
- Bentley DR, Balasubramanian S, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–9.
- Bergmann GT, Bates ST, et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem*. 2011;43(7):1450–5.
- Brown MV, Lauro FM, et al. Global biogeography of SAR11 marine bacteria. *Mol Syst Biol*. 2012;8:595.
- Caporaso JG, Kuczynski J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6.
- de la Bastide M, McCombie WR. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinforma*. 2007. Chapter 11: Unit11 14.
- Delmont TO, Malandain C, et al. Metagenomic mining for microbiologists. *ISME J*. 2011;5(12):1837–43.
- Delmont TO, Prestat E, et al. Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J*. 2012;6(9):1677–87.
- DeLong EF, Preston CM, et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*. 2006;311(5760):496–503.
- Dinsdale EA, Edwards RA, et al. Functional metagenomic profiling of nine biomes. *Nature*. 2008;452(7187):629–32.
- Droge J, McHardy AC. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform*. 2012;13(6):646–55.
- Dutilh BE, Huynen MA, et al. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics*. 2009;25(21):2878–81.
- Eid J, Fehr A, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133–8.
- Fan L, Reynolds D, et al. Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc Natl Acad Sci U S A*. 2012;109(27):E1878–87.
- Field D, Amaral-Zettler L, et al. The genomic standards consortium. *PLoS Bio*. 2011;9(6):e1001088.
- Fuhrman JA. Microbial community structure and its functional implications. *Nature*. 2009;459(7244):193–9.
- Fuhrman JA, Hewson I, et al. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci U S A*. 2006;103(35):13104–9.
- Fuhrman JA, Steele JA, et al. A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A*. 2008;105(22):7774–8.
- Gilbert JA, Field D, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One*. 2010a;5(11):e15545.
- Gilbert JA, Meyer F, et al. The earth microbiome project: meeting report of the "1 EMP meeting on sample selection and acquisition at Argonne National Laboratory October 6 2010". *Stand Genomic Sci*. 2010b;3(3):249–53.
- Gilbert JA, Bailey M, et al. The earth microbiome project: the Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2010. *Stand Genomic Sci*. 2011;5(2):243–7.
- Gilbert JA, Steele JA, et al. Defining seasonal marine microbial community dynamics. *ISME J*. 2012;6:298–308.
- Gill SR, Pop M, et al. Metagenomic analysis of the human distal gut microbiome. *Science*. 2006;312(5778):1355–9.
- Hess M, Sczyrba A, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*. 2011;331(6016):463–7.
- Iverson V, Morris RM, et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*. 2012;335(6068):587–90.
- Kanehisa M. The KEGG database. *Novartis Found Symp*. 2002;247:91–101. discussion 101–103, 119–128, 244–152.
- Knight R, Jansson J, et al. Designing better metagenomic surveys: the role of experimental design and metadata capture in making useful metagenomic datasets for ecology and biotechnology. *Nat Biotechnol*. 2012;30(6):513–2.
- Koren S, Schatz MC, et al. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat Biotechnol*. 2012;30(7):693–700.
- Li R, Li Y, et al. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24(5):713–4.
- Liu MY, Kjelleberg S, et al. Functional genomic analysis of an uncultured delta-proteobacterium in the sponge *Cymbastela concentrica*. *ISME J*. 2011;5(3):427–35.
- Loman NJ, Misra RV, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(5):434–9.
- Mackelprang R, Waldrop MP, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*. 2011;480(7377):368–71.
- Margulies M, Egholm M, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376–80.
- Markowitz VM, Ivanova NN, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*. 2008;36(Database issue):D534–8.
- Martiny JB, Bohannan BJ, et al. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*. 2006;4(2):102–12.

- Mavromatis K, Ivanova N, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*. 2007;4(6):495–500.
- McDonald D, Clemente JC, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*. 2012;1(1):7.
- McElroy KE, Luciani F, et al. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*. 2012;13:74.
- Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010;11(1):31–46.
- Meyer F, Paarmann D, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma*. 2008;9:386.
- Miller JR, Delcher AL, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*. 2008;24(24):2818–24.
- Miller JR, Koren S, et al. Assembly algorithms for next-generation sequencing data. *Genomics*. 2010;95(6):315–27.
- Morgan JL, Darling AE, et al. Metagenomic sequencing of an *in vitro*-simulated microbial community. *PLoS One*. 2010;5(4):e10209.
- Namiki T, Hachiya T, et al. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res*. 2012;40(20):e155.
- Nemergut DR, Costello EK, et al. Global patterns in the biogeography of bacterial taxa. *Environ Microbiol*. 2011;13(1):135–44.
- Ottesen EA, Marin R, et al. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J*. 2011;5(12):1881–95.
- Overbeek R, Begley T, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 2005;33(17):5691–702.
- Peng Y, Leung HC, et al. Meta-IDBA: a *de Novo* assembler for metagenomic data. *Bioinformatics*. 2011;27(13):i94–101.
- Prabakaran P, Streaker E, et al. 454 antibody sequencing – error characterization and correction. *BMC Res Notes*. 2011;4:404.
- Prosser JJ. Replicate or lie. *Environ Microbiol*. 2010;12(7):1806–10.
- Quail M, Smith ME, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012;13(1):341.
- Rho M, Tang H, et al. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010;38(20):e191.
- Riesenfeld CS, Schloss PD, et al. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*. 2004;38:525–52.
- Rothberg JM, Hinze W, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011;475(7356):348–52.
- Rusch DB, Halpern AL, et al. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*. 2007;5(3):e77.
- Schneider GF, Dekker C. DNA sequencing with nanopores. *Nat Biotechnol*. 2012;30(4):326–8. doi: 10.1038/nbt.2181.
- Salmela L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics*. 2010;26(10):1284–90.
- Seshadri R, Kravitz SA, et al. CAMERA: a community resource for metagenomics. *PLoS Biol*. 2007;5(3):e75.
- Simpson JT, Wong K, et al. ABySS: a parallel assembler for short read sequence data. *Genome Res*. 2009;19(6):1117–23.
- Trimble WL, Keegan KP, et al. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinforma*. 2012;13(1):183.
- Tringe SG, von Mering C, et al. Comparative metagenomics of microbial communities. *Science*. 2005;308(5721):554–7.
- Tyson GW, Chapman J, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 2004;428(6978):37–43.
- Venter JC, Remington K, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304(5667):66–74.
- Warnecke F, Luginbuhl P, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. 2007;450(7169):560–5.
- Whiteley AS, Jenkins S, et al. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) platform. *J Microbiol Methods*. 2012;91(1):80–8.
- Wilke A, Harrison T, et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinforma*. 2012;13:141.
- Wilkering J, Wilke A, et al. Using clouds for metagenomics: a case study. *IEEE Cluster 2009*. 2009
- Wommack KE, Bhavsar J, et al. Metagenomics: read length matters. *Appl Environ Microbiol*. 2008;74(5):1453–63.
- Yilmaz P, Kottmann R, et al. The “Minimum Information about an ENvironmental Sequence” (MIENS) specification. *Nat Biotechnol*. 2010. in print.
- Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821–9.
- Zhou R, Ling S, et al. Population genetics in nonmodel organisms: II. Natural selection in marginal habitats revealed by deep sequencing on dual platforms. *Mol Biol Evol*. 2011;28(10):2833–42.

A De Novo Metagenomic Assembly Program for Shotgun DNA Reads

Huaiqiu Zhu

Department of Biomedical Engineering, and
Center for Theoretical Biology, Peking
University, Beijing, China

Synonyms

MAP: metagenomic assembly program

Definition

Contig: a set of overlapping DNA segments that together represent a consensus region of DNA. Assembly (also genome assembly): the process of taking a large number of short DNA sequencing reads and putting them back together to create contigs from which the DNA originated.

Introduction

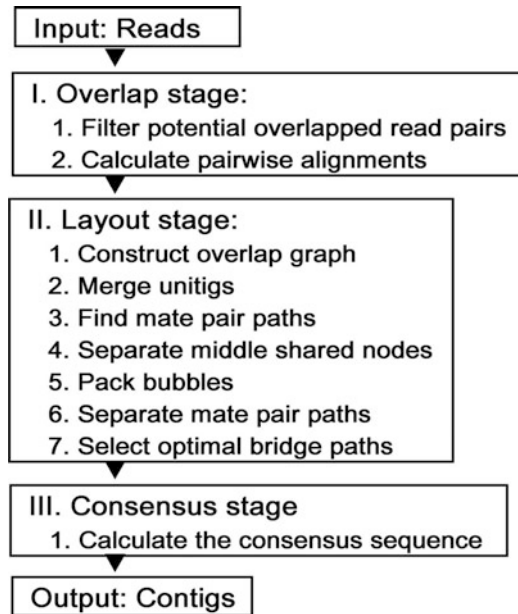
MAP (metagenomic assembly program) is a de novo assembler designed to be applicable to shotgun DNA reads (recommended as >200 bp) for metagenome sequencing project (Lai et al. 2012). The program focuses on the metagenomic assembly problem of longer reads produced by, for example, Sanger (typically 700–1,000 bp) and 454 sequencing (typically 200–500 bp). Meanwhile, mate-pair information from both ends of a DNA fragment for a given size (e.g., an insert in a vector plasmid in Sanger sequencing or a mate-pair template in 454 sequencing) in sequencing is introduced, which is commonly available in Sanger sequencing and most of the new sequencing technologies including 454 sequencing.

Although processing of shotgun metagenomic sequence data usually does not have a fixed end point to recover one or more complete genomes as for isolated microbial genomes, the assembly tools, which aim to combine sequence reads into contigs, are still expected to play an important

role in sequence processing, due to more valuable genomic content they can provide (Tyson et al. 2004; Venter et al. 2004). In the past decade, a good many assembly algorithms have been proposed to deal with the sequence assembly problem, among of which are the early algorithms targeted to the Sanger sequencing technology, such as Phrap (<http://www.phrap.org>), Celera (Myers et al. 2000; Miller et al. 2008), and PCAP (Huang et al. 2003), and the up-to-date algorithms targeted to the next-generation technology, such as Velvet (Zerbinor and Birney 2008) and SOAPdenovo (Li et al. 2010). However, these methods are not targeting the metagenome sequencing in spite of the situation that they are still usually employed to undertake assembling of the metagenomic sequencing reads.

Compared to isolated genome assembly problem, the metagenomic assembly problem is more complicated due to two challenges (Kunin et al. 2008): (1) the genomic repeats may originate from either the same genome or the different genomes; therefore, large numbers of mixed short DNA reads belong to many different species (we even know little about the population structure for some environmental samples); and (2) the inhomogeneous coverage distribution and the low abundance of organisms provide limited information to handle repeats. Due to the specific challenges of the metagenomic assembly problem, traditional assembly methods developed for single genome assembly problem usually generate poor quality draft assembly on metagenomic data (Mavromatis et al. 2007). Thus, it is in need to develop highly efficient assembly method specifically for metagenomic data.

Moreover, compared with Sanger and 454 sequencing, the current limitation of shorter reads (<200 bp, typically 25–100 bp) and higher errors by the new sequencing platforms does not allow a significant utility for metagenomic analyses for the difficulty in phylogenetic study or gene function inference. In fact, shorter reads technologies have not been widely used in metagenome sequencing, and meanwhile the sequencing technologies producing longer reads, such as Sanger (usually 700–1,000 bp)



A De Novo Metagenomic Assembly Program for Shotgun DNA Reads, Fig. 1 The flowchart of MAP algorithm

and 454 sequencing (usually 200–500 bp), are still the overwhelming recommendation and thus remain the major source of metagenomic sequence data. Therefore, it is never trivial to continue to emphasize the importance of longer reads to metagenomic analyses, clearly including the reads assembly tool designed specifically.

Algorithm of MAP

MAP designs an improved approach of the classical overlap/layout/consensus (OLC) strategy, in which several special algorithms are incorporated into its stages, to calculate correct contigs by connecting the fragments linked by mate pairs to prevent the false merge of unrelated reads. For the improved OLC strategy, MAP deploys a series of algorithms in three stages as shown in Fig. 1. In the overlap stage, the filter algorithm based on q-gram (Mullikin et al. 2003) is used to obtain the read pairs that are supposed to have the overlaps, and the seed and extend alignment approach, similar to that used by BLAST (Altschul et al. 1990), is employed in the pairwise alignment calculation. While in the consensus

stage, a consistency-based consensus algorithm is used (Rausch et al. 2009), which is based on a multi-read alignment algorithm aligning the reads with a consistency-enhanced alignment graph of shared sequence segments identified in advance. The most important innovation of MAP is the layout stage which applies mate-paired information to deal with repeat problem, which is described below.

In the OLC approach of MAP, the overlap graph is used to facilitate the assembly process. Conceptually, reads and overlaps are represented in the graph G by nodes and bidirected edges, respectively. The arrows of both ends of the edge are determined by the way how two reads overlap. Herein, a dovetail path is defined as an acyclic path with each node has only one arrow outward it and one arrow inward it. Thus, a dovetail path can determine a certain contig by means of threading the reads corresponding to the nodes in this path. Thus, the goal of the layout stage is to separate the graph into disconnected dovetail paths. However, since there may be quite many misleading edges in the graph that represent the false overlaps mainly originated from two repetitive DNA regions or similar

fragments of different genomes, this goal seems to be a formidable task. To this end, MAP is designed to determine the optimal dovetail paths with the aids of the clues given by mate pairs (Lai et al. 2012).

Compared with other assemblers, several distinct features of MAP algorithm should be pointed out. First, MAP does not refer to any other information such as genome length or sequencing coverage that is often used in the assemblers targeting the isolated genomes, because such information is clearly not applicable to the situation of metagenomic assembly. What is more important is that MAP employs mate-paired information different from other assemblers do. For example, the Celera Assembler (Myers et al. 2000) used mate-paired information in the scaffold construction. The Celera Assembler later developed a new pipeline CABOG, which finds the best overlap graph in the unitigger module (Miller et al. 2008). In this algorithm, mate pairs are used to correct the misassemblies by breaking the unitigs which are found violated with the mate-pair constrains. PCAP (Huang et al. 2003) used mate-paired information to correct contigs and to link contigs into scaffolds. Different from these assemblers, MAP uses mate pairs as a core measure to construct contigs when repeats hamper the assembly. Based on mate-paired information, MAP designs a series of procedures to implement the layout stage.

Performance of MAP

MAP is designed for metagenomic assembly on long reads data with mate pairs, such as Sanger reads (700–1,000 bp) and 454 reads (200–500 bp). MAP method was assessed on simulated data compared with widely used assemblers on long reads data. Specifically, the assessment test results on simulated dataset with 800 bp reads demonstrate that the total assembly performance of MAP can be superior to both Celera and Phrap for typical longer reads by Sanger sequencing, and the results on simulated dataset with 200 bp reads show that MAP has evident advantage over Celera, Newbler

(Margulies et al. 2005), and Genovo (Laserson et al. 2011), for typical shorter reads by 454 sequencing (Lai et al. 2012).

Availability

MAP is written in C++ and the source code is freely available under GNU GPL license. The MAP is freely available at <http://bioinfo.ctb.pku.edu.cn/MAP/>.

References

- Altschul SF, Gish W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
- Huang X, Wang J, et al. PCAP: a whole-genome assembly program. *Genome Res.* 2003;13:2164–70.
- Kunin V, Copeland A, et al. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev.* 2008;72:557–178.
- Lai B, Ding R, et al. A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics.* 2012;28(11):1455–62.
- Laserson J, Jojic V, et al. Genovo: de novo assembly for metagenomes. *J Comput Biol.* 2011;18:429–43.
- Li R, Zhu H, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 2010;20:265–72.
- Margulies M, Egholm M, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437:376–80.
- Mavromatis K, Ivanova N, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4:495–500.
- Miller JR, Delcher AL, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics.* 2008;24:2818–24.
- Mullikin JC, Ning Z, et al. The phusion assembler. *Genome Res.* 2003;13:81–90.
- Myers EW, Sutton GG, et al. A whole-genome assembly of *Drosophila*. *Science.* 2000;287:2896–204.
- Rausch T, Koren S, et al. A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads. *Bioinformatics.* 2009;25:1118–24.
- Tyson GW, Chapman J, et al. Genomic structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004;428:37–43.
- Venter JC, Remington K, et al. Environmental genome shotgun sequencing of Sargasso sea. *Science.* 2004;304:66–74.
- Zerbinor DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18:821–9.

Ab Initio Gene Identification in Metagenomic Sequences

Shiyuyun Tang¹ and Mark Borodovsky²

¹School of Biology, Biodiversity Research Center, Georgia Institute of Technology, Atlanta, GA, USA

²Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Center for Bioinformatics and Computational Genomics, Atlanta, GA, USA

Synonyms

Statistical or intrinsic methods of gene prediction

Definition

Computational inference of how a metagenomic sequence is divided into protein-coding and non-coding regions based on presence or absence of characteristic oligonucleotide frequency patterns.

Introduction

As of April 2013 sequences of 370 metagenomes were available in databases. On the other hand, Genomes Online Database (www.genomesonline.org) lists 186 complete archaeal and 3,956 complete bacterial genomes; also there are about 15,000 incomplete (draft) prokaryotic genomes. With the average size of a metagenome being 100 times larger than an average prokaryotic genome, the current volume of metagenomic sequences is twice as large as the total sequence in “genomic” data. Therefore, current metagenomes carry a larger wealth of genes than all the prokaryotic genomes, and this gap is growing.

Notably, gene prediction and annotation of gene and protein function is more challenging in metagenomes than in draft or complete genomes. To give a historic perspective, one can compare gene annotation of a metagenome with

annotation of the first completely sequenced archaeal genome, *Methanococcus jannaschii* (Bult et al. 1996). All the *M. jannaschii* genes were predicted by the ab initio statistical method (Borodovsky and McIninch 1993) while function of 2/3 of them was a mystery since the translated protein sequences did not show sequence similarity to proteins in databases.

The history repeats itself in metagenomes, since majority of protein-coding regions in a new metagenome may code for proteins that do not show similarity to already known proteins. “Evidence-based” or “similarity-based” methods of gene finding (Kunin et al. 2008) provide gene prediction along with valuable information about function of encoded proteins. Similarity-based gene finders possess high specificity, close to 100 % (Altschul et al. 1997; Badger and Olsen 1999; Frishman et al. 1998; Gish and States 1993). Still, the drawback of similarity-based methods is low sensitivity; they cannot predict novel genes.

The similarity-based methods are less useful for gene prediction in metagenomes that carry many novel genes, while the ab initio gene prediction methods, not depending on presence of homologs in protein databases, are both effective and efficient for annotating genes in metagenomic sequences (Kunin et al. 2008).

Ab Initio Gene Finding

Ab initio gene prediction tools have high sensitivity (above 90 % for the best tools) and high specificity (above 90 % as well). Ab initio gene finders use statistical pattern recognition methods (Wooley et al. 2010). Statistical models such as Markov models, hidden Markov models (HMM), and hidden semi-Markov models (HSMM, also called hidden Markov model with duration) proved to be very useful to model statistical patterns of nucleotide ordering in protein-coding and noncoding regions. Accurate ab initio gene finding in isolated genomes requires ample sequence data for estimation of algorithm parameters (model training).

Contrary to isolated (complete and draft) genomes metagenomic sequences are derived

from numerous genomes of heterogeneous microbial communities (microbiomes). A typical metagenomic sequence is short; its genomic context and the phylogenetic origin are rarely known. Gene identification is also affected by sequencing and assembly errors; for example, errors that lead to frameshifts (change of coding frame).

The major challenge for ab initio gene prediction in metagenomic sequences is that the metagenomic sequences are often too short for reliable estimation of parameters of species-specific models of coding and noncoding regions. Special training techniques have to be developed to address the challenging task of parameter estimation (see below). Similarly to gene prediction in isolated genomes, newly predicted genes are immediately translated into proteins and the similarity search is used in an attempt of function annotation.

Gene Finders Currently Available for Metagenomes

Current metagenomic gene-finding tools include FragGeneScan (Rho et al. 2010), Glimmer-MG (Kelley et al. 2012), MetaGene Annotator (Noguchi et al. 2008), MetaGeneMark (Zhu et al. 2010), and Orphelia (Hoff et al. 2009, 2008). Glimmer-MG and MetaGeneMark are extensions of gene finders for complete or draft genomes Glimmer3 (Delcher et al. 2007) and GeneMarkS (Besemer et al. 2001), respectively.

The MetaGeneMark algorithm uses HSMM architecture, originally developed in GeneMarkS (Besemer et al. 2001). The HSMM parameter derivation approach used in MetaGeneMark is to arrive to a large set of parameters (thousands of parameters related to oligonucleotide frequencies) from a small set (nucleotide frequencies determined in a short fragment) using the dependencies between oligonucleotide and nucleotide frequencies that have been formed in evolution. The original idea of this approach (Besemer and Borodovsky 1999) has been developed for small viral genomes before the start of “metagenomic era” (see below for more details).

Glimmer-MG is based on interpolated Markov models or IMM (Salzberg et al. 1998). Glimmer-MG scores metagenomic sequences and assigns them into clusters; then, the algorithm iteratively estimates the IMM parameters and reassigns sequences to clusters.

FragGeneScan (Rho et al. 2010), an HMM-based gene finder, has an additional ability to predict frameshifts caused by sequencing errors. Transition probabilities between coding frames are determined with respect to the error models of sequencing technologies used to derive the input sequence.

MetaGene Annotator (Noguchi et al. 2008) works in two steps: in the first step the program scores open reading frames (ORFs) with respect to base composition and lengths; in the second step, it connects high-scoring ORFs using dynamic programming.

Machine learning classification algorithms such as support vector machines and neural networks are also used for ab initio gene finding. In order to classify coding or noncoding ORFs, Orphelia (Hoff et al. 2009, 2008) uses an artificial neural network combining multiple features to get ORF's scores.

Parameter Estimation for Metagenomic Gene-Finding Algorithms

Patterns of oligonucleotide frequencies differ in coding and noncoding regions; these patterns are more pronounced when frequencies of longer oligomers are considered. Sequences with specific oligomer frequencies can be modeled by Markov chain models and in the important case of protein-coding sequences by three-periodic Markov chain models (Borodovsky et al. 1986). The number of parameters of a three-periodic Markov chain model increases exponentially with the model order; estimation of parameters of the practically useful fifth order model requires at least several hundred thousand nucleotide long sequence. Use of a shorter training sequence leads to over-fitting and will corrupt gene prediction. If the origin of the metagenomic sequence is

known, sequences from the whole parent genome could be used for training. Alternatively, if novel metagenomic sequences from a single species are assembled in sufficiently long contig the model parameters can be estimated by self-training on the contig sequence (Besemer et al. 2001; Kelley et al. 2012). Most frequently, however, metagenomic sequences are short and novel (of the order of a few hundred nucleotides). Therefore, new approach to the model parameter derivation is needed.

A novel approach for constructing parameters and making efficient models for gene prediction in short genomic sequences was proposed back in 1999 (Besemer and Borodovsky 1999). The idea was to use observed trends in the nucleotide frequencies in the three codon positions in genomes with various GC content. Use of these dependencies allows for reconstructing the species-specific codon usage pattern in the whole genome starting from a short fragment of this genome whose length is sufficient to estimate the genome GC content. This approach is based on the assumption of genome compositional uniformity that is largely valid for prokaryotic genomes. It was shown that parameters provided by this approach allow sufficiently accurate gene prediction in short metagenomic sequences. Later on, with more genomes becoming available, this idea was extended (Zhu et al. 2010) to longer oligonucleotides (e.g., hexamers). With GC content of a genome being an independent variable X , it could be shown that frequency of phased K -mers in any of three frames, variable Y , can be approximated by a polynomial of order K . Particularly, the mononucleotide frequencies in three codon positions can be approximated by linear functions. These dependencies indicate that GC content is a major driving factor that determines evolution of genome-wide codon usage pattern (Chen et al. 2004). In MetaGeneMark, the value of GC content determined for a short metagenomic sequence is used as an estimate of GC content of the whole genome the sequence originated from. This value allows immediate reconstruction of frequencies of phased oligonucleotides and, at the

next step, parameters of three-periodic Markov chain models of the heuristic model (Zhu et al. 2010).

Interestingly, the heuristic models can also be used for gene prediction in complete genomes or draft genomes. In comparison with the “native” models (models trained on a genome of interest), heuristic models are more sensitive to so-called “atypical” genes. Many atypical genes appear to be horizontally transferred genes with codon frequencies deviating from dominant codon usage pattern of the “host” genome.

Another approach to model parameter estimation is attempting to make a sufficiently large set of training sequences by linking anonymous sequences that appear to be taxonomically close. For example, Glimmer-MG assigns a taxon for a metagenomic sequence by a classification method called Phymm (Brady and Salzberg 2009) and then searches databases for genomes that belong to this taxon. Since such type of training is executed in real time, the running time of gene-finding algorithm may increase significantly in comparison with the algorithm selecting a heuristic model from a set of models precomputed for possible values of GC contents.

Additional Sequence Features Used by Metagenomic Gene Finders

Besides function-specific patterns in oligonucleotide composition, gene identification algorithms can use additional features that help discriminate protein-coding and noncoding regions. Such features include empirical length distributions of coding and noncoding regions, mutual orientation of neighboring coding regions, and sequence patterns related to functional sites such as ribosomal binding sites (RBS). The two-component model of RBS, containing positional frequency matrix as a model of the RBS motif and the length distribution of a “spacer,” the sequence between RBS and gene start, carries important additional information for improving accuracy of gene start prediction. In prokaryotic genomes an average spacer length is 5–7 nt. The RBS positional

Ab Initio Gene Identification in Metagenomic Sequences, Table 1 Gene prediction accuracy for five ab initio gene finders. Sn stands for sensitivity and Sp stands for specificity

Programs	Test set	Sequence length (bp)	Sn (%)	Sp (%)	(Sn + Sp)/2 (%)	Publication
Orphelia	Fragments from 12 test species	300	82.1	91.7	86.9	Hoff et al. (2009)
FragGeneScan	Simulated short reads of 9 genomes	400	91.3	86.1	88.7	Rho et al. (2010)
MetaGeneMark	Fragments from 50 microbial chromosomes	400	97.0	94.6	95.8	Zhu et al. (2010)
Glimmer-MG	Simulated 454 sequences	535	98.4	71.8	85.1	Kelley et al. (2012)
MetaGeneAnnotator	Subsequences of 13 genomes	700	95.1	91.0	93.1	Noguchi et al. (2008)
FragGeneScan	Simulated reads with 1 % sequencing error rate	400	85.4	79.5	82.5	Rho et al. (2010)
Glimmer-MG	Simulated 454 reads with 1 % sequencing error rate	535	83.6	62.5	73.1	Kelley et al. (2012)

frequency matrix can be derived by algorithms such as MCMC (Markov chain Monte Carlo)-based Gibbs sampler (Lawrence et al. 1993) or EM (Expectation Maximization)-based MEME (Bailey and Elkan 1994); detection of the RBS motif is done by finding the most conserved set of ungapped sequence fragments within the multiple alignment window. The structure of two-component RBS model is convenient for incorporation into HMM-based framework of several algorithms such as MetaGeneMark and FragGeneScan

Another feature, the prokaryotic gene length distribution, is approximated for complete or draft genomes by the gamma distribution with mean value about 900 nt; yet another one, the distribution of length of noncoding region is approximated by exponential distribution. These two distributions, as well as the RBS spacer length distribution, are used as in the HSMM-based algorithms (Besemer et al. 2001). Since short metagenomic sequences are more likely to contain partial genes than complete genes, length distributions of partial genes are used in HSMM-based metagenomic gene finders (Rho et al. 2010; Zhu et al. 2010).

About 70 % of neighboring genes in prokaryotic genomes have the same orientation

(Noguchi et al. 2006), and many of them make co-transcribed “chains” or operons. Genes in an operon are located on a close distance or even overlap. Four base-pair overlap ATGA is very common in adjacent genes as an overlap of stop and start codons ATG and TGA. Average distance between adjacent genes having the same orientation is shorter than that between neighbor genes residing in complementary strands, especially in gene start-to-gene start configuration where additional space has to be available for promoters.

All these features are incorporated in metagenomic gene finders, e.g., MetaGeneMark. Tests of ab initio gene finders on simulated metagenomic sequences have shown that these algorithms are quite accurate, with average values of sensitivity and specificity above 90 %; see Table 1. However, the sensitivity drops if the sequence length goes below 200 nt (Yok and Rosen 2011; Zhu et al. 2010).

An Initio Gene Finding in Metagenomic Sequences with Errors

Real metagenomic sequences contain errors: substitutions, insertion, and deletions (indels), as well

Ab Initio Gene Identification in Metagenomic Sequences, Table 2 Frameshift prediction accuracy

Programs	Sequence length (bp)	Sn (%)	Sp (%)	(Sn + Sp)/2	Test set	Publication
FragGeneScan	400	81.0	43.2	62.1	Fragments from 18 prokaryotic genomes with 20 % containing frameshifts	Tang et al. 2013
	600	81.9	35.1	58.5		
	800	82.8	29.4	56.1		
MetaGeneTrack	400	75.8	70.2	73.0		
	600	80.1	61.7	70.9		
	800	81.5	51.9	66.7		

as chimerisms, when two reads from different species are joined due to assembly error. Indels can cause frameshifts in coding regions; thus gene prediction accuracy is affected by sequencing errors. The overall effect on accuracy depends on error rates specific to sequencing and finishing technologies; for example, the error rates reported for Sanger sequencing may be as low as 0.001 % while sequencing errors in NGS technologies can go above 1 %. In both simulated Sanger reads and simulated 454 reads significant decrease of gene prediction sensitivity is observed when error rate exceeds 1 % (Hoff 2009). Still, in assembled sequences, the per-nucleotide error rate of 0.5 % in raw reads can be reduced to as low as 0.005 %. This error rate is still large enough to affect ~3–4.5 % of genes in assembled sequences (Luo et al. 2012).

To identify frameshift errors in metagenomic sequences, gene-finding algorithms have to model frame transitions that occur due to sequencing errors. In HSMM-based gene finders, e.g., FragGeneScan, new hidden states designating transitions between coding frames in the same strand were incorporated into the HSMM architecture. Another recent tool able to detect frameshift in metagenomic coding regions is MetaGeneTack (Tang et al. 2013). It combines the original HSMM-based MetaGeneMark with an ab initio frameshift finding program GeneTack (Antonov and Borodovsky 2010). Several filters of false-positive predictions were employed in MetaGeneTack to achieve higher accuracy. MetaGeneTack is reported to have higher frameshift prediction specificity than FragGeneScan

(Table 2) in reads with error rate typical for metagenomic projects (Tang et al. 2013).

Yet another approach was used in Glimmer-MG, which, to trace possible indel errors, splits an ORF into three branches (frames), starting from the position of a nucleotide called with low confidence (Kelley et al. 2012). This approach was reported to have higher gene prediction accuracy on error-contained reads than FragGeneScan. Methods that account for sequencing errors generally perform better in real error-prone metagenomic sequences than “idealistic” approaches. The accuracy of sequencing error detection, however, depends on how accurate is the modeling of sequencing errors is.

Summary

Accurate ab initio gene prediction in metagenomic sequences is necessary for reliable functional annotation. Ab initio algorithms identify genes in metagenomic sequences by detecting intrinsic statistical patterns of coding and noncoding regions. Being independent of data stored in databases, these methods are especially useful for discovering novel genes. Special techniques have been developed for derivation of parameters of the ab initio algorithms working with short anonymous metagenomic sequences. We have reviewed several ab initio gene finders developed for metagenomic sequences including the latest tools that take into account possible sequencing errors (frameshifts).

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [FragGeneScan: Predicting Genes in Short and Error-Prone Reads](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)
- ▶ [Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies](#)
- ▶ [Proteomics and Metaproteomics](#)
- ▶ [RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data](#)

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
- Antonov I, Borodovsky M. Genetack: frameshift identification in protein-coding sequences by the viterbi algorithm. *J Bioinforma Comput Biol.* 2010;8(3):535–51. PubMed PMID: 20556861.
- Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol.* 1999;16(4):512–24.
- Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for Molecular Biology, Vol. 2; 1994; p. 28–36.* PubMed PMID: 7584402.
- Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* 1999;27(19):3911–20. PubMed PMID: 10481031. Pubmed Central PMCID: 148655.
- Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 2001;29(12):2607–18. PubMed PMID: 11410670. Pubmed Central PMCID: 55746.
- Borodovsky M, McIninch J. GENMARK: parallel gene recognition for both DNA strands. *Comp Chem.* 1993;17(2):123–33.
- Borodovsky MY, Sprizhitskii Y, Golovanov E, Aleksandrov A. Statistical patterns in primary structures of functional regions in the *E. coli* genome. III. Computer recognition of coding regions. *Mol Biol.* 1986;20:1145–50.
- Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods.* 2009;6(9):673–6. PubMed PMID: 19648916. Pubmed Central PMCID: 2762791.
- Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, et al. Complete genome sequence of the methanogenic archaeon. *Methanococcus jannaschii*. *Science.* 1996;273(5278):1058–73. PubMed PMID: 8688087.
- Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci U S A.* 2004;101(10):3480–5. PubMed PMID: 14990797. Pubmed Central PMCID: 373487.
- Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics.* 2007;23(6):673–9. PubMed PMID: 17237039. Pubmed Central PMCID: 2387122.
- Frishman D, Mironov A, Mewes H-W, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* 1998;26(12):2941–7.
- Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet.* 1993;3(3):266–72.
- Hoff KJ. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics.* 2009;10:520. PubMed PMID: 19909532. Pubmed Central PMCID: 2781827.
- Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, Meinicke P. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinforma.* 2008;9:217. PubMed PMID: 18442389. Pubmed Central PMCID: 2409338.
- Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 2009 Jul 37(Web Server issue): W101-5. PubMed PMID: 19429689. Pubmed Central PMCID: 2703946.
- Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 2012;40(1):e9. PubMed PMID: 22102569. Pubmed Central PMCID: 3245904.
- Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev.* 2008;72(4):557–78. Table of Contents. PubMed PMID: 19052320. Pubmed Central PMCID: 2593568.
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science.* 1993;262(5131):208–14. PubMed PMID: 8211139.
- Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE.* 2012;7(2): e30087.
- Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun

sequences. *Nucleic Acids Res.* 2006;34(19):5623–30. PubMed PMID: 17028096. Pubmed Central PMCID: 1636498.

Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res Int J Rapid Publ Rep Genes Genomes.* 2008;15(6):387–96. PubMed PMID: 18940874. Pubmed Central PMCID: 2608843.

Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010;38(20):e191. PubMed PMID: 20805240. Pubmed Central PMCID: 2978382.

Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 1998;26(2):544–8. PubMed PMID: 9421513. Pubmed Central PMCID: 147303.

Tang S, Antonov I, Borodovsky M. MetaGeneTack: ab initio detection of frameshifts in metagenomic sequences. *Bioinformatics.* 2013;29(1):114–6. PubMed PMID: 23129300. Pubmed Central PMCID: 3530910.

Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol.* 2010;6(2):e1000667. PubMed PMID: 20195499. Pubmed Central PMCID: 2829047.

Yok NG, Rosen GL. Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinforma.* 2011;12:20. PubMed PMID: 21232129. Pubmed Central PMCID: 3042383.

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38(12):e132. PubMed PMID: 20403810. Pubmed Central PMCID: 2896542.

AbundanceBin, Metagenomic Sequencing

Yuzhen Ye

Indiana University, School of Informatics and Computing, Bloomington, IN, USA

Definition

Binning is unsupervised clustering of metagenomic sequences into an unknown set of species.

AbundanceBin is a binning tool utilizing the different abundances of the species in a community.

Introduction

Binning is one of the challenging problems in the metagenomics field. It has two main applications. One application is for studying the structure of microbial communities. The other application is for improving the downstream analysis of metagenomic sequences, including metagenome assembly (which has shown to be extremely difficult), considering that assembling reads one bin at a time significantly reduces the complexity of the metagenome assembly problem.

Composition-based methods have been the main approaches to unsupervised classification of reads. The basis of these approaches is that the genome composition (G + C content, dinucleotide frequencies, and synonymous codon usage) vary among organisms and are generally characteristic of evolutionary lineages. Tools in this category include TETRA (Teeling et al. 2004), TACOA (Diaz et al. 2009), and MetaCluster (Leung et al. 2011). Due to the substantial variance in sequence properties along a genome, the main limitation of composition-based approaches is that they require relatively long reads (at least 800 bp), although it is shown that MetaCluster (Leung et al. 2011) can bin reads of 300 bp by employing a different distance metric (Spearman Footrule Distance) to reduce the local variations for 4-mers.

Note a large collection of methods have been developed to classify sequencing reads in a supervised manner. MEGAN (Huson and Mitra 2012) is a representative approach of this kind. These methods either use composition information (as in NCB, a naïve Bayes classifier to metagenomic sequence classification (Rosen et al. 2011)) or employ similarity searches of metagenomic sequences against a database of known genes/proteins (as in MEGAN) and assign metagenomic sequences to taxa accordingly, with or without using phylogeny. They also differ in the algorithms used for classification: MEGAN pioneers the lowest common ancestor (LCA) algorithm (Huson et al. 2007), MTR (Gori et al. 2011) improves on LCA algorithm considering multiple taxonomic ranks, and MetaPhyler (Liu et al. 2011) achieves better classification

results by tuning the taxonomic classifier to each matching length, reference gene, and taxonomic level. Note that some tools in this category can only classify a subset of the metagenomic sequences instead of all. MLTreeMap (Stark et al. 2010) uses phylogenetic analysis of 31 marker genes for taxonomic distribution estimation. CARMA (Krause et al. 2008) searches for conserved Pfam domains and protein families in raw metagenomic sequences and classifies them into a higher-order taxonomy. RDP classifier is designed for classification of 16S rRNA genes, and later extended to classification of 18S rRNA genes using a naïve Bayes classifier (Cole et al. 2009).

AbundanceBin

AbundanceBin (Wu and Ye 2011) is the first unsupervised clustering algorithm that utilizes abundance information of the species in the same microbial community to group reads into bins. The fundamental assumption of the AbundanceBin algorithm is that reads are sampled from genomes following a Poisson procedure, such that the sequencing reads can be modeled as a mixture of Poisson distribution.

An expectation–maximization (EM) algorithm is used in AbundanceBin to find parameters for the Poisson distributions (i.e., the means), which reflect the relative abundance levels of the source species. AbundanceBin then assigns reads to bins based on the fitted Poisson distributions. AbundanceBin gives an estimation of the genome size (or the concatenated genome size of species of the same or very similar abundances) and the coverage (which reflects the abundances of species) of each bin in an unsupervised manner without requiring prior knowledge of the structure of the microbial communities. The EM algorithm needs an important parameter, the number of bins, which is typically unknown, as for most metagenomic projects. AbundanceBin solves this problem by using a recursive binning approach to determine the total number of bins automatically. The recursive binning approach works by separating a dataset into two bins and proceeds by

further splitting bins. The recursive procedure continues if (1) the predicted abundance values of two bins differ significantly; (2) the predicted genome sizes are larger than a certain threshold; and (3) the number of reads associated with each bin is larger than a certain threshold proportion of the total number of reads classified in the parent bin.

AbundanceBin achieves accurate classification of even very short sequences sampled from species with different abundance levels, as tested on simulated and real metagenomic datasets. The software is available for download at <http://omics.informatics.indiana.edu/AbundanceBin>.

Integrated Binning Methods

MetaCluster 3.0 is an integrated binning method based on the unsupervised top–down separation and bottom–up merging strategy, which can bin metagenomic fragments of species with very balanced abundance ratios to very different abundance ratios (Leung et al. 2011). MetaCluster 4.0 further improves the binning algorithm and is able to handle datasets with large number of species (e.g., 100 species) (Wang et al. 2012). MetaCluster is available for download at <http://i.cs.hku.hk/~alse/MetaCluster/>.

Joint Analysis of Multiple Metagenomic Samples

Baran and Halperin proposed an abundance-based (also termed as coverage-based) binning algorithm (MultBin) that operates on multiple samples of the same environment simultaneously, assuming that the different samples contain the same microbial species, possibly in different proportions (Baran and Halperin 2012). MultBin employs a k -medoids clustering algorithm to cluster reads according to their coverage across the samples. Testing of MultBin on simulated metagenomic datasets shows that integrating information across multiple samples yields more precise binning on each of the samples.

Summary

Abundance-based (or coverage-based) binning approaches achieve an accurate performance even for extremely short reads – when there exist species abundance differences, an ability that cannot be achieved by composition-based approaches which suffer from the variances of the compositions of short reads. Approaches that integrate abundance and composition information and approaches that utilize multiple samples have shown promising binning results.

References

- Baran Y, Halperin E. Joint analysis of multiple metagenomic samples. *PLoS Comput Biol.* 2012;8(2):e1002373.
- Cole JR, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 2009;37(Database issue):D141–5.
- Diaz NN, Krause L, Goesmann A, et al. TACO: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics.* 2009;10:56.
- Gori F, Folino G, Jetten MS, et al. MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics.* 2011;27(2):196–203.
- Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol.* 2012;856:415–29.
- Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17(3):377–86.
- Krause L, Diaz NN, Goesmann A, et al. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* 2008;36(7):2230–9.
- Leung HC, Yiu SM, Yang B, et al. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics.* 2011;27(11):1489–95.
- Liu B, Gibbons T, Ghodsi M, et al. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics.* 2011;12 Suppl 2:S4.
- Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics.* 2011;27(1):127–9.
- Stark M, Berger SA, Stamatakis A, et al. MLTreeMap—accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics.* 2010;11:461.
- Teeling H, Waldmann J, Lombardot T, et al. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics.* 2004;5:163.
- Wang Y, Leung HC, Yiu SM, et al. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *J Comput Biol.* 2012;19(2):241–9.
- Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using 1-tuples. *J Comput Biol.* 2011;18(3):523–34.

Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads

Fengzhu Sun and Li Charlie Xia

Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Synonyms

Genome Relative Abundance estimation using Mixture Model theory (GRAMMy)

Introduction

Accurate estimation of microbial community composition based on metagenomic sequencing data is fundamental for subsequent metagenomic analysis. However, it is also a challenging computational problem because of the mixed nature of metagenomes and the fact that only a small fraction of them get sequenced.

With the advents of next-generation sequencing (NGS) technologies, there has been significant increase in sequencing capacity yet reduction in single read length. This paradigm shift in sequencing technologies has impacted downstream analyses. Specifically, the identification of the origin of a read becomes more difficult for several reasons. First, a large number of short reads cannot be uniquely mapped to a specific location of one genome. Instead, they map to multiple locations of one or multiple genomes.

These ambiguities are directly associated with the read length reduction in NGS technologies. Second, communities usually consist of many microbes with similar genomes, different only in some parts, making it indeed impossible to determine the origin of a particular short read based solely on its sequence.

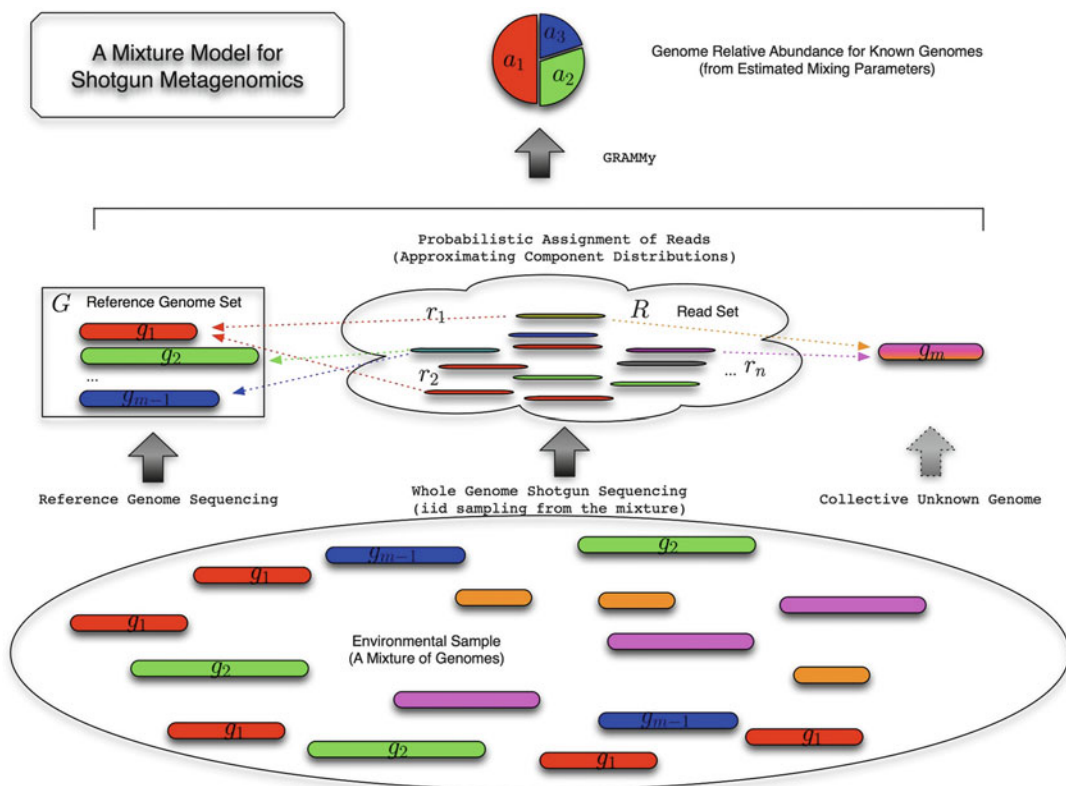
Despite these difficulties, NGS read sets have brought in richer abundance information of microbial communities than traditional datasets because of the significant increase in the number of reads. Along with the increase of read set size, efforts to assemble more reference genomes are ongoing. In addition, new experimental techniques, such as single-cell sequencing approaches, are being developed to sequence reference genomes directly from environmental samples. In face of the challenges from short reads and the opportunities from fast-expanding reference genome databases,

GRAMMy is a statistical framework developed to accurately and efficiently estimate the relative abundance of microbial organisms within the community (Xia et al. 2011).

Description

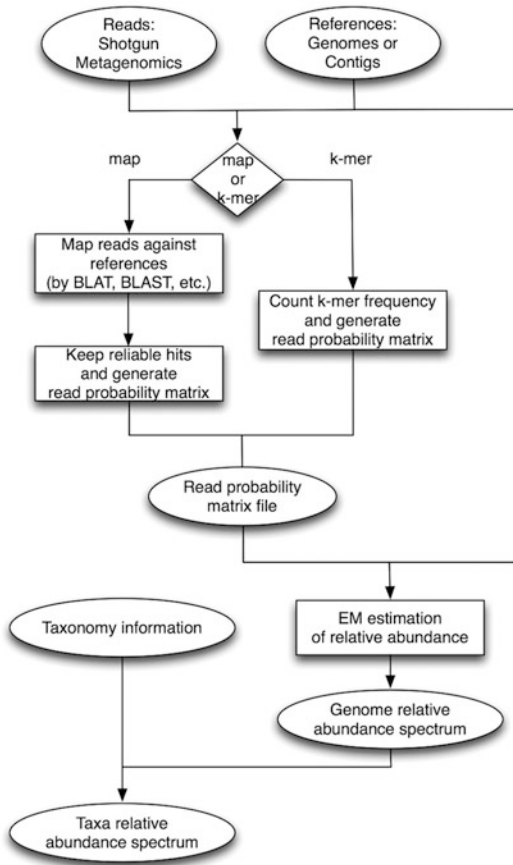
The GRAMMy Framework

The GRAMMy framework is based on a mixture model for the short metagenomic sequencing and an expectation-maximization (EM) algorithm, as outlined in the model schema and the analysis flowchart in Figs. 1 and 2. GRAMMy accepts a set of shotgun reads as well as external references (e.g., genomes, scaffolds, or contigs) as inputs and subsequently performs the maximum-likelihood estimation (MLE) of the genome relative abundance (GRA) levels.



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 1 The GRAMMy model. A schematic diagram of the finite

mixture model underlies the GRAMMy framework for shotgun metagenomics. In the figure, "iid" stands for "independent identically distributed"



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 2 The GRAMMy flowchart. A typical flowchart of GRAMMy analysis pipeline employs “map” and “k-mer” assignment

In the typical GRAMMy workflow, which is shown in Fig. 2, the end user starts with the metagenomic read set and reference genome set and then chooses between mapping-based (“map”) and k-mer composition-based (“k-mer”) assignment options (He and Xia 2007). In either option, after the assignment procedure, an intermediate matrix describing the probability that each read is assigned to one of the reference genomes is produced. This matrix, along with the read set and reference genome set, is fed forward to the EM algorithm module for estimation of the GRA levels. After the calculation, GRAMMy outputs the GRA estimates as a numerical vector, as well as the log-likelihood and standard errors for the

estimates. If the taxonomy information for the input reference genomes is available, strain (genome) level GRA estimates can be combined to calculate high taxonomic level abundance, such as species- and genus-level estimates.

Accurate GRAMMy Estimates with EM Algorithm

Formally, GRA is defined as the normalized abundance for m unique genomes, where the relative abundance for the j th known genome is

$$a_j = \frac{\text{\#}j\text{-th genome}}{\text{\#}known\ genomes}$$

Note that g_m is a collective surrogate for unknown genomes and cannot be estimated in the model. Knowing length l_j , a_j is one-to-one related to the corresponding mixing parameter π_j by

$$a_j = \frac{\pi_j}{l_j \sum_{k=1}^{m-1} \frac{\pi_k}{l_k}}$$

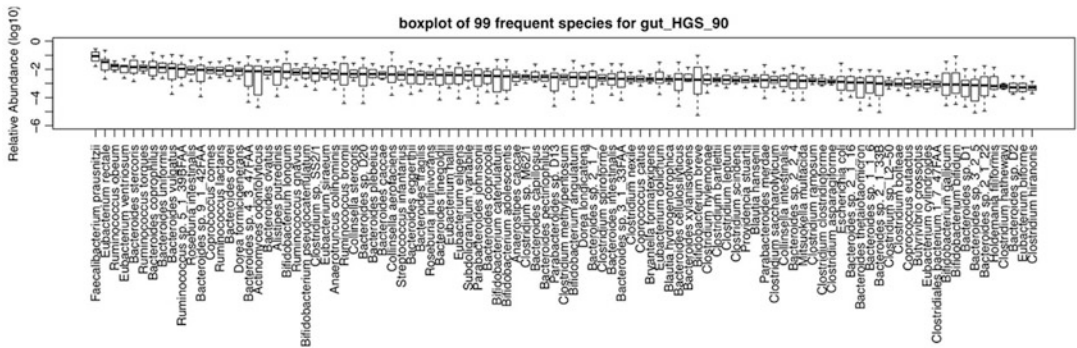
Mixing component distributions are needed to solve for mixing parameter π , which are $p(r_i|z_{ij} = 1; \mathbf{g})$ ’s – i.e., the probabilities of generating a read r_i from g_j . They are approximated empirically. The first approach is to use the number of high-quality hits s_{ij} from BLAST, BLAT, or other mapping tools and approximate by $\frac{s_{ij}}{l_j}$; the second approach is to use k-mer composition as detailed in the original study (Xia et al. 2011). The EM algorithm to calculate π iterates between E-step

$$z_{ij}^{(t)} = \frac{p(r_i|z_{ij} = 1; \mathbf{g})\pi_j^{(t)}}{\sum_{k=1}^m p(r_i|z_{ik} = 1; \mathbf{g})\pi_k^{(t)}}$$

and M-step

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n z_{ij}^{(t)}}{n}$$

until convergence, where n is the total number of reads and z_{ij} ’s are entries in the missing read



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 3 Frequent species of human gut microbiome. The

99 species occurring in at least 50 % of the 33 human gut samples with a minimum relative abundance of 0.05 % were selected

origin matrix Z . The estimated mixing parameters π are then converted back to GRA estimates \mathbf{a} .

GRAMMy Estimates for Human Gut Metagenomes

The human gastrointestinal tract harbors the largest group of human symbiotic microbes. Figure 3 shows the 99 most frequent species of human gut based on the GRAMMy analysis of the 33 metagenomic samples collected from three human gut metagenome projects (Gill et al. 2006; Kurokawa et al. 2007; Turnbaugh et al. 2009). The medians of estimated average genome lengths for these metagenomes range from 2.8 to 3.7 Mbp. Among the top ten most frequent species, there are eight from the *Firmicutes* phylum including members of *Faecalibacterium*, *Eubacterium*, and *Ruminococcus* genera, and two from the *Bacteroides* genus of the *Bacteroidetes* phylum. *Firmicutes* and *Bacteroidetes* dominate the human gastrointestinal tract. Species' relative abundance displays a long-tail distribution, suggesting that many are detected across samples, though most of them are not highly abundant. The abundance levels of some species are highly variable (with larger box size), while most others are relatively constant.

Conclusions

GRAMMy is a rigorous probabilistic framework for accurately and efficiently estimating genome relative abundance (GRA) based on shotgun metagenomic reads. Users have a wide choice of mapping and alignment tools to assign reads to references. The method is particularly suitable for NGS short read datasets due to its better handling of read assignment ambiguities. GRAMMy tools are packaged as a C++ extension to Python, which can be downloaded freely from GRAMMy's homepage: <http://meta.usc.edu/softs/grammy>.

Cross-References

- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Extended Local Similarity Analysis \(eLSA\) of Biological Data](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)
- ▶ [Molecular Ecological Network of Microbial Communities](#)

References

- Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. *Science*. 2006;312(5778):1355–9.
- He PA, Xia L. Oligonucleotide profiling for discriminating bacteria in bacterial communities. *Comb Chem High Throughput Screen*. 2007;10(4):247–55.
- Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*. 2007;14(4):169–81.
- Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457(7228):480–4.
- Xia LC, Cram JA, Chen T, et al. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*. 2011;6(12):e27992.

All-Species Living Tree Project

Pablo Yarza¹, Raul Munoz², Jean Euzéby³, Wolfgang Ludwig⁴, Karl-Heinz Schleifer⁴, Rudolf Amann⁵, Frank Oliver Glöckner^{6,7} and Ramon Rosselló-Móra²

¹Ribocon GmbH., Bremen, Germany

²Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Illes Balears, Spain

³Society of Systematic Bacteriology and Veterinary (SBSV) & National Veterinary School de Toulouse (ENVT), Toulouse, France

⁴Lehrstuhl Für Mikrobiologie, Technische Universität München, Freising, Germany

⁵Molecular Ecology Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

⁶Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

⁷Jacobs University Bremen gGmbH, Bremen, Germany

Synonyms

16SrRNA(SSU) and 23SrRNA(LSU) gene sequence databases; Alignments; LTP project; Manual curation; “Orphan” species; Taxa boundaries; Taxonomy/classification/phylogeny of Bacteria and Archaea; Type strains

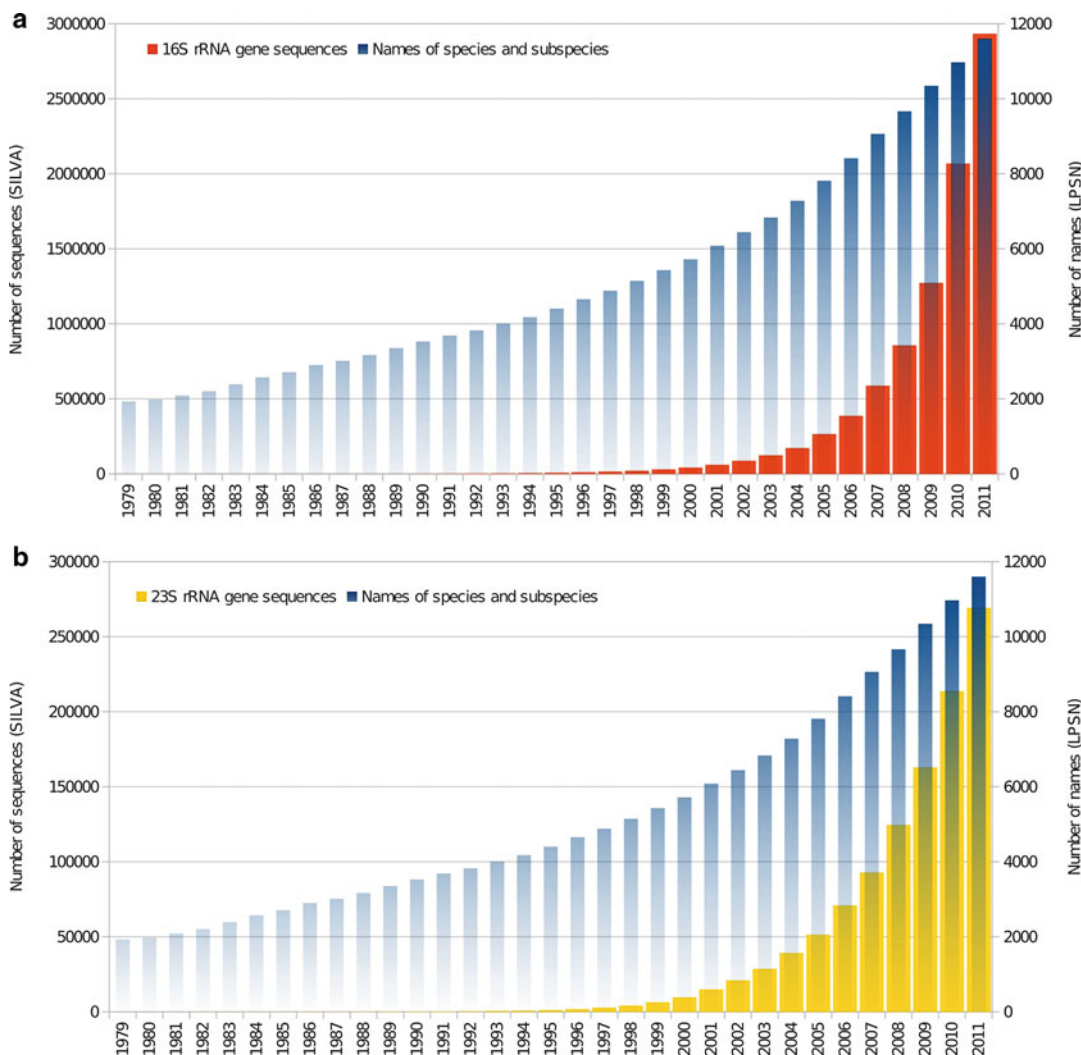
Definition

The All-Species Living Tree Project (LTP) is an international initiative for the creation and maintenance of highly curated 16SrRNA and 23SrRNA gene sequence databases, alignments, and phylogenetic trees for all the type strains of *Bacteria* and *Archaea*.

Introduction

Classification and identification of *Bacteria* and *Archaea* came across to a turning point around 35 years ago. It was the time when Carl Woese and co-workers demonstrated that ribosomal markers were appropriate to infer genealogical relationships by means of phylogenetic reconstructions (Fox et al. 1977). Rapidly, comparative analysis of rRNA gene sequences became a standard procedure with mature implications in microbial ecology and taxonomy: culture-independent exploration of ecosystems' diversity (Amann et al. 1995) and settlement of the phylogenetic backbone (i.e., our current accepted classification of *Bacteria* and *Archaea*; Garrity 2001). As a result, the total amount of ribosomal RNA entries in the public DNA databases has grown exponentially since early 1990s, currently comprising at least 3,500,000 small (SSU) and 300,000 large (LSU) ribosomal subunit gene sequence entries. On the other hand, the number of bacterial and archaeal species with validly published names has followed arithmetic trends with a ratio of around 500–700 annual descriptions during the last 7 years (Fig. 1), currently (December 2012) exceeding the total number of 10,300 species and subspecies. A comparative overview of these trends until December 2011 is shown in Fig. 1.

As from early 1990s, the 16S rRNA has been, by orders of magnitude, the most often sequenced gene, there is no alternative phylogenetic marker with such a high coverage in public repositories. However, abundance is not the single requisite for a proper phylogenetic inference and other single molecules (e.g., 23S rRNA) or combinations of them might perform better at reflecting



All-Species Living Tree Project, Fig. 1 Annual growth of ribosomal 16S rRNA (a) and 23S rRNA (b) gene sequence databases and species and subspecies names with standing in nomenclature until December 2011. SILVA SSU-Parc111 and LSU-Parc111 databases (<http://www.arb-silva.de/documentation/release-111/>) were filtered by submission date until December 2011 and its cumulative annual growth was plotted in red (SSU, 1A)

and yellow bars (LSU, 1B). The cumulative growth of published species and subspecies names (according to LPSN; <http://www.bacterio.cict.fr/number.html>) since 1980 until December 2011 is plotted in blue. Note that the total number of names is around 2,000 above the total number of distinct type strains due to homotypic synonyms, new combinations, nomina nova, later heterotypic synonyms, or illegitimate names

genealogies of certain groups given the higher information content (Ludwig and Klenk 2001). Although far from reaching 16S rRNA levels, submission of alternative markers is growing fast, mostly because (i) the number of metagenomes and complete genomes is growing exponentially due to the reduction on sequencing and analysis costs and (ii) the recent initiative to

complete the genome sequence of all type strains (GEBA initiative). Undoubtedly, comparative genomics will involve a new breakthrough for microbial taxonomy and the current phylogenetic backbone based on ribosomal sequences will be carefully reviewed (Coenye et al. 2005). Nevertheless, at this point, the number of sequenced genomes of type strains is still low and therefore

the current possibilities for an in-depth taxonomic study are sparse.

The responsible teams of the ARB, SILVA, and LPSN projects (www.arb-home.de, www.arb-silva.de, and www.bacterio.net) together with the journal Systematic and Applied Microbiology (SAM) started the “All-Species Living Tree Project” (LTP; <http://www.arb-silva.de/projects/living-tree>), a project conceived to provide a tool especially designed for the microbial taxonomist scientific community (Yarza et al. 2008). The main objectives considered so far are (1) provide a curated 16S and 23S rRNA gene database for the type strains of all species with validly published names; (2) set up an optimized and universally usable alignment; (3) reconstruct reliable phylogenetic trees with all the type strains; (4) maintain the database, alignments, and trees through regular updates including the new validly published taxa and their respective 16S and 23S rRNA gene sequences; and (5) investigate, with the use of the database, fundamental aspects about taxonomy of *Bacteria* and *Archaea* such as phylogenetic thresholds in new taxa circumscriptions, coherence of current taxonomy by means of phylogenetic schemes, and relevance of the ribosomal RNA genes in taxonomic studies.

Creation and Maintenance of LTP Releases

LTP Datasets

First LTP datasets (release LTPs93 for SSU (Yarza et al. 2008), release LTPs102 for LSU (Yarza et al. 2010)) were prepared following six main steps:

1. Set up a list of candidate sequences. An initial sequence dataset consisted on a subsample of the SILVA database, filtering by “type” (T) or “cultured” (C) strains; this information mainly came from StrainInfo.
2. Set up a list of species names. In parallel we built a comprehensive, updated, and nonredundant (i.e., free of synonyms and according to latest valid nomenclature) list of validly published species and subspecies

names from LPSN. When a species is divided into subspecies, we substituted the original species name by that of the subspecies (e.g., *Staphylococcus sciuri* subsp. *sciuri* instead of *Staphylococcus sciuri*). We avoided the “Candidatus” names (e.g., “*Candidatus Aciduliprofundum boonei*”), *Cyanobacteria* not validly published under the Bacteriological Code (e.g., *Anabaena oscillatorioides*), and later heterotypic synonyms (e.g., *Pseudomonas chloritidismutans*).

3. Manual cross-check. Then, each entry from our initial list of sequences was assigned to a species name by manually examining the companion contextual metadata. This process had to be done manually given the often outdated, mistaken, or absent taxonomic information such as the organism name or the strain numbers.
4. Quest for missing type strains. We realized that not all species names were represented in the list of sequences. Then, we inverted the process by searching in resources like EMBL, Bergey’s Outlines, issues of the International Journal of Systematic and Evolutionary Microbiology (IJSEM), etc. with the aim to find a good-quality sequence entry for each missing type strain.
5. “Orphan” species recognition. Finally, we got a group of type strains whose 16S/23S rRNA genes had never been sequenced or that the existing sequences were of too low quality to be considered for the project (i.e., in terms of sequence length, number of ambiguities, etc.). We called them “orphan” species. The LTP project together with eleven international culture collections has driven the sequencing of these “orphan” species through the SOS initiative (Yarza et al. 2013).
6. Keep one sequence per species. On the other hand, the list of type-strain sequences was redundant in the sense that one single type strain could be represented by multiple sequence entries. This is the case of multiple independent sequencings and submissions, or the existence of several sequences due to multiple copies of the ribosomal operon. The aim of the LTP is, whenever possible, to keep one

All-Species Living Tree Project, Table 1 Summary of LTP releases. “Sync” fields correspond to IJSEM and EMBL release dates. “Net increase” of a release is the number of new entries minus the number of deleted entries. “% incorrect” refers to the percentage of new entries whose INSDC records carried incorrect information in the organism name field. Averages include standard deviation

Release	Type	IJSEM sync	EMBL sync	Total entries	New entries	Deleted entries	Net increase	% incorrect ^a	Average length ^a	Average ambig. ^b
LTPs93	SSU	Dec. 2007	Dec. 2007	6,728	6,728	0	6,728	22	1,465.0 ± 51.2	0.10 ± 0.26
LTPs95	SSU	Jun. 2008	Jun. 2008	7,006	299	21	278	45	1,446.0 ± 46.3	0.04 ± 0.11
LTPs100	SSU	Aug. 2009	Jun. 2009	7,710	750	46	704	50	1,448.0 ± 54.2	0.03 ± 0.11
LTPs102	SSU	Feb. 2010	Nov. 2009	8,029	363	44	319	58	1,453.6 ± 52	0.33 ± 0.12
LTPs102	LSU	Feb. 2010	Nov. 2009	792	792	0	792	6	2,866.1 ± 177.6	0.02 ± 0.11
LTPs104	SSU	Dec. 2010	May 2010	8,545	545	29	516	74	1,444.6 ± 62	0.27 ± 0.11
LTPs106	SSU	May 2011	Dec. 2010	8,815	279	9	270	77	1,445.9 ± 51.1	0.03 ± 0.12
LTPs108	SSU	Dec. 2011	Jun. 2011	9,279	490	26	464	60	1,455.4 ± 51.9	0.02 ± 0.09

^aAverage length for the “new entries”

^bAverage percentage of ambiguities for the “total entries”

sequence per type strain in order to maintain simplicity, avoid confusion, and improve tree navigation and database usability. In general, the best quality available (including manual inspection of the alignment) was selected for the project and, in case of doubt, the earliest submission to an INSDC partner (www.insdc.org). From release LTPs102 (Yarza et al. 2010), when multiple paralogues exist due to rRNA operon copy number, several copies are kept if they show less than 98 % sequence identity (see below for further details).

LTP is maintained by a scrutiny of the new described species, nomenclatural changes, taxonomic notes, and opinions that are monthly published in the IJSEM journal. Their respective 16S and 23S rRNA gene sequence entries are acquired from the latest SILVA release and appended to the existing LTP database. Therefore, SILVA’s Reference (Ref) ARB databases (<http://www.springerreference.com/docs/html/chapterdbid/304116.html>) serve as template for the new LTP-ARB databases. Until now (December 2012) one LSU-based and seven SSU-based LTP releases have been produced (Table 1). New species are incorporated into the database only if they account a good-quality sequence existing in the respective SILVA release. Certain entries can be deleted

if their corresponding species names are seen to be later heterotypic synonyms, if they become rejected, or as a matter of taxonomic opinions. Sequence entries existing in an LTP database can also change by means of their metadata. Thus, for example, new combinations (i.e., a type strain which changes its name due to reclassification) or subdivision of a species into subspecies leads to an entry modification at its taxonomic information fields.

Inaccurate or Mistaken Metadata

Inaccurate sequence-associated metadata tend to happen in more than 50 % of the new added 16S rRNA entries (Table 1). Often, these “mistakes” consist on a lack of entries’ updating tasks at the time of their first appearance in a scientific publication. It mainly occurs in taxonomy-associated information fields. To prove the uniqueness of a new species and to name it take time and, in the meanwhile, sequences are quickly produced and easily submitted to nucleotide databases. Most often, these submissions only show genus specifications, for example, sequence entry GU808562 appears as “*Hymenobacter* sp. HMD1010” but its real name is *Hymenobacter yonginensis*. Indeed, a Bacteriological Code-compliant (Lapage et al. 1992) nomenclature may be somewhat tricky and is frequent to

consider several Latin terms and derivations until one species name is finally accepted by authors and reviewers. Unavoidably, this bad-quality information is propagated from INSDC databases (primary sources) to other technological services like dedicated ribosomal databases (e.g., SILVA). Although extensive data curation is not a task of primary sources of information, it would be very beneficial that authors enhance their commitment with the correctness of the metadata provided (e.g., like the species name) or that authors are forced to update their INSDC entries prior to manuscript acceptance (recommended action for scientific journals). Successively, this rough data arrives finally to resources like LTP, which have no choice but checking it carefully to provide new informational fields with corrected information; curated information can return back to other resources of information.

Multiple Copies of the Ribosomal Operon

In 2010, a comprehensive study was conducted to evaluate the intra-genomic variability of the 16S rRNA gene on complete type-strain genomes (Yarza et al. 2010). We observed that in very unusual exceptions, the intra-genus (94.5 %; Yarza et al. 2008) or intraspecies (98.7 %; Stackebrandt and Ebers 2006) boundaries could be exceeded within a single genome. In such cases, the selection of one or another sequence might seriously affect the interpretation of a phylogenetic inference. However, despite the fact that the vast majority of strains contain multiple copies of the *rrn* operon, only 2 % of them reveal divergences beyond 2 % (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not affect the phylogenetic reconstructions. Consequently, starting from release s104 (Munoz et al. 2011), the LTP database includes all paralogues with higher divergences than 2 %. By now, it is the case of three species: *Haloarcula marismortui* ATCC 43049^T, accession number AY596297, with 5.7 % of maximum inter-operonic divergence; *Thermoanaerobacter pseudethanolicus* ATCC 33223^T, accession number CP000924, with 3.66 % of maximum inter-operonic divergence; and *Desulfitobacterium hafniense*

DCB-2^T, accession number CP001336, with 4.34 % of maximum inter-operonic divergence.

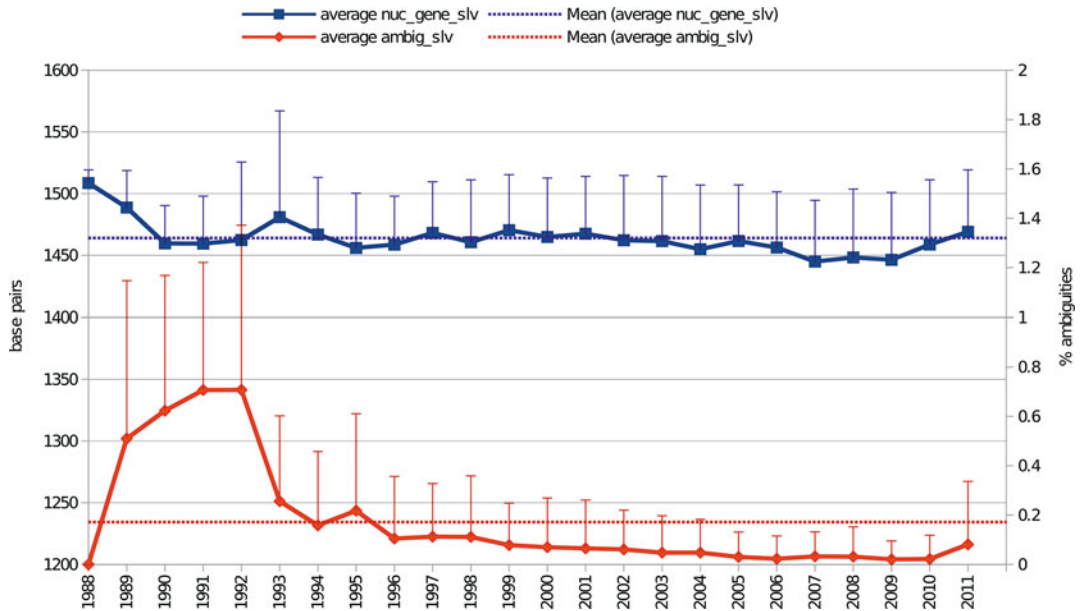
Sequence Quality in LTP Datasets

It has been suggested that sequences produced for taxonomic purposes should be equal or larger than 1,450 bases with less than 0.5 % ambiguities (Stackebrandt et al. 2002). Reason is that informative content of a molecular clock is linked to the total number of its variable positions (Ludwig and Klenk 2001). Statistics derived from LTP datasets indicate that in general, sequence quality is acceptable for in-depth phylogenetic studies (~1,455 bases and 0.02 % ambiguities for LTPs108; Table 1). Figure 2 shows annual variation of gene sequence length and percentage of ambiguities. Quality increase is mainly observed in terms of ambiguities reduction, probably related to amelioration of sequencing techniques. In any case, the completion of more full genome sequences of type strains will substantially increase the sequence quality (indicated by these two parameters) in the LTP database. Researchers should be encouraged to complete 5' ends of 16S rRNA gene sequences, as first 250 bases contain hypervariable regions V1 and V2 which play an important role in comparisons between highly related organisms (Chakravorty et al. 2007).

Curated Metadata Introduced by the LTP

In addition to regular fields provided by the ARB-SILVA databases, sequence entries include now the following LTP-specific metadata fields:

1. *fullname_ltp*: corrected species name according to LPSN (<http://www.bacterio.net>).
2. *rel_ltp*: name of the LTP release where a sequence entry appeared for the first time.
3. *hi_tax_ltp*: name of the family where the taxon is classified. For unclassified genera, the name of the next available higher taxon above genus (e.g., “*Acidobacteria*” for *Bryobacter aggregatus*).
4. *type_ltp*: type species receive the label “type sp.” in this field.
5. *riskgroup_ltp*: risk-group classification of microorganisms risk-group classification of microorganisms obtained from the DSMZ



All-Species Living Tree Project, Fig. 2 Annual distribution of the 16S rRNA gene sequence length and % of ambiguities in the 9,279 type-strain sequences corresponding to LTP release s108. Gene sequence length

is given by the SILVA parameter “nuc_gene_slv” which cuts off the bases at the extremes when beyond the *E.coli*'s 16S rRNA gene limits. Percentage of ambiguities is given by the SILVA descriptor “ambig_slv”

(Deutsche Sammlung von Mikroorganismen und Zellkulturen), according to the Federal Institute for Occupational Safety and Health (BAuA) in Germany.

6. *tax_ltp*: taxonomic classification into higher taxonomic ranks according to LPSN (<http://www.bacterio.cict.fr/classifphyla.html>).
7. *url_lpsn_ltp*: it contains the variable part of the URL leading to the LPSN's species file (e.g., <http://www.bacterio.net/bryobacter.html>).

Alignments and Phylogenetic Trees

Setting up universal alignments is a key step in order to achieve optimal and comparable phylogenetic reconstructions. It has been one of the constant motivations of Wolfgang Ludwig and co-workers who dealt with the huge task of preparing common and reliable alignment of ribosomal SSU and LSU sequences of *Bacteria*, *Archaea*, and *Eukarya* (Ludwig and Schleifer 1994). They found out that secondary structure formations such as loops and helices occurred at the same relative positions along the molecule. This helped to refine the alignments because

variable stretches, with low sequence similarities, could be optimally positioned by recognizing functional homology (due to evolutionary pressure) and functional stability of helices (due to chemical stability of base pairs' bounds). A core dataset of sequences with highly curated alignments was incorporated into the SILVA system so new added sequences can be automatically aligned using this “seed alignment” as a reference (Ludwig et al. 2004; Pruesse et al. 2007). Periodically more and more manually curated sequences are added into the seed which improves its quality over time.

Although all new sequences incorporated into the LTP come from an ARB-SILVA database, they are again manually revised to further correct misplaced bases and to check highly variable regions. Before tree calculation, the complete alignment is shifted using maximum frequency filters (Table 2) that remove dubious orthologous positions caused by sequencing errors and hypervariability. Typically, LTP phylogenetic trees are calculated using the 40 % maximum frequency filter.

All-Species Living Tree Project, Table 2 Maximum frequency filters implemented into the LTPs 108ARB database

Filter name	Start position	Stop position	%min ^a	%max ^a	No. of positions ^b
LTPs108_ssu_10	0	50,000	10	100	1,433
LTPs108_ssu_20	0	50,000	20	100	1,433
LTPs108_ssu_30	0	50,000	30	100	1,432
LTPs108_ssu_40	0	50,000	40	100	1,390
LTPs108_ssu_50	0	50,000	50	100	1,288

^aMinimum and maximum sequence identity. For tree reconstructions, only columns are taken into account if they have a positional conservation above the respective minimum values

^bNumber of homologous positions (columns) taken into account for tree reconstructions

The first 16S rRNA-based phylogenetic tree was calculated for the release LTPs93 (Yarza et al. 2008). The sequence dataset consisted of 6,728 type-strain sequences plus 3,247 supporting sequences belonging to non-type strains used to reinforce underrepresented groups and to stabilize the topology. The multiple alignment of 9,975 16S rRNA gene sequences was submitted to different treeing methodologies including neighbor-joining, maximum likelihood, and maximum parsimony, all tested with several filters (30 %, 40 %, and 50 % maximum frequency filters) and all implemented in the ARB software package (Ludwig et al. 2004). A high degree of congruence was observed among them. The tree considered as optimal was a 40 %-filtered maximum likelihood reconstruction calculated using the RAxML algorithm (Stamatakis 2006), with the GTRGAMMA correction, with 100 bootstrap replicates, in a 5-node and 20-processor parallel environment. The last de novo phylogenetic reconstruction appears in the release LTPs108 and was similarly calculated; tree calculation was run with a dataset of 12,166 16S rRNA gene sequences.

The phylogenetic tree calculated using the 23S rRNA gene was particularly challenging due to data shortage in many groups. In order to set up a reliable phylogeny based on 23S rRNA data, we defined a core dataset made of high-quality sequences (type and non-type strains). The stringent quality filtering approach ended with around 2,000 high-quality and nonredundant LSU sequences. This dataset was submitted to a maximum likelihood reconstruction in combination with a 50 % maximum frequency filter allowing 2,463 positions of the entire alignment.

The missing partial or lower-quality type-strain sequences were added to the tree using the ARB parsimony tool with the option for keeping the initial topology while inserting additional data.

The groups shown in the trees are defined by recognizing the type members and according to the taxonomic classification. The trees are carefully compared against previously reported topologies and current taxonomic classifications (Yarza et al. 2010). All the additional supporting sequences used to reconstruct the phylogeny are removed from the final tree by keeping its topology intact. Within the ARB database, the type species are labeled with a distinct color for easy recognition and tree handling.

Files Provided by the LTP

As a taxonomic tool, the LTP must be understood as a collection of reference materials, all publicly available at the project's Web page (<http://www.arb-silva.de/projects/living-tree>), including:

1. Release documentation: (I) readme file with a release description and (II) PDF document describing the metadata fields introduced by the LTP
2. Tables: (I) new entries with outdated submission names and (II) list of changes in the dataset: added/deleted/modified entries
3. Export filter: ARB-export filter (.eft format) to extract data from LTP-ARB databases
4. Databases: (I) complete ARB databases including sequences, alignments, metadata, filters, and trees and (II) datasets in CSV format including LTP metadata

5. Alignments: (I) gapped exports in multi-FASTA format and (II) compressed exports in multi-FASTA format
6. Phylogenetic trees: (I) collapsed overviews in PDF format showing the distinct phyla, (II) full SSU (more than 80 pages long) and LSU trees in PDF format, and (III) full trees in NEWICK format, including group names and branch lengths

Side Research

Sequencing the Orphan Species Initiative (SOS)

The understanding that around 6 % of all classified species were missing from the ribosomal SSU sequence catalogues motivated us to start the “Sequencing the Orphan Species” (SOS) initiative (Yarza et al. 2013). During 3 years of work, the LTP team coordinated a network of 12 partner researchers and culture collections (ATCC, BZF, CECT, CIP, CCUG, DSMZ, JCM, ICMP, BCCM/LMG, MMG, NBRC, NCCB) in order to improve this situation by (re)sequencing the 16S rRNA gene of the “orphan” species. As a result, 351 type strains appear represented now by a good-quality SSU gene sequence in the databases. They comprise representatives of 14 bacterial and archaeal phyla, 76 type species, and 78 pathogenic species. However, 201 type strains could not be accessed as cultivable strains were not available at recognized culture collections. They represent 10 phyla and 17 type species.

Taxonomic Boundaries

In order to understand how the higher taxonomic categories could be circumscribed by means of a sequence identity threshold, we performed a statistical procedure to get the lowest similarity found within the members of a certain taxon (Yarza et al. 2008, 2010). By taking into account all the taxa at a particular taxonomic rank, we obtained general lower cutoff values of sequence identity for genus, family, and phylum based on 16S rRNA and 23S rRNA. In general, minimum 16S rRNA gene sequence identities of

94.9 % \pm 0.4, 87.5 % \pm 1.3, and 78.4 % \pm 2.0 lead to the circumscription of a new genus, family, and phylum, respectively. For 23S rRNA genes, these values are slightly different: 93.2 % \pm 1.3 (genus), 87.7 % \pm 2.5 (family), and 75.3 % (phylum). As shown by the low errors, historically used criteria for genera, families, and phyla are quite homogeneous and do not lead to unambiguous circumscriptions. These cutoffs should be used with caution and always as a complementary approach. They are especially recommended for prospective studies in clone libraries and as additional support for the circumscription of new taxa or emendation of existing ones.

Summary

SSU and LSU databases made by the All-Species Living Tree Project (LTP; <http://www.arb-silva.de/projects/living-tree>) provide high-quality nearly full-length sequences of the type strains of all *Archaea* and *Bacteria* with validly published names. Setting up a type-strain sequences database included the sieving of the public DNA databases whose sequence entries often appeared outdated or mistaken at their taxonomic metadata. It involved the initial manual cross-check of nearly 14,000 SSU and 6,000 LSU sequence entries against the catalogue of distinct species with validly published names retrieved from LPSN. Databases are complemented with manually curated metadata, manually curated alignments, and state-of-the-art phylogenetic reconstructions (in contrast to other similar resources like the EzTaxon (Santamaria et al. 2012)). The LTP team wants to remark that the aim of the project is not to reconstruct the currently described species genealogy with total fidelity but to provide a curated taxonomic tool for the scientific community. Our small but very representative SSU and LSU datasets may be used as a reference for identification and classification purposes in several fields of application, for example, facilitating the collection of sequences for the reconstruction of taxa genealogies (Cousin et al. 2012), enabling fast and

reliable taxonomic affiliations in rRNA surveys (Santamaria et al. 2012), or serving as reference datasets for testing bioinformatic procedures (Mizrahi-Man et al. 2013).

Cross-References

- ▶ [Culture Collections in the Study of Microbial Diversity, Importance](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [SILVA Databases](#)

References

- Amann R, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev.* 1995;59:143–69.
- Chakravorty S, Helb D, Burday M, et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods.* 2007;69:330–9.
- Coenye T, Gevers D, Van de Peer Y, et al. Towards a prokaryotic genomic taxonomy. *FEMS Microbiol Rev.* 2005;29:147–67.
- Cousin S, Gulat-Okalla ML, Motreff L, et al. *Lactobacillus gigeriorum* sp. nov., isolated from chicken crop. *Int J Syst Evol Microbiol.* 2012;62:330–4.
- Fox GE, Pechman KR, Woese CR. Comparative cataloguing of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. *Int J Bacteriol.* 1977;27:44–57.
- Garrity GM. *Bergey's manual of systematic bacteriology*. 2nd ed. New York: Springer; 2001.
- Lapage SP, Sneath PHA, Lessel EF, et al. *International code of nomenclature of bacteria (1990 revision)*. Washington, DC: American Society for Microbiology; 1992. p. 295.
- Ludwig W, Klenk HP. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, Garrity GM, editors. *Bergey's manual of systematic bacteriology*. 2nd ed. New York: Springer; 2001. p. 49–65.
- Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol Rev.* 1994;15:155–73.
- Ludwig W, Strunk O, Westram R, et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* 2004;32:1363–71.
- Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One.* 2013;8:e53608.
- Munoz R, Yarza P, Ludwig W, et al. Release LTPs104 of the all-species living tree. *Syst Appl Microbiol.* 2011;34:169–70.
- Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 2007;35:7188–96.
- Santamaria M, Fosso B, Consiglio A, et al. Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform.* 2012;13:682–95.
- Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today.* 2006;33:152–5.
- Stackebrandt E, Frederiksen W, Garrity GM, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol.* 2002;52:1043–7.
- Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006;22:2688–90.
- Yarza P, Richter M, Peplies J, et al. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol.* 2008;31:241–50.
- Yarza P, Ludwig W, Euzéby J, et al. Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. *Syst Appl Microbiol.* 2010;33:291–9.
- Yarza P, Spröer C, Swiderski J, et al. Sequencing Orphan Species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. *Syst Appl Microbiol.* 2013;36:69–73.

antiSMASH

Eriko Takano¹, Rainer Breitling¹ and Marnix H. Medema²

¹Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

²Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Definition

antiSMASH (Medema et al. 2011) is a web server and a stand-alone software to identify, annotate, and compare gene clusters that encode the biosynthesis of secondary metabolites in bacterial and fungal genomes. antiSMASH offers a wide

range of options to identify and analyze biosynthetic gene clusters, including protein domain analysis of the large multi-domain enzymatic assembly lines involved, prediction of core chemical structures of their end compounds, and multiple cluster alignments to a database of all currently sequenced gene clusters.

The antiSMASH web server can be found at <http://antismash.secondarymetabolites.org>.

Introduction

Microbial secondary metabolites are of great interest to society because of their diverse biological activities that are interesting starting points for drug development. Many of them are already used as antibiotics, antitumor agents, or cholesterol-lowering drugs (Hutchinson and McDaniel 2001; Fischbach and Walsh 2009). Automated computational identification of gene clusters in newly sequenced genomes is becoming a cornerstone of genome-based drug discovery, due to the affordability of sequencing large numbers of genomes from microorganisms that potentially produce novel secondary metabolites (Walsh and Fischbach 2010).

Functionalities

Gene Cluster Detection

antiSMASH detects a wide range of different types of biosynthetic gene clusters, including those encoding the pathways toward polyketides (PKs), nonribosomal peptides (NRPs), terpenoids, ribosomal peptides, aminoglycosides, and non-NRP siderophores. The detection is performed by screening the gene sequences from the input against a library of profile Hidden Markov Models (pHMMs) (Eddy 2011), each of which is specific for genes characteristic for a certain gene cluster type, and passing the results through a hierarchical logic filter. A second detection algorithm is also run, which detects genomic regions that are enriched in Pfam domains (Finn et al. 2010) linked to secondary metabolism.

Protein Domain Analysis of Polyketide Synthases and Nonribosomal Peptide Synthetases

PKs and NRPs are synthesized by large megasynthase enzymes containing a multitude of protein domains, such as condensation (C) and adenylation (A) and PCP-binding domains in nonribosomal peptide synthetases (NRPSs), ketosynthase (KS), and acyltransferase (AT) and ACP-binding domains in polyketide synthases (PKSs) (Fischbach and Walsh 2006). antiSMASH contains a library of pHMMs that can recognize all these protein domains as well as distinguish between various subtypes of these domains. In the antiSMASH output, the domain structures of any NRPSs or PKSs encoded in a gene cluster are visualized, and several downstream analysis options are provided for each domain (Fig. 1).

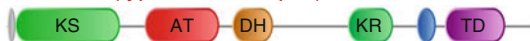
Core Chemical Structure Prediction

When a secondary metabolite biosynthesis gene cluster is detected, one of the key questions of course is what kind of chemical structure it produces. For NRPs and PKs, antiSMASH is able to already give a first approximation of the core chemical structure of the end compound (Fig. 2). To do so, it uses several substrate specificity prediction methods (Yadav et al. 2003; Minowa et al. 2007; Röttig et al. 2011) that are based on the amino acid sequences of the A domains of NRPSs and the AT domains of PKSs. To infer the sequential arrangement of the predicted substrates of the A/AT domains in the resulting polyketide or peptide, the order of the PKS enzymes in a multimodular assembly line is predicted using their estimated docking domain binding affinities (Yadav et al. 2009) or, alternatively, colinearity of the PKS or NRPS genes with their enzymes is assumed.

Comparative Analysis of Gene Clusters

In order to understand the architecture and function of a secondary metabolite biosynthesis gene cluster, much is gained by examining it within its evolutionary context through the

SCO6273 (type I modular pks)



SCO6274 (type I modular pks)

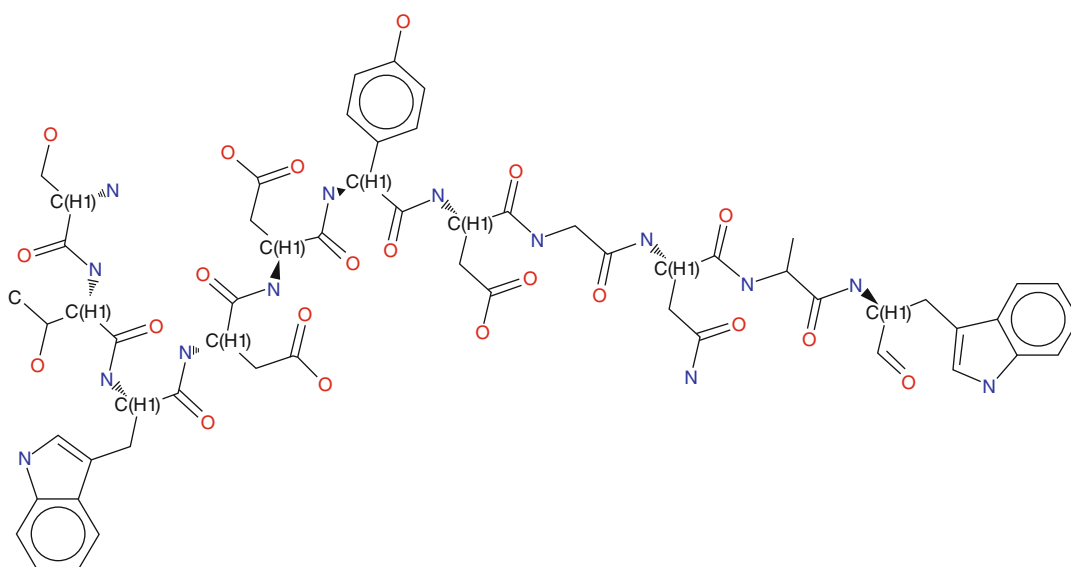


SCO6275 (type I modular pks)



antiSMASH, Fig. 1 Domain structure of multi-domain enzymes such as PKSs and NRPSs as visualized by antiSMASH, offering several options for analysis when

the mouse is positioned over a domain: one can, for example, run a BlastP search specifically with the sequence of this domain



antiSMASH, Fig. 2 Prediction of the core chemical structure of an NRP by antiSMASH. The residues are based on a consensus between three prediction methods

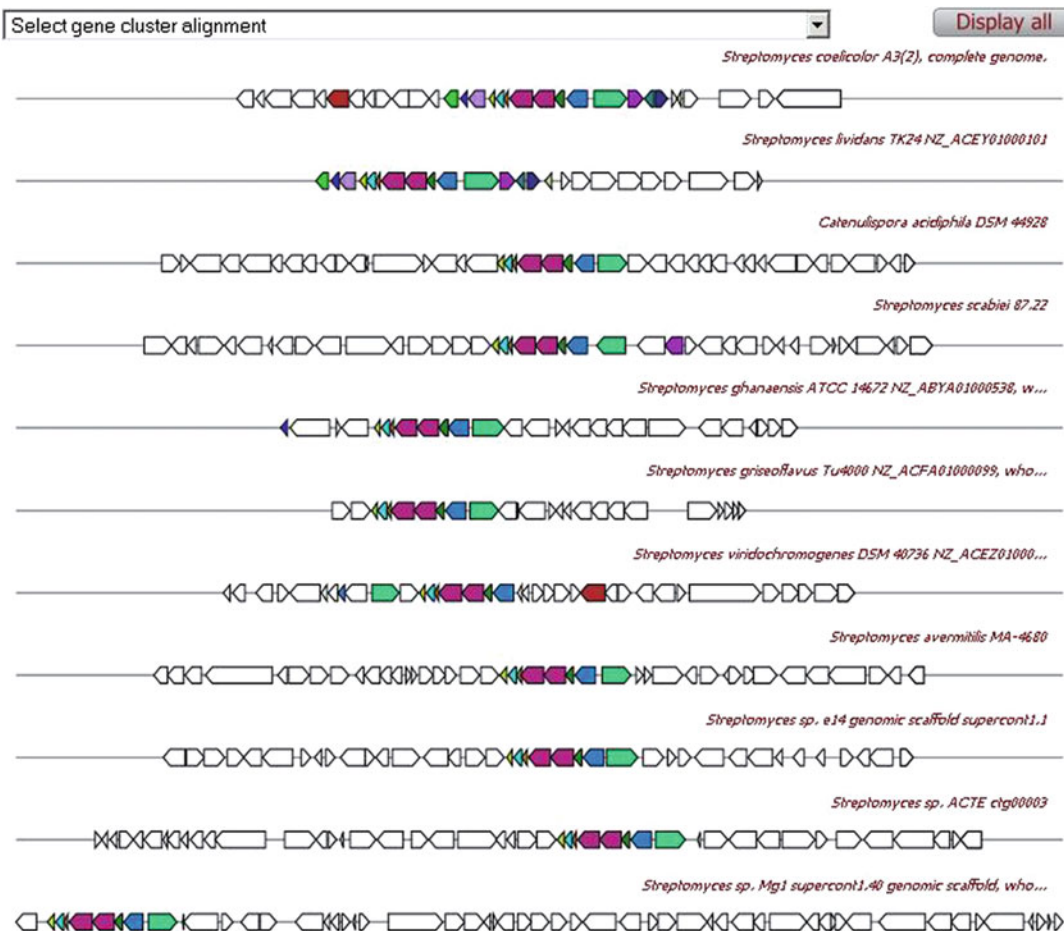
for the substrate specificities of the NRPS adenylation domains in the gene cluster

comparison with related gene clusters from species across the tree of life. To facilitate this, antiSMASH hosts a regularly updated database of gene clusters it has detected in all nucleotide sequences present in GenBank. antiSMASH then combines multiple BlastP runs into a comparative search of every identified gene cluster against all other known gene clusters. This is used to generate a multiple gene cluster alignment (Fig. 3), which can aid the biologist in assessment of the novelty of the gene cluster,

detecting the borders of the gene cluster and identifying the conserved multigene modules that constitute its building blocks.

Secondary Metabolism-Specific Gene Family Analysis

Most genes involved in the biosynthesis of secondary metabolite have (close) homologues with similar functions in other secondary metabolite biosynthesis gene clusters. This can be used to infer the functions of the genes



antiSMASH, Fig. 3 Example of a multiple gene cluster alignment by antiSMASH, showing identified homologue clusters of the query gene cluster

residing in the biosynthetic gene cluster based on sequence homology. antiSMASH simplifies this process by categorizing the genes of every identified gene cluster into secondary metabolism-specific gene families and automatically generating approximate phylogenetic trees of each gene in the context of its gene family.

Genome-Wide Pfam and Blast Analysis

Finally, antiSMASH also offers the possibility (transferred from CLUSEAN; Weber et al. 2009) to do a comprehensive analysis of all genes within a submitted genome, identifying

Pfam matches and running Blast for each gene against a database of all bacterial and fungal protein sequences.

Stand-Alone Version

Stand-alone versions of antiSMASH are available for download for Windows, Mac OS X, and Ubuntu Linux. Additionally, several related scripts are available from the antiSMASH website. An EMBL formatting script can be downloaded to format raw FASTA sequences together with a text file containing gene

annotations into an EMBL file that can be submitted to antiSMASH. Also, a script is available which allows running antiSMASH on multiple files, in batch mode.

Development

antiSMASH is still under active development. Some features projected for the next release are batch input on the web server, protein sequence input, and subclass prediction for enzyme classes like terpene synthases and trans-AT PKSs. Feature requests, bug reports, or other questions/suggestions can be sent to the development team via the online contact form on the antiSMASH website.

Related Tools

Several other software tools for the study of secondary metabolism have been published. For example, ClustScan (Starcevic et al. 2008) and NP.searcher (Li et al. 2009) can both be used to detect bacterial polyketide and NRP biosynthesis gene clusters. The same is the case for CLUSEAN (Weber et al. 2009), the pipeline which has now been integrated entirely into antiSMASH. For the analysis of fungal sequences, SMURF (Khaldi et al. 2010) offers a gene cluster detection potential similar to that of antiSMASH. Structural analysis of polyketide synthases can be performed with the SBSPKS suite (Anand et al. 2010). Finally, draft genomes with many small contigs and metagenomes with fragments too small for gene cluster detection can be scrutinized with NaPDoS (Ziemert et al. 2012) in order to find protein domains related to secondary metabolite biosynthesis and analyze these phylogenetically.

Summary

antiSMASH is an easy-to-use web server for the detection of secondary metabolite biosynthesis

gene clusters. Various functionalities – comparative, phylogenomic, enzymatic, etc. – are integrated in one single pipeline, making it straightforward for genomicists and natural product researchers to study the biosynthetic potential of any organism.

Cross-References

- ▶ [Bacteriocin Mining in Metagenomes](#)
- ▶ [CLUSEAN, Overview](#)
- ▶ [Mining Metagenomic Datasets for Antibiotic Resistance Genes](#)
- ▶ [Phylogenetics, Overview](#)

References

- Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D. SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* 2010;38:W487–96.
- Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7:e1002195.
- Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, et al. The Pfam protein families database. *Nucleic Acids Res.* 2010;38: D211–22.
- Fischbach MA, Walsh CT. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev.* 2006;106:3468–96.
- Fischbach MA, Walsh CT. Antibiotics for emerging pathogens. *Science.* 2009;325:1089–93.
- Hutchinson CR, McDaniel R. Combinatorial biosynthesis in microorganisms as a route to new antimicrobial, antitumor and neuroregenerative drugs. *Curr Opin Investig Drugs.* 2001;2:1681–90.
- Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol.* 2010;47:736–41.
- Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. *BMC Bioinformatics.* 2009;10:185.
- Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 2011;39:W339–46.

- Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J Mol Biol.* 2007;368:1500–17.
- Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* 2011;39:W362–7.
- Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res.* 2008;36:6882–92.
- Walsh CT, Fischbach MA. Natural products version 2.0: connecting genes to molecules. *J Am Chem Soc.* 2010;132:2469–93.
- Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol.* 2009;140:13–7.
- Yadav G, Gokhale RS, Mohanty D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J Mol Biol.* 2003;328:335–63.
- Yadav G, Gokhale RS, Mohanty D. Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput Biol.* 2009;5:e1000351.
- Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One.* 2012;7:e34064.

Approaches in Metagenome Research: Progress and Challenges

Heiko Nacke and Rolf Daniel
Institute of Microbiology and Genetics,
Georg-August-University of Göttingen,
Göttingen, Germany

Synonyms

Function-based screening, Metagenomic biomolecule, Metagenomic library, Metagenomics, Next-generation sequencing, Sequence-based screening

Definition

Metagenomics comprises the culture-independent and DNA-based analysis of entire microbial communities and complements cultivation-based analysis of microorganisms. Metagenomic approaches allow comprehensive insights into phylogenetic and functional diversity of complex microbial consortia present in moderate as well as extreme environments on Earth. The introduction of next-generation sequencing technologies enabled cost-effective high-throughput sequencing of metagenomic DNA molecules resulting in increased resolution of microbial community analysis. In addition, screening of metagenomic libraries led to the identification of numerous novel biomolecules from various environments such as soil, seawater, or glacial ice.

Introduction

The immensely manifold microbial niches on Earth comprise an extraordinarily high abundance and diversity of prokaryotic and eukaryotic microorganisms. The human body is colonized by a wide variety of microbes representing all three domains of life. The entirety of these microbial cells (the human microbiome) that is often described as an additional organ exceeds the number of human cells by at least an order of magnitude and outnumbers human genes by more than 100-fold (Li et al. 2012; Weinstock 2012). Also in extreme environments such as hydrothermal vents, sea ice, or deep inside the Earth's crust, various microorganisms could be detected. For example, a phylogenetically diverse and metabolically active microbial assemblage was identified in the brine of an ice-sealed Antarctic lake (Murray et al. 2012). The microorganisms existing in this aphotic ecosystem withstand a temperature of $-13\text{ }^{\circ}\text{C}$, anoxic conditions, and high salinity.

Currently, less than 1 % of the microorganisms on Earth are readily culturable under laboratory conditions. To investigate the high percentage of uncultured microbes, different

metagenomic approaches can be routinely applied. Metagenomics allows the direct study of the collective genomes present in microbial ecosystems (Handelsman 2004). This approach significantly expanded our knowledge on microbial phylogenetic and functional diversity and enabled the discovery of numerous previously unknown biomolecules. In the recent history of metagenomics, especially next-generation sequencing techniques, allowing cost-effective and rapid decoding of metagenomic DNA, were applied to analyze microbial populations. As a consequence, a number of bioinformatic tools to evaluate and compare comprehensive high-throughput metagenomic data have been developed in the last few years.

In this review, an overview of traditional and recent metagenomic research approaches, associated future challenges, and a short description of related meta-omic studies will be given.

Microbial Phylogenetic and Functional Diversity Determination

Small-subunit rRNA genes, universally distributed across prokaryotic and eukaryotic organisms, can be considered as evolutionary clocks enabling phylogenetic analysis. Most commonly, metagenome-derived 16S rRNA and 18S rRNA genes are used to phylogenetically characterize microbial communities. Furthermore, other conserved genes such as *recA*, *rpoB*, *HSP70*, or *EF-Tu* allow phylogenetic assignments (Ludwig and Klenk 2001). These genes can be investigated by applying traditional molecular approaches including fingerprinting methods such as denaturing gradient gel electrophoresis and terminal restriction fragment length polymorphism analysis or Sanger sequencing. A significant drawback of the Sanger sequencing-based analysis of microbial communities is the time-consuming and labor-intensive nature of this approach, as well as the required construction of clone libraries.

More recently, next-generation sequencing platforms were used to decode metagenomic DNA. Currently, the following next-generation

sequencing technologies are available: sequencing by ligation (SOLiD – Applied Biosciences/Life Technologies), sequencing by synthesis (Solexa/Illumina), semiconductor chip sequencing (Ion Torrent/Life Technologies), pyrosequencing (454/Roche), and single-molecule sequencing (Oxford Nanopore Technologies, SMRT – Pacific Biosciences). Compared to Sanger sequencing, these cloning-independent techniques allow the generation of far more sequence data per run. Thus, microbial diversity comparisons between different environmental samples, requiring replicated data and statistical analysis, as well as deep analysis of highly complex microbial community structures, are possible. Currently, often tens to hundreds of thousands partial metagenomic small-subunit rRNA gene sequences are produced using next-generation sequencing platforms. In a recent pyrosequencing-based 16S rRNA gene survey, a total of 41,141 bacterial and 30,651 archaeal sequences were analyzed to investigate prokaryotic diversity in Yunnan and Tibetan hot springs (Song et al. 2013). To (pre-)process small-subunit rRNA gene sequence datasets, various tools, software packages, analytical web servers, and virtual instances can be used (Gonzalez and Knight 2012). The QIIME package (Caporaso et al. 2010) provides workflows to extensively analyze high-throughput amplicon-based sequence data starting with raw sequences. Nevertheless, the avoidance of marker gene amplification bias by applying direct sequencing of metagenomic DNA instead of amplicon-based sequencing allows the most exact taxonomic assessment (Simon and Daniel 2011). For further improvement of microbial diversity and abundance estimation, Kembel et al. (2012) recently introduced an approach, which incorporates 16S rRNA gene copy number information.

To identify the taxonomic affiliation of all sequences derived from metagenomic DNA, a process called binning can be carried out. Within binning procedures, sequences of a metagenomic dataset sharing the same taxonomic origin are “binned” (grouped). Composition-based binning is based on conserved genomic features such as dinucleotide

frequencies, GC content, and synonymous codon usage, whereas similarity-based binning makes use of sequence homology. Among others, PhyloPythiaS, introduced by Patil et al. (2011), represents an appropriate application to perform composition-based binning. With respect to similarity-based binning, typically searches against reference databases (e.g., National Center for Biotechnology Information databases) are performed using alignment tools such as BLAST+ (Camacho et al. 2009). Subsequently, BLAST results can be interpreted by applying software such as MEGAN (Huson et al. 2011).

Due to the often very high diversity of microbial communities, assembly of metagenome-derived sequences is challenging. In a recent metagenomic survey of honey bee gut microbiota, de novo assembly of 81,343,096 Illumina paired-end reads resulted in 54,700 scaffolds of contigs (total length, 76.6 Mb) (Engel et al. 2012). Similar to the approach conducted by Engel et al. (2012), single-genome assemblers were used for metagenome assembly with modified settings. Recently, a single-genome assembler (Velvet) has been extended to enable the assembly of short metagenomic reads (Namiki et al. 2012). This new de novo assembler (MetaVelvet) generated significantly higher N50 scores, a parameter that evaluates assembly quality, than analyzed single-genome assemblers for simulated datasets.

Based on assemblies or individual metagenomic sequence reads, gene prediction, annotation, and reconstruction of pathways can be carried out to assess the functional potential encoded by metagenomes. Consecutive processing of these steps is provided by a number of web-based tools like MG-RAST (Meyer et al. 2008). These tools utilize resources of reference databases such as SEED (Overbeek et al. 2005) and KEGG (Kanehisa et al. 2008) to link biological information to predicted genes. In a recent survey including metagenomic methods, the functional potential of Arctic *Thaumarchaeota* was investigated (Alonso Sáez et al. 2012). By analyzing a metagenome derived from a Southeast Beaufort Sea sample collected

during Arctic winter, Alonso Sáez et al. (2012) identified thaumarchaeal pathways for ammonia oxidation. A number of other *Thaumarchaeota* are also capable of ammonia oxidation, but unexpectedly these Arctic thaumarchaeal organisms harbored a high abundance of genes involved in urea transport and degradation.

Metagenomic Biomolecule Discovery

To access the large pool of unexplored biomolecules, microbial community DNA has been extracted and metagenomic libraries have been constructed. Small-insert and large-insert metagenomic libraries can be screened to identify novel biomolecules. For the construction of small-insert libraries containing metagenomic DNA ≤ 15 kb, plasmids are appropriate vectors, whereas cosmids, fosmids, and bacterial artificial chromosomes (BACs) can be used for cloning of large metagenomic DNA molecules (cosmids and fosmids, ≤ 40 kb; BACs, 100–200 kb). Metagenomic libraries from different microbial habitats such as glacier ice, digestive tracts of animals, soil, hot springs, or seawater have already been constructed and successfully screened for novel biomolecules (see, e.g., Nacke et al. 2012). Some of these biomolecules exhibit valuable characteristics for industrial applications such as thermal stability, halotolerance, and activity under acidic or alkaline conditions. In a recent metagenomic approach, Sulaiman et al. (2012) isolated a gene encoding a novel cutinase homolog designated LC-cutinase with polyethylene terephthalate-degrading activity from a leaf-branch compost fosmid library. The enzyme showed higher specific polyethylene terephthalate-degrading activity than previously reported bacterial and fungal cutinases. Thus, LC-cutinase is a potent candidate for industrial applications, i.e., in textile industry. In general, two different metagenomic screening approaches for the identification of novel biomolecules can be distinguished: function-based screening and sequence-based screening.

Principle and Variations of Function-Driven Screens

To perform function-driven screening, the construction of small-insert or large-insert metagenomic libraries is required. A broad array of different function-based screening approaches can be applied using these libraries. The phenotypic insert detection (PID) is the most frequently applied screening strategy. Metagenomic library-containing clones expressing target genes are identified based on phenotypic characteristics. This method has been applied to identify novel lipolytic genes and gene families from German forest and grassland soil samples using tributyrin as a screening substrate (Nacke et al. 2011). A total of 37 lipolytic clones, encoding novel lipases and esterases, which could be assigned to five different known families and two putatively new families of lipolytic enzymes, were identified by halo formation on indicator agar plates. The potential to identify entirely novel target genes is an important advantage of function-driven screening approaches. Modulated detection (MD) represents another commonly applied strategy to perform function-based screening. Only if a certain gene product is expressed by a metagenomic library-containing host strain, it can grow under selective conditions. Recently, novel acid resistance genes were derived from planktonic and rhizosphere microbial communities of the Tinto River (Spain) using this strategy (Guazzaroni et al. 2013). Fifteen genes, mainly encoding putative proteins of unknown function, conferred acid resistance to the host strain *Escherichia coli*. Moreover, substrate-induced gene expression (SIGEX), product-induced gene expression (PIGEX), and metabolite-regulated expression (METREX) screening strategies allow the identification of target genes from metagenomic libraries (Simon and Daniel 2009). Recently, Wang et al. (2012) suggested biosensor-based genetic transducer (BGT) systems as an alternative and sensitive approach to screen for gene clusters whose expression produce small molecules that activate the employed

biosensors. Nevertheless, all of these function-based screening approaches share one significant disadvantage: the dependence of target gene production on the expression machinery of the metagenomic library host.

Principle and Variants of Sequence-Based Screening

Conserved regions of genes or proteins enable sequence-driven screening approaches. Based on these regions degenerate primers can be designed and fragments of target genes amplified. For example, novel biphenyl dioxygenase DNA segments encoding active site residues were obtained from polychlorobiphenyl-contaminated soils using this strategy (Standfuß-Gabisch et al. 2012). After sequencing of an amplified partial target gene, it can be decoded completely using primer walking and extracted environmental DNA or a metagenomic library as a template. In this way, an entire xylose isomerase gene (*xymI*) has been derived from a soil metagenomic library (Parachin and Gorwa-Grauslund 2011). The gene product of *xymI* consisted of 443 amino acids and was most similar (83 % identity) to a xylose isomerase from *Sorangium cellulosum*. Additionally, novel complex polyketide and nonribosomal peptide biosynthesis gene cluster that often exceed average insert sizes of large-insert metagenomic libraries can be discovered by using degenerate primers and subsequent chromosome walking (Piel 2011). The potential to identify genes of interest even if they are not expressed in a metagenomic library host represents a major advantage of sequence-based screening, but only novel variants of already-known gene or protein families can be detected by this method.

Future Challenges in Metagenomic Research and Related Meta-omic Approaches

One of the major requirements to combine and compare metagenomic studies conducted by

research groups worldwide is the definition and acceptance of minimum standards in experimental design. The same applies to metatranscriptomics, metaproteomics, and metabolomics. In this way, comparison and combination of results obtained from the different meta-omic approaches are feasible. Metatranscriptomics, metaproteomics, and metabolomics comprise the study of the collective gene transcripts, expressed proteins, and metabolites, respectively, generated by the microorganisms within an ecosystem (Nacke et al. 2014; Hettich et al. 2012; Patti et al. 2012). The consequent application and combination of appropriate meta-omic approaches will lead to an enormous extension of knowledge on the gene structure, diversity, activity, and responses of microbial communities on an ecosystem level. Furthermore, the rapid growth of meta-omic technologies will continuously demand for progress in the field of bioinformatics. Thus, further development and linkage of meta-omic analysis tools will be important in the future. In addition, the application and improvement of culture-based methods will be still valuable in the future to extend the number of available reference genomes allowing mapping of metagenomic data. In this context, the young discipline of single cell genomics has potential to play a complementary role by continuously contributing novel reference genomes.

Summary

The introduction of metagenomics allowed culture-independent analysis of microbial populations in complex ecosystems. Subsequently, other culture-independent meta-omic disciplines including metatranscriptomics, metaproteomics, and metabolomics were established. Metagenomics provided insights into the enormous phylogenetic and functional diversity of microbial communities within various environments on Earth. The increasing number of next-generation sequencing technologies led to a more comprehensive and cost-effective assessment of the information encoded by metagenomic DNA. Metagenomic approaches comprising the construction and screening of

metagenomic libraries resulted in identification of previously unknown biomolecules, including biomolecules with industrially relevant characteristics.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Extraction Methods, Variability Encountered in](#)
- ▶ [Fosmid System](#)
- ▶ [Genome Portal, Joint Genome Institute](#)
- ▶ [Microbial Diversity, Bar-Coding Approaches](#)
- ▶ [Microbial Ecosystems, Protection of](#)
- ▶ [Phylogenetics, Overview](#)

References

- Alonso Sáez L, Waller AS, Mende DR, et al. Role for urea in nitrification by polar marine Archaea. *Proc Natl Acad Sci USA*. 2012;109:17989–94.
- Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinforma*. 2009;10:421.
- Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7:335–6.
- Engel P, Martinson VG, Moran NA. Functional diversity within the simple gut microbiota of the honey bee. *Proc Natl Acad Sci USA*. 2012;109:11002–7.
- Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Curr Opin Biotechnol*. 2012;23:64–71.
- Guazzaroni ME, Morgante V, Mirete S, et al. Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. *Environ Microbiol*. 2013;15:1088–1102.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68:669–85.
- Hettich RL, Sharma R, Chourey K, et al. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Curr Opin Microbiol*. 2012;15:373–80.
- Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. *Genome Res*. 2011;21:1552–60.
- Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and environment. *Nucleic Acids Res*. 2008;36:D480–4.
- Kembel SW, Wu M, Eisen JA, et al. Incorporating 16S gene copy number information improves estimates of

- microbial diversity and abundance. *PLoS Comput Biol.* 2012;8:e1002743.
- Li K, Bihan M, Yooshep S, Methé BA. Analyses of the microbial diversity across the human microbiome. *PLoS ONE.* 2012;7:e32118.
- Ludwig W, Klenk HP. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In: Garrity GM, Boone DR, Castenholz RW, editors. *Bergey's manual of systematic bacteriology*, Vol. 1. 2nd ed. New York: Springer; 2001. p. 49–65.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9:386.
- Murray AE, Kenig F, Fritsen CH, et al. Microbial life at –13 °C in the brine of an ice-sealed Antarctic lake. *Proc Natl Acad Sci USA.* 2012;109:20626–31.
- Nacke H, Will C, Herzog S, et al. Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German biodiversity exploratories. *FEMS Microbiol Ecol.* 2011;78:188–201.
- Nacke H, Engelhaupt M, Brady S, et al. Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnol Lett.* 2012;34:663–75.
- Nacke H, Fischer C, Thürmer A, et al. Land use type significantly affects microbial gene transcription in soil. *Microb Ecol.* 2014;67:919–30.
- Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012;40:e155.
- Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33:5691–702.
- Parachin NS, Gorwa-Grauslund MF. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenomic library. *Biotechnol Biofuels.* 2011;4:9.
- Patil KR, Haider P, Pope PB, et al. Taxonomic metagenome sequence assignment with structured output models. *Nat Methods.* 2011;8:191–2.
- Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol.* 2012;13:263–69.
- Piel J. Approaches to capturing and designing biologically active small molecules produced by uncultured microbes. *Annu Rev Microbiol.* 2011;65:431–53.
- Simon C, Daniel R. Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol.* 2009;85:265–76.
- Simon C, Daniel R. Metagenomic analyses: past and future trends. *Appl Environ Microbiol.* 2011;77:1153–61.
- Song ZQ, Wang FP, Zhi XY, et al. Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environ Microbiol.* 2013;15:1160–75.
- Standfuß-Gabisch C, Al-Halbouni D, Hofer B. Characterization of biphenyl dioxygenase sequences and activities encoded by the metagenomes of highly polychlorobiphenyl-contaminated soils. *Appl Environ Microbiol.* 2012;78:2706–15.
- Sulaiman S, Yamato S, Kanaya E, et al. Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Appl Environ Microbiol.* 2012;78:1556–62.
- Wang Y, Chen Y, Zhou Q, et al. A culture-independent approach to unravel uncultured bacteria and functional genes in a complex microbial community. *PLoS ONE.* 2012;7:e47530.
- Weinstock GM. Genomic approaches to studying the human microbiota. *Nature.* 2012;489:250–6.

Arbuscular Mycorrhizal Fungi Assemblages in Chernozems

Chantal Hamel, Luke D. Bainard and Mulan Dai
Semi-arid Prairie Agricultural Research Centre,
Agriculture and Agri-Food Canada, Swift
Current, SK, Canada

Synonyms

Diversity, arbuscular mycorrhizal fungi, Canadian Prairie, Chernozem, land use.

Definition

AM fungi are obligate plant symbionts that form the phylum Glomeromycota. These fungi contribute to plant nutrient uptake, influence soil biotic and abiotic environments, and provide important ecosystem services. 454-pyrosequencing of amplicons from metagenomic DNA revealed the distribution of AM fungi in major Canadian Chernozem great groups as influenced by land use and crop management.

Introduction

AM fungi form a mycorrhizal symbiosis with the roots of the majority of land plants. They have

coevolved with plants over 450 Ma to produce today's mycorrhiza, which is an organ specialized in the extraction of soil nutrients. As such, AM fungi are seen as a key stone of agricultural sustainability (Garg and Chandel 2010).

World grain, pulse, and biofuel crop production mainly occurs on deep (typically >18–25 cm) warm-colored soils rich in humus (>0.6 % organic carbon) and weatherable minerals, with high levels of base saturation (>50 %) and calcium as the main exchangeable cation (Durán et al. 2011). These soils have similar properties but have different names in other soil classification systems. They are Chernozems in Canada, Ukraine, and Russia; Mollisols in the USA and South America; Isohumosols or Black Soils in China; and Chernozems, Kastanozems, and Phaeozems according to the FAO (Liu et al. 2012). These soils have typically developed under condition of moisture deficit and grassland vegetation in temperate regions around the globe. They mainly occur in a band across Eastern Europe and Central Asia, in northeast China, from south-central Canada down to the Gulf of Mexico, and over most of Uruguay and part of Argentina.

Tackling the Complexity of Soil Biodiversity

Soil hosts an extremely high level of microbial diversity (Young and Crawford 2004). However, high-throughput next-generation sequencing now allows generation of the massive sequence data required to characterize soil microbial diversity.

Amplicon sequencing is preferred over whole genome sequencing for the study of the taxonomic diversity of targeted microbial groups. The 454 FLX and 454 FLX + technologies allow the sequencing of DNA amplicons up to 400 and 800 bp in length, respectively. Such long sequences contain sufficient taxonomic information for the characterization of microbial communities and their use conveniently eliminates the need for sequence assembly.

Pyrosequencing of amplicons and bioinformatic analysis of sequence data yield the profile

of operational taxonomic units (OTU) of the target microbial group in a soil sample. The concept of an OTU is useful in soil microbiology as the majority of microbial species are still undescribed. OTUs serve as a proxy for species making it possible to measure and describe soil microbial diversity. In addition, OTUs can be identified by comparison with known sequences in public databases such as GenBank and MaarjAM. AM fungi have been difficult to study due to their obligate biotrophy and inability to grow in pure culture. However, polymerase chain reaction (PCR) made possible the amplification of DNA from their spores and enabled the molecular characterization and classification of taxa within the Glomeromycota (Schuessler 2013).

Fungal diversity is commonly assessed based on the internal transcribed spacer (ITS) of the ribosomal RNA gene. However, abundant SSU rRNA gene sequences of AM fungi are found in databases due to the traditional use of this region for the Glomeromycota. Several primers sets producing taxonomically informative amplicons short enough for use with first- and next-generation molecular techniques have been used in ecological studies of AM fungi.

The AM fungi have a patchy distribution in soil (Hart and Klironomos 2003). Thus in order to capture their diversity, multiple samples must be taken at a study site. A composite sample is usually made by pooling and homogenizing all the samples from a sampling site. The distribution of organisms varies with soil depth, thus sampling depth also matters. The AM fungi are normally found within the rooting depth.

Arbuscular Mycorrhizal Fungi in the Canadian Chernozems

AM fungal communities in the Canadian Prairie Chernozem soils are composed of a few dominant and a large number of subordinate taxa. Less than 6 % of the AM fungal OTUs accounted for half of all AM fungal reads (Dai et al. 2013). Across the Canadian prairie landscape, the Glomeraceae were the most abundant family, accounting for

65 % of all AM fungal OTUs and 54 % of the AM fungal reads. The Claroideoglomeraceae is second in abundance with 25 % of all AM fungal OTUs and 39 % of the AM fungal reads. Diversisporaceae accounted for 8 % of the OTUs and 7 % of the AM fungal reads. Paraglomaceae, Gigasporaceae, and Archeosporaceae are poorly distributed across the prairie landscape, and Gigasporaceae and Archeosporaceae are rare.

In other regions, spore counts in grazed Kastanozems of Inner Mongolia revealed that the AM fungal communities resembled those observed in Canadian Chernozems (Tian et al. 2009). The Gigasporaceae are susceptible to disturbance and largely absent in croplands, which explains their greater abundance in the Kastanozems than in the Canadian Prairie Chernozems (Dai et al. 2012, 2013). Poorer AM fungal diversity is reported from American spore-based surveys of Mollisols under tallgrass prairie cover where Paraglomaceae and Archeosporaceae were undetected (Eom et al. 2001; Bentivenga and Hetrick 1992). Tallgrass prairies managed with fire were found to be very highly dominated by the Glomeraceae (Bentivenga and Hetrick 1992), underlining the importance of land use in the structuring of AM fungal communities.

AM fungi share root occupation with fungal endophytes belonging to different taxonomic groups. Non-AM fungal endophytes are particularly abundant in temperate grasslands (Porrás-Alfaro et al. 2011). This observation triggered the question as to whether AM fungi are at the end of their range in dry areas.

This hypothesis was explored in the Canadian Prairie using primers Glo1/NS31, which produced 18S rDNA amplicons of about 230 bp (Yang et al. 2010). A succession of AM fungi was detected as the soil dried from early to late summer, suggesting that the adaptation of AM fungi to soil moisture availability varies with species. *Glomus viscosum*, *Funneliformis mosseae*, and *Glomus hoi* were dominant in early summer, under conditions of moisture sufficiency, whereas the dominant AM fungal OTUs in late season conditions (i.e., dry soil) belonged to *Glomus iranicum* and *Glomus macrocarpum*.

This concurs with the previous observation of differences in the seasonal pattern of sporulation of different AM fungal species (Dhillon and Anderson 1993). Seasonal variation of AM fungi in the North American Great Plains was also described as the replacement of the fungi of the order Helotiales by AM fungi as the season unfolds in the North American Great Plains (Jumpponen 2011).

The Chernozem great groups are distributed along a gradient of precipitation radiating outward from the US border in eastern Alberta, i.e., from the Brown soil zone through Dark Brown and Black soils up to the Gray soil zone at the fringe of the boreal forest. The lowest abundance, richness, and diversity of AM fungi were observed in the driest soil zone (Brown Chernozem), which supported a negative impact of moisture deficit on these fungi.

Soil moisture appears to be just one of several factors that influence the composition of AM fungal communities in Chernozem soils. Despite the highest levels of precipitation in the Gray soil zone, the highly productive Black soils harbor the most abundant and diverse AM fungal communities (Dai et al. 2012). Black, Gray, Dark Brown, and Brown soils had an average of 10.2, 7.1, 7.0, and 6.2 AM fungal OTUs, respectively, and the Shannon diversity index of these soil groups follows a similar trend. AM fungal communities in Brown soils are characterized by a reduced relative abundance of Claroideoglomeraceae compared to Black and Dark Brown soils. Other important factors that influenced the abundance of AM fungal OTUs were A horizon thickness and physicochemical properties of the soils, such as bulk density, Zn level, pH, electrical conductivity, and sulfur level.

Soils are classified based on their physical and chemical properties. A soil type represents a living environment inhabited by different AM fungal communities. American Mollisols and Alfisols contain distinct AM fungal spore assemblages (Ji et al. 2012). Similarly, Canadian Chernozems and Podzols and even different great groups of Chernozems contained distinct assemblages of AM fungal rRNA gene sequences (Dai et al. 2013).

Land use modifies the conditions of the soil environment and the impact of land use on the structure of AM fungal communities exceeds that of soil type. In the Canadian Prairie, roadsides host a higher level of AM fungal diversity than cropland or natural areas (Dai et al. 2013). Roadsides have higher soil moisture levels than cropland and most natural areas, further indicating that water availability is an important determinant of the abundance and structure of AM fungal communities. Seven percent of the AM fungal OTUs found across the prairie soil zones are unique to croplands, whereas 14 % of the AM fungal OTUs are specific to roadsides. Roadsides and natural areas are dominated by an OTU closely related to *Claroideoglossum lamellosum*, *C. etunicatum*, and *C. claroideum*, which account for 14 % and 19 % of all AM fungal reads. In cropland, an OTU closely related to *Funneliformis mosseae* accounted for as much as 17 % of all AM fungal reads. The dominance of *F. mosseae* in croplands of the Canadian prairie is supported by studies based on metagenomic methods (Ma et al. 2005; Sheng et al. 2012; Dai et al. 2012, 2013) and on spore counts (Talukdar and Germida 1993).

Crop management systems also have a strong influence on the composition of AM fungal communities in Chernozem soils. Organic systems have been shown to support more abundant and diverse AM fungal communities compared to conventional systems (Dai et al. 2014). Organic systems also promote greater proliferation of *Claroideoglossum* and of *incertae sedis* taxa of the Glomeraceae, currently referred to as *Glomus iranicum* and *Glomus indicum*. However, these Glomeraceae *incertae sedis* are seemingly parasitic as they were associated with reduced crop growth and N and P uptake efficiency.

Summary

Metagenomic studies on the distribution of AM fungi in Chernozems are extremely useful to understand how the living soil provides ecological services and supports the production of food and bioproducts. Brown Chernozems are

relatively poor in symbiotic AM fungi and are less hospitable to the *Claroideoglossum* than other Chernozems, whereas Black Chernozems are rich in AM fungal resources. The influence of soil type on the composition of AM fungal communities is relatively small compared to that of land use type. *Funneliformis* have a competitive edge and proliferate in conventional crop production systems, whereas *Claroideoglossum* and Glomeraceae *incertae sedis* are favored in organic production systems. These Glomeraceae *incertae sedis*, currently known as the *G. iranicum*/*G. indicum* group, are associated with reduced crop productivity and nutrient uptake.

References

- Bentivenga SP, Hetrick BAD. The effect of prairie management practices on mycorrhizal symbiosis. *Mycologia*. 1992;84:522–7.
- Dai M, Bainard LD, Hamel C, Gan Y, Lynch D. Impact of land use on arbuscular mycorrhizal fungal communities in rural Canada. *Appl Environ Microbiol*. 2013;79:6719–29. doi:10.1128/aem.01333-13.
- Dai M, Hamel C, Bainard LD, St. Arnaud M, Grant CA, Lupwayi NZ, Malhi SS, Lemke R. Negative and positive contributions of arbuscular mycorrhizal fungal taxa to wheat production and nutrient uptake efficiency inorganic and conventional system in the Canadian prairie. *Soil Biol Biochem*. 2014;74:156–166.
- Dai M, Hamel C, St. Arnaud M, He Y, Grant C, Lupwayi N, Janzen H, Malhi SS, Yang X, Zhou Z. Arbuscular mycorrhizal fungi assemblages in Chernozem great groups revealed by massively parallel pyrosequencing. *Can J Microbiol*. 2012;58:81–92.
- Dhillon SS, Anderson RC. Seasonal dynamics of dominant species of arbuscular mycorrhizae in burned and unburned sand prairies. *Can J Bot*. 1993;71:1625–30.
- Durán A, Morrás H, Studdert G, Xiaobing L. Distribution, properties, land use and management of Mollisols in South America. *Chin Geogr Sci*. 2011;21:511–30.
- Eom AH, Wilson GWT, Hartnett DC. Effects of ungulate grazers on arbuscular mycorrhizal symbiosis and fungal community structure in tallgrass prairie. *Mycologia*. 2001;93:233–42.
- Garg N, Chandel S. Arbuscular mycorrhizal networks: process and functions. A review. *Agron Sustain Dev*. 2010;30:581–99.
- Hart MM, Klironomos JN. Diversity of arbuscular mycorrhizal fungi and ecosystem functioning. In: van der Heijden MGA, editor. *Mycorrhizal ecology, Ecological studies*, vol. 157. Berlin: Springer; 2003. p. 225–42.

- Ji B, Bentivenga SP, Casper BB. Comparisons of AM fungal spore communities with the same hosts but different soil chemistries over local and geographic scales. *Oecologia*. 2012;168:187–97.
- Jumpponen A. Analysis of ribosomal RNA indicates seasonal fungal community dynamics in *Andropogon gerardii* roots. *Mycorrhiza*. 2011;21:453–64.
- Liu X, Lee Burras C, Kravchenko YS, Duran A, Huffman T, Morras H, Studdert G, Zhang X, Cruse RM, Yuan X. Overview of Mollisols in the world: distribution, land use and management. *Can J Soil Sci*. 2012;92:383–402.
- Ma WK, Siciliano SD, Germida JJ. A PCR-DGGE method for detecting arbuscular mycorrhizal fungi in cultivated soils. *Soil Biol Biochem*. 2005;37:1589–97.
- Porras-Alfaro A, Herrera J, Natvig DO, Lipinski K, Sinsabaugh RL. Diversity and distribution of soil fungal communities in a semiarid grassland. *Mycologia*. 2011;103:10–21.
- Schuessler A. *Glomeromycota*. *Taxonomy*. 2013. Accessed 6 Nov 2013. <http://schuessler.userweb.mwn.de/amphylo/>
- Sheng M, Hamel C, Fernandez MR. Cropping practices modulate the impact of glyphosate on arbuscular mycorrhizal fungi and rhizosphere bacteria in agroecosystems of the semiarid prairie. *Can J Microbiol*. 2012;58:990–1001.
- Talukdar NC, Germida JJ. Occurrence and isolation of vesicular-arbuscular mycorrhizae in cropped field soils of Saskatchewan, Canada. *Can J Microbiol*. 1993;39:567–75.
- Tian H, Gai JP, Zhang JL, Christie P, Li L. Arbuscular mycorrhizal fungi in degraded typical steppe of Inner Mongolia. *Land degrad dev*. 2009;20:41–54.
- Yang C, Hamel C, Schellenberg MP, Perez JC, Berbara RL. Diversity and functionality of arbuscular mycorrhizal fungi in three plant communities in semiarid Grasslands National Park. *Can Microb Ecol*. 2010;59:724–33.
- Young IM, Crawford JW. Interactions and self-organization in the soil-microbe complex. *Science*. 2004;304:1634–7.

B

Bacterial Diversity in Tree Canopies of the Atlantic Forest

Marcio R. Lambais¹ and David E. Crowley²

¹Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba, SP, Brazil

²Environmental Sciences, University of California, Riverside, Riverside, CA, USA

Synonyms

Bacterial communities in the phyllosphere of the Atlantic forest

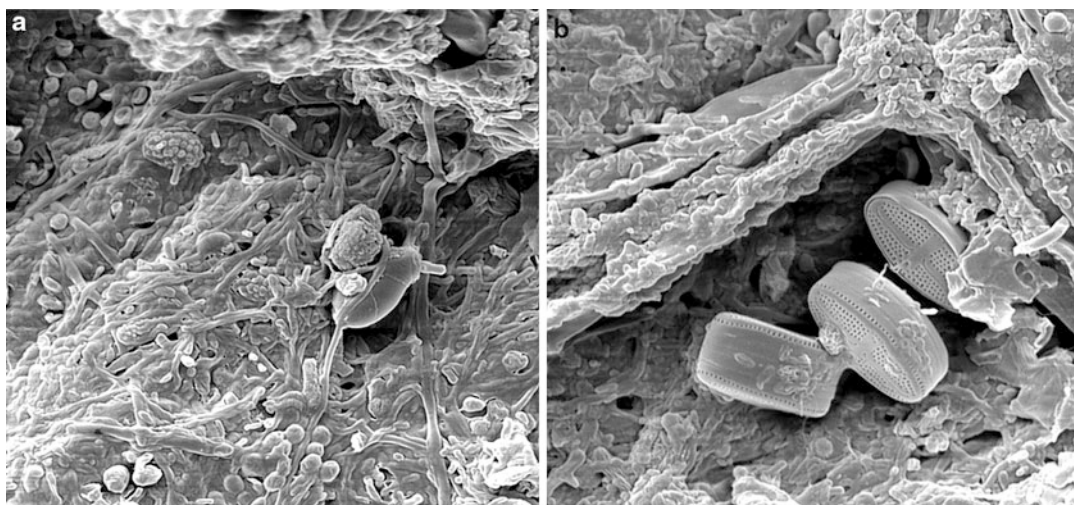
Definition

16S rRNA gene profiling is one of the main approaches used for the study of microbial communities that are associated with plants and animals, which are mostly comprised of species unable to grow under laboratory conditions. Even though plants harbor an enormous microbial diversity on their various surfaces, the functions of these microorganisms, except for a few that are pathogens or symbionts, are largely unknown, but are speculated to modify plant chemical signals, alter root exudation patterns, and provide protection against pathogens. Understanding of the factors that shape the structure of microbial communities, and the functions of

microorganisms that are associated with plants, will likely be essential for establishing conservation strategies for protecting endangered plant species. The large reservoir of microbial diversity on plant surfaces also represents a largely untapped bank of microbial products that may be of interest for pharmaceutical, agricultural, and environmental applications.

Introduction

Plant surfaces in natural and agricultural ecosystems are colonized by a variety of epiphytic microorganisms that have been examined in relation to their diversity, ecology, and genetics using culture-dependent and culture-independent approaches. Among the various surfaces that are presented by plants, the leaf surface, also known as the phyllosphere (Ruinen 1956), is one of the most common habitats for terrestrial microorganisms. The phyllosphere may be colonized by bacterial cells at an average density of 10^6 – 10^7 cells cm^{-2} on plants from temperate regions (Lindow and Brandl 2003) and may be even higher on tropical plants where dense canopies and a moist shaded environment are conducive for bacterial growth. Considering that the estimated total leaf area of terrestrial plants is approximately 6.4×10^8 km^2 (Morris and Kinkel 2002), the number of bacterial cells on leaf surfaces globally has been estimated to be as high as 10^{26} cells. Despite the importance of



Bacterial Diversity in Tree Canopies of the Atlantic Forest, Fig. 1 Microbial biofilm on the leaf surface of trees of the Atlantic forest. (a) Biofilm with multiple

microbial species, based on morphology of cells. (b) Diatom cells embedded in the microbial biofilm on the leaf surface

plant-microbe interactions in plant disease, almost nothing is known about the indigenous, nonpathogenic bacteria that colonize plant leaf surfaces and their functions in terrestrial ecosystems.

The Phyllosphere Habitat

Due to the harsh conditions and the highly competitive environment on plant leaves, microorganisms that live in the phyllosphere almost certainly have evolved specific traits that enable them to grow in such environments. Diurnal variations in UV light incidence, temperature, water availability, osmotic conditions, the concentration of reactive oxygen species, as well as the low availability of nutrients make the phyllosphere an extreme environment for microbial growth (Vorholt 2012). All of these factors, together with the specific morphological traits of the leaves, may contribute to the selection of specific microbial populations of bacteria, fungi, archaea, and protozoa that will colonize the phyllosphere and interact at different levels with the plant host. In addition, the microbial populations will interact with each other through

metabolic and signaling networks, leading to the self-organization of highly complex communities that have been selected by long-term coevolution with their plant host. In general, the bacterial populations in the phyllosphere occur as multi-species biofilms (Fig. 1) mostly located at the base of trichomes and nutrient-rich locations along the veins and junctions of epidermal cells (Morris et al. 1998; Monier and Lindow 2004). Communication between microbial cells, and between microbial and plant cells, may be an important factor controlling the dynamics of leaf colonization and biofilm growth and development.

One of the major selection factors for microbial colonization of leaf surfaces is the ability to tolerate or grow on the myriad chemical substances that are released from plant leaf tissues and/or produced by other microorganisms. This includes many thousands of secondary metabolites, such as monoterpenes that serve as signal factors and defense compounds, as well as chemical attractants and deterrents for insects, herbivores, and pathogens. However, the specific secondary metabolites driving the structure of bacterial communities in the phyllosphere are unknown.

Bacterial Communities in the Phyllosphere

Many early surveys of phyllosphere communities have relied on descriptions of bacteria that can be cultivated on agar media and isolated as individual colonies. Using various types of growth media, 85 species of culturable microorganisms from 37 genera have been reported in the phyllospheres of rye, olive, sugar beet, and wheat (Ercolani 1991; Legard et al. 1994; Thompson et al. 1993). While this is an impressive number of species, studies using molecular methods have revealed that the actual microbial species richness in the phyllosphere of agricultural plants is much greater than this and suggest that different plant species harbor unique communities that are similar for individuals of the same plant species (Yang et al. 2001). The discovery of high levels of bacterial species richness associated with different agronomic plants has prompted many questions about the true extent of microbial diversity that may be associated with the phyllosphere of different plants in natural ecosystems around the world. It has been speculated that since bacteria can be transported across the globe in dust (Griffin et al. 2002), only a small number of bacterial species may be adapted to grow on leaf surfaces. On the other hand, if each plant species selects for its own microbial community, the microbial species diversity that is associated with all of the different plant species on earth could be enormous. This question can only be answered by systematic surveys of phyllosphere microbial diversity in different ecosystems. Considering the current rate of extinction of plant species, it is especially urgent to begin surveys of phyllosphere microorganisms that are associated with endangered biomes.

Bacterial Community in the Phyllosphere of the Atlantic Forest

Many tropical forests and biodiversity hotspots contain endemic plant species that are preserved only in a few remnant areas. The Atlantic forest of Brazil is an example of a forest with high levels

of biodiversity that is struggling to survive. The Atlantic forest used to be the second largest tropical forest in South America and represented 1.3 million km² in the 1500s, when the Portuguese first arrived in Brazil. Today, approximately 7% of the original Atlantic forest remains, since most of it has been converted to agricultural or urban areas, leaving a patchwork of fragmented remnants. The remnants of the Atlantic forest are considered to be some of the oldest undisturbed forests on the planet, containing approximately 20,000 plant species, of which nearly half are endemic (Tabarelli et al. 2003). Several research projects have been developed in the Atlantic forest as part of the ongoing BIOTA-FAPESP (São Paulo Research Foundation) program, which has been successfully established to examine the biodiversity of the São Paulo State (Brazil).

Different approaches can be used to survey the microbial diversity in the phyllosphere. The first approach is using DNA fingerprinting methods. A low-resolution DNA fingerprinting method referred to as PCR-DGGE (polymerase chain reaction-denaturing gradient gel electrophoresis), through which amplified fragments of highly variable regions of the bacterial 16S rRNA gene are separated by electrophoresis in a denaturing gradient polyacrylamide gel, has been used for studying the bacterial community structures in the phyllosphere of tree species of the Atlantic forest. This methodology generates a distinctive fingerprint that can be used to compare the relative similarities of communities, but does not provide information on the identities of the bacterial species within the communities. To compare the phylogenetic diversity in the phyllosphere and generate diversity indices for different phyllosphere communities, sequencing of specific regions of the bacterial 16S rRNA gene is normally used.

With these combined approaches, it has been shown that the 16S rRNA gene band patterns for the bacterial communities from different tree species of the Atlantic forest are distinct from each other (Lambais et al. 2006). Communities from replicates for different individuals of the same tree species showed some expected variation, but overall are highly similar to each other.

The similarities between the leaf bacterial communities within and between species were further measured statistically and showed that the trees could be segregated into groups according to tree species, family, and order, suggesting a coevolution between trees and microbial populations associated with the phyllosphere (Lambais et al. data not published). Evidence of coevolution of microbial populations associated with the bark (dermosphere) and rhizosphere of trees of the Atlantic forest also has been observed, suggesting that plants coevolved with specific microbiomes (Lambais et al. data not published). An estimate of the bacterial species richness associated with the phyllosphere of trees in the Atlantic forest suggests the existence of 2–13 million undescribed bacterial species that colonize the collective phyllosphere of the Atlantic forest (Lambais et al. 2006). Interestingly, studies of the phyllosphere of different individuals of the same tree species in the Atlantic forest over a range of distances and at different times show that the similarities between bacterial community structures in the phyllosphere of the same plant species decrease with the increasing distance between individual trees, even though they still share high levels of similarity (Lambais et al. data not published). Over larger scales, such as when the bacterial communities of the individuals of the same plant species are separated by hundreds of kilometers, significant differences in community structure are observed. These data suggest that the bacterial diversity in the phyllosphere of plants of the Atlantic forest may be even higher than the predicted 2–13 million species estimate that does not take into account beta diversity.

While still in an early phase, research aimed at measurements of beta diversity includes a survey of *Tamarix* trees in Mediterranean and Dead Sea regions in Israel and two locations in the USA (Finkel et al. 2011). These studies suggest that besides the plant genetic component driving the bacterial community structure in the phyllosphere, environmental conditions associated with particular geographical locations are also important. On the other hand, the high levels of similarity of the bacterial communities in the phyllosphere of *Pinus ponderosa* over

transcontinental distances (Redford et al. 2010) suggest a strong genetic component in the regulation of the phyllosphere associated microbiome.

The majority of bacterial OTUs in the phyllosphere of the trees of the Atlantic forest have been assigned to the phylum Proteobacteria. Based on a survey of several tree species in the Atlantic forest, including *Ocotea dispersa*, *Ocotea teleiandra*, *Mollinedia schottiana*, *Mollinedia uleana*, *Eugenia cuprea*, *Eugenia melanogyna*, and *Tabebuia serratifolia*, it has been shown that, in general, approximately half of the bacteria in the phyllosphere are phylogenetically related to Gammaproteobacteria, whereas 20 % are related to Alphaproteobacteria and 5 % to Flavobacteria, even though interspecific variation may occur (Lambais et al. data not published). For instance, in the phyllosphere of *Ocotea teleiandra*, a high frequency of Alphaproteobacteria and a low frequency of Gammaproteobacteria have been detected, in contrast to other tree species.

Altogether, these results show that every tree species that has been examined in the Atlantic forest contains its own unique bacterial community and that spatially separated individuals of the same tree species have similar bacterial communities, within the same environment (forest physiognomy). The variations in bacterial community structures in the phyllosphere that were observed using the PCR-DGGE and sequencing approaches to compare similarities among individuals indicate that the community compositions may vary on different leaves. This may correspond with different leaf ages, location in the canopy, light incidence, and microclimate conditions that influence the leaf environment and types of chemical substances that are secreted by the plant leaves. The bacteria may also interact with various fungi and algae that colonize the leaf surfaces and change the chemical and physical environment of the leaf habitat. In future studies, it will be necessary to examine the microbial communities on leaf surfaces at the microsite scale to determine changes in species composition and the ecology of different habitats on the leaf surface, for example, on the adaxial

and abaxial leaf surfaces or within biofilms and microcolonies at distinct physical locations on the leaf surface.

Drivers of Community Structure in the Phyllosphere

The development of different bacterial communities in the phyllosphere of different tree species demonstrates the strong effect of leaf surface environment as a selection factor. The initial inoculation of leaves of different trees very likely begins with the growth of opportunistic microorganisms that are transported in dust, by insects, or that are splashed from adjacent trees by rain. Inheriting a minimal microbiome through the seeds may also be a possibility. Further selection then occurs depending on differences in the types of carbon substrates that are available for growth, as well as various physical and environmental factors and interactions within the microbial community. The primary carbon substrates that are used for microbial growth include carbohydrates, amino acids, and organic acids. The composition and amounts of these substances may vary for different plant species, but may also vary over time depending on leaf age, insect damage, and rainfall, for instance. Another potentially important selective factor is the production of different types and quantities of monoterpenes and other volatile substances that are released from the leaf surfaces. These substances may be both toxic to some microorganisms and used as growth substrates by others. Phytochemistry research has shown that tree species have species-specific differences in their biochemical signatures for volatile molecules (Arey et al. 1995). If terpenes act as selective substances, certain types of bacteria may be predicted to occur in relation to the biochemical signatures of volatile organic compounds released by the leaves. Very little work has been conducted on this research topic, but bacteria are known to contain enzymes that convert terpenes to derivative substances. In this manner, the phyllosphere bacteria may influence chemical signaling to insects and other microorganisms or

between plants. Terpenes and other plant secondary metabolites produced in plant leaves are also important feedstocks for various biochemicals that are used in the industry and for pharmacology. Future studies should investigate the genomes and genes encoding enzymes in the phyllosphere that may have broad application for industrial biotechnology, as in the work described by Delmotte et al. (2009), which used proteogenomics to study the microbial community associated with the phyllosphere of soybean, clover, and *Arabidopsis*.

Conclusion

Recent studies have provided only a glimpse into the microbial diversity that is associated with the tree canopies in the Atlantic forest, and there are many new questions that arise from this research. For example, to what degree do soil, nutritional, and other environmental factors affect the composition and structure of microbial communities in the phyllosphere? What is the diversity of fungi and Archaea on the plant leaf surfaces, and how do these microorganisms interact? Future research should also examine the functional aspects of phyllosphere microbial communities and the interactions that occur between phyllosphere bacteria and their host plants using metagenomics, metaproteomics, and metabolomics. As we begin to survey these bacterial communities through systematic study of different plant species, there will be exciting opportunities for studies of the metabolic capabilities and ecological functions of phyllosphere microorganisms in terrestrial ecosystems.

Summary

Each plant species is able to select its own bacterial community, and probably its own microbiome, which may be affected by plant genomic components and the environment. Altogether, the phyllosphere of plant species of the Atlantic forest may harbor several million species of bacteria that remain to be described. The roles

of the microbial communities of the phyllosphere in forest ecology are not yet known, but are likely to include chemical signaling, nitrogen fixation, and plant protection, among other functions. This immense microbial diversity may also provide new biomolecules of interest for pharmaceutical, agricultural, and environmental applications.

Cross-References

► [New Computational Methodologies to Understand Microbial Diversity](#)

References

- Arey J, Crowley DE, Crowley M, Resketo M, Lester J. Hydrocarbon emissions from natural vegetation in California South-coast-air-basin. *Atmos Environ*. 1995;29:2977–88.
- Delmotte N, Knief C, Chaffron S, et al. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc Natl Acad Sci USA*. 2009;106:16428–33.
- Ercolani GL. Distribution of epiphytic bacteria on olive leaves and the influence of leaf age and sampling time. *Microb Ecol*. 1991;21:35–48.
- Finkel OM, Burch AY, Lindow SE, Post AF, Belkin S. Geographic allocation determines the population structure in phyllosphere microbial communities of a salt-excreting desert tree. *Appl Environ Microbiol*. 2011;77:7647–55.
- Griffin DW, Kellogg CA, Garrison VH, Shinn EA. The global transport of dust – an intercontinental river of dust, microorganisms and toxic chemicals flows through the Earth’s atmosphere. *Amer Sci*. 2002;90:228–35.
- Lambais MR, Crowley DE, Cury JC, Büll RC, Rodrigues RR. Bacterial diversity in tree canopies of the Atlantic forest. *Science*. 2006;312:1917.
- Legard DE, McQuilken MP, Whipps JM, Fenlon JS, Fermor TR, Thompson IP, Bailey MJ, Lynch JM. Studies of seasonal changes in the microbial populations on the phyllosphere of spring wheat as a prelude to the release of a genetically modified microorganism. *Agric Ecosyst Environ*. 1994;50:87–101.
- Lindow SE, Brandl MT. Microbiology of the phyllosphere. *Appl Environ Microbiol*. 2003;69:1875–83.
- Monier JM, Lindow SE. Frequency, size, and localization of bacterial aggregates on bean leaf surfaces. *Appl Environ Microbiol*. 2004;70:346–55.
- Morris CE, Kinkel LL. Fifty years of phyllosphere microbiology: significant contributions to research in related fields. In: Lindow SE, Hecht-Poinar EI, Elliott V, editors. *Phyllosphere microbiology*. St Paul: APS Press; 2002. p. 365–75.
- Morris CE, Monier JM, Jacques MA. A technique to quantify the population size and composition of the biofilm component in communities of bacteria in the phyllosphere. *Appl Environ Microbiol*. 1998;64:4789–95.
- Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N. The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environ Microbiol*. 2010;12:2885–93.
- Ruinen J. Occurrence of *Beijerinckia* species in the phyllosphere. *Nature*. 1956;177:220–1.
- Tabarelli M, Pinto LP, Silva JMC, Costa CMR. Endangered species and conservation planning. In: Galindo-Leal C, Câmara IG, editors. *The Atlantic forest of South America: biodiversity, status, threats and outlooks*. Washington, DC: Island Press; 2003. p. 86–94.
- Thompson IP, Bailey MJ, Fenlon JS, Fermor TR, Lilley AK, Lynch JM, McCormack PJ, McQuilken MP, Purdy KJ. Quantitative and qualitative seasonal changes in the microbial community from the phyllosphere of sugar beet (*Beta vulgaris*). *Plant Soil*. 1993;150:177–91.
- Vorholt JA. Microbial life in the phyllosphere. *Nat Rev Microbiol*. 2012;10:828–40.
- Yang CH, Crowley DE, Borneman J, Keen NT. Microbial phyllosphere populations are more complex than previously realized. *Proc Natl Acad Sci USA*. 2001;98:3889–94.

Bacteriocin Mining in Metagenomes

Orla O’Sullivan^{1,2}, Colin Hill³, Paul Ross^{1,2} and Paul Cotter^{1,2}

¹Teagasc Food Research Centre, Moorepark, Fermoy, Co., Cork, Ireland

²Alimentary Pharmabiotic Centre, University College, Cork, Ireland

³Alimentary Pharmabiotic Centre, Department of Microbiology, University College, Cork, Ireland

Definition

Bacteriocins are heat-stable ribosomally synthesized peptides produced by one bacterium which are active against other bacteria and against which the producer has a specific immunity mechanism.

Introduction

Bacteriocins are ribosomally synthesized antimicrobial peptides that are produced by many bacteria and which kill or inhibit the growth of other bacteria. Bacteriocin producers are protected as a consequence of dedicated immunity (self-protective) systems (Cotter et al. 2005). Bacteriocins are of both academic and commercial interest, with several in use as food preservatives or as the active agent in clinical or veterinary antimicrobials. It is not surprising that there is significant interest in the identification and characterization of new bacteriocin gene clusters. The growing volume of metagenomic sequence data is an important resource which can be mined for the in silico discovery of novel bacteriocins.

A Background to Bacteriocins

Bacteriocins were first described in 1925 and since then bacteriocin producers have been identified in a myriad of different environments, bearing out a prediction by Klaenhammer in 1988 that bacteriocin production may be almost ubiquitous (Klaenhammer 1988). The spectrum of activity of these peptides can be narrow (lethal to bacteria in the same or closely related species) or broad (lethal to bacteria in other genera). Many bacteriocins function by depolarizing the cell membrane or through the inhibition of cell wall synthesis (Cotter et al. 2005). There are a number of different classification schemes. One approach, originally employed to classify bacteriocins produced by Gram-positive bacteria, has been to divide bacteriocins into two major

classes: those which are modified (Class I) and those which are unmodified (Class II) (Cotter et al. 2005; Rea et al. 2011) (Table 1). This approach to classification excludes larger proteins, such as the bacteriolysins and the colicin-type antimicrobials, which as a consequence of their larger size may be regarded as representing different classes of antimicrobials.

Further classification of the Class I and II peptides is possible, for example, Class I bacteriocins from Gram-positive bacteria can be divided into Class Ia, Class Ib, and Class Ic. Class Ia, the *lantibiotics*, harbor the unusual posttranslationally modified residues lanthionine (Lan) and/or β -methylanthionine (meLan); these are products of the interaction of cysteines with enzymatically dehydrated serines (dehydroalanine; Dha) and threonines (dehydrobutyrine; Dhb). Lantibiotics can be subdivided according to the enzyme responsible for lanthionine formation; subclass I use LanBC, subclass II use LanM, and subclass III use RamC-like, while subclass IV are modified by LanL enzymes. It should be noted, however, that subclass III and IV peptides identified to date have not been shown to possess antimicrobial activity and thus are referred to as lantipeptides. Class Ib, the *labyrinthopeptins*, have a labyrinthine structure and contain the posttranslationally modified amino acid labionin, formed through a series of serine phosphorylations, dehydrations of phosphoserines to didehydroalanines, and cyclizations. Class Ic, the *sactibiotics*, are cyclic peptides, generated from the posttranslation formation of intramolecular cross-linkages between the α -carbon and sulfur of amino

Bacteriocin Mining in Metagenomes, Table 1 Classification scheme for bacteriocins (Modified from (Rea et al. 2011))

Class	Divisions	Further subclasses	Examples
Class I	Ia: Lantibiotics	Subclass I–IV	Lacticin 3,147, nisin A, subtilin
	Ib: Labyrinthopeptins		
	Ic: Sactibiotics		
Class II	Iia: Pediocin-like	Subclasses I–IV	Pediocin PA-1, munticin
	Iib: Two-peptide bacteriocins		
	Iic: Circular bacteriocins		
	Iid: Linear non-pediocin-like		
	Single-peptide bacteriocins		

acids within the peptide. Class Ic bacteriocins can be further subdivided depending on whether they are single- or two-peptide bacteriocins.

Class II bacteriocins can be divided into Class IIa, Class IIb, Class IIc, and Class IId. Class IIa, pediocin-like bacteriocins, are typically highly active against the food pathogen *Listeria monocytogenes* and contain a conserved hydrophilic, cationic region in the N-terminal region termed the “pediocin box.” They can be subdivided into subclasses I–IV based on sequence homology. Class IIb are two-peptide unmodified bacteriocins. Both peptides are required for activity and both possess a conserved GxxG motif. These are further subdivided into two subclasses based on sequence homology. Class IIc are cyclic peptides, resulting from the covalent linkage of their N- and C-termini and tend to contain numerous α -helical structures. Class IIc can also be further divided into two subclasses based on sequence identity. Finally, the Class IId, unmodified linear, non-pediocin-like bacteriocins are in essence bacteriocins which do not fit into any of the other subclasses. (See Fig. 1 for an example of bacteriocin structure.)

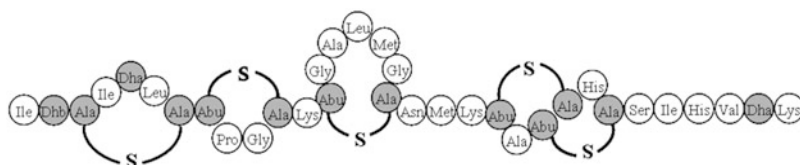
In addition to the requirement for a precursor bacteriocin peptide, bacteriocin activity is also dependent on the production of several other proteins encoded within the corresponding bacteriocin gene cluster. This gene cluster may encode proteins responsible for bacteriocin transport, processing, regulation, immunity, and, in the case of the Class I bacteriocins, peptide modification enzymes. The highly conserved accessory proteins encoded by bacteriocin gene clusters can serve as useful driver sequences for downstream analysis as the bacteriocin peptides themselves can be very diverse in their primary sequences.

Application of Bacteriocins

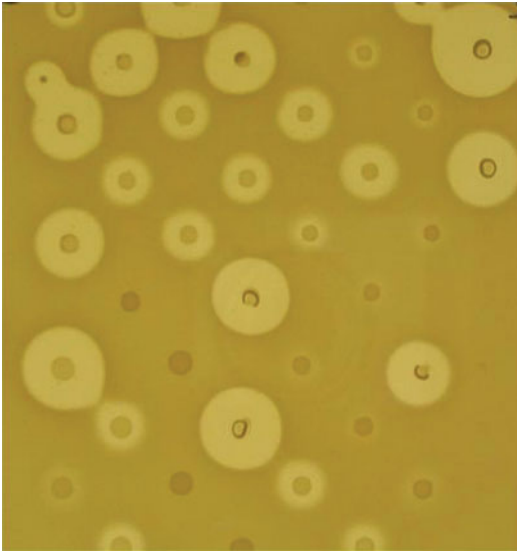
Bacteriocins have proved useful as antimicrobial compounds in the food and health industries. In the food industry, bacteriocins such as nisin and pediocin PA-1 can improve food safety and food quality. Bacteriocins produced by lactic acid bacteria (LAB) are of particular interest to the food industry since LAB have been awarded GRAS status (Generally Regarded As Safe) and can therefore be used in food preparations (Cotter et al. 2005). More recently, the contribution of bacteriocin production to the efficacy of certain probiotics has been recognized, suggesting another route via which bacteriocins can be of value within the food for health arena (Dobson et al. 2012). In the health industry, the use of bacteriocins as an alternative to antibiotics has long been mooted (Piper et al. 2009). The potential benefits of employing bacteriocins in this arena have been particularly apparent in recent times as a consequence of an appreciation of the “collateral damage” which antibiotics can inflict on the commensal microbiota. Narrow spectrum bacteriocins may well address this issue in view of their target specificity. In the area of veterinary medicine, bacteriocins have proven useful in the control of mastitis in cattle and as an additive to animal feed with a view to improving general animal health (Abriouel et al. 2011). It has also been suggested that bacteriocins or bacteriocin-producing microbes could be employed as bio-control agents which, for example, could be added to soil to control plant pathogens (Abriouel et al. 2011).

Identification of Novel Bacteriocin Gene Clusters

Traditionally, the identification of novel bacteriocin gene clusters has involved using classical



Bacteriocin Mining in Metagenomes, Fig. 1 Structure of nisin A; the prototypical Gram-positive-modified bacteriocin (Modified residues in gray)



Bacteriocin Mining in Metagenomes, Fig. 2 Representative agar plate depicting the outcome of a culture-based screen for bacteriocin activity

microbiology to screen for large collections of strains, using a culture-based assessment of their ability to produce novel antimicrobials (Fig. 2). This is then followed by the subsequent identification of the responsible genes through subcloning, mutagenesis, reverse genetics, or, more recently, sequencing of the corresponding genome. However, in spite of constant improvements in culturing techniques, it is still estimated that just 10–50 % of bacteria are culturable. Fortunately, metagenomic DNA sequencing provides an alternative with respect to identifying novel bacteriocin gene clusters by facilitating an unbiased characterization of entire microbial communities. In particular, recent improvements in sequencing technologies have resulted in a massive increase in sequence data, leading to the development of valuable public databases and annotation pipelines (<http://camera.calit2.net/>, <http://img.jgi.doe.gov/>, <http://metagenomics.anl.gov/>). The generation of vast quantities of DNA sequence data from metagenomics-based projects from varying environments across the globe represents a considerable resource from which new bacteriocin gene clusters can be identified. There are a number of ways in which this information can be harnessed. One example is

BACTIBASE, a bacteriocin database and suite of analysis tools established to archive known bacteriocin sequences and enhance the discovery of bacteriocins in genomic data (Hammami et al. 2010). The current release of BACTIBASE contains 177 bacteriocin sequences against which one can test the homology of a query bacteriocin sequence, perform sequence alignments, and predict peptide structure (Hammami et al. 2010). Searches are limited to the known sequences already in the database, and the usefulness of the tool is also affected by the fact that bacteriocin peptides themselves often share little or no homology. A specific bacteriocin mining tool, BAGEL2 (BACTERIOCIN GENOME LOCATION), was established to search for novel bacteriocin sequences in genomic data (de Jong et al. 2010). BAGEL2 has a built-in database of bacteriocin and bacteriocin-related sequences and, in addition to genes encoding the structural bacteriocin peptide, uses genes involved in bacteriocin biosynthesis, regulation, export, and immunity to reveal related genes in novel clusters. Additionally searches can be implemented against finished genome sequences or against novel genomes uploaded by the user. The fact that genes involved in the modification of Class I bacteriocins, such as those generically named *lanM*, *lanB*, and *lanC* or those encoding radical SAMs associated with sacitibiotic production, are frequently more highly conserved than the structural genes themselves has also been utilized in recent years to identify Class I gene clusters in genomic and metagenomic databases. During this period targeted searches for bacteriocins in genomic data have resulted in the discovery of several novel active bacteriocins, such as lichenicidin (Begley et al. 2009), and a *Streptococcus*-associated lantibiotic (Majchrzykiewicz et al. 2010), among others. This strategy parallels similar genome-based approaches which have identified gene clusters encoding other ribosomally synthesized natural products (Velásquez and van der Donk 2011). In addition to the identification of novel bacteriocins, the screening of genomes using the LtnM1 protein of lactacin 3147 (Begley et al. 2009; O’Sullivan et al. 2011) or the radical SAM proteins of thuricin CD (Murphy et al. 2011) as drivers has also revealed several

potential bacteriocin-encoding clusters. It is anticipated that many of these will be the focus of further investigation in the coming years.

Identification of Bacteriocins in Metagenomes

The identification of bacteriocins within metagenomic DNA can be performed via laboratory-based or *in silico*-based approaches. A recent example of the former involved a PCR-based screen to establish the bacteriocin-producing potential within metagenomic DNA sourced from 40 Polish cheeses (Więckowicz et al. 2011). In this case, PCR-primers were designed to exploit conserved sequence motifs within the four anti-listeral bacteriocin peptides, divercin V41, enterocin P, mesenteric in Y105, and bacteriocin 423. It was established that metagenomic DNA for each one of the 40 cheeses yielded a PCR product thereby highlighting the bacteriocin-producing potential of the cheese microbiota (Więckowicz et al. 2011). While laboratory-based screens have considerable potential, the vast information present in metagenomic DNA databases suggests that *in silico* screening for bacteriocin gene clusters can be a more successful approach.

Recently, two studies have carried out basic homology searches against metagenomes to identify clusters containing *lanM* genes and potentially encoding novel type II lantibiotics (O'Sullivan et al. 2011), or those possessing *trnC*/*trnD*-like genes which potentially encode novel sactibiotics

(Murphy et al. 2011). In both studies, a simple BLAST search against the CAMERA (<http://camera.calit2.net>) metagenomic databases was implemented and homologous proteins were identified. The *lanM* search revealed homologs in 11 metagenomes (Table 2). Three of these came from an Indian Ocean metagenome, four from hypersaline lagoon metagenomes from the Galapagos Islands, and one each from a coastal sea water metagenome from the Gulf of Mexico, a farm soil metagenome, a whale fall carcass rib bone metagenome, and a coral reef metagenome (Table 2). Further phylogenetic analysis with the 11 homologs and previously identified bacteriocin-like gene clusters revealed that the homologs from the metagenomes were related to other *lanM* genes from a wide variety of different microorganisms, thus highlighting the diverse nature of the metagenome-associated genes (O'Sullivan et al. 2011). The search that used *trnC*/*trnD*-like genes as driver sequences yielded 365 TrnC homologs and 151 TrnD homologs in metagenomes from environments as diverse as Waseca soil, a coral reef, and the ocean surrounding the Galapagos Islands (Murphy et al. 2011), again highlighting the presence of bacteriocin-associated genes in metagenomic data.

Despite the valuable insights provided by these analyses, they failed to identify complete bacteriocin gene clusters. A more suitable analysis tool would allow a homology search with multiple genes (or even an operon) and therefore enhance the possibility of identifying a true

Bacteriocin Mining in Metagenomes, Table 2 LanM homologs in metagenomic databases from (O'Sullivan et al. 2011)

Protein function	Metagenome	Location	% identity	E-value
Lantibiotic-modifying enzyme	Sea water	Indian Ocean	29	1.07E-16
Hypothetical protein	Soil sample	Waseca County, USA	25	2.85E-16
Lantibiotic-modifying enzyme	Whale fall rib carcass	Santa Cruz Basin, USA	27	4.65E-12
Lantibiotic-modifying enzyme	Hypersaline lagoons	Galapagos Islands	30	1.83E-08
Hypothetical protein	Coastal sea water	Gulf of Mexico	25	2.39E-08
Hypothetical protein	Hypersaline lagoons	Galapagos Islands	24	1.55E-07
Hypothetical protein	Open ocean	Indian Ocean	36	4.51E-07
Hypothetical protein	Coral reef	Cook's Bay, French Polynesia	24	5.89E-07
Hypothetical protein	Open ocean	Indian Ocean	29	1.71E-06
Mersacidin-modifying enzyme	Open ocean	Galapagos Islands	25	3.81E-06
Hypothetical protein	Hypersaline lagoons	Galapagos Islands	24	4.94E-06

bacteriocin cluster. Existing tools for metagenome analysis are in two formats: functional key word search engines, such as those available through the MG-RAST (Glass et al. 2010) and IMG/M (Markowitz et al. 2008) platforms, and homology search engines, such as JCoast (Richter et al. 2008), MetaMine (Bohnebeck et al. 2008), and CAMERA (Sun et al. 2011). Functional searches rely heavily on searching among annotated genes. This is inherently reliant on accurate annotation, and due to the small size and heterogeneous nature of bacteriocin peptides, the corresponding genes are often overlooked or mis-annotated. Homology search tools such as CAMERA and JCoast are single gene search-driven, although JCoast does have a graphical user interface that allows visualization of the surrounding gene neighborhood which would prove particularly useful for screening for the presence of other genes in the bacteriocin operons (Richter et al. 2008). Metamine allows homology searches with “gene neighborhoods”; again this would prove particularly useful for bacteriocin clusters. Metamine searches are, however, restricted to marine metagenomic databases (Bohnebeck et al. 2008). It should also be noted that, as a consequence of the evolution of DNA sequencing technologies, longer stretches of contiguous metagenomic DNA will become available which will further enhance our ability to identify complete bacteriocin gene clusters. Despite this, it must also be noted that the presence of bacteriocin homologs alone is not an indicator of function. Clearly *in silico* analysis is not sufficient to determine functional presence of a bacteriocin. However, the likelihood that even a proportion of bacteriocin homologues will be deemed functional is an intriguing prospect.

Harnessing Bacteriocin Gene Clusters

While the *in silico* analysis of newly identified bacteriocin gene clusters within metagenomic DNA can be of great value from a fundamental perspective, the harnessing of the antimicrobial potential of these clusters will undoubtedly become a priority in the future. In the majority of instances, the specific strain from which the fragment of metagenomic DNA has originated

will not be available, or may not be culturable, and other strategies will be required. The genetics-based options available can be divided into *in vivo* and *in vitro* approaches. Regardless of the approach, specific genes within the cluster will need to be regenerated through DNA synthesis technology. In the case of *in vivo* harnessing, the DNA fragment(s) will be cloned and expressed heterologously, using approaches such as those employed to facilitate the production of a *Streptococcus*-associated lantibiotic cluster by *Lactococcus lactis* (Majchrzykiewicz et al. 2010) and by *Escherichia coli*. Alternatively, when dealing with modified bacteriocins, one can clone and express individual genes heterologously but then purify them to facilitate the *in vitro* reconstitution of biosynthesis using the corresponding modification proteins or related enzymes originating from other sources (Knerr and van der Donk 2012). Finally, an alternative non-genetics-based approach, which is available when gene clusters predicted to encode unmodified residues are identified, is to employ peptide synthesis with a view of generating a synthetic equivalent of the natural antimicrobial. It is anticipated that these various options will be widely used in the years to come.

Summary

In order to effectively mine metagenomes for bacteriocins, accurate annotation of the datasets is essential. As the volume of data grows, it is anticipated that the precision of annotation tools will improve in tandem. The number of bacteriocin-associated gene homologs present in diverse metagenomic environments suggests the presence of multiple corresponding gene clusters. The further expansion of metagenomic DNA databases will undoubtedly further increase our appreciation of just how widespread, and diverse, these clusters are. As the commercial application of bacteriocins becomes more common (for review see (Cotter et al. 2005)), we can anticipate that we will reap the benefits of *in silico* screening and harnessing of this untapped reservoir of novel bacteriocins.

Cross-References

- ▶ [Ab Initio Gene Identification in Metagenomic Sequences](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)

References

- Abriouel H, Franz CMAP, Omar NB, Gálvez A. Diversity and applications of Bacillus bacteriocins. *FEMS Microbiol Rev.* 2011;35(1):201–32. doi:10.1111/j.1574-6976.2010.00244.x.
- Begley M, Cotter PD, Hill C, Ross RP. Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Appl Environ Microbiol.* 2009;75(17):5451–60. doi:10.1128/aem.00730-09.
- Bohnebeck U, Lombardot T, Kottmann R, Glockner FO. MetaMine – a tool to detect and analyse gene patterns in their environmental context. *BMC Bioinforma.* 2008;9:459. doi:10.1186/1471-2105-9-459.
- Cotter PD, Hill C, Ross RP. Bacteriocins: developing innate immunity for food. *Nat Rev Micro.* 2005;3(10):777–88.
- de Jong A, van Heel AJ, Kok J, Kuipers OP. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.* 2010;38 suppl 2:W647–51. doi:10.1093/nar/gkq365.
- Dobson A, Cotter PD, Ross RP, Hill C. Bacteriocin production: a probiotic trait? *Appl Environ Microbiol.* 2012;78(1):1–6. doi:10.1128/aem.05576-11.
- Glass EM, Wilkening J, Wilke A, Antonopoulos D Meyer F Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc.* 2010(1), pdb prot5368, doi:2010/1/pdb.prot5368 [pii] 10.1101/pdb.prot5368.
- Hammani R, Zouhir A, Le Lay C, Ben Hamida J, Fliess I. BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.* 2010;10:22. doi:10.1186/1471-2180-10-22.
- Klaenhammer TR. Bacteriocins of lactic acid bacteria. *Biochimie.* 1988;70(3):337–49. 0300-9084(88)90206-4.
- Knerr PJ, van der Donk WA. Discovery, biosynthesis, and engineering of lantipeptides. *Annu Rev Biochem.* 2012. doi:10.1146/annurev-biochem-060110-113521.
- Majchrzykiewicz JA, Lubelski J, Moll GN, Kuipers A, Bijlsma JJ, Kuipers OP, et al. Production of a class II two-component lantibiotic of *Streptococcus pneumoniae* using the class I nisin synthetic machinery and leader sequence. *Antimicrob Agents Chemother.* 2010;54(4):1498–505. doi:10.1128/AAC.00883-09.
- Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008;36 suppl 1:D534–8. doi:10.1093/nar/gkm869.
- Murphy K, O’Sullivan O, Rea MC, Cotter PD, Ross RP, Hill C. Genome mining for radical SAM protein determinants reveals multiple sactibiotic-like gene clusters. *PLoS One.* 2011;6(7):e20852. doi:10.1371/journal.pone.0020852 PONE-D-11-04704[pii].
- O’Sullivan O, Begley M, Ross R, Cotter P, Hill C. Further identification of novel lantibiotic operons using LanM-based genome mining. *Probiotics Antimicrob Protein.* 2011;3(1):27–40. doi:10.1007/s12602-011-9062-y.
- Piper C, Cotter PD, Ross RP, Hill C. Discovery of medically significant lantibiotics. *Curr Drug Discov Technol.* 2009;6(1):1–18.
- Rea M, Cotter P, Hill C, Ross R. Classification of bacteriocins from gram-positive bacteria. In: Drider D, Rebuffat S, editors. *Prokaryotic antimicrobial peptides - from genes to applications.* New York: Springer; 2011. p. 29.
- Richter M, Lombardot T, Kostadinov I, Kottmann R, Duhaime MB, Peplies J, et al. JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinforma.* 2008;9:177. doi:10.1186/1471-2105-9-177.
- Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* 2011;39(Database issue):D546-551. doi:10.1093/nar/gkq1102.
- Velásquez JE, van der Donk WA. Genome mining for ribosomally synthesized natural products. *Curr Opin Chem Biol.* 2011;15(1):11–21.
- Więckowicz M, Schmidt M, Sip A, Grajek W. Development of a PCR-based assay for rapid detection of class IIa bacteriocin genes. *Lett Appl Microbiol.* 2011;52(3):281–9. doi:10.1111/j.1472-765X.2010.02999.x.

Binning Sequences Using Very Sparse Labels Within a Metagenome

Ching-Hung Tseng¹, Chon-Kit Kenneth Chan², Arthur L. Hsu², Saman K. Halgamuge² and Sen-Lin Tang³

¹Bioinformatics Program, Taiwan International Graduate Program, Biodiversity Research Center, Institute of Information Science, Academia Sinica, Taipei, Taiwan

²Department of Mechanical Engineering, The University of Melbourne, Melbourne, VIC, Australia

³Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan

Synonyms

Binning using seeded GSOM

Definition

Binning is the process to categorize sequences into different groups based on compositional features or sequence similarity or both of them.

Introduction

As metagenomes are typically composed of sequences from various species, how to categorize these sequences into groups can radically affect the accuracy and sensitivity of downstream analyses. Thus, the sequence binning is a critical step in the early process of metagenomic analysis pipeline. Several binning methods employing different strategies have been proposed. For example, BLAST homology search helps to identify sequences of related species; *kmer* (Sandberg et al. 2001), self-organizing map (SOM) (Abe et al. 2003), and TETRA (Teeling et al. 2004b) cluster sequences by similar compositional features, i.e., oligonucleotide frequency; PhyloPythia (McHardy et al. 2007), a support vector machine implementation, categorizes sequences based on both pattern similarity and oligonucleotide frequency. The above listed methods, either supervised or unsupervised, have their own limitations. Supervised learning methods, including BLAST, *kmer*, SOM, and PhyloPythia, require prior knowledge, like completed genomes or labeled long contigs, as training datasets; the unsupervised learning method, TETRA, may become intractable for huge metagenomes because of the required computation on all-versus-all pair-wise comparison. Although these methods have demonstrated great feasibility, a solution to resolve bins without identifiable labels makes the binning task more applicable to typical metagenomes. In our context, we introduce a semi-supervised learning method that couples a seeding strategy with the growing self-organizing map (GSOM), called the seeded GSOM (S-GSOM), for sequence binning.

Self-Organizing Map and Growing Self-Organizing Map Algorithms

The self-organizing map (SOM) (Kohonen 1990) is an unsupervised clustering algorithm. It can visualize the clustering of unlabeled feature vectors on a static lattice grid map, which has pre-defined grid shape and map size, during the entire training process. Within the map, every node (or lattice) has a weight vector of the same dimension as the input vector. The SOM algorithm separates training into three phases: initialization, ordering, and fine-tuning. In the initialization phase, the weight vector of each initial node can be either generated from random values or, generally, using the principal component analysis (PCA) to position a fully unfolded map on the plane formed by the first two principal vectors in the input space (Kohonen 1999). The number of initial nodes needs to be determined by the user. In the ordering and fine-tuning phases, each input identifies a winning node, which is of the smallest Euclidean distance to the input, on the map. Then the weight vectors of the winning node and its neighboring nodes are updated by

$$w(t+1) = w(t) + \alpha \times h \times [x(k) - w(t)],$$

where w is the weight vector of the node, x is the input vector ($w, x \in R^D$ where D is the dimension), k is the index of the current input vector, α is the learning rate, and h is the neighborhood kernel function.

The Growing SOM (GSOM) (Alahakoon et al. 2000; Hsu and Halgamuge 2003) is an extension of SOM. It is a dynamic SOM, which overcomes SOM's weakness of the static map structure, i.e., GSOM initiates its training with minimum single lattice grid, depending on whether the rectangular or hexagonal network topology is used, to facilitate the dynamic growth of the map in training process. GSOM employs the same weight adaptation and neighborhood kernel function as SOM. The map size of a perfectly trained GSOM map is controlled by a global

parameter of growth, which is called Growth Threshold (GT) and defined as

$$GT = -D \times \ln(SF),$$

where D is the data dimension and SF is the user-defined Spread Factor that takes value $(0, 1]$, with 0 representing minimum and 1 representing maximum growth.

There are four phases in GSOM training: initialization, growing, and two smoothing phases. In the initialization phase, weight vectors of initial nodes in the minimum single lattice grid are initialized by random values and the GT is calculated according to data dimension and user-defined SF. During the growing phase, every node keeps an accumulated error counter and the counter of the winning node (E_{winner}) is updated by

$$E_{\text{winner}}(t + 1) = E_{\text{winner}}(t) + |x(k) - w_{\text{winner}}(t)|.$$

When E_{winner} exceeds GT, the winning node that is at the boundary of current map will grow new nodes to its neighboring vacant lattice and initialize a weight vector by interpolating or extrapolating weight vectors of existed neighboring nodes around the winning node. If the winning node is not a boundary node, the accumulated error (E_{winner}) is evenly distributed outwards to its neighbors. The two smoothing phases are for fine-tuning the weights of nodes. The hexagonal lattice was used for GSOM in this study as the hexagonal lattice yields better data topology preservation (Hsu et al. 2003).

Seeding Sequence and Metagenomic Dataset Preparation

From the NCBI Archaea/Bacteria genome database, we randomly selected 10, 20, and 40 species to generate metagenomes of different complexity. Three sets were drawn for the 10 and 20 species datasets, and only one set for the 40 species dataset due to the limitation imposed by the available computing resources. Simulated metagenomes were denoted by “XSp_SetY”

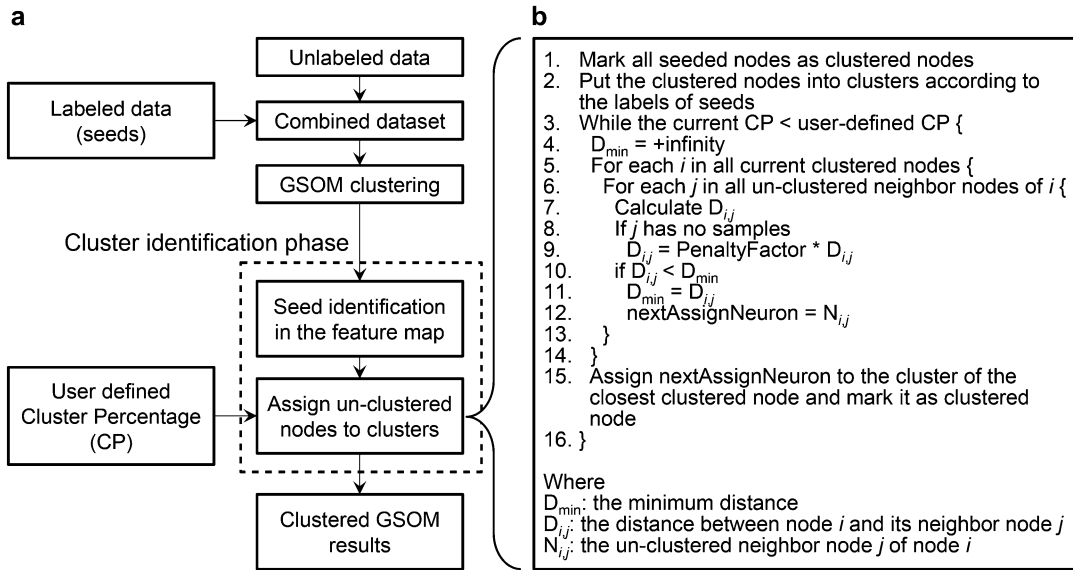
where X is the number of species included in the metagenome and Y represents the serial number of metagenome. For example, “10Sp_Set1” is the first metagenome containing 10 random species.

In each genome, the seeding sequences were firstly identified as the flanking region of 16S rRNA genes of the length ranging from 8 to 13 kilobase (kb). The seeding sequences that overlapped with other rRNA and tRNA genes were excluded to avoid possible interferences caused by highly similar sequence compositions. After removing the tRNA, rRNA, and seeding sequences, the remaining genomic regions were randomly chopped into simulated metagenomic fragments of the length from 8 to 13 kb. The length restriction of 8–13 kb is used to provide a standardized rule for either seeding or metagenomic sequences (Mavromatis et al. 2007), but with the outlook for single-molecule sequencing techniques on the horizon (Clarke et al. 2009), these are definitely achievable length for metagenomes in the near future.

The tetranucleotide frequency of metagenomic sequences is the training feature we used in our implementation for binning because it has a better resolution in species separation (Abe et al. 2003) and is highly similar between intragenomic fragments compared to intergenomic fragments. The tetranucleotide frequencies were computed using a four-base sliding window and normalized by the length of the corresponding sequence (frequency per base). A total of 256 (4^4) combinations of nucleotide usages, i.e., AAAA, AAAT, AAAG, AAAC ... CCCG, and CCCC, are represented in the feature vector of 256 dimensions.

Seeded GSOM Algorithm

Metagenomic sequences that belong to closely related species are likely to have homologous sequences between the clusters (bins), and this fact makes the identification of clustering boundaries much more difficult. Therefore, a modified strategy is needed to identify clusters so that GSOM can be improved as a more practical solution for binning.



Binning Sequences Using Very Sparse Labels Within a Metagenome, Fig. 1 The S-GSOM algorithm. (a) Schematic diagram of the clustering process of

S-GSOM. (b) The pseudo-code for node assigning process in S-GSOM

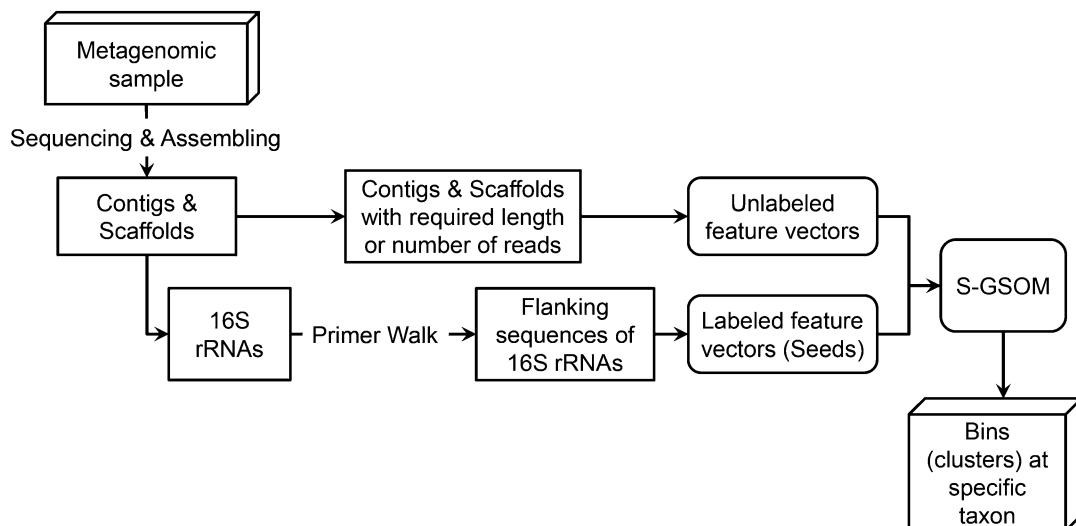
The seeded GSOM (S-GSOM), which allows identifying clusters automatically in the feature map using seeds (labeled data), is our proposed modification of GSOM. There are three core steps in S-GSOM. Firstly, the very small amounts of labeled seeds (labeled feature vectors) are combined with unlabeled data (unlabeled feature vectors). Secondly, the combined input vectors are fed into GSOM training, which treats the seeds as the unlabeled data. Finally, after the normal phases of GSOM training, S-GSOM identifies clusters based on the location of seeds in the final map and the specified amount of nodes in the cluster (Fig. 1a).

In the last step of S-GSOM training, the cluster identification phase, the nodes that have seeds are identified and labeled as clustered nodes. Then the S-GSOM is going to assign other un-clustered nodes, one by one, to clusters iteratively until the specified clustering percentage (more details in Clustering Percentage (CP) Determination section) is reached. In each iteration, a set of un-clustered nodes that are adjacent to the clustered nodes is identified. The node in the set of the shortest Euclidean distance to the adjacent clustered node will be assigned to the same cluster with the clustered node. However, nodes not

containing any sample are most likely representing a cluster boundary. So a penalty factor greater than one is multiplied to the actual distance when calculating the distance between empty nodes and clustered nodes. This will lead the S-GSOM not to label empty nodes to any cluster (Fig. 1b). According to the empirical observation that the clustering results are not very sensitive to the penalty factor value between 2 and 5, the penalty factor value of 2 was used in all our experiments.

Before the initiation of the taxonomy-assigning process, the seeded nodes must be assigned to a specific taxon. When all seeds in one node are coming from the same taxon or there is only a single seed, it is trivial for S-GSOM to assign the seeded node to the same taxon as contained seeds. If the seeds in one node belong to multiple taxa, the seeded node will be assigned to the major taxon. However, when seeds are of multiple taxa and have equal amounts, e.g., two seeds are in one node and belonging to taxon A and B, respectively, all seeds are discarded.

To illustrate the role of S-GSOM in binning, Fig. 2 depicts the schematic diagram that explains how S-GSOM fits into the whole binning process.



Binning Sequences Using Very Sparse Labels Within a Metagenome, Fig. 2 An overview of binning process using S-GSOM

Clustering Percentage (CP) Determination

Because metagenomic sequences of closely related species occur frequently at the cluster boundary (Abe et al. 2003; Chan et al. 2008b) that is very likely to disrupt the binning accuracy, an appropriate control of how many nodes assigned to bins is necessary for S-GSOM to have a trade-off between the amount of binned sequences and the binning accuracy. Hence, the clustering percentage (CP) value is introduced, which is defined as the percentage of the number of clustered nodes relative to the total nodes on the map. It was noted that the performance of S-GSOM declined when CP was higher than 55 % (Fig. 3). However, S-GSOM binned more than 80 % of sequences at CP = 55 % in most cases. Thus, the 55 % CP was used throughout the following experiments.

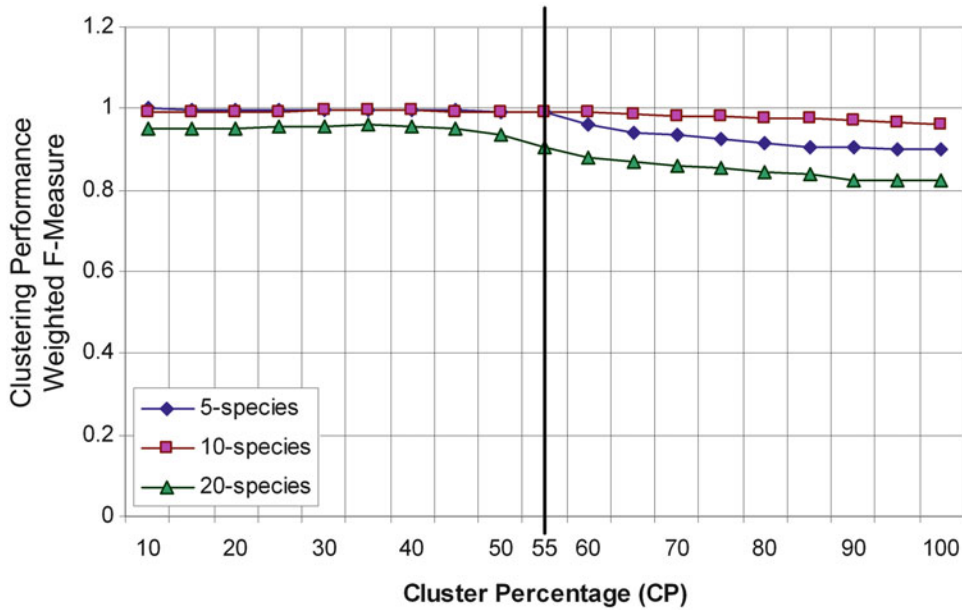
Comparison of Semi-supervised Algorithms for Binning

To test the feasibility of semi-supervised methods for binning, other four notable semi-supervised clustering algorithms, COP K-means,

Constrained K-means, Seeded K-means, and Transductive Support Vector Machine (TSVM) were used alongside S-GSOM. Among above methods, different runs of random initiation of the COP K-means and S-GSOM can lead to diverse results, which is not an issue for Constrained K-means and Seeded K-means because they use the labeled data for initiation. So the best results of COP K-means in 100 runs of random initiations were reported, and to ensure repeatability, all the feature vectors in S-GSOM's initialization were fixed with the mid value 0.5 in all dimensions.

Two indices were used to measure clustering performance: adjusted Rand index (ARI) (Hubert and Arabie 1985) and weighted F-measure (WF) (Van Rijsbergen 1979). The higher index indicates the better performance.

S-GSOM manifested consistently superior performance on both measures, ARI and WF, with the exception of Constrained K-means on the ARI measure for the 10Sp_Set3 dataset (Table 1). We suspect the considerable worse performance of TSVM as resulting from insufficient labeled data. The superior performance of S-GSOM, which accurately assigned 75–90 % of all sequences at CP = 55 %, clearly demonstrates that the adjustable CP value effectively



Binning Sequences Using Very Sparse Labels Within a Metagenome, Fig. 3 Identification of an appropriate clustering percentage (CP). Five datasets for each of 5, 10, and 20 species are randomly samples. The average of S-GSOM’s performance for the datasets are plotted

against CP. A trend of decreasing in performance with increasing in CP can be noted. A compromised value of CP = 55 % is marked where both the number of assigned nodes and performance are high

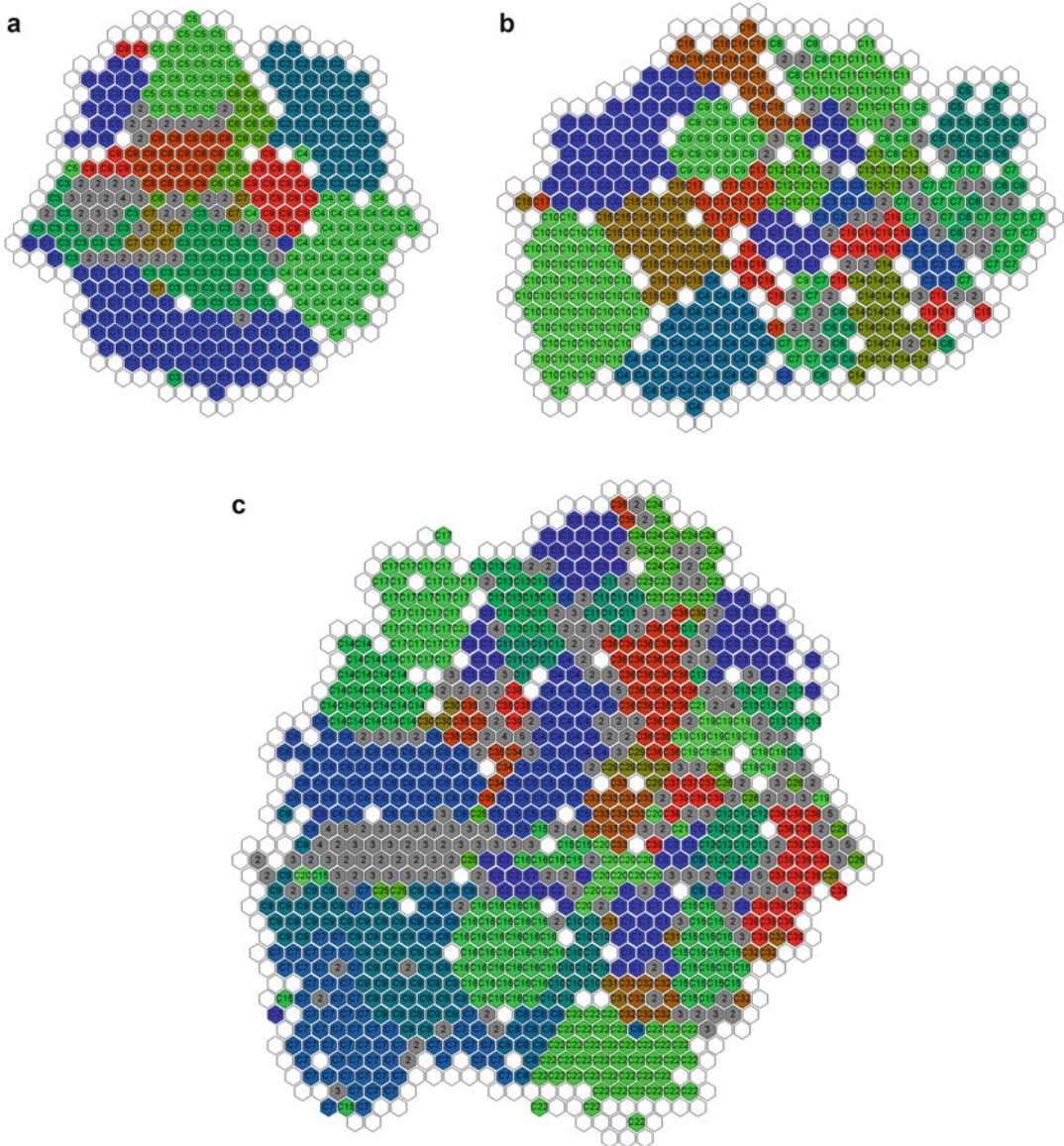
Binning Sequences Using Very Sparse Labels Within a Metagenome, Table 1 Clustering performance of semi-supervised algorithms. Performance is measured by the adjusted Rand index (ARI) and weighted F-measure (WF)

	COP K		Constrained K		Seeded K		TSVM		S-GSOM-55	
	ARI	WF	ARI	WF	ARI	WF	ARI	WF	ARI	WF
10Sp_Set1	0.84	0.94	0.84	0.94	0.84	0.93	0.25	0.59	0.85	0.95
10Sp_Set2	0.89	0.96	0.79	0.90	0.78	0.90	0.41	0.69	0.93	0.97
10Sp_Set3	0.58	0.83	0.85	0.93	0.84	0.93	0.27	0.62	0.83	0.93
20Sp_Set1	0.91	0.90	0.77	0.82	0.76	0.82	0.45	0.65	0.97	0.96
20Sp_Set2	0.76	0.82	0.70	0.79	0.67	0.79	0.43	0.62	0.83	0.89
20Sp_Set3	0.81	0.89	0.75	0.86	0.75	0.86	0.46	0.67	0.97	0.98
40Sp	0.58	0.76	0.71	0.85	0.68	0.84	0.24	0.56	0.83	0.91

helps S-GSOM to achieve better clustering by not assigning those ambiguous sequences. The S-GSOM visualization of binning sequences of 10Sp_Set1, 20Sp_Set1, and 40Sp is provided in Fig. 4.

We considered the 20-species metagenomes as examples to analyze the resolution of binning with S-GSOM. At CP = 55 %, an average of 82 % sequences were assigned with 92 % accuracy to their source species. The distribution of

binning result is shown in Fig. 4b. Nodes containing seeds from multiple species were colored in grey with the label of species number. A significantly higher abundance of grey nodes around “C6” and “C7,” respectively representing *Haemophilus influenzae* 86-028NP and *Haemophilus somnus* 129PT, indicates that metagenomic sequences with similar tetranucleotide frequencies, resulted from closely related species, tend to be clustered without



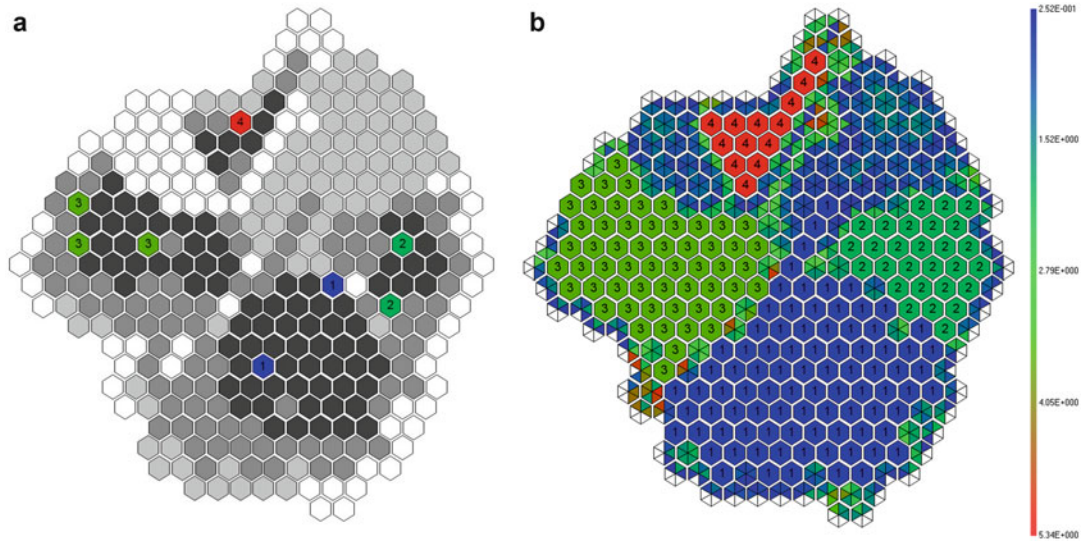
Binning Sequences Using Very Sparse Labels Within a Metagenome, Fig. 4 Resulted growing self-organizing maps (GSOM) of (a) 10Sp_Set1, (b) 20Sp_Set1, and (c) 40Sp metagenomes. Each hexagon represents a single node. If the node contains a single

species, it is displayed in a color that uniquely identifies the species. The node without a letter indicates that there is no data (sequences) located in it. The grey nodes represent multiple species in the node, and the exact number is as labeled

a clear boundary. This further highlights the importance of obtaining seeds in non-boundary regions.

In addition to the distinguished clustering performance, S-GSOM possesses a prominent advantage brought by the seeding method to

cluster sequences of unseeded species, i.e., the unknown species. To demonstrate this advantage, an iso-CP (constant CP) contour is delineated in Fig. 5a, generated with a five-species metagenome with only four seeds. By applying different CP values, a group of nodes were found



Binning Sequences Using Very Sparse Labels Within a Metagenome, Fig. 5 Illustration of exploring an unseeded cluster. (a) The five-species S-GSOM map. The seeded nodes are shown with unique colors and labels. Nodes in charcoal color represent nodes that will

be assigned when CP = 27 % and *dark grey* nodes at CP = 55 %, *light grey* at CP = 77 %, and *white* at CP = 100 %. (b) Internode distance map with nodes assigned at CP = 55 %

only clustered at CP = 77 % (on the top-right region). This situation is most likely when a species is relatively abundant, but does not have a seed. Figure 5b shows the allocation of nodes to seeds at CP = 55 %. However, a protrusion of species “1” into the unassigned region, which belongs to species “5,” is an incorrect assignment that sometimes happens to nodes without a correct seed.

Comparison of Binning Fidelity Using S-GSOM

In this section, we compared the binning performance of S-GSOM with three other methods: BLAST, *kmer*, and PhyloPythia, reported on the metagenomes of different complexities (Mavromatis et al. 2007) after assembly by Arachne (Batzoglou et al. 2002), Phrap (Green, 1996), and JAZZ (Aparicio et al. 2002). However, JAZZ produced small number of contigs compared to Arachne and Phrap (Mavromatis et al. 2007), so contigs assembled by JAZZ were excluded. In addition, because the simHC,

a community without any dominant species, has sparse long contigs required by the composition-based analysis (Mavromatis et al. 2007; Teeling et al. 2004a), we also excluded the simHC dataset from our analysis.

For the purpose of fair comparison, all methods need to be compared at the same taxonomic level of binning. Binning at a very high level, e.g., kingdom, clearly has no significance; therefore, the results are compared at the order level here and results for comparing at other taxonomic levels are included in the supplementary materials of original publication (Chan et al. 2008a). At the order level, the results for simLC (low complexity) and simMC (medium complexity) metagenomes are shown in two separated tables, one for binning contigs larger than 8 kb and the other for contigs composed of at least 10 reads. To evaluate the performance, rather than using simple averages of all bins (Mavromatis et al. 2007), we used weighted average that gives higher weights to larger bins to better reflect the amount of correctly binned contigs.

In both simLC and simMC, S-GSOM performed reasonably for binning contigs larger than 8 kb, where it is more accurate than all

Binning Sequences Using Very Sparse Labels Within a Metagenome, Table 2 Binning summary for low complexity metagenome for contigs larger than 8 kb

Assembler	Method	Binned		Total#Contigs	%ofBinContigs	#ofPredNotInAct	wSp	wSn
		Bins	Contigs					
Arachne	kmer (7 mer)	0	0	202	0	85	–	0.000
Arachne	kmer (8 mer)	0	0	202	0	149	–	0.000
Arachne	BLAST distr 1	0	0	202	0	0	–	0.000
Arachne	BLAST distr 2	0	0	202	0	0	–	0.000
Arachne	S-GSOM (CP = 55 %)	1	141	202	69.8	0	1.000	0.698
Arachne	gen PhyloPythia (p:0.85)	1	168	202	83.17	0	1.000	0.832
Arachne	ssp. PhyloPythia (p:0.85)	1	186	202	92.08	0	1.000	0.921
Arachne	S-GSOM (CP = 75 %)	1	180	202	89.11	0	1.000	0.891
Arachne	gen PhyloPythia (p:0.5)	1	201	202	99.5	0	1.000	0.995
Arachne	ssp. PhyloPythia (p:0.5)	1	201	202	99.5	0	1.000	0.995
Phrap	kmer (7 mer)	0	0	229	0	129	–	0.000
Phrap	kmer (8 mer)	0	0	229	0	154	–	0.000
Phrap	BLAST distr 1	0	0	229	0	0	–	0.000
Phrap	BLAST distr 2	0	0	229	0	0	–	0.000
Phrap	S-GSOM (CP = 55 %)	1	157	229	68.56	0	1.000	0.686
Phrap	gen PhyloPythia (p:0.85)	1	185	229	80.79	0	1.000	0.808
Phrap	ssp. PhyloPythia (p:0.85)	1	205	229	89.52	0	1.000	0.895
Phrap	S-GSOM (CP = 75 %)	1	204	229	89.08	0	1.000	0.891
Phrap	gen PhyloPythia (p:0.5)	1	227	229	99.13	0	1.000	0.991
Phrap	ssp. PhyloPythia (p:0.5)	1	227	229	99.13	0	1.000	0.991

Total#Contigs total number of contigs in the dataset, *%ofBinContigs* the percentage of contigs binned, *#ofPredNotInAct* the number of contigs predicted as a taxon that is not present in the dataset, which are treated as the un-binned contigs, *wSp* weighted specificity, *wSn* weighted sensitivity

Binning Sequences Using Very Sparse Labels Within a Metagenome, Table 3 Binning summary for medium complexity metagenome for contigs larger than 8 kb

Assembler	Method	Binned		Total#Contigs	%ofBinContigs	#ofPredNotInAct	wSp	wSn
		Bins	contigs					
Arachne	kmer (7 mer)	0	0	301	0	47	–	0.000
Arachne	kmer (8 mer)	0	0	301	0	191	–	0.000
Arachne	BLAST distr 1	0	0	301	0	0	–	0.000
Arachne	BLAST distr 2	0	0	301	0	0	–	0.000
Arachne	S-GSOM (CP = 55 %)	2	220	301	73.09	0	1.000	0.731
Arachne	gen PhyloPythia (p:0.85)	2	242	301	80.4	0	1.000	0.804
Arachne	ssp. PhyloPythia (p:0.85)	2	242	301	80.4	0	1.000	0.804
Arachne	S-GSOM (CP = 75 %)	2	279	301	92.69	0	1.000	0.927
Arachne	gen PhyloPythia (p:0.5)	2	301	301	100	0	1.000	1.000
Arachne	ssp. PhyloPythia (p:0.5)	2	301	301	100	0	1.000	1.000
Phrap	kmer (7 mer)	0	0	401	0	84	–	0.000
Phrap	kmer (8 mer)	0	0	401	0	271	–	0.000
Phrap	BLAST distr 1	0	0	401	0	0	–	0.000
Phrap	BLAST distr 2	0	0	401	0	0	–	0.000
Phrap	S-GSOM (CP = 55 %)	2	318	401	79.3	0	1.000	0.793
Phrap	gen PhyloPythia (p:0.85)	2	301	401	75.06	0	1.000	0.751
Phrap	ssp. PhyloPythia (p:0.85)	2	295	401	73.57	0	1.000	0.736
Phrap	S-GSOM (CP = 75 %)	2	367	401	91.52	0	1.000	0.915
Phrap	gen PhyloPythia (p:0.5)	2	399	401	99.5	1	1.000	0.995
Phrap	ssp. PhyloPythia (p:0.5)	2	399	401	99.5	1	1.000	0.995

settings of *kmer* and BLAST methods, but was outperformed by PhyloPythia in both confidence settings (CP = 75 % vs. *p*-value = 0.5 and CP = 55 % vs. *p*-value = 0.85) regardless of the assembler used (Tables 2 and 3). Nevertheless,

S-GSOM still outperformed PhyloPythia for the *simMC*, particularly in terms of sensitivity, i.e., having a higher true positive rate, at the family level (refer to the supplementary materials of original publication).

Binning Sequences Using Very Sparse Labels Within a Metagenome, Table 4 Binning summary for low complexity metagenome for contigs with at least 10 reads

Assembler	Method	Binned		Total#Contigs	%ofBinContigs	#ofPredNotInAct	wSp	wSn
		Bins	Contigs					
Arachne	kmer (7 mer)	0	0	367	0	168	–	0.000
Arachne	kmer (8 mer)	0	0	367	0	312	–	0.000
Arachne	BLAST distr 1	0	0	367	0	0	–	0.000
Arachne	BLAST distr 2	0	0	367	0	0	–	0.000
Arachne	S-GSOM (CP = 55 %)	3	295	367	80.38	0	1.000	0.798
Arachne	gen PhyloPythia (p:0.85)	2	214	367	58.31	0	1.000	0.583
Arachne	ssp. PhyloPythia (p:0.85)	2	236	367	64.31	0	1.000	0.638
Arachne	S-GSOM (CP = 75 %)	3	343	367	93.46	0	0.950	0.926
Arachne	gen PhyloPythia (p:0.5)	2	292	367	79.56	0	1.000	0.796
Arachne	ssp. PhyloPythia (p:0.5)	2	296	367	80.65	0	1.000	0.798
Phrap	kmer (7 mer)	2	3	482	0.62	159	1.000	0.000
Phrap	kmer (8 mer)	3	17	482	3.53	281	1.000	0.000
Phrap	BLAST distr 1	0	0	482	0	0	–	0.000
Phrap	BLAST distr 2	0	0	482	0	1	–	0.000
Phrap	S-GSOM (CP = 55 %)	8	381	482	79.05	9	1.000	0.728
Phrap	gen PhyloPythia (p:0.85)	3	236	482	48.96	0	1.000	0.488
Phrap	ssp. PhyloPythia (p:0.85)	3	272	482	56.43	0	1.000	0.560
Phrap	S-GSOM (CP = 75 %)	8	443	482	91.91	9	1.000	0.840
Phrap	gen PhyloPythia (p:0.5)	4	368	482	76.35	1	1.000	0.759
Phrap	ssp. PhyloPythia (p:0.5)	5	387	482	80.29	1	1.000	0.797

At the order level, while PhyloPythia performed best for all binning tests on contigs larger than 8 kb, our S-GSOM was the best-performing method when used to bin contigs that contain at least 10 reads (Tables 4 and 5).

Discussion

By including sequences with taxonomic information, i.e., seeds, S-GSOM exhibits more feasibility in binning task for metagenomes containing many unknown species. The visualization

Binning Sequences Using Very Sparse Labels Within a Metagenome, Table 5 Binning summary for medium complexity metagenome for contigs with at least 10 reads

Assembler	Method	Binned		Total#Contigs	%ofBinContigs	#ofPredNotInAct	wSp	wSn
		Bins	Contigs					
Arachne	kmer (7 mer)	1	2	1,372	0.15	133	1.000	0.000
Arachne	kmer (8 mer)	0	0	1,372	0	1,241	–	0.000
Arachne	BLAST distr 1	0	0	1,372	0	0	–	0.000
Arachne	BLAST distr 2	0	0	1,372	0	1	–	0.000
Arachne	S-GSOM (CP = 55 %)	5	1,061	1,372	77.33	0	0.998	0.768
Arachne	gen PhyloPythia (p:0.85)	3	562	1,372	40.96	0	1.000	0.409
Arachne	ssp. PhyloPythia (p:0.85)	3	657	1,372	47.89	0	1.000	0.478
Arachne	S-GSOM (CP = 75 %)	5	1,253	1,372	91.33	0	0.983	0.897
Arachne	gen PhyloPythia (p:0.5)	4	1,036	1,372	75.51	6	1.000	0.753
Arachne	ssp. PhyloPythia (p:0.5)	4	1,102	1,372	80.32	4	1.000	0.802
Phrap	kmer (7 mer)	1	1	1,980	0.05	163	1.000	0.000
Phrap	kmer (8 mer)	2	391	1,980	19.75	1,457	1.000	0.000
Phrap	BLAST distr 1	0	0	1,980	0	2	–	0.000
Phrap	BLAST distr 2	0	0	1,980	0	3	–	0.000
Phrap	S-GSOM (CP = 55 %)	8	1,409	1,980	71.16	9	0.995	0.686
Phrap	gen PhyloPythia (p:0.85)	3	799	1,980	40.35	1	1.000	0.404
Phrap	ssp. PhyloPythia (p:0.85)	3	844	1,980	42.63	1	1.000	0.426
Phrap	S-GSOM (CP = 75 %)	8	1,708	1,980	86.26	9	0.991	0.816
Phrap	gen PhyloPythia (p:0.5)	5	1,484	1,980	74.95	6	1.000	0.745
Phrap	ssp. PhyloPythia (p:0.5)	5	1,524	1,980	76.97	4	1.000	0.767

property of S-GSOM further allows the identification of unseeded clusters. However, the sequence number of unseeded species should be at least as many as in the seeded clusters; otherwise, S-GSOM may wrongly assigned the unseeded species to an unrelated species at low

CP value or be considered as part of the boundary of neighboring clusters and thus become hardly detectable.

It is very likely that the 16S rRNA fragments of some species were not or difficult to be sampled. In such circumstances, we can still obtain

those metagenomic sequences in the possible bins, which have been identified by using the iso-CP contour map, then comparing the sequences with existing databases by BLAST search. If any conserved marker gene is detected, such as elongation factors and cytochrome oxidase, then we may assess the clusters of these sequences by phylogenetic analysis.

Even though these composition-based binning methods have shown good results, currently they are hindered by the requirement of long sequence length. This limitation of length is partially due to the occurrence of chimeric sequences from cloning procedures of experiments and from the incorrect assembly of sequences. The former source of chimeric sequences can be reduced by advanced cloning-free sequencing, e.g., Roche 454 genome sequencer FLX. However, the latter source of chimeric sequence is derived from the incompatible design of current assembler, which assembles all reads into one single genome and may not satisfy the requirement of metagenomes of poor sequencing coverage or of high species complexity. Therefore, if the number of chimeric sequences is reduced, the required sequence length in S-GSOM can also be reduced. To help the reduction of chimeric sequences, we suggest including the compositional information in the assembling level.

Summary

S-GSOM enables the clustering (binning) of metagenomic sequences by incorporating sparse sequence fragments, with phylogenetic labels, around highly conserved genes as seeds. The application of seeds makes S-GSOM more feasible when dealing with metagenomes containing many unknown species, which can be visualized using CP contour display. In addition, S-GSOM is also an efficient algorithm in terms of the training time. By adjusting the CP value, users can retrieve different clustering results without retraining. The nature of self-organizing indeed forms S-GSOM an automated process that can be improved when new seeds are available.

References

- Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. Informatics for unveiling hidden genome signatures. *Genome Res.* 2003;13:693–702.
- Alahakoon D, Halgamuge SK, Srinivasan B. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Trans Neural Netw.* 2000;11:601–14.
- Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science.* 2002;297:1301–10.
- Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: a whole-genome shotgun assembler. *Genome Res.* 2002;12:177–89.
- Chan CK, Hsu AL, Halgamuge SK, Tang SL. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics.* 2008a;9:215.
- Chan CK, Hsu AL, Tang SL, Halgamuge SK. Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *J Biomed Biotechnol.* 2008b;2008(513701):p 10. doi:10.1155/2008/513701
- Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol.* 2009;4:265–70.
- Green P. Documentation for PHRAP. 1996; <http://bozeman.mbt.washington.edu/>
- Hsu AL, Halgamuge SK. Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation. *Int J Approx Reason.* 2003;32:259–79.
- Hsu AL, Tang S-L, Halgamuge SK. An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. *Bioinformatics.* 2003;19:2131–40.
- Hubert L, Arabie P. Comparing partitions. *J Classif.* 1985;2:193–218.
- Kohonen T. The self-organizing map. *Proc IEEE.* 1990;78:1464–80.
- Kohonen T. Analysis of processes and large data sets by a self-organizing method. *Intell Process Manuf Mater.* 1999;1:27–36.
- Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4:495–500.
- McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4:63–72.
- Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J. Capturing whole-genome characteristics in

- short sequences using a naive Bayesian classifier. *Genome Res.* 2001;11:1404–9.
- Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol.* 2004a;6:938–47.
- Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics.* 2004b;5:163.
- Van Rijsbergen CJ. *Information retrieval.* London: Butterworths; 1979.

Biological Treasure Metagenome

Geun-Joong Kim and Ho-Dong Lim
Department of Biological Sciences, College of
Natural Sciences, Chonnam National University,
Gwangju, Republic of Korea

Synonyms

The economic and academic values of metagenome resources

Definition

Given that the possibility and frequency of finding novel genes, enzymes, and metabolites through conventional pure culture technology are decreasing gradually, exploration of resources hidden in the metagenome as a treasure of new resources is expected to provide a new breakthrough. The metagenome is currently the most promising candidate for exploring new biological resources, and therefore there will be continuous efforts for refining strategies and developing new protocols. Through research using the metagenome, we can measure microbial diversity, understand ecosystems through a window into the microbial genome contents in a specific environmental habitat and explore useful resources, and then ultimately incorporate them into the process of practical uses.

Introduction

The coming of a new era – the “metagenome age” – that accesses the genomes of all microbes retrieved directly from environmental samples paves a new way for the understanding and practical application of microbial resources (Hunter-Cevera 1998). The metagenome will provide a revolutionary solution to offer powerful tools for understanding the microbial world that has the potential to uncover constituents of the entire living organism for valuable use in various fields such as agricultural, medicinal, and industrial biotechnology.

Microbes have currently recognized as being possessed the most extensive genetic biodiversity. They have proliferated in the ecosystem on Earth for a long age (3.5–4.2 billion years) and thus evolved fittingly to various habitats seemingly incompatible with life (from a conventional point of view) such as the animal gut, desert, Antarctic ice, and hot springs. Accordingly, their taxonomical species and metabolic functions are also more diverse than expected. Therefore, microbial consortia are a treasure of resources with infinite value in basic research and practical applications. Since the appearance of mankind on Earth, microbes and humans have maintained a close relationship through both direct and indirect interactions. In the view of long direct interaction, the roles of microbes in human nutrition and health are established through integrated research – the Human Microbiome Project – that aims to characterize the microbial communities of the human body, including nasal passages, oral cavities, skin, gastrointestinal tract, and urogenital tract (Lewis et al. 2012).

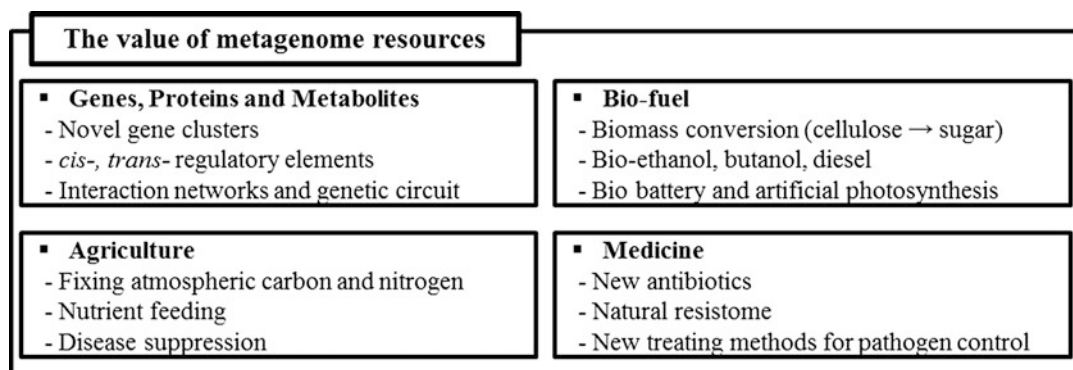
Accessing the microbial diversity from various niches through the approach using metagenome has been presented to provide valuable resources and clues for the applications of microbes in human health and industry, as well as for scientific research on the ecosystem function, the global biogeochemical cycle, and the origin of life. However, microbial species known as potential candidates in industry and those with

elucidated functions in ecosystems are only those that can be cultured, which are estimated to be about <1 % of microbes existing in nature; thus, most microbes are recognized as unculturable species (Handelsman 2004). Accordingly, the strategy to overcome the limitation of pure culture and thus to explore the entire microbial resource have attempted, which has induced a new paradigm shift.

Metagenomics is a research area that studies the metagenome which is the total genomes of all organisms existing in a certain habitat. It extracts DNA from a complex microbial community and analyzes the information of genomes using molecular biological tools mainly based on direct sequencing. Therefore, metagenomics is a microbial community analysis method to access all contents of microbial genomes, which goes beyond the limited scope of cultivated cells. Metagenomic research has been revolutionized by the development of genome-manipulating technologies, and despite its short history, new functional genes, proteins, and biomaterials have been mined successfully (Xu 2006). Comprehensive understanding has also been attained on the ecology and physiology of microbes. In addition, a huge amount of sequence information derived from metagenome is integrated using bioinformatic tools. Accordingly, the scope of application in the entire range of biotechnology has altered based on the potential value of the metagenome (Fig. 1).

The Value of Metagenome Resources Exploited

The results of gene prediction, annotation on the sequence, and metabolic assembly through (individual) genome reconstruction of metagenome give not only the understanding of microbial ecology and physiology but also the expectation that useful genetic resources and the whole synthetic pathway of specific compounds *in vivo* are readily explored. With the development of the amplification tools for rare DNAs and technology related to high-throughput sequencing of DNA, it is now possible to analyze and understand the function of individual species in the whole community of natural strains. As an example, the elucidation of broad distribution of non-extreme ammonia-oxidizing archaea, AOA, as dominant species in a wide range of ecological niches clarified a major provider of energy flow and nitrogen cycling in ecosystem (Erguder et al. 2009). Thus, a fundamental reconsideration of the geochemical cycle of nitrogen is demanded. In line with this, environmental shotgun sequencings of specific samples from ocean, soil, plant, and animal stimulated interest in the diversity of microorganism and indwelling metabolic gene clusters, enabling the elucidation of species and community functions in specific niche. With the introduction of high-throughput screening that can detect extremely weak activity and signal, new methods have been developed for rapid detection of target libraries with a small



Biological Treasure Metagenome, Fig. 1 A value of metagenome resource. Current metagenomic studies result in various fields of applications that include, but

are not limited to, environmental, agricultural, medical, and industrial needs

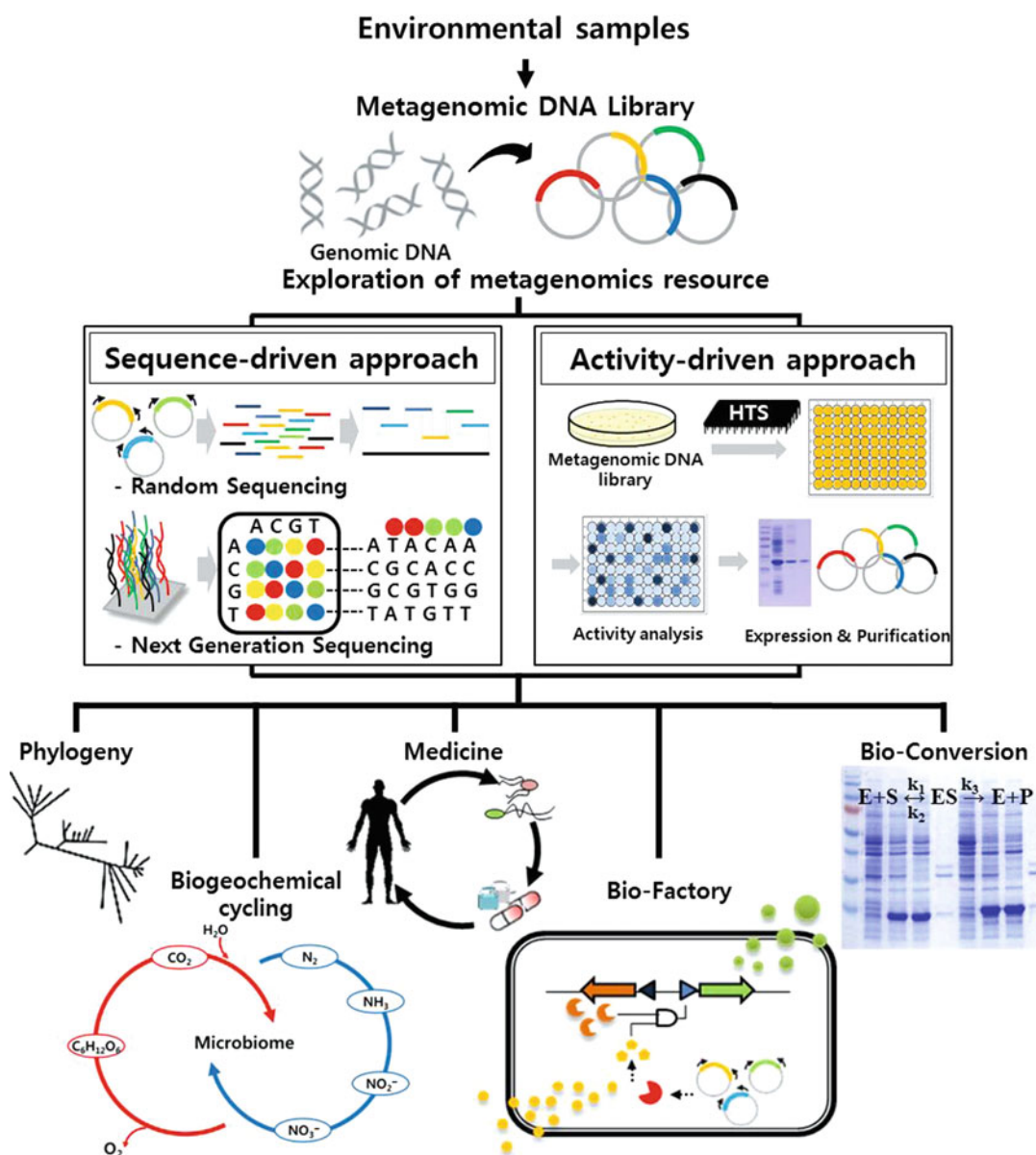
amount of sample, thus facilitating screening with more positive hits (genes and proteins). Candidates captured through this process are compared with other known resources in shared pattern of functional or sequence signature, which predicts the functional roles of the candidates *in silico*. With the derived functional roles, metabolic pathway and/or capacity of the whole microbiome is constructed. Currently, the microbiome in a specific environment is reliably established through reconstitution of the genome data of all organisms by bioinformatic tools (Kunin et al. 2008). Metagenome data provide the understanding of complex biological systems through online public databases and data-integration tools. Currently, assembled genomes analyzed in an integrated logics are providing a window for forming more complete genomes, and this is expected to reduce time and cost in finding new resources considerably.

The Value of Metagenome Resources Remain to be Elucidated

Besides the elucidated physiological and ecological values by using metagenomic approaches, there is an obvious reason for obtaining useful genes or physiologically active substances from the metagenome. This is because the access to screening the library of pure-cultured cells has been limited, and it is very difficult to sustain the novelty of resources originating from culturable strains. According to what is known, enzymatic degradation or synthesis is possible for almost every organic compound that can be found naturally or synthesized chemically. However, regardless of the existence of related enzymes with promising activities, the functional and sequence spaces of metagenome resources from the whole living organisms in ecosystem are still left mostly unexplored. Therefore, if hurdles in the screening process, due to the approach based on the homology of protein sequence, can be overcome, it will be possible to find resources in new areas. Of course, it is generally known that the approaches of screening from the metagenome compete with protein engineering technologies that mutated or fine-tuned existing genetic resources or induced forced evolution

in vitro. However, the limitation of engineering processes in the exploration of sequence space and the innate weak point of the stepwise screening process that cannot gather effectively the effect of beneficial mutation in the alternative landscape may reasonably explain the strength of the exploration of the metagenome from microbial community that has already possessed various functional space (including biologically permitted sequence space) by evolutionary experience. Therefore, the metagenome can play a significant role as a resource to provide new alternatives and get desired products from the highly precise and specific enzyme reactions of thousands of substrates used in industry (Fig. 2).

One of the major trends of research on biologically mediated processes is white biotechnology, which is to find alternatives to petrochemical compounds using renewable resources. Therefore, attention is paid to the production of fossil fuels by bioconversion or fermentation using biomass. To this end, the acquisition of regulatory genes, potential enzymes, and gene clusters related to the production of organic acids, alcohols, solvents, and diesels is also obtainable from metagenome. What is more, organic compounds, which have been out of people's attention for economical reasons, are again spotlighted along with their application to improved price competitiveness, low risk of environmental pollution, and innovative tools of systems biology. We also expect critical roles of the metagenome in increasing agricultural productivity and the utilization of biomass. Besides, research on the human metagenome can provide clues to causes of diseases, acquired immune system, and new methods of treating pathogen through the analyses of microbiome database from microbial communities (Gill et al. 2006). Also, in response to the serious side effects of synthetic drugs and increasing drug-resistant pathogens, finding new natural inhibitors or suppressors, including quorum-sensing blocker, as antibiotics in the metagenome could be possible. In this respect, there are many attempts to approach the new potential of metagenome resources through analyzing the resistome formed naturally by biological species existing around the ecological



Biological Treasure Metagenome, Fig. 2 Exploring value creation from integrative research activities of metagenome. Information concerning the application fields of metagenome resources is gathered and processed by systematically integrative systems. This information

results in various fields of further applications that solve global problems such as fine chemical, environmental, medical services, and future energies. Basically, metagenome information also provides a clue for the origin and minimal genome of living organisms

producers of these substances (D'Costa et al. 2007). It is generally believed that such an expectation can be realized by research on the metagenome evolved in ecosystem through countless mutations, suppression, and

competition in tens of millions of microbial species for billions of years.

Genomic data collected through the metagenome will be used ultimately in creating synthetically engineered species (with minimally

synthesized genomes; Gibson et al. 2008; Lartigue et al. 2007; Dymond et al. 2011) solving global problems such as medical services and energies. The goal of synthetic biology is to produce cell-level bio-factories, aiming at the biological production of valuable chemicals and drugs. In this research area, attempt to create a platform for the engineering of orthogonal factor (function without any interference in vivo), such as switches, circuits, and logic gates, was made to control independently multiple genes in host systems. In fact, the related tools and methods have already been successfully applied in various studies, giving rise to orthogonal DNA or RNA-protein pairs. As example of one such effort, orthogonal ribosome-mRNA pairs are composed of an mRNA containing a ribosome-binding site that does not recognize by the endogenous ribosome and an orthogonal ribosome that specifically translates the orthogonal mRNA and thus function independently without severe effects on cell physiology and metabolism when required (Wang et al. 2007). This result provides a possibility that useful components of cells can be synthesized efficiently for making microbes as cell factories equipped with minimal but plentiful genome and then adding more genes for specific purposes. The assignment of speciality and/or orthogonal function may partly be attainable through novel parts (genes and proteins) and genetic circuits (signaling cascades and metabolic pathways) to be mined from the metagenome.

Summary

Through metagenomics, scientists have obtained a new view to the microbial world that is different from traditional concepts and are working to overcome difficulties in future society. The exhaustion of natural resources such as fossil fuels will increase people's interest in biological

resources using renewable resources, and this alone makes the metagenome highly worthy of study. Microbial diversity is so extensive that it is not easy to estimate their history in the ecosystem of the planet, and even now at all of ecosystem they may continue to mutate in order to resist or adapt themselves to unceasing changes. Thus, the metagenome provides a huge potential as a resource with novel activity, which may be used for any purpose.

References

- D'Costa VM, Griffiths E, et al. Expanding the soil antibiotic resistome: exploring environmental diversity. *Curr Opin Microbiol.* 2007;10:481–9.
- Dymond JS, Richardson SM, et al. Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature.* 2011;477:471–6.
- Erguder TH, Boon N, et al. Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. *FEMS Microbiol Rev.* 2009;33:855–69.
- Gibson DG, Benders GA, et al. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science.* 2008;319:1215–20.
- Gill SR, Pop M, et al. Metagenomic analysis of the human distal gut microbiome. *Science.* 2006;312:1355–9.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev.* 2004;68:669–85.
- Hunter-Cevera JC. The value of microbial diversity. *Curr Opin Microbiol.* 1998;1:278–85.
- Kunin V, Copeland A, et al. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev.* 2008;72:557–78.
- Lartigue C, Glass JI, et al. Genome transplantation in bacteria: changing one species to another. *Science.* 2007;317:632–8.
- Lewis Jr CM, Obregon-Tito A, et al. The human microbiome project: lessons from human genomics. *Trends Microbiol.* 2012;20:1–4.
- Wang K, Neumann H, et al. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat Biotechnol.* 2007;25:770–7.
- Xu J. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol.* 2006;15:1713–31.

C

Carbohydrate-Active Enzymes Database, Metagenomic Expert Resource

Brandi Cantarel¹, Pedro Coutinho² and Bernard Henrissat²

¹Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

²Centre National de la Recherche Scientifique & Aix-Marseille Université, Marseille, France

Definitions

Carbohydrate-active enzymes (CAZymes) designate the ensemble of the enzymes that catalyze the assembly, breakdown, or modification of oligosaccharides, polysaccharides, and glycoconjugates. They are usually comprising glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and glycosyltransferases (GTs).

Introduction

Carbohydrates

Carbohydrates, in the form of mono-, di-, oligo-, and polysaccharides, as well as glycoconjugates, play important roles in all areas of biology. Beyond simple energy storage, carbohydrates underpin diverse biological processes, such as

host-pathogen interactions, signal transduction, inflammation, intracellular trafficking, diseases and tumor metastasis, and differentiation/development. Importantly, carbohydrates represent about 75 % of the structural components of photosynthetically produced biomass. Sugar-rich plant cell walls, seeds, and tubers thus represent a major source of nutrients for herbivorous and omnivorous animals and for humans. These carbohydrates have also significant potential to address energy and material needs.

A striking feature of carbohydrates is their global structural variety, which results from a large diversity of monosaccharide building blocks, and the possibility of numerous stereo- and regiospecific linkages (Laine 1994), which give rise to a myriad of structures that can be attached to proteins, lipids, nucleic acids, etc. Any biological molecule can be glycosylated, including proteins, lipids, nucleotides, and carbohydrate themselves, the level of such modifications often varying extensively. In fact, glycosylation of proteins is the most common posttranslational modification in eukaryotes but also present in prokaryotes, strongly influencing many of their functional aspects, including cellular localization, turnover, and protein quality. Proteoglycans mediate cell communication, growth factor sequestration, microbial recognition, chemokine and cytokine activation, tissue morphogenesis during embryotic development, cell migration, and proliferation. Nature has exploited the tremendous possibilities offered by the sugar code by elaborating and breaking

down very specific complex carbohydrates in a highly specific manner. Exquisite details of complex carbohydrates create immense functional differences. For instance, cellulose and amylose, two simple polymers of glucose residues linked between their position 1 and their position 4, only differ by the equatorial vs. axial orientation of the glycosidic bond (β for cellulose and α for amylose). This minute difference gives rise to two massively different polysaccharides: cellulose, whose mechanical properties rival those of steel, is synthesized by plants as a structural polysaccharide notoriously recalcitrant to hydrolysis while amylose is a component of starch and, as a reserve carbohydrate, is readily converted to glucose.

Carbohydrate-Active Enzymes and Their Classification

Carbohydrate-active enzymes (CAZymes) catalyze selective reactions to assemble and break down complex carbohydrates and glycoconjugates for a large array of biological functions globally underpinning glycobiology. These enzymes, which comprise glycoside hydrolases (GH), polysaccharide lyases (PL), carbohydrate esterases (CE), and glycosyltransferases (GT), have gradually evolved from a limited number of primordial carbohydrate-active enzymes coding genes by acquiring novel specificities at substrate and product level. In addition, these enzymes often display a modular structure with a catalytic module appended to one or several other domains, such as carbohydrate-binding modules, allowing for increased specificity and/or specific targeting to a particular substrate/region (Boraston et al. 2004).

The sequence-based classification of CAZymes was initiated in 1991 (Henrissat 1991; Henrissat and Bairoch 1993, 1996) as a complement to the long-standing Enzyme Commission (EC) number system (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>), which is based solely on enzyme activities. Given the prevalence of convergent evolution of enzymes that cleave glycosidic bonds, as well as the demonstrable catalytic promiscuity of individual enzymes, sequence-based classification has proven to be a robust way to unify information on enzyme

structure, specificity, and mechanism, which provides significant predictive power. Initially motivated by a need to delineate cellulases (EC 3.2.1.4) into distinct structural families, the first incarnation of the GH family classification, as such, comprised 35 GH families (Henrissat 1991). Five years later, the number of GH families grew to 57 families (Henrissat and Bairoch 1996), and has continuously expanded to reach 113 in 2009 (Cantarel et al. 2009). As of March 2012, 130 sequence-based families of GHs have been defined and are presented in the continuously updated CAZy database (<http://www.cazy.org/>). In parallel with the development of the classification of GH families, sequence-based classifications of the glycosyltransferases (GTs) (Campbell et al. 1997), polysaccharide lyases (PLs) (Lombard et al. 2010), carbohydrate esterases (CEs) (Cantarel et al. 2009), and carbohydrate-binding modules (CBMs) (Boraston et al. 2004) have similarly been developed and added to the CAZy database.

Functional Prediction of Carbohydrate-Active Enzymes

The immense variety of carbohydrate structures and their involvement in extremely different biological functions make that functional annotations such as “putative carbohydrate-active enzyme” or “putative glycosidase” have very limited information value. Instead, a useful functional prediction for a CAZyme should indicate the likely nature of sugar being cleaved or transferred, with a description of the exact connectivity between the sugar undergoing catalysis and the molecule it is attached to or detached from.

A feature that was recognized very early on was that the sequence-based families of carbohydrate-active enzymes group together enzymes of differing substrate specificity and hence group together enzymes with different EC numbers (Henrissat 1991; Campbell et al. 1997). Because of the multifunctional nature of these enzymes, it is believed that a limited number of catalytic and binding progenitors (protein domain families), which can be found in different combination, gave rise to the vast number of enzymes and of carbohydrate structures that exist in



modern organisms, resulting in the gradual and simultaneous acquisition of exquisite substrate specificity for both carbohydrate biosynthesis and carbohydrate degradation. Since most CAZyme protein domain families are multifunctional, prediction of functional roles for uncharacterized carbohydrate-active enzyme encoding genes simply by family assignment can lead to erroneous annotations, especially at high sequence divergence. Additionally, the universe of known carbohydrate structures with the same types of linkage bonds is smaller than the universe of possibility; therefore, even when functions are known, there are potentially more possible substrates. As a result the number of sequences that can be assigned to CAZy families increases very rapidly, but the number of CAZymes whose substrate specificity has been established (even roughly) grows at a much lower pace. As sequencing data grows with increasing genomic and metagenomic characterization, this proportion of characterized enzymes continues to decrease. In spite of limitations due to the presence of different substrate specificities in many CAZyme families, it is often possible to assign a broad substrate category (for instance, pectin, cellulose, xylan) to a number of CAZyme families (Cantarel et al. 2012) even if the precise substrate or product specificity (for instance, to distinguish between endo- and exo-acting enzymes or to distinguish between β -D-xylosidase and α -L-arabinofuranosidase) cannot be predicted accurately based on simple family assignment. In order to improve functional prediction, the partition of CAZyme families into subfamilies based on phylogenetic analysis has been explored. Significantly subfamily classification of several families of GHs and PLs has shown that the majority of the defined subfamilies were monospecific, thus indicating a better correlation of substrate specificity between sequences at the subfamily level than the family level (Lombard et al. 2010; St. John et al. 2010; Stam et al. 2006).

The advent of low cost DNA sequencing has revolutionized biology, and the central question is no longer how to obtain nucleotide sequence, but how to make sense of it. In the vast majority

of cases, inferences are done by detecting the similarity of sequence between the newly generated DNA sequence (or putatively encoded protein) and sequences already in databases. This approach does not perform equally with different classes of proteins in terms of the biological inference that can be derived. For instance, the assignment to families of protease/peptidases has often limited predictive power: the prediction are often only based on the fold the most informative information being essentially that of the catalytic machinery – for instance, “serine protease” – and little predictive power in terms of what is the specific peptide substrate targeted by the enzyme.

Thus, the very difficulty with CAZymes (huge structural and functional variety of substrates) is also at the origin of their intrinsic advantage: these enzymes had to evolve to achieve the exquisite specificity necessary to carry out their function in a selective manner. The high information content of complex carbohydrates has therefore translated into the proteins that assemble and deconstruct then by leaving evolutionary signals/traces that can be recognized in the sequence. While experimental developments in the field of glycomics are still slow in comparison to the boom in sequencing technologies, carbohydrate-active enzymes are perhaps the most adapted to functional inference from genomic and metagenomic data.

The direct genetic sequencing of microbial communities (metagenomics) is beginning to explore the great gene diversity in the microbial world. Environmental samples from diverse environment are being studied to better understand the role of microbes in various habitats from the human body to the ocean floor. This technology has allowed scientist to begin to answer questions not possible with studying only cultivable species. Here we review the burgeoning exploration of carbohydrate-active enzymes in metagenomic samples.

Glycobiology in Microbial Communities

Microbial communities isolated from human fecal material are the most well studied in the usage of CAZymes. CAZyme diversity in human gut microbiota studies (Gill et al. 2006;

Mahowald et al. 2009; Turnbaugh et al. 2010) showed 81 glycoside-hydrolase families, making the human gut one of the richest sources of CAZymes. Some of these studies aim to determine the relationship between CAZyme utilization and disease (Turnbaugh et al. 2009) and diet including a vegetarian (Kabeerdoss et al. 2011) or differential fiber intake (Tasse et al. 2010). Taxonomic and genetic differences were found between omnivores and vegetarians suggesting that energy intake, complex carbohydrate degradation, and butyrate production, a product of dietary fiber fermentation, was higher in omnivores with these functions associated with an increase in certain *Clostridiales*, such as *Clostridium*, *Roburia*, and *Eubacterium rectale* (Kabeerdoss et al. 2011). Gene clusters involved in dietary fiber degradation have been shown to be larger than clusters involved in starch degradation and often contain genes involved in carbohydrate transports and binding (Tasse et al. 2010). Studies of the human gut have also revealed possible lateral gene transfer between marine and human microbes interacting in the human digestion system (Hehemann et al. 2010), in order to increase algae cell wall degradation.

While the human gut microbiota has been the subject of most studies, studies in other animal gut microbiota reveal an evolutionary-driven composition of gut microbiota. Thus mammalian gut microbiome composition and functional capabilities is likely driven by diet, such that carnivores, no matter the mammalian phylogeny, contain organisms and have bacterial functions for using proteins as an energy source, compared to herbivores, whose gut microbiota aims to convert complex plant carbohydrates into energy (Pope et al. 2010; Muegge et al. 2011; Zhu et al. 2011). The digestive microbiota of herbivores is being actively explored for the discovery of novel enzymes for the conversion of plant biomass to biofuels (Brulc et al. 2009; Duan et al. 2009; Suen et al. 2010). In a similar vein, the termite gut microbiota revealed a wealth of CAZymes involved in degrading wood polysaccharides (Matteotti et al. 2011; Warnecke et al. 2007).

In a comparison of carbohydrate active enzymes in all human body sites (Cantarel et al. 2012), differences in abundance of CAZymes were identified between all major sites. In general, digestive sites and particularly stool contained the highest number of CAZY families and the highest abundance of CAZymes. These sites have a higher abundance of CAZymes involved in plant and algae degradation. GH94 (*cellobiose*, *cellodextrin*, and *chitobiose* phosphorylases) and GH30 (β -1,6-glucanase, β -xylosidase, β -D-fucosidase, β -glucosidase, and β -1,6-galactanase) are statistically more abundant in stool compared to the other four major body habitats. Oral sites appear to specialize in starch and glycogen degradation, as these functions are enriched in oral habitats compared to the stool. Vaginal microbial communities are enriched in CAZymes involved in sucrose cleavage and polymerization to fructans, potentially important for biofilm formation. Overall, the functional profiles are more similar within a body habitat than between habitats, even when the taxonomic profiles differ, suggesting functional adaptation of the community to the carbohydrates prevalent in the environment.

Practical Issues in Mining Metagenomes for CAZymes

Practically annotation of CAZymes is not completely trivial. First, for historical reasons some CAZY families are distantly related, meaning there are sequences that share statistically significant sequence similarity with multiple families in the same region. These families are therefore grouped into “clans,” similarly to PFAM clans. Therefore, family assignment of these proteins is ambiguous, suggesting these proteins are general members of the clan with broad functional predictions. Secondly CAZymes are modular, meaning these enzymes are often composed multiple protein domains connected by linker regions. These domains can combine in a variety of permutations to form or give rise to diverse functions, e.g., a carbohydrate binding domain can be attached to multiple catalytic domains to facilitate substrate specificity. Therefore, accurate annotation requires



comparison to a database of domains, rather than whole proteins. The length of the domains can vary greatly from as little as 30 amino acids to several hundred residues, so when strict expectation value (calculated by BLAST or HMMER) thresholds, such as E-value $< 1e-6$, are imposed, the false-negative rates increase for the identification of the distantly related short domain family members. For metagenomics, these challenges are amplified since metagenomic gene predictions are (i) often fragmented, (ii) are too short to contain multiple domains, and (iii) could be from organisms with little to no close evolutionary relatives.

As previously discussed, as sequencing coverage of microbial communities (in metagenomes) increases, these challenges of metagenomics will diminish and CAZy family assignment as well as domain structure prediction will gradually improve. The hardest problem to resolve will thus be with the precise prediction of the substrate/product specificity of CAZymes. Such predictions require close relatedness between the query and at least one biochemically characterized CAZyme. Accelerating the pace of exploration of the sequence to specificity space of CAZymes is key to a leap toward accuracy, and this will require new experimental innovations in the field of carbohydrates and coupling computer-guided high-throughput functional investigations to structural genomics initiatives. Yet, notwithstanding the accuracy of functional prediction, the interpretation of carbohydrate-active enzymes profiles resulting from metagenomic investigations will require expertise in complex carbohydrate assembly and breakdown.

Summary

The advent of low cost DNA sequencing has revolutionized biology, and the central question is no longer how to obtain nucleotide sequence, but how to make sense of it. Functional predictions start by sequence comparisons against databases of known annotated genes and proteins. However, there are two major caveats to this

approach: (i) sequence similarity does not equal same function and (ii) annotations on known sequences have varying degrees of accuracy depending on the level and quality of evidence, which ideally relies on experimental validation, but is often based on sequence similarity. Creating accurate annotations becomes complex in multifunctional protein families. Because the diversity in carbohydrate structure is large and the number of protein families acting on sugar limited, carbohydrate-active enzyme gene families are often multifunctional and specificity is mediated by additional structural or carbohydrate binding modules. The Carbohydrate Active Enzyme Database (Cantarel et al. 2009) (CAZy) provides an expert curated resource for the glycobiology community, whereby annotations and their underlying evidence are documented.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Human Gut Microbial Genes by Metagenomic Sequencing](#)
- ▶ [Mining Metagenomic Datasets for Cellulases](#)

References

- Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J.* 2004;382(Pt 3):769–81.
- Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A.* 2009;106(6):1948–53.
- Campbell JA, Davies GJ, Bulone V, Henrissat B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J.* 1997;326(Pt 3):929–39.
- Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 2009;37(Database issue):D233–8.
- Cantarel BL, Lombard V, Henrissat B. Complex carbohydrate utilization by the healthy human microbiome. *PLoS One.* 2012;7(6):e28742.

- Duan CJ, Xian L, Zhao GC, Feng Y, Pang H, Bai XL, et al. Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J Appl Microbiol.* 2009;107(1):245–56.
- Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. *Science.* 2006;312(5778):1355–9.
- Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature.* 2010;464(7290):908–12.
- Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J.* 1991;280(Pt 2):309–16.
- Henrissat B, Bairoch A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J.* 1993;293(Pt 3):781–8.
- Henrissat B, Bairoch A. Updating the sequence-based classification of glycosyl hydrolases. *Biochem J.* 1996;316(Pt 2):695–6.
- Kabeerdoss J, Shobana Devi R, Regina Mary R, Ramakrishna BS. Faecal microbiota composition in vegetarians: comparison with omnivores in a cohort of young women in southern India. *Br J Nutr.* 2011;20:1–5.
- Laine RA. A calculation of all possible oligosaccharide isomers both branched and linear yields $1.05 \times 10(12)$ structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology.* 1994;4(6):759–67.
- Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B. A hierarchical classification of polysaccharide lyases for glycomics. *Biochem J.* 2010;432(3):437–44.
- Mahowald MA, Rey FE, Sedorf H, Turnbaugh PJ, Fulton RS, Wollam A, et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci U S A.* 2009;106(14):5859–64.
- Matteotti C, Haubruge E, Thonart P, Francis F, De Pauw E, Portetelle D, et al. Characterization of a new beta-glucosidase/beta-xylosidase from the gut microbiota of the termite (*Reticulitermes santonensis*). *FEMS Microbiol Lett.* 2011;314(2):147–57.
- Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science.* 2011;332(6032):970–4.
- Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, et al. Adaptation to herbivory by the tamar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci U S A.* 2010;107(33):14793–8.
- St John FJ, Gonzalez JM, Pozharski E. Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups. *FEBS Lett.* 2010;584(21):4435–41.
- Stam MR, Danchin EG, Rancurel C, Coutinho PM, Henrissat B. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng Des Sel.* 2006;19(12):555–62.
- Suen G, Scott JJ, Aylward FO, Adams SM, Tringe SG, Pinto-Tomas AA, et al. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet.* 2010;6(9):e1001129.
- Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, et al. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* 2010;20(11):1605–12.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. *Nature.* 2009;457(7228):480–4.
- Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F, et al. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci U S A.* 2010;107(16):7503–8.
- Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature.* 2007;450(7169):560–5.
- Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc Natl Acad Sci U S A.* 2011;108(43):17714–9.

Challenge of Metagenome Assembly and Possible Standards

Matthew B. Scholz¹, Chien-Chi Lo¹ and Patrick Chain²

¹Genome Science Group, Los Alamos National Laboratory, Los Alamos, NM, USA

²Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Introduction

As technology and methodology have allowed for more advanced assemblies of metagenomes, the need for commensurate assignment of quality to these assemblies has become evident. There are currently no set standards for describing the quality of sequencing, assembly, or analysis of metagenomic assemblies. Uncorrected, this may

lead to faulty conclusions based on assumptions that the assembly is more or less accurate, or representative of the sample, than it truly is. This need is similar to, but far more complex than, the dilemma that faced the microbial sequencing and assembly community as more and more genomes were sequenced with new technologies and assembled with novel algorithms.

For bacterial genomes, the quality of assembly and finishing efforts has been standardized for several years, resulting in a much better understanding of the types of analyses that can be performed on each level of finished genome and the resulting value. While the need for standards in metagenomics has been very clear in terms of metadata (Yilmaz et al. 2011), less attention has been focused on the genomic data itself. As the field continues to advance and mature, it is clear that efforts in standardizing assemblies as well as functional and phylogenetic classification are sorely needed. Given the flux in application of various sequencing technologies (with different and sometimes variable sequence qualities) to genome reconstruction, the most recent version of this standard for microbial genomes is to divide sequences into broad levels of completeness and quality, from draft to completely finished (Chain et al. 2009). While these standards are valuable, it is difficult to apply similar standards to metagenomic assemblies, where the effort is to reconstruct the genomes (or parts of genomes) of many organisms present within a sample. The numbers of different species in a community selected for sequencing and metagenomic assembly can vary from two to millions of individual genomes from many species, with varying frequency of each genome. Additionally, each genome may be different in size, G+C content, and repetitiveness as well as have other genome-specific issues that make it impossible to assemble all genomes equally well.

Additionally, community genomics are complicated by the potential existence of many strains of the same species (species of the same genus, etc.), of recombination, of horizontal gene transfer events among members of the community, and of other factors that further complicate

analysis and assembly. All of these factors underlie the highly variable nature of metagenomes, making it difficult to generate accurate assemblies and also difficult to define standards or otherwise grade the effectiveness of an assembly. It can be reasonably stated that metagenomic assembly is still in its infancy and generally produces what can only be described as draft assemblies of metagenomic data, though it has certainly been possible in some rare cases to recover full and near-complete genomes from some environments (Huttenhower et al. 2012).

The utility of sequencing a community sample is based solely on the ability of the researcher to garner useful information from the data (assembly and annotation). This ability is, more often than not, reflective of the “quality” of a metagenome assembly. Additionally, the goals of a given project can also affect this question, by altering the types of analysis needed or the depth of sequencing required, among many factors. The gross differences and determinations of sequence diversity between two or more metagenomes can be typically analyzed using simple comparative tests, whereas a more in-depth analysis for gene content or for application to proteomic (peptide mass prediction) or metabolic pathway (function and operon prediction) analysis requires larger assembled regions of contiguous sequence (contigs), with low error rates.

While it is not possible to set de facto standards for metagenomic analysis or assembly, this entry is an attempt to discuss a number of the potential impediments to adequate or good metagenome assembly. Additionally, several possible methods for improvement and validation of assembled contig sets from metagenomic assemblies, as well as potential methods for generating higher quality draft metagenomes from an individual sample, will be discussed.

Barriers to Metagenomic Assembly

As has been addressed previously, metagenome assembly is difficult, requiring ever-increasing computational resources at a rate fast outpacing

“Moore’s law” (Miller et al. 2010; Scholz et al. 2012). This is a product of the limitations of current sequencing technologies coupled with the available assembly algorithms that can falter when running into the massive scale of data produced by current next-generation sequencers (NGS). For metagenomic sequencing and assembly, there is a paradoxical problem with data. While the relatively low cost and highest throughput sequencers produce hundreds to thousands of gigabytes of data per run, the short read lengths of these technologies limit the types of assembly procedures that can be applied (Scholz et al. 2012). Due to the diversity of genomes and the variation between members of a community, a good assembly of a metagenome requires significantly more sequencing (potentially terabases of data per sample for some environments). In direct opposition to this requirement, the current state-of-the-art assemblers for NGS data are limited by available computational memory, meaning that, currently, computers are only capable of assembling as little as 1 % of the data required for the most complex of samples. The computational time, processing power, and required system memory for assembling any genome using state-of-the-art assembly algorithms (Miller et al. 2010) are directly proportional to the size and complexity of the genome(s) to be assembled (and partly coupled with errors introduced during the sequencing process). In the case of metagenomes, then, the first approximation would be that the requirements for assembly are a function of the number of unique bases (or unique “words” or *Kmers*) in all genomes contained within the community to be sequenced. This easily overshadows even the largest and most complex eukaryotic genomes, making assembly of all microbial genomes within a single metagenomic sample, given today’s infrastructure and algorithms, infeasible. This variation and computational limitation leads to a variable amount of data that can be incorporated into any given assembly. It can be expected that read incorporation into metagenome assemblies will follow a logarithmic curve, with the amount of available sequence data covering more of the diversity and complexity of the community being

sequenced and increasing the percentage of reads incorporated, affecting the slope of the curve. In short, the more complex communities, such as those found in soils and sediments, will require much greater sequencing inputs to allow for assembly of a significant proportion of the data. Conversely, simpler communities, such as bioreactors, enrichment cultures, and naturally simple environments can achieve nearly 100 % incorporation of data even with relatively few reads (<200 million Illumina reads).

Assembling Subsets of Data

To allow current assemblers to better process the mountains of data, it is generally believed that dividing reads into smaller, categorical bins may enable improved, or “targeted,” assembly. While this partitions the data into manageable parcels for assembly, it has also been used as a filtering method, to remove extraneous reads from the dataset pre-assembly (Godoy-Vitorino et al. 2012). There have been several very thorough methods developed for binning of reads or contigs. Binning can be performed as a function of nucleotide frequencies, or abstractions of that (*Kmer*-based filtering, etc.), on statistical analysis of read relationships (read topology) or on similarity to known genomes or genome signatures (homology, etc.). Additionally, HMMs or other learning algorithms may eventually be developed to allow rapid binning of reads. However, once binning is performed, many of the issues surrounding assembly of many sequences again become relevant and require in-depth analysis and work. As each binning method will invariably introduce both false positives and false negatives, it is not clear what effect these algorithms may have on a “final” assembly or if the effect will be consistent among different samples.

Whole Sample Assembly

Full metagenome sequence runs, or bins of metagenomic data, are run through an assembly methodology or program; however most current

algorithms are designed for isolate genome assembly (Miller et al. 2010; Scholz et al. 2012). While isolate genome assembly assumes that there are a limited number of solutions to the assembly, as the complexity of the genome increases and concomitant amount of sequence data are required, these decisions become more difficult for algorithms to make. For metagenomes, there are additional complications, such as strain-level variation within a species, varying levels of similarity among the multiple species within the population, including horizontal gene transfer, and the ever present problem of variability of organism frequency/abundance within the community. Several recent attempts have been made (e.g., MetaVelvet (Namiki et al. 2012), RAY (Boisvert et al. 2010), or Meta-IDBA (Peng et al. 2011)) to solve one or more of these metagenome-specific issues. However, there is not yet a perfect algorithm, and all can benefit from improved understanding of the inherent complexities within metagenomes as well as from improved algorithms for determining which data are to be examined and how. Given the varied nature of the complexities that exist in communities, it is likely that the perfect assembly algorithm will have to evaluate the data and make decisions during metagenome assembly.

Assembly Validation and Metrics

Validation of metagenomic assemblies is not currently a standard process. Some efforts have focused on validation using tools adapted from single-genome assemblies which, due to the differences in complexity, can vary from being simply an inefficient method for validation at best to being misleading and based on incorrect assumptions at worst. Validation of assembly completeness (good assemblies provide large contigs with more of the raw data) and accuracy (good assemblies harbor few errors such that it is a close representation of the target organism) is a nuanced and nebulous process even with isolate genomes. A typical series of statistical properties of contigs is often used to describe the goodness

of single-genome assembly (N50, total assembly size, etc.). To improve accuracy, this can be combined with manual inspection of the data underlying the contigs, using tools such as Consed (Gordon 2003), Hawkeye (Schatz et al. 2011), or other alignment viewers. It bears noting here that these tools require sequential attention to each individual contig, making validation of metagenomic assemblies of many thousands to millions of contigs prohibitively time-consuming.

Additional validation can be gained by read mapping input sequence data to contigs to identify errors or areas with unexplained variances in coverage. None of the tools or processes available for validations of single-genome assemblies is directly applicable to metagenomes and requires either significant alterations in method or a completely new approach. This is due in part to the much larger amount of data required for a metagenome assembly as well as the intrinsic complexities associated with metagenomes, mentioned above.

How, then, does one assess whether a metagenomic assembly is good or valid? It is possible to examine statistics of an assembly to determine if it has value and assess the quality of assembled contigs to give a measure of what analysis can be performed on the assembled data (e.g., longer contigs allow for more annotation analysis). Additionally, it is easy to calculate the total number of bases assembled, allowing for a rough estimate of how many genomes may be captured in an assembly. However, it is also important to validate that the assembly is an accurate representation of the input data by use of read mapping or other comparative tools. For metagenomes, with the stated issues of lack of uniformity, it is likely that a valuable tool for obtaining improved assemblies will be to perform several assemblies in parallel and compare inter-sample assemblies to each other. This will also be an important method to compare the results of binning and of different assembly methods to combined, or iterative, assemblies of the entire dataset. For environments that have been amply studied and for which there are a number of pertinent reference sequences, such as for

human microbiome samples (Lampe 2008; Huttenhower et al. 2012; Methe et al. 2012), it is possible to use these to validate assemblies. Recent work with isolation and sequencing of single cells from within environmental samples raises the possibility of using reference-based validation tools on metagenome samples as well (Kant et al. 2011; Leung et al. 2012).

Statistical Comparisons

As mentioned above, the first approximation of the quality of any assembly is an examination of the metrics associated with the assembly. For metagenomes, these metrics should be different from those used for isolate genome work. Statistics that are linked to the total assembly size (e.g., N50) have little value, as the size of the metagenome, the assembly (and the assembled number of contigs), and the choices made for assembly (binning, filtering, assembler algorithm, Kmer size, etc.), which can affect the number and types of bases included in a metagenomic assembly, can all result in drastically different interpretations. The evaluation of metagenome assemblies is often conducted in a holistic manner, utilizing a number of important statistics and validation metrics. This can be used to assess the completeness of various assembly methods. Table 1 shows a selection of assembly statistics for a single sample (MH0001) from the MetaHIT

project using the assembly program SOAPdenovo with different Kmers as an input parameter. What is important to note is that it is difficult to select the best assembly based on any single metric, even given the same assembler with a single parameter change. In fact, it is the rule rather than the exception that no single assembly of the data can provide the best statistics for every metric.

Read Mapping as Contig Validation

It is important that any assembly be verified by methods beyond those utilizing simple statistical methods. It is also important that validation algorithms be independent from those utilized to perform the assembly. Currently, Burrows-Wheeler (Langmead et al. 2009; Li and Durbin 2010) read mapping can serve as an independent approach of validating the contigs assembled based on the raw sequencing data. This approach has the ability to validate assembled contigs by basis of coverage of every base contained within the contig (Fig. 1) as well as based on the variation of coverage within the contig (Fig. 2).

It may not always be the case that coverage along a contig will appear as even as with isolate genomes, due to the issues of strain (allele) variations, of gene duplication, of ribosomal gene similarities between species, and of horizontal gene transfer. Additionally, because read

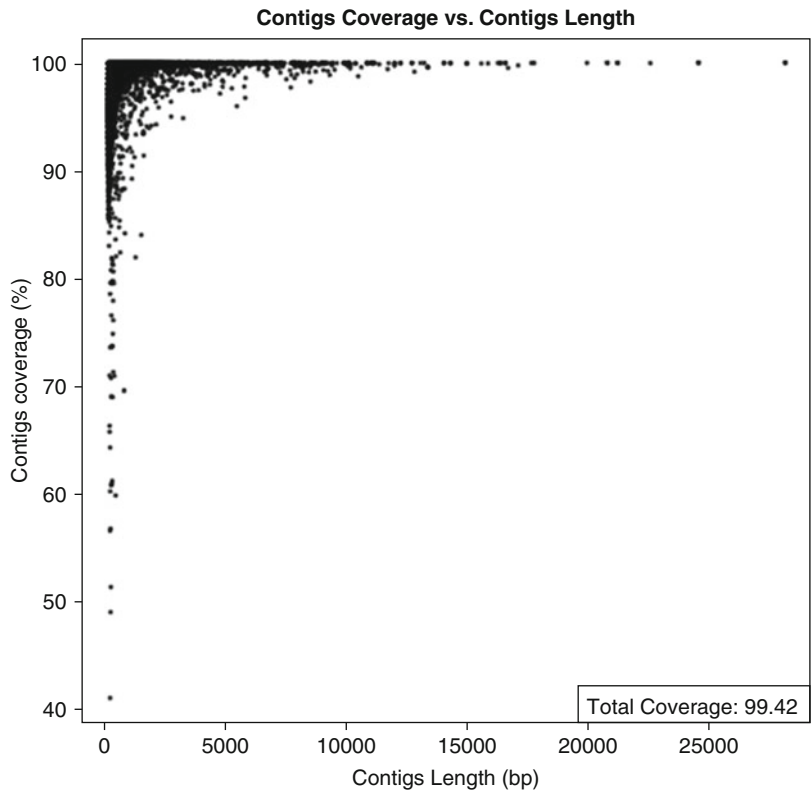
Challenge of Metagenome Assembly and Possible Standards, Table 1 Statistical metrics of metagenome assembly

Assembly type	Number of contigs	Maximum contig size	Total bases	Bases in largest 100 contigs	Bases in contigs > 10 kb	% read incorporation
SOAPdenovo-Kmer 21	378,624	18,148	63,350,050	1,025,623	438,734	60.8
SOAPdenovo-Kmer 23	303,536	18,150	55,682,346	1,155,330	839,420	61.1
SOAPdenovo-Kmer 25	244,200	25,192	47,972,706	1,220,072	949,421	60.6
SOAPdenovo-Kmer 27	188,074	23,935	40,311,428	1,162,160	843,499	59.9
SOAPdenovo-Kmer 29	140,502	28,068	33,228,335	1,177,230	804,055	58.9
SOAPdenovo-Kmer 31	109,722	28,068	27,463,402	1,245,286	918,627	57.8



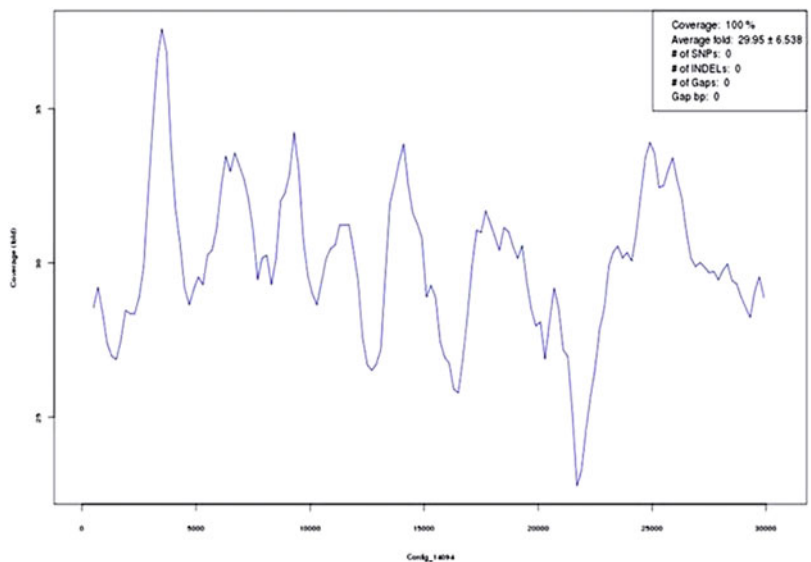
Challenge of Metagenome Assembly and Possible Standards, Fig. 1

Coverage histogram of metagenome assembly. Displays percentage coverage of every contig as a function of the contig length



Challenge of Metagenome Assembly and Possible Standards, Fig. 2

Base-by-base coverage histogram of a single contig generated within a metagenome assembly. Areas where coverage varies from the mean may be identified as regions of low quality or confidence



mapping is fundamentally different from Kmer-based assemblies, short contigs will generally have poorer coverage, when considering the percentage of total bases in the contig. This is due to

a so-called edge-effect that prevents a read from mapping to a contig if the read-to-contig alignment ends in the middle of the read yet at the end of the constructed contig. However, due to the



speed and accuracy of Burrows-Wheeler style aligners, this method of validation is both rapid and sufficiently accurate to allow reasonable certainty that an assembly is valid and that the contigs represent the genomes present within the sample. Finally, read mapping can be combined with a number of other tools such as SAMtools (Li et al. 2009) to locate possible population differences such as single nucleotide polymorphisms (SNPs), insertions or deletions (indels), and other assembly errors within the contigs. This allows assemblies to be validated and potentially improved in an unsupervised manner based on the alignment of reads as well as to make empirical judgments of assembly quality.

Comparisons of Multiple Assemblies

Beyond statistical comparisons of multiple assemblies and evaluation using the raw input data, it is also possible to determine how similar two assemblies are using the same initial data. For example, for the entries listed in Table 1, there is no guarantee that the largest contigs from each sample are the same or that the contigs have been recapitulated in the various other assemblies. The mechanisms for comparing two contig sets to each other are evolving and can range from full assembly alignments using BLAST- (McGinnis and Madden 2004) or NUCmer-based comparisons (Delcher et al. 2002) to protein coding content-based analyses to more sophisticated methods. In the future, training of better assembly pipelines may involve evaluating differences among several results in terms of possible rearrangements, SNPs, indels, and errors in joining repetitive regions to determine if one methodology can be considered consistently better than another.

Generalized References and Site-Specific References for Validation

The recent explosion in sequencing capacity has resulted in an ever-increasing number of draft and

finished reference bacterial genomes. These genomes are useful both for phylogenetic and functional classification and for validation of assembly. When it is known or suspected that a particular organism is present within a sample (e.g., *Rhizobium spp.* are expected in rhizosphere samples, while *Escherichia coli* are generally found in fecal samples), alignments against such references can be used to validate contigs that are generated from the metagenomic sample in question.

In the future, reference-based approaches may be best utilized in a sample-specific manner to both contribute to and help validate metagenomic assemblies by using draft reference genomes generated via single-cell (or microcolony) isolation from the same site, followed by amplification and sequencing. The advent of multiple displacement amplification to allow for the sequencing of minute quantities of DNA, including single cells or clusters of cells, shows great promise for metagenomic projects by allowing the inclusion of sample-specific genomes to be used in reference-based assembly methods.

Metagenome Assembly Standards: A Proposed Tiered System

As a nascent field, the methodology for metagenome assembly is still under great flux. Currently available tools are able to produce valid, useful assemblies of some fraction of any metagenomic sample. However, these assemblies must be considered as a set of draft contigs only, particularly if no form of validation has been performed. Read-based validation can be used to inform and improve on assemblies; however this is a time-consuming process and should not be expected to be a long-term, high-throughput solution for metagenome assemblies. However, this does not obviate the need for validation protocols; it merely highlights the lack of algorithmic approaches to the technique.

There are several promising areas of assembly investigation that could produce assemblies distinguishable from draft or validated draft metagenomic assemblies. These include the use



Challenge of Metagenome Assembly and Possible Standards, Table 2 Proposed statistical reporting metrics for metagenome assembly

Proposed metric	Description
Percent of read incorporation	Percentage of read mapping or incorporated into assembly. This serves as a metric as to how much additional sequencing may be required for better assembly
Size of metagenome assembly	Number of base pairs included in the final assembly. This is a measure of how many genomes may have been assembled and can be utilized to determine what additional sequencing will be allowed in terms of additional sequence data incorporation
Largest contig size	This is typically a measurement of how well the most abundant organism assembled
Number of bases in large contigs	This measurement is similar to largest contig size but also allows depth of analysis to potentially include less well-assembled species
Fold coverage histogram	A histogram describing the number of bases covered at a given fold coverage. This will indicate the variation between abundant and non-abundant organisms

of reference genome datasets to improve assemblies, the inclusion of long read technologies to help generate longer contigs and scaffolds as well as to allow linkage of genetic differences among haplotypes, and the use of iterative and combined assembly methods to correct “invalid” contig and scaffold regions and to find previously unreported overlaps among contigs and reads.

In order to provide a complete assembly overview, the standardized reporting of two important pieces of information for any assembly of metagenomic sample is proposed. Tables 2 and 3 describe a first approximation of reporting that would help disseminate information regarding the quality (Table 2) and assembly levels (Table 3) of metagenome assemblies to a broader audience. The first and most important level of reporting is an accurate and consistent description of the assembly metrics as discussed above and in Table 2. These metrics should include, at a bare minimum, the percentage of reads

Challenge of Metagenome Assembly and Possible Standards, Table 3 Classification of assembly methods for metagenomes. Reporting would ideally describe both classification and statistics described in Table 2

Quality	Description
Draft	One assembler, one parameter
Quality draft (QD)	Multiple assemblers, multiple parameters, merging-based final assembly
Binning assisted (HQD)	Multiple parameter assemblies performed by binning of reads into subsets, followed by merging-based final assembly
Reference-guided RHQD	Binning based on reference sequences, followed by HQD assembly
Location-specific reference-guided assembly	Reference-guided assembly including sequencing and assembly of individual isolate, single cells, or microcolony-based organisms isolated from the same environment as the metagenome sample in question

incorporated in a sample, the total number of bases in the resulting assembly, the size of the largest contig, and the number of bases in the largest 100, 1,000 and 100,000 contigs. Additional options can include a histogram of fold coverage of assembled contigs and alternative measures of assembly. The second level of reporting requires a community acceptance of assembly types, similar to isolate genome assemblies. The current default methodology for metagenome assembly (use of a single assembler, with a single or best parameter selected) is proposed to be called a Draft Metagenome Assembly. Iterative and multiple assemblies coupled with the merging of contigs and validation/correction of contigs, such as that utilized at the DOE Joint Genome Institute and Los Alamos National Laboratory, could be considered high-quality draft. Additional levels of quality require technologies that are not currently adopted, including the use of general reference genomes to perform reference-guided assemblies. Finally, the best assembly possible will require sequencing and assembly of genomes gathered by use of



single-cell or microcolony isolation techniques for organisms present in the study site. This best-case scenario would allow binning, assembly, and in-depth analysis of both the reads and the contigs assembled.

The levels of validation, including read mapping or correction of reads, should be reported. Using current technologies, this is the most likely end point for metagenome assemblies for the foreseeable future. The final stages of improvement will result in near-total incorporation of all generated reads into a final assembly set. Finally, for all assembly classifications, it is also important that metadata, including the sample type, the amount and type of sequencing technologies as well as any modifications (trimming, filtering, binning) to the data, the numbers and types of reference genomes used to guide assembly, as well as the percent of reads incorporated, be attached to all assemblies, for future analysis to be applicable to the same samples.

Summary

This entry discusses the current difficulties associated with metagenomic assembly and presents a path for systematic, universally understood and accepted methods for validation and a classification system for metagenomic assemblies. Each of these areas will require intensive research and tool development to approach the specified methods and to generate standard metrics for analysis and comparisons. There is still a strong need to develop universally applicable validation methods as well as a need to develop a panel of defined datasets for new techniques to be validated against. Validation in this manner, coupled with active development of new methods of assessing and reporting the quality of assembly techniques, will maximize the possibility of generating broadly applicable and accurate assembly tools that not only perform well using a single method of validation. In all, these proposed reporting mechanisms (metagenome metadata) will improve the ability of researchers to effectively and confidently utilize metagenome assembly data.

References

- Boisvert S, Laviolette F, et al. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol.* 2010;17(11):1519–33.
- Chain PSG, Grafham DV, et al. Genome project standards in a New Era of sequencing. *Science.* 2009;326(5950):236–7.
- Delcher AL, Phillippy A, et al. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 2002;30(11):2478–83.
- Godoy-Vitorino F, Goldfarb KC, et al. Comparative analyses of foregut and hindgut bacterial communities in hoatzins and cows. *Isme J.* 2012;6(3):531–41.
- Gordon D. Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinforma.* 2003. Chapter 11: Unit11 12.
- Huttenhower C, Gevers D, et al. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402):207–14.
- Kant R, van Passel MWJ, et al. Genome sequence of “*Pedospaera parvula*” Ellin514, an aerobic verrucomicrobial isolate from pasture soil. *J Bacteriol.* 2011;193(11):2900–1.
- Lampe JW. The human microbiome project: getting to the guts of the matter in cancer epidemiology. *Cancer Epidemiol Biomarkers Prev.* 2008;17(10):2523–4.
- Langmead B, Trapnell C, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3).
- Leung K, Zahn H, et al. A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities. *Proc Natl Acad Sci.* 2012;109(20):7665–70.
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589–95.
- Li H, Handsaker B, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.
- McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 2004;32(Web Server issue):W20–5.
- Methé BA, Nelson KE, et al. A framework for human microbiome research. *Nature.* 2012;486(7402):215–21.
- Miller JR, Koren S, et al. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010;95(6):315–27.
- Namiki T, Hachiya T, et al. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012.
- Peng Y, Leung HCM, et al. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics.* 2011;27(13):194–101.
- Schatz MC, Phillippy AM. et al. Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. *Brief Bioinform.* 2011.

Scholz MB, Lo CC, et al. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol.* 2012;23(1):9–15.

Yilmaz P, Gilbert JA, et al. The genomic standards consortium: bringing standards to life for microbial ecology. *Isme J.* 2011;5(10):1565–7.

CLUSEAN, Overview

Tilman Weber and Kai Blin

Interfakultäres Institut für Mikrobiologie und Infektionsmedizin Tübingen, Mikrobiologie/Biotechnologie, Eberhard-Karls Universität, Tübingen, Germany

Synonym

CLUster SEquence ANalyzer

Definition

CLUSEAN, the *CLUster SEquence ANalyzer*, is a BioPerl-based software pipeline for the annotation of secondary metabolite biosynthetic gene clusters encoding the biosynthesis of molecules with, e.g., antibiotic or anticancer activities. CLUSEAN contains modules for automated homology search, protein domain identification, and, in case of modular polyketide synthases and non-ribosomal peptide synthetases-containing pathways, substrate prediction for the biosynthetic enzymes.

Introduction

A majority of antimicrobials used in human medicine to combat infectious diseases, e.g., tetracycline, penicillin, vancomycin, or erythromycin, many anticancer drugs, and other bioactive molecules, e.g., the immunosuppressant rapamycin, are derived from microbial secondary metabolites, also denoted as natural products. These compounds are mainly synthesized by bacteria

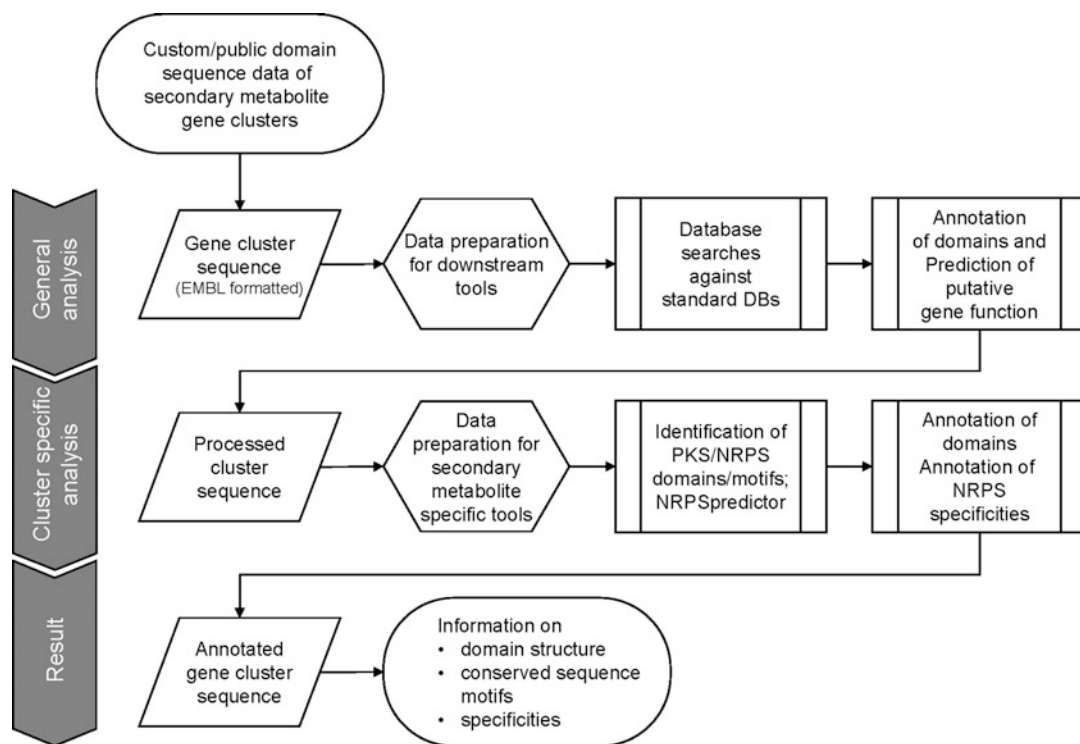
and fungi. Traditionally, screening for such compounds is performed by isolating potential producers from diverse sources, testing many different growth conditions, chemically isolating and purifying the produced compounds, and subsequently determining their structure and testing their bioactivities. The progress in the development of novel high-throughput sequencing technologies that allow cost-effective sequencing of microbial genomes, and the increased knowledge on the biosynthetic pathways of natural product formation, recently led to the availability of genome-mining methods as an alternative approach to the time- and cost-effective biological/chemical screening approach. In genome mining, DNA sequence information is used to assess and evaluate the genetic potential of the investigated strain. This approach is possible as the molecular principles underlying secondary metabolite biosynthesis – despite the vast diversity and number of compounds – are highly conserved.

Aim and Scope of CLUSEAN

CLUSEAN, the *CLUster SEquence Analyzer* (Weber et al. 2009), is a BioPerl (Stajich et al. 2002)-based tool collection that allows a semiautomatic annotation and analysis of secondary metabolite gene clusters. A typical CLUSEAN analysis run is carried out in two stages: in the first stage, the gene products of whole genomes or biosynthetic gene clusters are compared against standard databases. In the second stage, secondary metabolite-specific analyses are carried out (Fig. 1).

During the first analysis stage, similar proteins of all annotated gene products are identified using BLAST (Altschul et al. 1990) against the non-redundant protein database, and conserved protein domains are identified with the HMMER (Eddy 2001) software searching against the Pfam protein family database (Bateman et al. 2002).

In the second stage, protein domains commonly observed in the context of secondary metabolism are identified using HMMER on



CLUSEAN, Overview, Fig. 1 Data processing within the CLUSEAN annotation pipeline (Reprinted from Weber et al. 2009 with permission from Elsevier)

a custom HMM profile database. This analysis leads to the identification of the conserved functional domains in modular polyketide synthases and non-ribosomal peptide synthetases (NRPS). Amino acid specificities of NRPS adenylation domains are predicted with an integrated NRPS predictor (Rausch et al. 2005; Röttig et al. 2011).

All annotation is provided as annotation tags in EMBL-formatted sequence flat files which can be imported in standard sequence analysis tools, e.g., the Artemis sequence editing software (Rutherford et al. 2000) or the ACT sequence comparison tool (Carver et al. 2005). The CLUSEAN annotation can be exported in tabulator, or comma-separated text files, or as MS Excel tables.

In addition to the prediction modules integrated into the automated pipeline script, additional tools exist to define KS types of trans-AT PKS according to Nguyen et al. (2008) and to check the presence of conserved amino acids in

the catalytic domains of modular PKS and NRPS, which can indicate functionality of the enzymatic domain and thus has an influence on the synthesized product.

CLUSEAN has been included as an integral part into antiSMASH, antibiotics, and secondary metabolites analysis shell, <http://antismash.secondarymetabolites.org> (Medema et al. 2011), where most analysis results can be accessed interactively or downloaded on a user-friendly web page.

Availability and System Requirements

CLUSEAN is freely distributed under a GNU GPL and can be downloaded from <https://bitbucket.org/tilmweber/clusean>.

CLUSEAN has the following software requirements: BLAST + 2.2.24 (or later), HMMer 2, HMMer 3, BioPerl 1.6.9 (or later),



and Perl libraries `Sort::ArrayOfArrays` and `Spreadsheet::WriteExcel::Simple`.

Summary

Mining microbial and fungal genome data is a successful novel strategy to identify producers of novel drug candidates. CLUSEAN is a widely used tool to provide automated annotation of secondary metabolite gene clusters and to extract information from the sequence data which can be basis for the deduction of the putative biosynthetic products.

Cross-References

- ▶ [antiSMASH](#)
- ▶ [Bacteriocin Mining in Metagenomes](#)

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
- Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, et al. The Pfam protein families database. *Nucleic Acids Res.* 2002;30(1):276–80.
- Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the artemis comparison tool. *Bioinformatics.* 2005;21(16):3422–3.
- Eddy SR. HMMER: profile hidden Markov models for biological sequence analysis. 2001. Available from: <http://hmmer.janelia.org/>.
- Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. AntiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 2011;39(Web Server issue):W339–46.
- Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T, et al. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol.* 2008;26(2):225–33.
- Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* 2005;33(18):5799–808.
- Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* 2011;39(Web Server issue):W362–7.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. *Bioinformatics.* 2000;16(10):944–5.
- Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: perl modules for the life sciences. *Genome Res.* 2002;12(10):1611–8.
- Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol.* 2009;140(1–2):13–7.

Computational Approaches for Metagenomic Datasets

Colin Davenport

Hannover Medical School, Hannover, Germany

Synonyms

Bioinformatic analysis; Metagenome data analysis

Definition

The process of gaining information about a metagenomic community from sequence data using a variety of interdisciplinary techniques and approaches.

Introduction

The history of computational approaches to metagenomic data analysis is brief given the rapid development of the field. In 1998, a visionary paper described techniques for investigation of the molecular diversity of environmental communities and coined the term metagenome (Handelsman et al. 1998). Focus was placed on screening clone libraries for interesting biological activities, a mainly laboratory-based endeavor

which has been continually successful at identifying relevant novel genes with novel functionality. Other researchers took a more technology-driven approach by randomly sequencing metagenomic DNA from an acid mine biofilm and the well-known Sargasso Sea projects. These sequence-based approaches required considerable computational capacity for assembly and similarity searches. The Sargasso Sea project in particular provided researchers with considerable headaches with data analysis due to its sheer size. Microbiomes of humans and mice quickly followed and have remained a major source of metagenomic data to date, particularly with respect to diet, health, and disease. As sequencing has become cheaper so has the demand for multiple groups of samples and detailed comparative analyses of time courses. Study design with control groups has in turn become more complex and critical. Some groups have even flirted with the next stage in community analysis and investigated metatranscriptomes and metaproteomes of environmental communities.

While targeted sequencing of single genes has been the norm for most projects to date (2012), many groups are becoming interested again in true metagenomics *sensu stricto*, i.e., the investigation of microbial community structure and function using whole genome shotgun metagenome datasets. Large studies such as Metahit (Qin et al. 2010) and a comprehensive cow rumen analysis (Hess et al. 2011) have driven the acceptance of this approach. Storage requirements and computational resources can quickly become limiting in these types of analyses, though many of the state-of-the-art algorithms described below do a good job at mitigating these factors.

In the following sections we concentrate on the principles, advantages, and problems of the main approaches to computational metagenome analysis. We highlight existing approaches and mention some of the most widely applied software in the field, which is then listed with web links in Table 1. We first deal with 16S rDNA profiling, before describing the state of the art in metagenome assembly and taxonomic assignment algorithms. Subsequently, we discuss

programs intended for annotation of metagenomic sequences before making some comments on relevant statistical analyses. Lastly, we briefly review the current state of affairs in metadata collection and standards.

16S rDNA Profiling

Targeted sequencing of the ubiquitously present 16S SSU bacterial and archaeal ribosomal gene has become a common technique in deriving estimates of microbial diversity in a community. Despite its popularity, with approximately 90 % of all datasets having been produced according to this method (Davenport and Tümmler 2012), this approach is not metagenomics in the strict sense. 16S rDNA profiling completely ignores functional diversity such as gene content and accessory genome elements while also overlooking potentially important viral and eukaryotic taxa. However, this approach provides consistent qualitative estimates of bacterial and archaeal members of the community, although care must be taken with quantitative aspects. In addition, several capable software packages are available for analysis. Errors can occur for a number of reasons, including copy number variations of ribosomal RNA operons in prokaryotic genomes, the lack of coverage of “universal” primers, and multi-template PCR biases. A recent effort has incorporated copy number information of the 16S gene and reported improvements in microbial diversity estimates (Kembel et al. 2012).

In the past, 16S genes were sequenced using long Sanger reads, and only fully covered genes were used for analysis. Later, the long-read 454 sequencing technology made targeting of one or more of the shorter so-called hypervariable regions of 16S genes possible at a much reduced cost. In turn, others have investigated the use of overlapping paired end Illumina short-read technologies to sequence hypervariable regions. There is still debate about whether only targeting regions of the 16S gene leads to similar results as using the full length gene and if this leads to biases for some phylogenetic groups (Pinto and Raskin 2012).



Computational Approaches for Metagenomic Datasets, Table 1 A non-exhaustive list of software used directly or indirectly in metagenomics and mentioned in the article

Program	Availability (online tool or standalone)	Purpose	URL
Allpaths LG	Standalone	Read assembly	http://www.broadinstitute.org/software/allpaths-lg/blog/?page_id=12
PE-Assembler	Standalone	Read assembly	http://www.comp.nus.edu.sg/~bioinfo/peasm/PE_manual.htm
SSPACE	Standalone	Contig scaffolding	http://www.baseclear.com/landingpages/sspacev12/
AMOS (AMOScmp)	Standalone	Assisted read assembly	http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOScmp
Velvet (Columbus)	Standalone	Assisted read assembly	http://www.ebi.ac.uk/~zerbino/velvet/
Newbler (runMapping)	Standalone	Assisted read assembly	http://454.com/products/analysis-software/index.asp
VAAL	Standalone	Assisted read assembly, polymorphism discovery	ftp://ftp.broadinstitute.org/pub/crd/VAAL/VAAL_manual.doc
MetaVelvet	Standalone	Metagenome assembly	http://metavelvet.dna.bio.keio.ac.jp/
Meta-IDBA	Standalone	Metagenome assembly	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba/
Cross_match	Standalone	Masking of vector sequences	http://www.phrap.org/phredphrapconsed.html
Phrap	Standalone	Long-read assembly	http://www.phrap.org/phredphrapconsed.html
CAP3	Standalone	Long-read assembly	http://seq.cs.iastate.edu/cap3.html
Glimmer-MG	Standalone	Ab initio gene finding in metagenomic samples	http://www.cbcb.umd.edu/software/glimmer-mg/
MetaGeneMark	Online and standalone	Ab initio gene finding in metagenomic samples	http://exon.gatech.edu/metagenome/Prediction/ http://exon.gatech.edu/license_download.cgi
FragGeneScan	Standalone	Ab initio gene finding in metagenomic samples	http://omics.informatics.indiana.edu/FragGeneScan/
MetaGeneAnnotator	Standalone	Ab initio gene finding in metagenomic samples	http://metagene.cb.k.u-tokyo.ac.jp
Orphelia	Standalone	Ab initio gene finding in metagenomic samples	http://orphelia.gobics.de/
Prodigal	Standalone	Ab initio gene finding in metagenomic samples	http://prodigal.ornl.gov/
BLAST	Standalone and online	Homology search	http://blast.ncbi.nlm.nih.gov/
BLAT	Standalone and online	Homology search	http://genome.ucsc.edu/FAQ/FAQblat.html
HMMer	Standalone and online	Homology search	http://hmmer.janelia.org/
MG-RAST	Online	Metagenomic analysis pipeline	http://metagenomics.anl.gov/
IMG/M	Online	Metagenomic analysis pipeline	http://img.jgi.doe.gov/cgi-bin/m/main.cgi
CAMERA	Online	Metagenomic analysis pipeline	http://camera.calit2.net/
WebMGA	Online	Metagenomic analysis pipeline	http://weizhong-lab.ucsd.edu/metagenomic-analysis/

(continued)

Computational Approaches for Metagenomic Datasets, Table 1 (continued)

Program	Availability (online tool or standalone)	Purpose	URL
QIIME	Standalone and online	Metagenomic analysis pipeline	http://qiime.org/
Mothur	Standalone	Metagenomic analysis pipeline	http://www.mothur.org/
Uclust	Standalone	Sequence fragment clustering	http://drive5.com/usearch/manual/uclust_algo.html
tRNAscan-SE	Standalone and online	tRNA detection	http://lowelab.ucsc.edu/tRNAscan-SE/
InterProScan	Standalone and online	Protein functional analysis	http://www.ebi.ac.uk/Tools/pfa/iprscan/
MEGAN	Standalone	Comparative metagenomic analysis	http://ab.inf.uni-tuebingen.de/software/megan/
Vegan (R package)	Standalone	Major ordination methods	http://cc.oulu.fi/~jarioksa/softhelp/vegan.html
Picard (CollectGcBiasMetrics)	Standalone	GC bias metrics	http://picard.sourceforge.net/
MetaPhlAn	Standalone	Taxonomic classification	http://huttenhower.sph.harvard.edu/metaphlan
PhyloPythiaS	Online	Taxonomic classification	http://phylopythias.cs.uni-duesseldorf.de/index.php?phase=wait
Genometa	Standalone	Taxonomic classification	http://genomics1.mh-hannover.de/genometa/
PhymmBL	Standalone	Taxonomic classification	http://www.cbcb.umd.edu/software/phymm/

A typical 16S rDNA profiling analysis would include the following steps. Artifacts and errors can be excluded for the most part by rigorously filtering sequence reads according to empirical experience. Commonly used filter steps include rigorous checking of sequencing barcodes, read average Phred quality scores of 20 or more, and exclusion of reads with uncalled nucleotides (Ns). Software tools such as QIIME (Table 1) facilitate the computational processing of the sequence data. The next step is alignment to a reference sequence of mostly full length 16S genes such as the Greengenes or Silva databases (Table 2) using the naïve Bayesian Classifier from the Ribosomal Database project (Table 1). Lastly, software tools such as QIIME, Mothur, or the R Vegan package (Table 1) allow calculation of diversity metrics, rarefaction curves and various statistical analyses (see below) which provide further information about the community under study.

Assembly

Longer DNA sequences extracted from metagenomic samples afford more precise taxonomic assignment and annotation by providing more information for homology and composition analyses at the cost of losing quantitative information on the number of reads attributed to taxa. Also, full length genes may be recovered from resulting assemblies when using a gene prospecting approach investigating phylogenies created with single-copy genes. Therefore, depending on available coverage, sequence read assembly can be considered. Estimation of the proportion of bacteria of interest present in a metagenomic sample and the total number of microbial (nonhost) reads can give a general idea of how successful the assembly step may be. A good discussion on estimation of probability to assemble a whole genome or achieve a specific average contig length for a particular microbe of

Computational Approaches for Metagenomic Datasets, Table 2 A non-exhaustive list of online databases used directly or indirectly in metagenomics and mentioned in the article

Online database	Data	URL
Greengenes	16S dRNA	http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
Silva	rRNA	http://www.arb-silva.de/
PFAM	Protein families	http://pfam.sanger.ac.uk/
TIGRFAM	Curated multiple sequence alignments, HMMs for protein sequence classification	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi
KEGG	Genomic, chemical, and systemic functional information	http://www.genome.jp/kegg/kegg1.html
EggNOG	Orthologous groups of genes	http://eggnog.embl.de/version_3.0/
COG	Clusters of orthologous groups of proteins	http://www.ncbi.nlm.nih.gov/COG/
SEED	Functional classification	http://www.theseed.org/wiki/Main_Page
GenBank	Annotated DNA	http://www.ncbi.nlm.nih.gov/genbank/
RefSeq	Genomic DNA, transcripts, and proteins	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProt	Protein sequence and functional information	http://www.uniprot.org/
GO	Gene ontology	http://www.geneontology.org/
PATRIC	Protein families	http://www.patricbrc.org/portal/portal/patric/Home
PROSITE	Protein domains, families, and functional sites	http://prosite.expasy.org/
PRINTS	Protein fingerprints	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php
Pfam	Protein families	http://pfam.sanger.ac.uk/
ProDom	Protein domain families	http://prodom.prabi.fr/prodom/current/html/home.php
SMART	Proteomes	http://smart.embl-heidelberg.de/
PIR superfamily	Protein families, functions and pathways, interactions, structures and structural classifications, genes and genomes, ontologies, literature, and taxonomy	http://pir.georgetown.edu/pirwww/dbinfo/iproclass.shtml
Superfamily	Structural and functional annotation of proteins and genomes	http://supfam.cs.bris.ac.uk/SUPERFAMILY/
Gene3D	Protein domains	http://gene3d.biochem.ucl.ac.uk/Gene3D/
Panther	Gene functions	http://www.pantherdb.org/
HAMAP	Microbial proteomes	http://pbil.univ-lyon1.fr/help/HAMAP.html

interest in a metagenomic sample is given in Wendl et al. (2012). Several sequencing biases are likely to differentially affect the quality of assembly for bacterial genomes due to their DNA composition. For example, reads from GC-rich genomes are likely to be underrepresented due to sequencing issues caused by the secondary structures and higher melting temperatures of GC-rich sequences (Frey et al. 2008) interfering with polymerization and ligation reactions. Polymerization errors occurring during PCR amplification will also result in sequencing errors. Presence of homopolymer runs is likely to

introduce frameshift errors during sequencing (especially with the 454 and Ion Torrent platforms). Genome length would also affect the number of sequenced reads for a particular bacterium. These biases should be taken into account in quantitative metagenomics analysis. In summary, the dominant (or more common) species in the sample should produce more reads and are more likely to assemble better, though depending on their DNA composition and the sequencing technology used the higher rate of sequencing errors can cause poor-quality assemblies for these dominant species.

Read length plays an important role in achieving accuracy and high genomic coverage of an assembly. There is an ongoing debate whether it is sufficient to use short reads for metagenomic analysis (Luo et al. 2012) or alternatively only use reads that are as long as possible for proper annotation of genes, which ideally should include their promoters, riboswitches, co-operonic genes, and signature protein domains (Temperton and Giovannoni 2012). Regardless, using longer paired end reads will always result in more accurate assemblies better covering the length of the assembled genomes.

Due to sequencing costs, next-generation sequencing technologies offering relatively long, paired end reads and yet providing high coverage (enough to assemble underrepresented bacterial genomes) are ideal for metagenomics projects. For example, the Illumina MiSeq instrument can produce 8 Gb of 2×250 bp reads in one run at a lower per base cost than Sanger or 454 sequencing technologies, while still offering substantial sequence length. Another reason to avoid short single-end reads in the assembly step is that aside from problems with assembly of repetitive regions these reads are also likely to produce misassembled chimeric contigs. Generally, repetitive regions of any length can be assembled by using multiple libraries of paired end reads with varying insert sizes. Libraries with shorter insert sizes can be used to build initial contigs avoiding misassembly of repetitive reads into pseudo contigs. Longer insert libraries can be used for scaffolding and gap filling of the initial contigs. For these reasons programs like Allpaths LG and PE-Assembler (Table 1) require paired end libraries with different insert sizes. Alternatively, standalone scaffolding tools such as SSPACE (Table 1) can be utilized to scaffold already existing contigs using long insert libraries. It is recommended to filter out poor-quality reads and analyze average base quality of the remaining reads. Based on this analysis a minimal required read length can be determined for uniform or adaptive (quality-based) trimming. When references of closely related organisms are available, it is possible to perform assisted assemblies using the reference genome

as a guide. Programs such as AMOScmp, Velvet (Columbus), Newbler (runMapping), and VAAL (Table 1) can be used for assisted assembly. Most short-read assemblers are designed to assemble a single genome and, thus, not optimal for assembly of metagenomic samples where reads from homologous regions of less represented genomes can be treated as error reads. Development of metagenomic assemblers, such as MetaVelvet and Meta-IDBA (Table 1) should address this problem. In these assemblers the de Bruijn graph (Flicek and Birney 2009) for the entire assembly is analyzed for presence of subgraphs corresponding to multiple bacterial genomes in the sample.

Due to high costs Sanger shotgun sequencing of metagenomic samples is less attractive. However, it can be considered for low-diversity samples. Assembly of long Sanger reads can generate nearly complete bacterial genome sequences, ideal for subsequent annotation efforts. Cloning vectors offer large insert sizes, e.g., bacterial artificial chromosomes (BACs) (up to 200 Kb), yeast artificial chromosomes (YACs) (up to 1.5 Mb), and fosmids (up to 90 Kb). Therefore, it is possible to amplify, sequence, and assemble manageably large stretches of DNA sequence randomly positioned within a genome and overlapping each other, thus leading to assembly of nearly complete genomes. Vector sequences should be excluded from the assembly using vector-masking software such as Cross_match (Table 1). The two most commonly used long-read assemblers are Phrap and CAP3 (Table 1). Trimming of poor-quality 5'- and 3' ends should be implemented to improve the assembly.

Taxonomic Assignment (Binning)

Assignment of derived sequence reads to their taxon of origin is a key goal of most metagenomic studies. This process is also referred to as binning, as sequences are placed into "bins" representing the various taxa. Two types of assignment have been largely utilized to date, compositional and sequence similarity based.

Compositional signals depend on the concept of the genome signature. This relies on the simple idea that the composition of oligomers such as tetramers from closely related genomes is more similar than those from distantly related genomes (Mrázek 2009). There is a significant body of research on this topic involving research into identification of genomic islands, genes of aberrant composition, genome evolution, and classification of metagenome sequences. The main advantage of compositional classifiers is that they can determine associations in the absence of alignment by assessment of normalized oligomer counts. Furthermore, unsupervised machine learning techniques such as self organizing maps are not biased by the availability of fully sequenced reference sequences. The main drawback of these classifiers is the long sequences needed to derive robust oligomer statistics. For example, the program PhyloPythiaS (Table 1) and more recent frameworks typically require more than 1,000 bp of input sequence. As such, they are not able to assign the numerous short reads from modern Illumina and SOLiD sequencers to various taxa, which is certainly possible with other techniques (see next section), but they do work well on assembled contigs. This leads to problems, as contigs do not reflect the distributions of raw reads initially observed in the metagenome. Also, some distantly related organisms may not have sufficiently divergent genome signatures for assignment.

Compositional data has been used in a number of metagenomic studies. Willner et al. (2009) analyzed the compositions of 86 microbial and viral metagenomes sequenced with 100 bp 454 reads. They found that dinucleotides explained more of the variance observed than higher order nucleotides such as tetramers, although this is probably due to the short length of the read sequences used, which leads to non-robust statistics for higher order oligomers. Another advantage of oligomers is their ability to detect contamination in contigs due to the divergent oligomer profile and their relatively modest computational burden.

A more widely-used method of binning sequences is to find sequences by similarity,

given that a reference sequence is available. The key advantages of these methods are that they are an accurate and widely accepted robust method and also can give direct knowledge of gene content following alignment. The main disadvantage is the lack of available reference sequence for some taxa, which can lead to false overrepresentations of somewhat related taxa in the estimates. Also, computation tends to be more demanding than the compositional approach. This is especially so in the case of the BLAST algorithm used in the popular software MEGAN (Table 1). MEGAN uses a lowest common ancestor approach to assign reads with two database hits to a taxon. If the reads hit unrelated bacteria from different phyla, the lowest common ancestor will be that prior to phylum, such as Bacteria. However, if the reads hit different species of say Burkholderia, the algorithm will appoint a hit to the genus Burkholderia. BLAST is effective since it allows alignments against the well characterized metagenomic protein space, as well as the less well-known nucleotide space.

Another popular solution is the web-based analysis toolbox MG-RAST (Table 1). MG-RAST allows taxonomic binning, but is more focused on functional investigation and comparison of metagenomes. It is further detailed in the Annotation section below. WebMGA (Table 1) is an alternative very capable metagenomics web server which uses efficient algorithms such as FR-HIT and CD-HIT for flexible read alignment and highly efficient clustering, respectively. MetaPhlAn (Table 1) attempts to optimize unique clade-specific marker genes as a reduced reference sequence of about 400 thousand genes most representative of each taxonomic unit and map reads to it. This kind of mapping potentially allows assignment of reads to higher taxonomic levels such as species and has the advantage of being extremely rapid. A further solution which seeks to use curated reference sequences is Genometa (Table 1). This GUI program puts emphasis on finding the mapping coordinates of even very short reads in a genome to check if a taxon is actually present, or if it is more likely to be either contamination or just a related ORF or genomic island.

Other programs aim to combine compositional and similarity based tools. A well-known approach is PhymmBL (Table 1). This program uses both BLAST and compositional attributes to assign even reads as short as 100 bp. The authors found this technique to be more accurate than either of the methods alone and have continued to improve their software. In general, binning is still a difficult task, and algorithms which work very well on one dataset may be extremely limited on the next. As such, we recommend using at least two binning approaches on the sample to gain the maximum possible information.

Annotation

Annotation of metagenomics samples requires identification of features of interest in the assembled fragments or reads binned into their Operational Taxonomic Units (OTUs). For ab initio identification of potential gene sequences entire ORFs should be located. Various software exists to perform this task in metagenomic projects, e.g., Glimmer-MG, MetaGeneMark, FragGeneScan, MetaGeneAnnotator, Orphelia (Table 1). These programs utilize various types of Markov models for analysis of codon usage or frequency of other genome composition elements in the binned genomes. However, instead of a single model the analysis is based on multiple Markov models trained with data from a large variety of bacterial species. The trained model providing the best fit is then selected for gene prediction. Given the complexity of metagenomic assemblies, especially when only short reads are utilized, it is expected that a large proportion of assembled contigs may only have partial ORFs. These sequences can still be included in homology analysis using BLAST, BLAT, or HMM (Table 1) searches against gene or protein nonredundant databases. There are a number of online pipelines, e.g., MG-RAST, IMG/M, and CAMERA (Table 1), available for ab initio- and homology-based DNA structure annotation as well as functional analysis of identified genes using a battery of

publically available databases that can be used for functional annotation, such as PFAM, TIGRFAM, KEGG, EggNOG, COG, SEED, GenBank, RefSeq, UniProt, GO, and PATRIC (Table 2).

MG-RAST allows the users to upload their sequence data (in FASTA, FASTQ, and SFF format) and metadata. The uploaded data are preferred to be shared, but this is not mandatory. The data are quality controlled with the QC pipeline based on the settings provided by the user. The QC pipeline features include read quality filtration and trimming, dereplication, model organism screening, demultiplexing and merging mate pairs. Currently, the minimal accepted read length is 75 bp. Assemblies can also be submitted. Starting from version 4.0 the pipeline will also support read assembly. Submissions to this pipeline are queued and submitted for feature prediction using FragGeneScan, which identifies the most likely reading frame and performs homology search on translated features. A program called Uclust is then used to cluster 90 % identical protein fragments. The number of reads in each cluster is identified to estimate abundances. The pipeline also provides various visualization tools to view the results and to perform comparative analysis with over 590 public metagenomes.

IMG/M concentrates on comparative analysis of microbial genomes. The pipeline accepts assembled or unassembled reads. Unassembled reads are quality controlled, trimmed, and dereplicated; their low-complexity regions are masked. Aside from protein coding genes, ab initio gene finding also includes detection of CRISPRs and noncoding RNA. RNA detection is performed using tRNAscan-SE for tRNAs and HMM models for rRNAs. Coding sequences are predicted using a combination of Prodigal, Metagene, MetaGeneMark, and FragGeneScan (Table 1). Longer sequences are also searched against a local nonredundant protein database using BLASTX. IMG/M provides functional annotation of the entire metagenome and supports functional comparisons to other stored annotated metagenomes. Various visualization tools facilitate this kind of comparison,

e.g., Phylogenetic Distribution of Genes or Radial Phylogenetic Tree.

CAMERA provides a collection of online tools for metagenomics analysis. The provided tools allow the following analysis steps: sequence QC, sequence assembly, ORF prediction, RNA prediction, BLAST, clustering, functional annotation, and viral diversity estimation.

InterProScan (Table 1) is one of the most advanced programs for protein functional analysis. It incorporates BLAST and HMM searches against an array of protein domain and functional site databases (PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR superfamily, SUPERFAMILY, Gene3D, PANTHER, and HAMAP (Table 2)). Online and locally installed versions are available. Due to highly CPU-intensive nature of the BLAST and HMM searches, it is recommended to run this program on a computer cluster.

MEGAN is a standalone tool for visualization of BLAST search results as taxonomic dendrograms, functional dendrograms using the SEED classification, pathways using KEGG orthology, comparative visualization, etc. A good collection of general information about other available metagenomics software and resources can be found on <http://seqanswers.com/wiki/Metagenomics>.

Statistical Analysis

Early metagenomic datasets, such as the Sargasso Sea, were relatively simple surveying projects by design. Attempts were made to quantitate species abundances using relative abundance of reads and presence of 16 S rRNA and single-copy genes. Later studies then focused more on comparative spatial or temporal variation of the microbial community. Due to this increasing sophistication multiple metrics for characterizing the complexity of the community have been developed. As many projects in metagenome analysis are not based on strict hypothesis testing, exploratory data analysis techniques such as multivariate statistics are often employed (see below). Generally, the metric under study is the estimate of abundance of a taxon, which can be

obtained by multiple methods. An important step to ensure comparability is normalization. Normalization of these metrics can be undertaken using relative abundances, GC content, genome size, or prevalence of single-copy genes. However, particularly normalization of true metagenomic data is in a state of flux with little current consensus. Care must be taken with the GC content delivered by sequencers as a result of the different sample preparation and amplification schemes. It must be assumed that all sequencing runs have some form of quantitative bias against either or both low GC and high GC organisms, meaning that they will be underrepresented in the samples. This problem has not been widely considered in metagenomics to date. GC bias assessment programs, such as Picard's CollectGcBiasMetrics (Table 1), are particularly useful in observing and quantifying relative bias of read coverage at different GC values using just a reference sequence and a BAM alignment file. Larger genomes are more likely to be sampled in a randomly sheared metagenomic DNA sample. This can be compensated for by normalizing for genome length, if applicable for the taxonomic attribution method used.

Many metrics have been taken directly from the field of ecology. Alpha, beta, and gamma diversity summarize the species diversity in one habitat, species diversity across multiple habitats, and total diversity over total species diversity across a larger scale landscape, respectively. Species richness is simply the number of species found, while species diversity includes a measure of the abundance of members of each species. Other measures such as Shannon and Simpson indices are also available. One use case is from Dinsdale and coworkers (2008), where functional metagenomic diversity was characterized separately across a range of bacterial and viral genomes in many different habitats. Interestingly, functional metagenomics was reported by Dinsdale and coworkers to explain a larger proportion of the variance in each dataset (about 75 %) and thus be predictive of metabolic capacity within the taxa of an ecosystem, than analysis of taxa by 16S rRNA genes only (about 10 %).

The aforementioned indices attempt to characterize a highly multidimensional dataset into a single number, which can be useful as a summary but obscures the underlying data. Therefore, advanced ordination methods for multidimensional datasets such as principal components analysis (PCA) and multidimensional scaling (MDS) have been applied to differentiate communities and reveal associations with abiotic parameters. Whichever of the many ordination methods is chosen, it is of great importance to check the variance explained by the observed components or functions. Where the principal components or alternative statistics explain little of the variance in the data, this indicates the variation in the data cannot be explained by the variables measured, and caution must be taken in interpreting the results. Various clustering algorithms have also been demonstrably useful in grouping similar datasets based on measures from normalized read counts to oligomer content and differentiating them from controls in a manner identical to the clustering schemes popular in microarray group expression analysis (Mrázek 2009). Clustering can also be important for quality control and identification of outlier microbial communities, which may also be attributed to technical artifacts.

Further types of statistical community comparison metrics have been developed especially for metagenomics. One example is the UniFrac distance metrics used to calculate a distance measure between microbial communities using information from a supplied phylogenetic tree (Hamady et al. 2010). UniFrac uses a beta diversity measure detailing community membership over space and time, which has distinct advantages, and the phylogenetic tree method shows improvements over comparing simple lists of taxa.

Experimental design is of paramount importance in obtaining robust statistical results. Since estimates of microbial communities tend to be noisy, replicates are necessary to gain a reliable assessment of variance. As finance is usually the limiting factor, either samples can be sequenced at a lesser depth or cheaper sequencing technologies can be used.

As with any statistical analyses, care must be taken when performing multiple tests due to frequent generation of false positives. While Bonferroni corrections are extremely good at removing false positive test results, the extreme stringency of this method will certainly mask a number of biologically true associations (false negatives). As such, we advocate the use of less stringent tests such as the Benjamini-Hochberg false discovery rate method (FDR; van den Oord and Sullivan 2003). Lastly, it should be noted that extensive and high quality metadata is crucial to observing and quantitating trends in microbial community structure.

Metadata

Collection of metadata about metagenomes is essential for making the sequence data and analysis results meaningful and reusable by the scientific community. Moreover, properly collected and complete metadata can also help the scientists originally analyzing a metagenomic sample to draw conclusions about their findings that otherwise may be overlooked. A first step in this direction is development of the minimum information about a genome sequence (MIGS) specification and its extension to the minimum information about a metagenome sequence (MIMS) specification by the Genomic Standards Consortium (GSC). MIGS provides general information about a genomic sequence, similar to what is collected by the NCBI Trace Archive or NCBI Short Read Archive, extended to more detailed metadata about environment, nucleic acid sequence source, and assay preparation. MIMS extends this specification to also include metadata about the habitat, e.g., temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, and primary production (Field et al. 2008). An XML schema is used to implement the MIGS/MIMS checklist. This schema is the basis for ongoing development of the Genomic Contextual Data



Markup Language (GCDML). This language should support polymorphic validation of various taxa (requiring different checklists) and development of ontologies.

Another interesting resource that addresses the need for sharing standardized metagenomics data is the Genomes OnLine Database (GOLD, <http://www.genomesonline.org/>). This database contains a collection of completed and ongoing projects with the associated metadata, which are based on a controlled vocabulary coordinated with the GSC.

Another online resource that collects GSC-compliant metadata is CAMERA, already mentioned in the Annotation section of this review. CAMERA is involved in GSC activities and provides input for development of metagenomic metadata standards that are also used for submission of metagenomic data to CAMERA.

Summary

Recent improvements in next-generation sequencing technologies are providing new opportunities for metagenomics. While 16S rRNA gene profiling is still predominantly used for quantitative profiling of communities analysis, availability of long paired end reads produced by an Illumina MiSeq instrument or similar technology with high coverage and comparatively low cost should shift the focus of future metagenomics projects to genome assembly of microbes of interest and their functional annotation. As more microbial genomes and metagenomes are being sequenced and annotated, it is hard to overestimate the need for sharing these data and standardization of the associated metadata for comparative analysis. In this regard, development of minimum information standards (MIGS and MIMS) by the Genomic Standards Consortium and adaptation of these standards by the online analysis/storage resources, such as MG-RAST or IMG/M, are encouraging developments. As the increasing number of metagenomic projects becomes more detailed and complex, the need for more

advanced analysis software also increases. In conclusion, computational analysis of metagenomic samples is becoming more affordable and available to the research community and provides exciting research and software development opportunities.

Advances in metagenomic analysis of microbial communities also provide opportunities for metatranscriptomic and metaproteomic research.

Cross-References

- ▶ [Lessons Learned from Simulated Metagenomic Datasets](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)
- ▶ [Nucleotide Composition Analysis: Use in Metagenome Analysis](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [Silva Databases](#)

References

- Davenport CF, Tümmler B. Advances in computational analysis of metagenome sequences. *Environ Microbiol.* 2012. doi:10.1111/j.1462-2920.2012.02843.x.
- Dinsdale EA, Edwards RA, Hall D, et al. Functional metagenomic profiling of nine biomes. *Nature.* 2008;452:629–32.
- Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol.* 2008;26:541–7.
- Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods.* 2009;6: S6–12.
- Frey UH, Bachmann HS, Peters J, Siffert W. PCR-amplification of GC-rich regions: slowdown PCR. *Nat Protoc.* 2008;3:1312–7.
- Hamady M, Lozupone C, Knight R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* 2010;4:17–27.
- Handelsman J, Rondon MR, Brady SF, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:R245–9.
- Hess M, Sczyrba A, Egan R, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science.* 2011;331:463–7.
- Kembel SW, Wu M, Eisen JA, et al. Incorporating 16S gene copy number information improves estimates of



- microbial diversity and abundance. *PLoS Comput Biol.* 2012. doi:10.1371/journal.pcbi.1002743.
- Luo C, Tsementzi D, Kyrpides NC, et al. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* 2012;6:898–901.
- Mrázek J. Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol.* 2009;26:1163–9.
- Pinto AJ, Raskin L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One.* 2012;7:e43093.
- Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:59–65.
- Temperton B, Giovannoni SJ. Metagenomics: microbial diversity through a scratched lens. *Curr Opin Microbiol.* 2012;15:605–12.
- van den Oord EJCG, Sullivan PF. False discoveries and models for gene discovery. *Trends Genet.* 2003;19:537–42.
- Wendl MC, Kota K, Weinstock GM, et al. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens theorem. *J Math Biol.* 2012. doi:10.1007/s00285-012-0586-x.
- Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol.* 2009;11:1752–66.

Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes

Yong Wang¹ and Pei-Yuan Qian²

¹Division of Deep Sea Science, Sanya Institute of Deep Sea Science and Engineering, San Ya, Hainan, China

²KAUST Global Collaborative Program, Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China

Definition

The taxonomic classification of prokaryotic organisms based on morphological differences is difficult. A ribosomal RNA (rRNA) sequence has many polymorphic sites that can act as a genetic earmark to uncover the genetic background of prokaryotes (Fox 2010). Bacterial and archaeal genomes contain one to several copies of 16S

rRNA genes (rDNA), depending on the physiological condition of the microbes (Klappenbach et al. 2000; Liao 2000). Because the rRNA sequences are vertically delivered to the next generation, they cannot be inherited by a different species. Hence, 16S rRNA sequences are considered to be a stable marker of morphological difference and have been applied in the taxonomic classification of prokaryotes (Woese 1987). Since the 1980s, partial and full-length sequences have been obtained using polymerase chain reaction (PCR) technology (Lane et al. 1985). These sequences have been deposited in the Ribosomal Database Project (RDP), Greengenes (<http://greengenes.lbl.gov>), and SILVA rRNA public databases (Cole et al. 2009; Pruesse et al. 2007). The closest relative of a microbial organism of interest can be figured out by comparing the organism's rRNA sequence with the collected sequences of known species (DeLong 1992; Fuhrman et al. 1992). Moreover, with the development of next-generation sequencing techniques (Quail et al. 2012), rRNA sequences of microbial communities in environmental samples can be massively obtained in a short period (the efficiency is platform dependent). These advances in detection have dramatically improved our understanding of the communities of environmental microbes in different sites on Earth (Qian et al. 2011; Roussel et al. 2008).

The primer design is critical, regardless of which sequencing method is employed, the Sanger method or next-generation methods. With the rapid advances in metagenomics, environmental samples can now consist of thousands of microbial species (Tremaroli and Backhed 2012). Therefore, it is important to use primers that are suitable to most of the species to fully investigate the microbial community. If the primers fail to land on the matching parts of the rDNA of certain dominant microbes, then these species will be excluded from the PCR amplicons, resulting in a poor survey of the community. For example, in a study of the microbial communities in the Red Sea, the selection of primers almost failed to capture the entire SAR11 group belonging to alpha-Proteobacteria (Qian et al. 2011). Primer specification is also the

major concern in other studies (Huse et al. 2008; Huws et al. 2007; Klindworth et al. 2013).

The strategy for primer design is based on the conservation of the target sequences. Primers are designed to obtain the variant sequences between two conserved regions. The degree of conservation of the regions directly contributes to the coverage rate of the primers targeting a community in an environmental sample. The conserved regions in 16S rRNA sequences are involved in essential translational functions and interact with ribosomal proteins. For instance, universally conserved sites G530, A1492, and A1493 in 16S rRNA sequences are crucial for tRNA binding in the A site (Brimacombe and Stiege 1985; Demeshkina et al. 2012). Along with the neighboring conserved sites, these sites have been recognized to be ideal regions for primer design, as exemplified by the frequently used universal primers U519 and U1492 (Baker et al. 2003). Apart from the conserved regions, there are a total of nine variant regions that correspond to the species-specific structural sequences of ribosomal RNA (Huws et al. 2007; Wang and Qian 2009). The variant regions can be obtained through PCR, followed by sequencing and comparison for taxonomic assignment.

Conserved Regions in 16S rRNA Sequences

In a previous study, a method had been developed for the de novo identification of conserved regions in 16S rRNA sequences (Wang and Qian 2009). First, conserved sites are detected by checking the alignment file of the 16S rRNA sequences, and consecutive conserved sites are regarded as potential candidates as primers with a high coverage rate for all known species. Thus, all possible conserved regions can be located without having to understand every detail of the role of 16S rRNA sequences in ribosome and protein translation. The previous study examined long 16S rDNA sequences (>1,200 bp), but the overrepresentation of Firmicutes and Proteobacteria sequences skewed the results toward the dominant phyla (Wang and Qian 2009). The conserved

regions were searched again using nonredundant core sequences from the SILVA database. A total of 11 bacterial and seven archaeal segments with degeneration sites were obtained. Because the nonredundant sequences were used and many of them were incomplete, the identified conserved sequences had more polymorphisms and the conservation degree at both ends of the 16S rDNA sequences could not be evaluated (Table 1). However, three new conserved regions, located at the bacterial 252–275 and 547–575 regions and the archaeal 560–578 region, were found in these sequences. In the overlapping segment between 565 and 575, a universally conserved region was recognized for bacterial and archaeal sequences: 5'-TGGG[C/T][C/G/T]TAAAG-3'. This region has been used to design primers for the identification of clinical bacteria (Nikkari et al. 2002). The positions of the conserved regions are standardized to the approximate positions on *Escherichia coli* 16S rDNA. It is interesting that all the archaeal conserved segments have corresponding bacterial segments at the same standardized positions and share some conserved sites with the bacterial counterparts (Table 1).

Evaluation of Candidate Primers

Candidate primers were selected from the conserved regions and were subjected to further evaluation. The candidates were matched to the core 16S rDNA sequences, and only two mismatches were allowed between the primers and the targeting regions. The results in Table 2 show the coverage rates of the candidates on the SILVA core datasets. Archaeal primers in the 328–346, 340–357, 916–931, and 953–972 regions are associated with a low coverage rate. This was caused by the presence of short archaeal 16S rDNA in the core dataset, at least 7 % of which did not have the targeting regions for these primers. Therefore, the core sequences could have been recovered better by these candidates. In addition to these archaeal primers, the two primers initiate from position 683 in the bacterial and archaeal 16S rDNA sequences have the lowest coverage rates of

Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes, Table 1 Conserved regions in archaeal and bacterial 16S rRNA core sequences

Start	End	Conserved sequence
Bacteria		
252	275	TTGGYRRGGTAAHRGCYYACCAAG
311	365	CCACAHKGGVACTGAGAYACKGBCCACCTACGGGWGGCWGCAGTVRRGAAT
507	536	CTAACTHYGTGCCAGCAGCCGCGGTAAKAC
547	575	AGCGTTRYTCGGAWTYAYTGGGYKTAAG
683	707	GTGTAGVRGTGAAATBCGTWGAKAT
765	806	GAAAGCKWGGGKAGCRAACRGGATTAGATACCCBGGTAGTCC
883	932	CTGGGRAGTACGVYCGCAAGRBTRAAACTCAAAGGAATTGACGGGGRCYC
935	986	ACAAGCRGYGGAGYRTGTGGYTAAATTCGAHRMWAMGCGMRRACCTTACC
1,045	1,062	CAGGTGBTGCATGGYTGT
1,067	1,085	AGCTCGTGYCGTGAGRTGT
1,090	1,113	TTAAGTSCBRYAACGAGCGCAACC
Archaea		
325	359	CWRGYCCTACGGGRYGCAGCAGKCGCGAAAMCTYY
514	539	GGTGYCAGCCGCCCGGTAHACCCG
560	578	WTTAYTGGGYTAAAGCRT
679	701	GACRGTGAGGRAYGAARSCYDGG
781	806	CRAWCSGGATTAGACCSRGTAGTCC
883	931	CTGGGRAGTAYGRYCGCAAGRYTGAAACTTAARGGAATTGGCGGGGGAG
953	972	GGTYAATYGRABTCAACGC

The conserved regions were obtained by searching consecutive conserved sites in alignment file of 339 archaeal and 1,845 bacterial nonredundant core 16S rDNA sequences in SILVA database (release 108). Cutoff percentage of occurrence of a nucleotide at a conserved site is 90 %. The positions are according to *Escherichia coli* 16S rDNA positions. Abbreviations for degeneration sites are Y for C or T, R for A or G, W for A or T, K for G or T, M for C or A, S for C or G, V for not T, H for not G, B for not A, and D for not C

88.5 % and 81.1 %, respectively (Table 2). Two candidates, 683–700 and 691–707, were selected from the bacterial conserved region of 683–707 for the test; both were associated with low coverage rates. Obviously, more degeneration sites have been introduced in these bacterial primers compared with the same sets described previously (Wang and Qian 2009). This means that more polymorphisms emerged in the nonredundant dataset, which results in the low rates for the two primers. Thus, the primers from this rRNA region are not recommended due to their generally low coverage rate, although a previous study obtained a 90.5 % coverage rate for a similar primer (Wang and Qian 2009). The overall quality of the other candidate primers is high, with the average coverage rate being 92.7 % (Table 2). The best bacterial primers in this study were located at the *E. coli* positions of 547–568, 556–575, 907–928, and 1,046–1,062. These primers are able to recover

more than 95 % of the bacterial 16S rDNA sequences. For the archaeal candidates, the two candidate primers in the region of 514–539 have the highest coverage rates, 93.5 % and 93.8 %, respectively. Thus, the archaeal rDNA sequences appear to be more difficult to be fully covered than the bacterial sequences, considering the average coverage rate of 90.8 % with the low rates for the four primers at both ends ignored. In regard to universal primers, this study recommends the primers at the 515–533, 785–806, and 907–928 positions. High coverage rates for these primers were confirmed using the bacterial and archaeal datasets (Table 2).

Summary

A short list of 16S rDNA primers has been compiled using simplified nonredundant rDNA

Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes, Table 2 Evaluation of candidate primers

Position	Sequence	% coverage
Bacteria		
259–275	GGTAAHRGCYYACCAAG	93.6 %
321–338	ACTGAGAYACKGBCCACC	86.6 %
334–353	CCACCTACGGGWGGCWGCAG	94.1 %
515–533	GTGCCAGCAGCCGCGGTAA	93.6 %
547–568	AGCGTTRYCYCGGAWTYAYTGGG	95.1 %
556–575	CGGAWTYAYTGGGYKTAAG	96.9 %
683–700	GTGTAGVRGTGAAATBCG	88.5 %
691–707	GTGAAATBCGTWGAKAT	75.9 %
765–782	GAAAGCKWGGGKAGCRAA	82.1 %
785–806	GGATTAGATACCCBGGTAGTCC	94.7 %
907–928	AAACTCAAAGGAATTGACGGGG	96.6 %
946–964	AGYRTGTGGYTTAATTCGA	92.5 %
1,046–1,062	AGGTGBTGCATGGYTGT	95.8 %
1,067–1,085	AGCTCGTGYCGTGAGRTGT	93.0 %
1,090–1,113	TTAAGTSCBRYAACGAGC	90.8 %
Archaea		
328–346	GYCCTACGGGRYGCAGCAG	83.8 % ^a
340–357	GCAGCAGKCGCAAAMCT	80.2 % ^a
514–533	GGTGYCAGCCGCCGCGGTAA	93.8 %
519–539	CAGCCGCCGCGGTAHACCGC	93.5 %
560–578	ATTAYTGGGYTTAAAGCRT	90.3 %
683–701	GTGAGGRAYGAARSCYDGG	81.1 %
785–806	GGATTAGATACCCSRGTAGTCC	90.9 %
883–902	CTGGGRAGTAYGRYCGCAAG	92.6 %
897–914	CGCAAGRYTGAAACTTAA	91.4 %
907–928	AAACTTAARGGAATTGGCGGGG	92.9 %
916–931	GGAATTGGCGGGGGGAG	82.9 % ^a
953–972	GGTTYAATYGRABTCAACGC	79.9 % ^a

Degenerated nucleotides are referred to Table 1

^aCoverage percentages that need an adjustment due to incompleteness of some short 16S rDNA sequences at the region

datasets. These primers will be useful for identifying environmental microbes, as they are capable of detecting more than 90 % of the known bacteria and archaea. However, the number of prokaryotic organisms that resist being captured by these 16S rRNA primers cannot be estimated. It has been estimated that only about 1 % of the microbes on Earth are culturable. Moreover, the microbes colonizing extreme and geologically isolated environments are far from being completely explored (Pace 1997; Sogin et al. 2006). The conserved sites in the 16S rDNA sequences will be degenerated if more

sequences and polymorphisms are revealed. Alternatively, the conserved sites will be more evident if full-length 16S rRNA sequences from variant rare biospheres are found to exhibit the same conservation patterns. In return, these conservation patterns may help to understand the role of ribosome RNAs in protein translation.

The method proposed here is also useful for generating specific primers for interested taxa at lower taxonomic levels in an environment. In a previous report, the CHECK_PROBE program in the RDP database and the BLAST program were employed to predict cyanobacteria-specific

primers (Nubel et al. 1997). However, there are problems with these methods as demonstrated previously (Wang and Qian 2009). Hence, the method here is recommended since it may enable more specific primers to be generated for different taxonomic levels.

Cross-References

- ▶ [Binning Sequences Using Very Sparse Labels Within a Metagenome](#)
- ▶ [Challenge of Metagenome Assembly and Possible Standards](#)
- ▶ [I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments](#)
- ▶ [RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data](#)

References

- Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods*. 2003;55:541–55.
- Brimacombe R, Stiege W. Structure and function of ribosomal RNA. *Biochem J*. 1985;229:1–17.
- Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*. 2009;37:141–5.
- DeLong EF. Archaea in coastal marine environments. *Proc Natl Acad Sci U S A*. 1992;89:5685–9.
- Demeshkina N, Jenner L, Westhof E, Yusupov M, Yusupova G. A new understanding of the decoding principle on the ribosome. *Nature*. 2012;484:256–9.
- Fox GE. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol*. 2010;2:1–18.
- Fuhrman JA, McCallum K, Davis AA. Novel major archaeobacterial group from marine plankton. *Nature*. 1992;356:148–9.
- Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*. 2008;4:e1000255.
- Huws SA, Edwards JE, Kim EJ, Scollan ND. Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. *J Microbiol Methods*. 2007;70:565–9.
- Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of Bacteria. *Appl Environ Microbiol*. 2000;66:1328–33.
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2013;41:e1.
- Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*. 1985;82:6955–9.
- Liao D. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol*. 2000;51:305–17.
- Nikkari S, Lopez FA, Lepp PW, Cieslak PR, Ladd-Wilson S, Passaro D, Danila R, Relman DA. Broad-range bacterial detection and the analysis of unexplained death and critical illness. *Emerg Infect Dis*. 2002;8:188–94.
- Nubel U, Garcia-Pichel F, Muyzer G. PCR primers to amplify 16S rRNA genes from cyanobacteria. *Appl Environ Microbiol*. 1997;63:3327–32.
- Pace NR. A molecular view of microbial diversity and the biosphere. *Science*. 1997;276:734–40.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. 2007;35:7188–96.
- Qian P-Y, Wang Y, Lee OO, Lau SCK, Yang J, Lafi FF, Al-Suwailem A, Wong TYH. Vertical stratification of microbial communities in the Red Sea revealed by 16S rDNA pyrosequencing. *ISME J*. 2011;5:507–18.
- Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012;13:341.
- Roussel EG, Bonavita M-AC, Querellou J, Cragg BA, Webster G, Prieur D, Parkes RJ. Extending the sub-sea-floor biosphere. *Science*. 2008;320:1046.
- Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. Microbial diversity in the deep sea and the underexplored rare biosphere. *Proc Natl Acad Sci U S A*. 2006;103:12115–20.
- Tremaroli V, Backhed F. Functional interactions between the gut microbiota and host metabolism. *Nature*. 2012;489:242–9.
- Wang Y, Qian P-Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE*. 2009;4:e7401.
- Woese CR. Bacterial evolution. *Microbiol Rev*. 1987;51:221–71.

Culture Collections in the Study of Microbial Diversity, Importance

Martin Sievers

Zurich University of Applied Sciences, Institute of Biotechnology, Waedenswil, Switzerland

Introduction

Prokaryotes, which comprise the bacterial and archaeal domains, show very high biodiversity. Over 50 different phyla including candidate phyla with cultivable species and uncultivable representatives, which are only characterized via metagenomics, have been detected. Microbial strains are ubiquitous and are able to grow in extreme environments, and to determine their functions and activities in the environment is essential for our understanding of life. Culture collections can help in the cataloging and preservation of microbial strains and their genomic DNA.

Role of Culture Collections

Culture collections are important in the preservation of biodiversity and thus contribute to the objectives of the Convention on Biological Diversity (CBD; www.cbd.int) through the preservation of important genetic resources.

The primary function of microbial culture collections is to gather, maintain, and distribute strains which have unique properties and are of practical value in various applications like research, teaching, quality control assays, and biotechnology (Uruburu 2003; Emerson and Wilson 2009). Culture collections supply their users with well-characterized strains and replicable parts (plasmid, DNA) as well with the associated documentations relevant to these biological materials. Cultures, strain, and DNA from culture collections are distributed with a material transfer agreement (MTA) which provides users with all relevant handling information and regulations for commercial use of the supplied biological material.

Type strains, which constitute the name-bearing reference strain of a species and are often used in the study of bacterial systematics, are available from culture collections worldwide. Type strains must be deposited in two public collections in two different countries in order to have the name and thus the species validated (Stackebrandt 2010).

Culture collections are a valuable resource for the exploitation of biological diversity and can help countries rich in biodiversity to understand and utilize their microbial diversity more effectively (Arora et al. 2005). Culture collections also act as an interface between their providers and users of genetic resources to support fair and equitable sharing of the benefits based on documents like Prior Informed Consent (PIC) and mutually agreed terms (Sievers et al. 2010). In fulfilling these roles, culture collections have several responsibilities regarding biosafety requirements. These include compliance with international agreements and conventions on biodiversity, the support of researchers seeking intellectual property rights, and to implement new technologies and to find additional funding for their vital work.

In addition, microbial culture collections which are recognized as international depository authorities (IDA) offer deposition of microorganisms involved in inventions for patent purposes according to the Budapest Treaty (http://www.wipo.int/treaties/en/registration/budapest/trtdocs_wo002.html).

Importance of Microorganisms

Microbial strains are used for a wide range of scientific, industrial, and health-care applications, for example, as sources of enzymes, proteins, vitamins, organic acids, bioactive compounds, antimicrobial peptides, and biopolymers. Microorganisms are used in agriculture as bio-fertilizer, in wastewater treatment as agents for degradation of compounds with complex structures, for metal recovery to catalyze specific chemical reactions, for bioenergy production, as starter cultures in the

production of fermented food, as probiotics, and as reference material in diagnostics and development of new therapeutics.

In contrast to their beneficial relatives, pathogenic microbes cause severe diseases in humans, animals, and plants, resulting in significant economic loss and risk to global health. The useful products and processes provided by microorganisms can be grouped into four broad categories: fine chemicals, processes, commodities, and emerging technologies (Kuo and Garrity 2002). Thus, the use and the study of microorganisms contribute to further economic growth and health promotion and are of immense social and ecological value (Komagata 1999; Prakash et al. 2012; Smith 2003).

Microbial Diversity

Due to their myriad environmental roles and functions, microorganisms are important components of the world's biodiversity. Microbial diversity refers to the richness and degree of variability among species and strains within an ecosystem. Microbial communities of an investigated sample are composed of species which could be isolated as well as the "silent majority" species which are considered non-culturable under standard laboratory conditions and only their DNA is accessible for genetic characterization. The richness of bacterial species is highly variable in different environmental communities. Some environments like the upper atmosphere, glacial ice, and highly acidic stream waters have low numbers of bacterial species in comparison to soil, microbial mats, and marine water, which harbor vast numbers of bacterial species (Fierer and Lennon 2011). The estimation of the number of bacterial species per gram of soil is not a trivial task. Metagenomic approaches based on analysis of environmental DNA sequence data help to study microbial communities and to estimate their species richness. Based on high-throughput 16S rDNA pyrosequencing and phylogenetic analysis, the most abundant species of bacteria in different soil samples were assigned to the phyla *Proteobacteria* and *Bacteroidetes*

(Roesch et al. 2007). A large percentage of soil bacteria could not be isolated by cultivation (between 90 % and 99 % in a given sample). Soil bacteria of the phyla *Acidobacteria* and *Verrucomicrobia* are poorly represented in pure cultures, and members of the *Actinobacteria*, *Firmicutes*, and *Proteobacteria*, in contrast, are well represented in culture collections. For example, the most dominant genus of the American Type Culture Collection (ATCC) soil accessions is *Streptomyces*, belonging to the phylum *Actinobacteria*, reflecting their importance as producers of bioactive compounds and in soil ecology (Floyd et al. 2005).

Currently, culture collections cover only a fraction of the diversity of microorganisms and will benefit from the deposition of new strains which are suitable for industrial use since they represent rich and abundant source of novel molecules with various biological activities.

Identification of Strains at Species Level

Isolated strains are checked for purity microscopically, for morphological homogeneity by uniformity of colony form on agar plate, by distinct color formation on chromogenic agar, and conformation by denaturing gradient gel electrophoresis DGGE (single band obtained). Identification of pure strains at species level is usually performed using ribosomal RNA gene sequence analysis. Housekeeping genes encoding RNA polymerase beta subunit (*rpoB*), RNA polymerase sigma factor (*rpoD*), gyrase beta subunit (*gyrB*), recombinase A (*recA*), or heat shock protein (*hsp60*) provide in some cases better genetic resolution on the species level than the 16S rDNA sequence used in taxonomic studies. Combinations of housekeeping genes in multi-locus analyses provide a taxonomic tool for identification of prokaryotes at species and strain level (Moore et al. 2010). DNA sequences in combination with protein spectra obtained by MALDI-TOF-MS are very efficient to identify strains at species level. MALDI-TOF-MS used for species identification generates protein spectra in the size range between 2 and 20 kDa which is dominated by ribosomal

proteins. By use of this technology, the generated spectra of an unknown strain are compared with a reference data bank (Wieser et al. 2012).

DNA-DNA hybridization (DDH) values are used to determine relatedness between strains and strains belong to the same species when DDH values are approximately 70 % or greater (Wayne et al. 1987). Average nucleotide identity (ANI) of common genes is discussed to be an alternative method for replacing DDH. The cut-off value of 70 % DDH for species delineation correlates to 95 % ANI value (Goris et al. 2007). ANI can be calculated by partial sequencing of the genomes (at least 20 %) of the query strains (Richter and Rosselló-Móra 2009). Unique strains of one species can be identified by metabolic activities (sugar utilization, acid production), resistance to antibiotics, and genetic fingerprints obtained by rep-PCR.

Strains in a collection should undergo minimal passages before distribution to reduce genetic variations within these strains. This can be achieved by establishment of a two-tiered system composed of a master and working (distribution) bank for each organism (Day and Stacey 2008).

Future Tasks of Culture Collections

Environmental samples from habitats that harbor undiscovered microorganisms and are disappearing due to climate change or forest clearance should be a primary focus of culture collections to preserve its microbial diversity. The filamentous fungi *Penicillium clavariaeformis* producing an orange pigment penicillipsin occurred on fruits and seeds of *Diospyros* trees in Indonesia and Taiwan (Hsieh and Ju 2002; Oxford and Raistrick 1940), and strains of these species are lost for isolation, when the trees disappear (Colwell 2002). To prevent loss of microbial strains from disappearing habitats, preservation methods of ecosystems and natural communities have to be developed (Prakash et al. 2013).

Data generated from genomic and proteomic studies are useful to identify microbial species in communities and help to determine their function

in an ecosystem. These data sets can be applied to develop cultivation methods for ecologically important microorganisms which are not-yet cultivable (Prakash et al. 2012). Information based on DNA sequences is increasingly used in ecological research and in investigating microbial communities. Storage of extracted DNA for use of DNA barcoding technology should be deposited in a repository for further taxonomic and biotechnological studies (Vernooy et al. 2010). Handling of biological sequence data derived from “omics” (genomics, transcriptomics, proteomics, metabolomics) including storage and accessibility should be standardized. Culture collections developing to biological resource centers (BRC) meet the high standard of quality management and accreditation processes and are able to participate in networking initiatives to strengthen the collaboration between collections and their users (Janssens et al. 2010; Stackebrandt 2010).

Sequencing of complete bacterial genomes leads to the discovery and characterization of new gene families (Wu et al. 2009). The ongoing characterization of microbes will lead to new strains, microbial metabolites, and novel protein-coding genes suitable for use in many industrial and health applications.

Summary

Culture collections play a key role in preservation, taxonomic characterization, and supply of diverse microbial strains with associated informative documents. Collection organizations such as WFCC (World Federation for Culture Collections) and ECCO (European Culture Collections' Organisation) and activities of CABRI (Common Access to Biological Resources and Information) promote and support culture collections and their related services.

With recent advances in genomic analysis and molecular genetics, researchers are increasingly able to understand, harness, and engineer the vast biochemical potential of microorganisms. And thus, further activities are necessary to fully realize the huge scientific and economic potential of these rich and diverse nature resources.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [All-Species Living Tree Project](#)
- ▶ [Biological Treasure Metagenome](#)
- ▶ [Culturing](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)
- ▶ [Microbial Ecosystems, Protection of](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [Silva Databases](#)

References

- Arora DK, Saikia R, Dwivedi R, Smith D. Current status, strategy and future prospects of microbial resource collections. *Curr Sci*. 2005;89:488–95.
- Colwell RR. The future of microbial diversity research. In: James TS, Reysenbach A-L, editors. *Biodiversity of microbial life: foundations of earth's biosphere*, vol. 3. New York: Wiley-Liss; 2002. p. 521–34.
- Day JD, Stacey GN. Biobanking. *Mol Biotechnol*. 2008;40:202–13.
- Emerson D, Wilson W. Giving microbial diversity a home. *Nat Rev Microbiol*. 2009;7:758.
- Fierer N, Lennon JT. The generation and maintenance of diversity in microbial communities. *Am J Bot*. 2011;98:439–48.
- Floyd MM, Tang J, Kane M, Emerson D. Captured diversity in a culture collection: case study of the geographic and habitat distributions of environmental isolates held at the American type culture collection. *Appl Environ Microbiol*. 2005;71:2813–23.
- Goris J, Konstantinidis KT, Klappenbach JA, et al. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*. 2007;57:81–91.
- Hsieh HM, Ju YM. *Penicillium pseudocordyceps*, the holomorph of *Pseudocordyceps seminicola*, and notes on *Penicillium clavariaeformis*. *Mycol*. 2002;94:539–44.
- Janssens D, Arahall DR, Bizet C, Garay E. The role of public biological resource centers in providing a basic infrastructure for microbial research. *Res Microbiol*. 2010;161:422–9.
- Komagata K. Microbial diversity and the role of culture collections. 1999. <http://old.iupac.org/symposia/proceedings/phuket97/komagata.pdf>
- Kuo A, Garrity GM. Exploiting microbial diversity. In: James TS, Reysenbach A-L, editors. *Biodiversity of microbial life: foundations of earth's biosphere*, vol. 3. New York: Wiley-Liss; 2002. p. 477–520.
- Moore ERB, Mihaylova SA, Vandamme P, et al. Microbial systematics and taxonomy: relevance for a microbial commons. *Res Microbiol*. 2010;161:430–8.
- Oxford AE, Raistrick H. Studies in the biochemistry of micro-organisms: penicillins, the colouring matter of *Penicillium clavariaeformis* Solms-Laubach. *Biochem J*. 1940;34:790–803.
- Prakash O, Shouche Y, Jangid K, Kostka JE. Microbial cultivation and the role of microbial resource centers in the omics era. *Appl Microbiol Biotechnol*. 2012. doi:10.1007/s00253-012-4533-y.
- Prakash O, Nimonkar Y, Shouche YS. Practice and prospects of microbial preservation. *FEMS Microbiol Lett*. 2013;339:1–9.
- Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*. 2009;106:19126–31.
- Roesch LFW, Fulthorpe RR, Riva A, et al. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. 2007;1:283–90.
- Sievers M, Dasen G, Wermelinger T, et al. Culture collections and the biotechnology deal. *Chimia*. 2010;64:782–3.
- Smith D. Culture collections over the world. *Int Microbiol*. 2003;6:95–100.
- Stackebrandt E. Diversification and focusing: strategies of microbial culture collections. *Trends Microbiol*. 2010;18:283–7.
- Uruburu F. History and services of culture collections. *Int Microbiol*. 2003;6:101–3.
- Vernooy R, Haribabu E, Muller MR, et al. Barcoding life to conserve biological diversity: beyond the taxonomic imperative. *PLoS Biol*. 2010;8:e1000417.
- Wayne LG, Brenner DJ, Colwell RR, et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol*. 1987;37:463–4.
- Wieser A, Schneider L, Jung J, Schubert S. MALDI-TOF MS in microbiological diagnostics-identification of microorganisms and beyond (mini review). *Appl Microbiol Biotechnol*. 2012;93:965–74.
- Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*. 2009;462:1056–60.

Culturing

Sarah Highlander
Genomic Medicine, J. Craig Venter Institute,
La Jolla, CA, USA

Definitions

Microbiome: The microbes (bacteria, archaea, fungi, protists, and viruses) that inhabit a specific environment or host, such as all the microbes that live in and on the human body.

Mesophile: An organism that grows and thrives within a moderate temperature range, usually between 20° and 45 °C.

Fastidious Organism: An organism that has very specific and usually complex growth requirements.

Microbial Commensalism: Interaction between two species where one benefits but does not harm or affect the other.

Microbial Mutualism: Interaction between two species where both organisms benefit.

Introduction

Only about 1 % of all prokaryotic species in the biosphere are thought to be cultivatable (Handelsman 2004). A few thousand taxa are associated with the human microbiome. The taxa on the skin are mostly cultivatable (*ca.* 90 %) (Gao et al. 2007), about 50 % of the oral species are cultivatable (Dewhirst et al. 2010), and about 50 % of the taxa of the gut microbiome may be cultivatable (Goodman et al. 2011). The history of cultivation of microbes can be traced to the Egyptians who used yeasts and lactic acid bacteria in the production of breads, wines, and beers. Robert Koch first cultivated bacteria on solid medium in 1881. This permitted single colony isolation and production of pure cultures. While this seemed a logical approach, we are now learning that some organisms cannot grow in pure culture and these pure cultures are often not representative of the environment from which they were isolated.

Environmental Requirements for Growth

Bacteria can be separated into groups based on their growth requirements, which include specific nutrients, optimal temperature (usually mesophiles for members of the human microbiome) and pressure ranges, and sensitivity or tolerance to salinity or pH. A very important feature is the optimal oxygen concentration required for growth. There are five categories:

(1) obligate aerobes (require *ca.* 20 % O₂); (2) microaerophiles, which grow well at reduced O₂ concentration; (3) facultative anaerobes that can grow aerobically or can respire anaerobically or grow fermentatively; (4) aerotolerant anaerobes, which are not killed by O₂ but that cannot respire aerobically and only grow optimally under anoxic conditions; and (5) obligate anaerobes that are usually killed in the presence of O₂ and grow only in the absence of oxygen. Strict anaerobes must be manipulated in an anoxic chamber. Aerotolerant anaerobes can be briefly handled on the laboratory bench for standard bacteriological techniques, but they must be incubated anaerobically. Anaerobic organisms predominate in the human oral, gastrointestinal, and vaginal tracts. Examples of obligate genera in the human body are *Clostridium* and *Bacteroides*; aerotolerant members include *Propionibacterium* and *Lactobacillus*; *Escherichia* is a facultative anaerobe. Some organisms require high concentrations of CO₂ (5 %) in addition to O₂. These are capnophiles; examples are some streptococci, *Neisseria* spp., and *Haemophilus* spp. that are residents of the respiratory tract.

Nutritional Requirements: Media Types

All bacteria require carbon, nitrogen, and sulfur for metabolism. It is generally believed that the human body is colonized only by heterotrophic bacteria that obtain energy from oxidation of organic carbon substrates, although the presence of autotrophic cyanobacteria has been reported in some oral and fecal samples based on 16S rDNA sequencing. In complex bacteriological culture media, carbon, nitrogen, and sulfur for heterotrophic growth are usually provided by a peptone (digested protein) such as casamino acids, Lab-Lemco (Oxoid), tryptone, or soytone. Vitamins, amino acids, and carbohydrates are provided by the addition of yeast extract. Sodium chloride is sometimes added as an osmotic stabilizer. Since many bacteria have special requirements for vitamins and other trace minerals, media are often supplemented

with hydrolysates such as beef extract or yeast extract. To create a solid medium, agar, a polysaccharide product of seaweed, is added to the broth formulation prior to autoclaving.

To support the growth of fastidious organisms, fresh defibrinated blood (usually sheep's blood) is often added to cooled agar media prior to pouring into Petri dishes. *Haemophilus* and *Neisseria* require blood that has been lysed prior to use (chocolate agar). Strict anaerobes require agar media that can be pre-reduced (i.e., incubated in an anoxic environment to remove all O₂ prior to use), and some species require the addition of hemin and vitamin K. A good general agar for this use is Anaerobic Reducible Blood Agar (Remel, Lenexa, KS), which contains cysteine HCl, palladium chloride, and dithiothreitol to maintain a low redox potential of the agar. It also contains hemin and vitamin K and can be purchased with colistin and nalidixic acid as a selective medium for isolation of gram-positive organisms or with kanamycin, vancomycin, and neomycin as a selective medium for gram-negative organisms, particularly the *Bacteroides*.

Numerous preformulated specialty powdered and premade bacteriological media and plates are available for selection, identification, and cultivation of a wide variety of human bacterial species, with a focus on pathogens.

Methods to Enhance Growth of Uncultivable Organisms

The concept of isolation of the pure cultures has been challenged by groups that have shown that cocultivation of organisms can sometimes lead to the successful isolation of previously uncultivable organisms. New technologies have been applied to isolate and capture cells and then incubate them in an environment that simulates (or is) the natural one.

The groups of Slava Epstein and Kim Lewis have made key contributions to methods and discoveries leading to the cultivation of such

isolates. In 2002 they reported on the use of diffusion chambers to grow microbial colonies from an intertidal sandy flat in an aquarium containing seawater as the growth medium (Kaeberlein et al. 2002). They estimated that up to 40 % of the cells inoculated into the chamber could be cultivated, but attempts to grow these microcolonies in pure culture were very inefficient. One isolate, which grew poorly on the agar plates, grew well in coculture with any of three other isolates obtained from the chambers. They expanded this to create a high-throughput Ichip diffusion array that contains 192 chambers per array (Nichols et al. 2010). A clever application of the technology was the creation of an upper palate dental appliance that carried a 72-chamber Ichip diffusion array (Sizova et al. 2012). The appliance was worn by a subject for 48 h then recovered and placed in an anaerobic chamber. Bacterial cells from the chambers were plated on a “basic anaerobic medium” that was low in sugar concentration to prevent selection for fast-growing species. This method contributed 39 isolates, several of which represented taxa that had not been previously cultivated. The take-home lessons that these authors stressed were that “domestication” of uncultivated organisms from the human microbiome is more likely if the bacteria are first grown in vivo and that cell growth should be allowed to occur “unimpeded by neighbors,” for example, by growth in diffusion chambers, by dilution to extinction (Rappe et al. 2002), or by growth encapsulated in microdroplets (Zengler et al. 2002). They also stressed the requirements of strict anaerobic conditions (for oral samples) and the utilization of media low in readily utilizable carbohydrates (Sizova et al. 2012).

The commensalism and mutualism of some bacterial species have been exploited to stimulate growth of previously uncultivated organisms. Examples of commensalism in dental plaque are well established, such as the catabolism of sugars by streptococci to lactic acid, which is fermented by the veillonellae, which cannot utilize sugars. Vartoukian et al. were able to cultivate Cluster

A *Synergistetes*, which had not been previously accomplished, by growing human plaque samples in a complex cooked meat medium (Vartoukian et al. 2010). Using fluorescent in situ hybridization (FISH) directed against the *Synergistetes* 16S rRNA, they followed the presence of an isolate, *Synergistetes* SGP1, and observed that the *Synergistetes* cells formed aggregates with other bacteria. Ultimately, they showed that growth of SGP1 was stimulated by cross-streaks of *Staphylococcus aureus*, *Fusobacterium nucleatum*, *Parvimonas micra*, and *Treponema forsythia*, which were members of the cell aggregates. The mechanism of the effect has not yet been ascertained. Siderophore sharing, or “stealing,” is a theme common in bacterial pathogenesis and is the one that the Epstein and Lewis group observed as a mechanism that permitted coculture of some strains of bacteria isolated from sand biofilms (D’Onofrio et al. 2010). They observed that samples plated in high density yielded much higher numbers of colonies than expected compared to plates with diluted biofilm samples and hypothesized that adjacent pairs of species might have growth dependencies. One strain, *Micrococcus luteus* KLE1011, was shown to secrete 5 distinct but related siderophores, any one of which was able to induce growth of the uncultivated strain *Maribacter polysiphoniae* KLE1104. The *M. luteus* strain was then used as “bait” to capture additional uncultivated bacteria from the samples (D’Onofrio et al. 2010). It would be surprising if this phenomenon was not observed between members of the human microbiome.

Summary

The cultivation of prokaryotes continues to follow mostly traditional methods, although some groups are beginning to recognize that the cultivation of the uncultivable requires a better appreciation of the *in vivo*

environment of the species that is sought. The application of diffusion chambers and microdroplet technologies to human microbiome samples should accelerate cultivation of some species, and metabolic predictions from whole genome shotgun sequencing may, in future, permit rationale cultivation of new species of bacteria.

References

- D’Onofrio A, Crawford JM, Stewart EJ, Witt K, Gavriš E, Epstein S, et al. Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chem Biol.* 2010;17:254–64.
- Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, et al. The human oral microbiome. *J Bacteriol.* 2010;192:5002–17.
- Gao Z, Tseng CH, Pei Z, Blaser MJ. Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci U S A.* 2007;104:2927–32.
- Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, Dantas G, et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci U S A.* 2011;108:6252–7.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev.* 2004;68:669–85.
- Kaerberlein T, Lewis K, Epstein SS. Isolating “uncultivable” microorganisms in pure culture in a simulated natural environment. *Science.* 2002;296:1127–9.
- Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, et al. Use of ichip for high-throughput in situ cultivation of “uncultivable” microbial species. *Appl Environ Microbiol.* 2010;76:2445–50.
- Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature.* 2002;418:630–3.
- Sizova MV, Hohmann T, Hazen A, Paster BJ, Halem SR, Murphy CM, et al. New approaches for isolation of previously uncultivated oral bacteria. *Appl Environ Microbiol.* 2012;78:194–203.
- Vartoukian SR, Palmer RM, Wade WG. Cultivation of a *Synergistetes* strain representing a previously uncultivated lineage. *Environ Microbiol.* 2010;12:916–28.
- Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM, et al. Cultivating the uncultured. *Proc Natl Acad Sci U S A.* 2002;99:15681–6.

Customizable Web Server for Fast Metagenomic Sequence Analysis

Sitao Wu¹, Zhengwei Zhu¹, Limin Fu¹,
Beifang Niu¹ and Weizhong Li²

¹Center for Research in Biological Systems
(CRBS), University of California, San Diego,
La Jolla, CA, USA

²J. Craig Venter Institute, La Jolla, CA, USA

Synonyms

A customizable Web server for fast metagenomic sequence analysis

Definition

WebMGA is a Web server through which researchers can upload metagenomic sequence data and run various tools to analyze the data. The tools in WebMGA can also be accessed through the Web services using client-side scripts.

Introduction

Metagenomics is an approach that studies the environmental microorganism populations predominantly using the next-generation sequencing technologies developed during the last decade. Today, scientists have already studied the microbes under many different environments such as water, soil, air, human body sites, and many others.

Metagenomic data analysis from raw sequencing reads to biological discoveries is a very complicated process and includes many computational procedures such as sequence quality control, filtering, mapping, assembly, gene prediction, normalization, function and pathway analyses, visualization, and statistical studies. In the last several years, many computational tools have been developed to address the problems in metagenomic data analysis. For example,

Metagene (Noguchi et al. 2006) and FragGeneScan (Rho et al. 2010) predict ORFs from fragmented sequences. Meta-RNA (Huang et al. 2009) scans rRNA from short sequences. Mothur (Schloss et al. 2009), QIIME (Caporaso et al. 2010), and CD-HIT-OTU (Li et al. 2012) are software packages for estimating microbial diversities based on 16S rRNA tags. RAMMCAP (Li 2009) is an integrated annotation pipeline that provides gene prediction, clustering, function annotations, and several other functions.

These complicated processes plus limited availability of computational resources tend to overwhelm bench biologists from attempting to analyze their own metagenomic data. So, integrated bioinformatics systems specific to metagenomic data analysis, especially easy-to-use Web portals, are of great importance for researchers in various communities to fully utilize metagenomic approach.

WebMGA was developed as a fast, easy, and flexible solution for metagenomic data analysis. It is freely available at <http://weizhongli-lab.org/metagenomic-analysis> to all users.

Metagenomic Analysis Tools Provided in WebMGA

More than 20 different analysis tools specially designed for metagenomic data analysis are provided by WebMGA Web portal. Most of these tools are very fast. Also, they are implemented to be executed in parallel on a computer cluster.

Given the raw metagenomic sequencing reads, the following analyses can be completed:

- A quality control (QC) script filters and trims raw sequencing reads and yields high-quality reads.
- Software SolexaQA (Cox et al. 2010) can also be used as a QC tool.
- Program CD-HIT-454 (Niu et al. 2010) identifies and removes artificial duplicate.

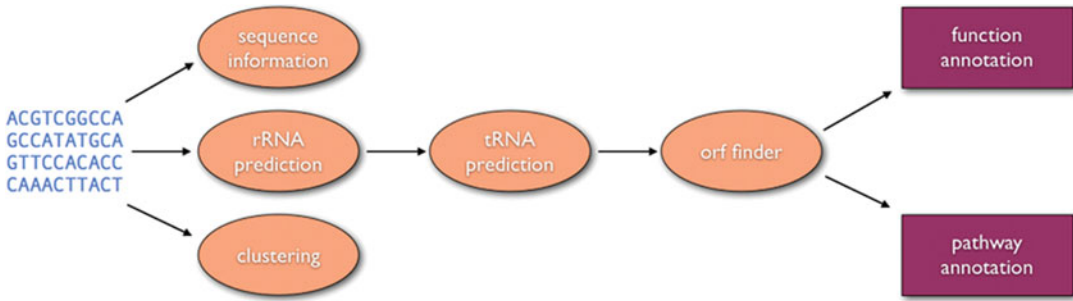
If the input sequences are filtered reads or DNA sequences, WebMGA provides the following analyses:

- tRNAscan (Lowe and Eddy 1997) finds tRNAs from the input sequences.



Customizable Web Server for Fast Metagenomic Sequence Analysis, Fig. 1 The web server page for DNA clustering

- Meta-RNA (Huang et al. 2009) identifies rRNAs from fragmented sequences using a hidden Markov model-based algorithm.
- A BLAST-based program identifies rRNAs by comparing the query against several rRNA reference databases.
- For metagenomic data from human subjects, WebMGA offers a tool that identifies human DNAs and RNAs and removes them from input metagenomic sequences. A fast mapping program FR-HIT (Niu et al. 2011) is used to align the input sequences against human reference sequences.
- CD-HIT-EST, an ultrafast sequence-clustering program, clusters the DNAs into groups or removes redundant sequences.



Customizable Web Server for Fast Metagenomic Sequence Analysis, Fig. 2 A simple workflow using tools in WebMGA

- WebMGA has a taxonomy-binning tool that maps the reads to reference genomes using FR-HIT and then assigns taxonomy annotations.
- ORF_finder (Li 2009) calls ORFs from input sequences by six-reading-frame translation.
- Metagene (Noguchi et al. 2006) identifies ORFs from fragmented sequences.
- FragGeneScan (Rho et al. 2010) identifies ORFs and also tries to correct frameshift errors.

Users can input protein or peptide sequences to run the following analyses:

- CD-HIT (Li et al. 2001, 2002; Li and Godzik 2006; Huang et al. 2010) clusters the input sequences into protein clusters or removes redundant sequences.
- A multistep clustering pipeline groups protein sequences into protein families.
- WebMGA uses HMMER3 program (Eddy 2009) to compare input peptides against Pfam and Tigrfam databases and assign the domain or protein families.
- WebMGA uses RPS-BLAST to compare NCBI's COG, KOG, and PRK databases and provide function annotations.
- WebMGA provides Gene Ontology (GO) annotations.
- WebMGA searches KEGG database and provides pathway annotations.

16S rRNA tags can also be analyzed through WebMGA:

- RDP Classifier (Wang et al. 2007) analyzes rRNA tags and assigns taxonomy annotations.

- CD-HIT-OTU (Li et al. 2012) is a pipeline that filters and processes the raw rRNA tags and clusters them into operational taxonomic units (OTUs). CD-HIT-OTU is available at <http://weizhongli-lab.org/cd-hit-otu>.

Each of the above tools has a Web interface where users can run them individually. Users with programming skills can even compose a script to run a customized multistep analysis workflow through WebMGA's Web services. As illustrated in Fig. 1, a user can upload a DNA dataset to run several analysis processes in parallel. The user can use HMM-based or BLAST-based method to find rRNAs and to produce a FASTA file with rRNA masked. The latter result file is then processed by an ORF calling program, and the ORFs are used for function and pathway annotation. This workflow is illustrated in Fig. 2.

Summary

WebMGA provides researchers the tools for rapid metagenomic sequence analysis through Web server and Web services. The tools and functions in WebMGA cover a large scope of metagenomic data analysis such as raw sequence quality control, human DNA filtering, OTU estimation, taxonomy binning, sequence clustering, and function and pathway annotation. By directly accessing the Web services with client-side scripts, users can customize and run their own workflows. The tools and data in WebMGA are



constantly being updated, and new tools for fast metagenomic data analysis will be continuously added.

Cross-References

- ▶ [Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences](#)
- ▶ [FR-HIT Overview](#)

References

- Caporaso JG, Kuczynski J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6.
- Cox MP, Peterson DA, et al. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinforma*. 2010;11:485.
- Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 2009;23(1):205–11.
- Huang Y, Gilna P, et al. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*. 2009;25(10):1338–40.
- Huang Y, Niu B, et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26(5):680–2.
- Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinforma*. 2009;10:359.
- Li WZ, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658–9.
- Li WZ, Jaroszewski L, et al. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;17(3):282–3.
- Li WZ, Jaroszewski L, et al. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 2002;18(1):77–82.
- Li W, Fu L, et al. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform*. 2012;13(6):656–68.
- Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25(5):955–64.
- Niu B, Fu L, et al. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinforma*. 2010;11:187.
- Niu B, Zhu Z, et al. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*. 2011;27(12):1704–5.
- Noguchi H, Park J, et al. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*. 2006;34(19):5623–30.
- Rho M, Tang H, et al. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010;38(20):e191.
- Schloss PD, Westcott SL, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537–41.
- Wang Q, Garrity GM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 2007;73(16):5261–7.

D

DACTAL

Tandy Warnow
Institute for Genomic Biology, University of
Illinois, IL, USA

Synonyms

Phylogeny = phylogenetic tree = tree; Multiple
sequence alignment = MSA

Definition

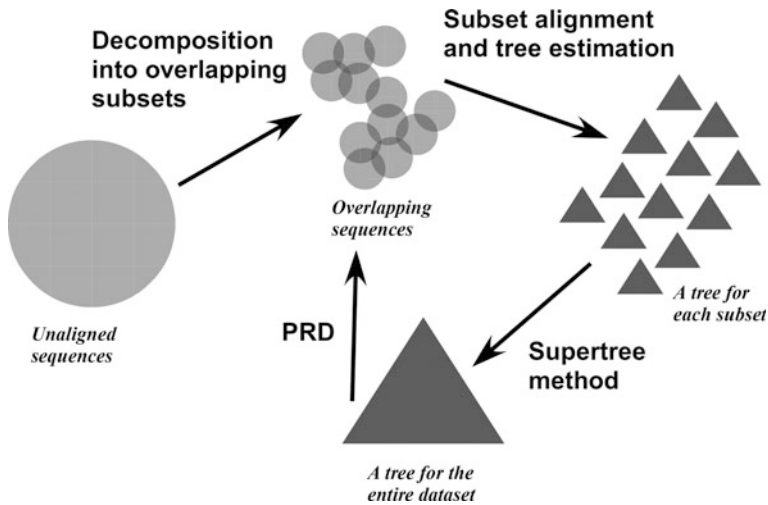
DACTAL = “Divide-and-conquer trees (almost
without alignments.”

Introduction

DACTAL (divide-and-conquer trees (almost
without alignments) is a method for estimating
very large phylogenetic trees which utilizes an
iterative divide-and-conquer technique to
“boost” the accuracy and speed of an existing
phylogeny estimation method. DACTAL con-
structs trees without needing to compute or use
a multiple sequence alignment on the full dataset.
This contribution describes the method and dem-
onstrates its performance on biological and sim-
ulated datasets.

Phylogeny estimation is a basic step in many
bioinformatics analyses, and there are many

methods for estimating phylogenies (Felsenstein
2003). Most frequently, phylogeny estimation for
a given set of taxa is performed in a sequence of
steps: (1) a gene is selected, (2) sequences for that
gene in the taxa are obtained, (3) a multiple
sequence alignment of the molecular data
(DNA, RNA, or amino acid) is estimated, and
(4) a tree is estimated on that resultant alignment.
Although many preferred methods for phylogeny
estimation are based on hard optimization prob-
lems (e.g., maximum likelihood), small datasets
are not that hard to analyze, and effective heuris-
tics are able to analyze small datasets quite well.
Maximum likelihood analysis of datasets with
many thousands of sequences is much more dif-
ficult, although some methods (e.g., RAxML
(Stamatakis 2006) and FastTree-2 (Price
et al. 2010) but see also Felsenstein 2003;
Warnow 2013) are highly effective even on
these datasets. The estimation of large multiple
sequence alignments (MSA) is itself very chal-
lenging (Kemena and Notredame 2009; Liu
et al. 2010; Blair and Murphy 2011): most current
MSA methods are unable to analyze very large
datasets (10,000 sequences and more) due to
computational issues, while those that can ana-
lyze datasets of this size (e.g., Clustal-Quicktree
and MAFFT-PartTree) produce alignments that
result in insufficiently accurate trees. Of the var-
ious methods available for large-scale multiple
sequence alignment, to date only SATe (Liu
et al. 2009, 2012) has been shown to be effective
at producing alignments that result in highly
accurate trees. However, SATe is limited to



DACTAL, Fig. 1 DACTAL algorithmic design. DACTAL can begin with an initial tree (*bottom triangle*), or through a technique that divides the unaligned sequence dataset into overlapping subsets. Each subsequent DACTAL iteration uses a novel decomposition strategy called “PRD” (padded recursive decomposition) to divide

the dataset into small, overlapping subsets, estimates trees on each subset, and merges the small trees into a tree on the entire dataset (figures included from a previous publication (Nelesen et al. 2012), with permission from the publisher).

datasets of about 50,000 sequences. Thus, the estimation of a large phylogenetic tree is a very challenging problem, and one of the biggest issues is the estimation of the multiple sequence alignment for the dataset.

Methods

DACTAL (Nelesen et al. 2012) is a method for estimating a very large phylogeny without needing to estimate a multiple sequence alignment on the entire dataset. The basic approach is a combination of divide-and-conquer plus iteration (see Fig. 1).

The input is a set of unaligned but homologous sequences, and each iteration produces a tree (but no alignment) on the full dataset. With the exception of the first iteration, each iteration begins with the tree from the previous iteration. In the first iteration, the method begins by dividing the dataset into overlapping subsets, each with at most some user-specified number of sequences; the default for this is 200. This division into subsets can be accomplished through the use of a technique that uses BLAST to form small sets

of sequences around each sequence with some overlap between the sequence subsets. Alternatively, the division can be performed by computing an alignment and tree on the dataset (using some fast and approximate methods) and then using the tree to produce a recursive decomposition of the sequence dataset. In either case, the decomposition that is produced produces subsets that overlap at least one other subset by some specified minimum amount (default 50) and that are themselves small (by default each subset has at most 200 sequences).

Once the decomposition is performed, trees are estimated on each subset, using some favored method; the default is a maximum likelihood analysis (default RAxML) on a good multiple sequence alignment, with the default being MAFFT (Katoh et al. 2005). These subsets are small (by default, they have at most 200 sequences in them), and as the experimental results show, this is sufficient even for datasets with about 28,000 sequences.

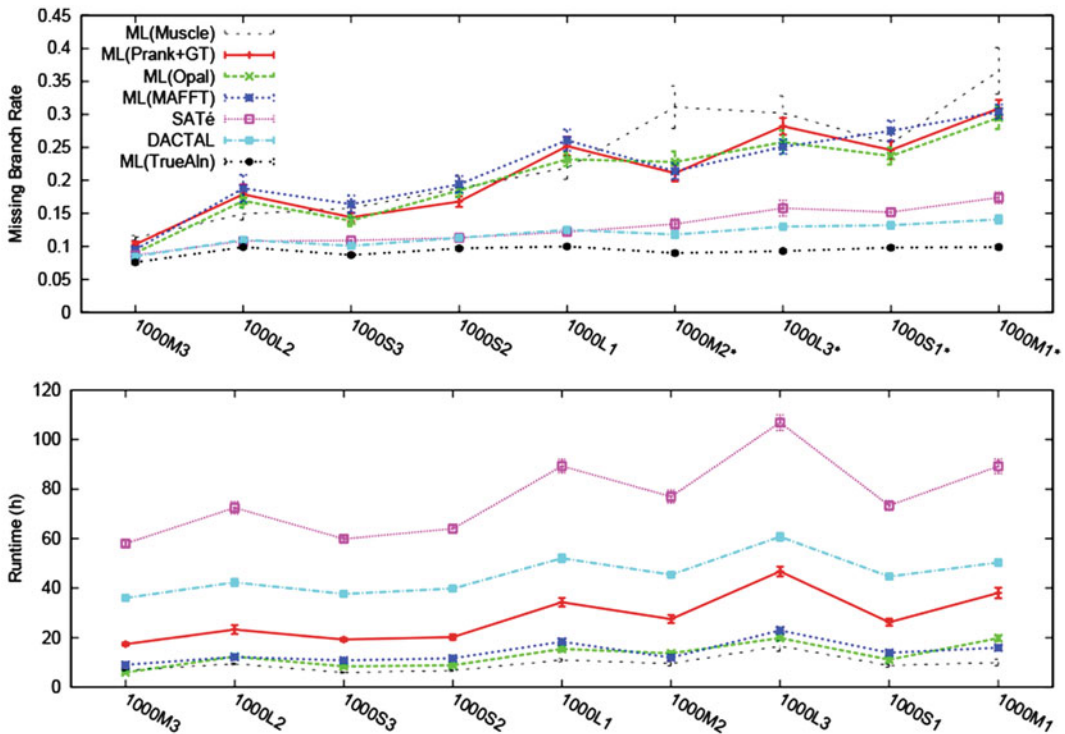
After the trees are computed, they can be merged together into a tree on the full set of taxa using a supertree method; the default is SuperFine+MRP (Swenson et al. 2012),

a supertree method that has excellent accuracy and which “boosts” the accuracy of MRP (another supertree method; see Bininda-Emonds 2004). Subsequent iterations begin with the tree estimated during the previous iteration and then decompose the dataset into overlapping subsets, compute trees on subsets, and merge the trees into a tree on the full dataset. The number of iterations is a parameter that is set by the user. Thus, DACTAL is a method that can be modified to enable different techniques for estimating trees on subsets and for combining subset tree into a full set of trees, and the target subset size and overlap between subsets are parameters that can be set by the user. The default settings were selected for accuracy and speed and provide good results, as the results section demonstrates.

Results

The performance of DACTAL was evaluated in comparison to maximum likelihood trees computed on SATE-I (Liu et al. 2009) and other alignment methods on simulated datasets with 1,000 sequences and on biological datasets with 6,000–28,000 sequences (Nelesen et al. 2012). The results of these experiments are shown in the figures below and demonstrate that DACTAL had accuracy comparable to that of SATE-I and could analyze larger datasets than SATE-I. These experiments also show that DACTAL was substantially more accurate than two-phase methods (i.e., methods that align sequences and then estimate trees on these alignments).

Figure 2 compares running time and tree accuracy on the 20 replicate datasets for 15 model

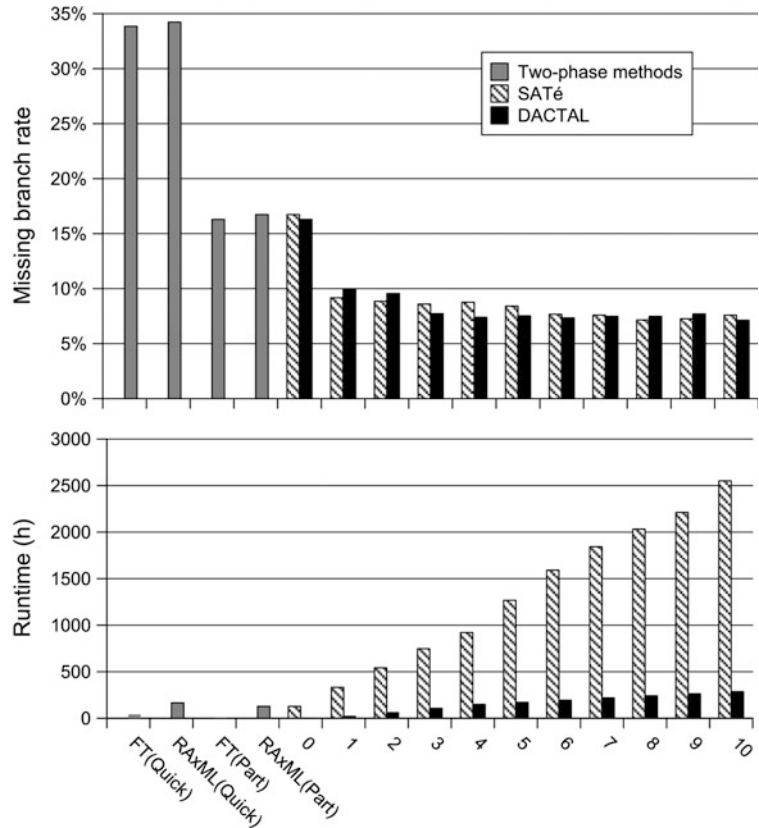


DACTAL, Fig. 2 Comparisons of ten iterations of DACTAL to SATE and RAxML trees estimated on different alignments on “moderate-to-difficult” simulated 1,000-taxon datasets. We show missing branch rates (top) and runtimes in hours (bottom); $n = 20$ for each model condition, and standard error bars are shown. DACTAL and SATE runtimes include the time to compute

RAxML(MAFFT) starting trees. Asterisks (*) denote model conditions for which DACTAL’s missing branch rate is a statistically significant improvement over the next best method, according to Benjamini-Hochberg-corrected pairwise t-tests ($n = 40$; $\alpha = 0.05$) (figures included from a previous publication (Nelesen et al. 2012), with permission from the publisher).

DACTAL,

Fig. 3 Comparisons of DACTAL and SATe iterations with two-phase methods on the 16S.T dataset with 7,350 sequences. The starting trees were RAxML on the MAFFT-PartTree alignment (RAxML(Part)) for SATe and FastTree-2 on the MAFFT-PartTree alignment (FT(Part)) for DACTAL. We show missing branch rates (*top*) and cumulative runtimes in hours (*bottom*); $n = 1$ for each reported value. Iteration 0 is used to compute the starting tree for DACTAL and SATe (figures included from a previous publication (Nelesen et al. 2012), with permission from the publisher).



conditions with 1,000 taxa, originally used to evaluate SATe-I (Liu et al. 2009). These model conditions vary in terms of rates of evolution, indel lengths (short, medium, or long), and relative rates of substitutions and indels (insertions and deletions).

The error in tree estimation is computed using the missing branch rate, which is the fraction of the nontrivial bipartitions in the true (model) tree that are missing in the estimated tree. In this experiment, DACTAL is run for ten iterations, while SATe-I runs for 24 h after it computes the RAxML(MAFFT) starting tree. The running time comparison shows that DACTAL is much faster than SATe-I on every model condition. The comparison with respect to accuracy shows that DACTAL has approximately the same accuracy as SATe-I and that both DACTAL and SATe-I are much more accurate than the two-phase methods on the difficult 1,000-taxon model conditions. Finally, this figure also shows that

DACTAL is faster than SATe, although it is slower than the two-phase methods.

Figure 3 shows performance on a single biological dataset, 16S.T, from the Comparative Ribosomal Webpage (CRW) (Cannone et al. 2002). This dataset has 7,350 sequences and a high rate of evolution and so represents a challenging phylogenetic dataset. The reference tree for this dataset is based on a curated structural multiple sequence alignment (Cannone et al. 2002). This figure gives four different two-phase methods (maximum likelihood computed using FastTree-2 or RAxML on either Clustal-Quicktree or MAFFT-PartTree alignments), but also shows trees obtained for each of ten iterations produced by SATe-I and DACTAL. Note how SATe-I and DACTAL both improve with each iteration, with the initial iterations producing the biggest reductions in tree error, and that they track each other iteration by iteration. However, note that each DACTAL

iteration is much faster than each SATe-I iteration, so that ten iterations of DACTAL finish in about 1/8 the time of ten iterations of SATe-I.

Discussion

DACTAL is a method for estimating trees from unaligned sequences. While it does not require the estimation of an alignment on the full dataset, it is not entirely alignment-free, since it estimates alignments on subsets. However, these subsets are small, containing only 200 sequences, which reduces the computational and analytical challenges to running DACTAL. These experiments show that DACTAL can produce highly accurate phylogenetic estimates on very large datasets, improving on the accuracy of both two-phase methods (that first align the sequences and then estimate the tree) and SATe-I.

Alignment-free methods (i.e., that do not use any multiple sequence alignment technique at all to compute trees) have also been designed; these are surveyed in Vinga and Almeida 2003 and Chan and Ragan 2013. Alignment-free methods typically compute trees in three steps: first, each sequence is characterized by some distribution (e.g., its k -mer distribution for some appropriately chosen k), then distances between sequences are computed, and finally a tree is computed on the distance matrix. Unlike DACTAL, these truly alignment-free methods have not, to our knowledge, been shown to produce trees of comparable accuracy to methods that estimate multiple sequence alignments and then compute maximum likelihood trees on these alignments. Furthermore, the alignment-free methods surveyed in these papers do not have any theoretical guarantees under Markov models of evolution. An interesting contrast to these methods is the recent result given in Daskalakis and Roch 2010. This technique is guaranteed statistically consistent under the TKF1 model (Thorne et al. 1991) and so represents an important advance in theory. However, this method has not yet been implemented, so it remains a theoretical contribution rather than a usable technique.

Unlike these truly alignment-free methods, DACTAL is not completely alignment-free, since it does compute alignments on subsets. However, the results shown here suggest that highly accurate trees are indeed possible without requiring a multiple sequence alignment on the full dataset.

Future Work

The phylogenetics research community has been developing improved methods for alignment and phylogeny estimation. These methods may well lead to improved estimations of larger trees and could reduce the need for methods like DACTAL. However, DACTAL may continue to be a useful tool for improving scalability of these methods to very large datasets, containing many tens of thousands of sequences, since these improved techniques could be used to estimate trees on subsets of taxa. This may be particularly relevant to the recent effort to develop methods that co-estimate sequence alignments and trees under complex models of sequence evolution (see Bouchard-Cote and Jordan 2013 for a recent paper and other methods surveyed in Warnow 2013). Most of these methods are computationally very intensive and limited to at most 200 sequences (and even then are computationally intensive), and DACTAL could potentially be used to improve their scalability to larger datasets. More generally, the phylogenetics research community has been developing sophisticated techniques for highly accurate estimations of alignments and trees, but these statistically based methods often use techniques (such as MCMC) that are computationally intensive and do not run on large datasets. DACTAL provides a basic tool for improving the scalability of these techniques and so complements these efforts. Thus, large-scale phylogeny estimation may well improve through a combination of efforts – some aimed at improving the estimation of trees and alignments on small datasets, using statistically informed but computationally intensive methods, and other efforts aimed at using divide-and-conquer to combine smaller trees into larger trees.

Summary

DACTAL is a method for estimating large trees from unaligned sequences that uses an iterative divide-and-conquer technique. By design, DACTAL does not produce a multiple sequence alignment, yet analyses on many datasets (both real and simulated) show that DACTAL produces trees with great accuracy, improving on existing two-phase methods that first align and then estimate the tree from the sequences. These analyses also show that DACTAL matches the accuracy of SATE while being much faster. With the increased interest in estimating very large trees, this type of approach could enable highly accurate and very large-scale phylogenetic estimation.

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [DACTAL](#)
- ▶ [MRL and SuperFine+MRL](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [SATE-Enabled Phylogenetic Placement](#)
- ▶ [Use of Viral Metagenomes from Yellowstone Hot Springs to study phylogenetic relationships and evolution](#)

References

- Bininda-Emonds O, editor. *Phylogenetic supertrees: combining information to reveal the tree of life*. Dordrecht: Kluwer Academic Publishers; 2004.
- Blair C, Murphy RW. Recent trends in molecular phylogenetics. *J Hered*. 2011;102(1):130–8.
- Bouchard-Cote A, Jordan MI. Evolutionary inference via the Poisson Indel Process. *Proc National Academy of Sciences*. 2013;110(4):1160–1166.
- Cannone J, Subramanian S, Schnare M, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron and other RNAs. *BMC Bioinforma*. 2002;3:2. doi:10.1186/1471-2105-3-2
- Chan CX, Ragan MA. Next-generation phylogenomics. *Biology Direct* 2013;8:3.
- Daskalakis C, Roch S. Alignment-free phylogenetic reconstruction. In: Berger B, editor. *Proc. RECOMB 2010*, volume 6044 of *Lecture Notes in Computer Science*. Berlin: Springer; p. 123–37.
- Felsenstein J. *Inferring phylogenies*. Sunderland: Sinauer Associates; 2003.
- Katoh K, Kuma K, Miyata T, et al. Improvement in the accuracy of multiple sequence alignment MAFFT. *Genome Inform*. 2005;16:22–33.
- Kemena C, Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*. 2009;25:2455–65.
- Liu K, Raghavan S, Nelesen S, et al. Rapid and accurate large-scale co-estimation of sequence alignments and phylogenetic trees. *Science*. 2009;324:1561–4.
- Liu K, Linder CR, Warnow T. Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLOS Currents Tree of Life*. 2010. doi: 10.1371/currents.RRN1198.
- Liu K, Warnow T, Holder M, et al. SATE-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol*. 2012;61:90–106.
- Nelesen S, Liu K, Wang LS, et al. DACTAL: divide-and-conquer trees (almost) without alignments. *Bioinformatics*. 2012;28:i274–82.
- Price MN, Dehal PS, Arkin AP. FastTree-2 – approximately maximum likelihood trees for large alignments. *PLoS ONE*. 2010;5:e9490. doi:10.1371/journal.pone.0009490.
- Stamatakis A. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
- Swenson M, Suri R, Linder CR, et al. SuperFine: fast and accurate supertree estimation. *Syst Biol*. 2012;61:214–27.
- Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*. 1991;33:114–24.
- Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics*. 2003;19(4):513–523.
- Warnow T. Large-scale multiple sequence alignment and phylogeny estimation. In: Chauve C, El-Mabrouk N & Tannier E, editors. *Models and Algorithms for Genome Evolution*. Springer: London; 2013. p. 85–146.

Diversity and Distribution of Marine Microbial Eukaryotes

Connie Lovejoy
Department of Biology, Laval University,
Québec, QC, Canada

Marine microbial eukaryotes (MME) are morphologically, phylogenetically, and functionally diverse. The term protist is often used but not

a valid taxonomic classification (Adl et al. 2005, 2007), and evolutionary relationships among MME at the highest taxonomic ranks remain controversial.

Functionally, they span all trophic levels, with phototrophic, heterotrophic, and mixotrophic taxa, where mixotrophic taxa are able to use both photosynthesis and heterotrophy as sources for carbon and energy. Taxonomically MME are found among all major branches of the eukaryotic tree of life, with the exception, at least up until now, of the Excavata (Adl et al. 2012). This taxonomic and functional diversity is also manifest in morphological diversity over several microscopic scales, depending on the group.

Unassignable taxa: Taxa that do not fit within descriptions based on standard taxonomy. In the case of gene sequences, this occurs when the sequence in question shows little homology with other sequences. The degree of difference is indicative of taxonomic level differences, e.g., domain, phyla, and order, to the level of genus. For the first 18S rRNA gene surveys using cloning and Sanger sequencing, several phyla-level novel sequences were discovered. The majority of these belong to uncultivated taxa and so there are no organisms available to infer functional roles.

The use of molecular tools to identify MME began some 10 years after the first reports of bacterial and archaeal diversity in the sea, in part because the existing taxonomic record from microscopy seemed complete. This changed when the first surveys were carried out with two publications in 2001 highlighting surprising diversity of MME in the deep sea and identification of unassignable taxa (Diez et al. 2001); (Moon-van der Staay et al. 2001). These studies triggered major programs that were mostly aimed at the so-called picoplankton operationally defined as cells passing through a 3 μm filter and collected on either 0.8 or 0.2 μm filters (Vaulot et al. 2008). Most studies

operationally defined picoplankton by size-fractionated filtration, and many of the novel groups originally thought to be picoplanktonic can in fact include cells $>3 \mu\text{m}$. Fragile cells are broken during filtration with cellular contents passing through the filter, and free DNA preserved in seawater can also be collected on the 0.2 μm filter. Such size fractionation however is useful since it enriched the proportion of smaller cells. More recently, surveys of picoplankton have been carried out using cells that were collected following flow cytometry (FCM) cell sorting. Using this technique, other novel photosynthetic taxa have been discovered (Not et al. 2008). Placing novel taxa into known phylogenies requires aligning sequences with known groups and determining their placement within phylogenetic trees. Using nearly full-length 18S rRNA gene sequences, most of the early novel MME have been found to be within some higher level taxonomic grouping, and as new environments are surveyed, the distribution and diversity of uncultivated groups can be documented.

In principle, it is possible to identify whole communities of MME using metagenomic approaches. In practice, high-throughput multiplexing of different samples with primers specific for hypervariable regions of the 18S rRNA gene can be used to identify MME in natural environments (Amaral-Zettler et al. 2009); (Comeau et al. 2011). These short sequences or reads are taxonomically assigned based on reference-curated 18S rRNA gene phylogenies. However, as with Bacteria and Archaea, the utility of identifying MME and their functional genes is directly related to the accuracy and completeness of reference databases. Many species that are found in the marine environment are the same as those reported using microscopy, and a major challenge is linking microscopy records with sequence data. An additional complication is the desirability to exploit historic data sets and taxonomic treatises where only morphological descriptions are provided with no voucher specimens or cultures in existence, and sequences cannot be matched to describe morphological species.

Classification of MME

Historically MME were divided into plants (algae) and animals (protozoa), with algal classification following the botanical nomenclature code and protozoa following the zoological code. The term algae is now considered non-taxonomic functional grouping of oxygenic C-photo-autotrophic (oxygen evolving photosynthesis) organisms that are neither bryophytes nor vascular plants. Cyanobacteria are sometimes referred to as algae, are the important phototrophs in much of the world ocean, but are not eukaryotes and not treated here. Below is a brief survey of major MME categorized by trophic roles.

Phototrophs

While diatoms, coccolithophores, and dinoflagellates are the most frequently mentioned phototrophs in the ocean, there are many other taxa that can contribute substantially to oceanic primary productivity. The eukaryotic algae include heterogeneous and evolutionarily different groups. The origin and development of the first eukaryotic algae is explained through an endosymbiotic event where a heterotrophic eukaryote acquired or enslaved an ancestral cyanobacterium (cf. Reys-Prieto et al. 2010). After genetic reduction and transformation, this event gave rise to primary plastids (chloroplasts) present in Glaucophyta, Rhodophyta (red algae), and Chlorophyta (green algae), and the three lineages are classified as Plantae (Raven et al. 2005) or more broadly as Archaeplastida (Adl et al. 2010) with the higher plants. Chlorophyta are ancestral to algal Streptophyta and are predominantly green with chlorophyll *b* as a secondary pigment; Prasinophyta and Mamiellaceae are the most common marine pelagic Chlorophyta.

Other algae are polyphyletic (lack an identifiable common ancestor), and for most, their chloroplasts originated as a secondary endosymbiotic event where a single-celled pre-rhodophyte alga was acquired or enslaved by another heterotrophic protist. Over time this lineage gave rise to other major algal phyla (Reyes-Prieto et al. 2010); (Keeling 2009). Chlorophyll *c* is a secondary pigment common to most of these other algae;

in the ocean, these include Diatomea, Pelagophyceae, Eustigmatales, Dictyochophyceae, Chrysophyceae, Raphidophyceae, and other stramenopiles. Among these are Parmales, which have siliceous walls and have been reported from electron microscopy (Kosman et al. 1993) and are closely related to or within the flagellated bolidophytes (Ichinomiya et al. 2011). About half of the living species of Dinophyceae (within the alveolates) are photosynthetic (Taylor et al. 2008).

Cryptophyta and Haptophyta, also with chlorophyll *c*, are now thought to have arisen through separate endosymbiotic events with different protists (Baurain et al. 2010), and their phylogenetic positions are uncertain.

Haptophyta in the sea include flagellated taxa mostly in the Prymnesiales, and Phaeocystales, and coccolith bearing taxa referred to as coccolithophores, which include Isochrysidales (e.g., *Emiliania*) and Coccolithales. Many coccolithophores also have flagellated stages, and some species lack chloroplasts (Adl et al. 2012).

There are two other algal phyla that arose from endosymbiotic events where single-celled green algae gave rise to the chlorophyll *b* containing chloroplasts in the photosynthetic Euglenophyta and the Chlorarachniophyta, both of which are common in marine waters. Several dinoflagellates from diverse lineages have lost their original secondary endosymbiotically acquired chloroplast and have acquired new chloroplasts directly from either green algae, cryptophytes, or even diatoms in what are termed tertiary endosymbiotic events (Keeling 2009).

Mixotrophs

The majority of the phototrophic groups named above are also more than likely mixotrophic, at least on some level. Mixotrophy can range from the ability to use dissolved organic matter (osmotrophy) to the capacity to engulf (phagotrophy) bacteria or other microbial eukaryotes, including species that are larger than themselves. These mixotrophs can compete with heterotrophs for nutrients, carbon, and even energy when taking up preformed organic material. It is important to reiterate that these

are not plants. For example, mixotrophic Chrysophyceae are particularly common in Arctic Sea ice and Arctic marine waters (Lovejoy et al. 2002); (Rozanska et al. 2008). Other stramenopile mixotrophs include members of the Dictyochophyceae, Pelagophyceae, and Raphidophyceae; Euglenozoa, Cryptophyceae, Haptophyceae, and photosynthetic dinoflagellates are also mixotrophic.

Heterotrophs

Heterotrophic MME are also phylogenetically, morphologically, and functionally diverse. Among these are several microscopically recognizable heterotrophic species with uncertain taxonomic affinities that are also frequently recovered in environmental gene surveys. These include *Telonema*, *Katablepharidae*, and *Centrohelida*. Other groups are well placed in phylogenies, for example, choanoflagellates which are included along with animals in the Opisthokonta. Among novel uncultivated MME are the biliphytes originally termed picobiliphytes (Not et al. 2007), which although still uncultivated have been sequenced using single-cell genome amplification technology on cells collected via FCM. The genome from these cells has confirmed that they branch apart from other eukaryotes as a sister group to Cryptophytes and that they are most likely strict heterotrophs (Yoon et al. 2011), and have been formally described as a new Phyla; the Picozoa by Seenivasan et al. (2013).

Among historically important protist heterotrophs in the sea are representatives from the large supergroup Rhizaria, which includes Cercozoa, Polycystinea, Acantharia, and Foraminifera. Cercozoa from microscopy studies that have also been retrieved using environmental gene surveys include *Cryothecomonas* (Thaler and Lovejoy 2012).

The first environmental 18S rRNA gene surveys revealed a number of distinct lineages of marine stramenopiles (MASTs) (Massana et al. 2004), which for the most part seem to be bacterivores, although cultured representatives are lacking (Massana et al. 2009). Also among groups that are mostly known from

environmental surveys are several clades of marine alveolates that are mostly related to parasitic taxa, including *Amoebophyra*, which infect dinoflagellates (Grosillier et al. 2006) and others most closely related to zooplankton parasites (Skovgaard et al. 2005).

References

- Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor M. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol.* 2005;52:399–451.
- Adl SM, Leander BS, Simpson AGB, Archibald JM, Anderson OR, Bass D, Bowser SS, Brugerolle G, Farmer MA, Karpov S, Kolisko M, Lane CE, Lodge DJ, Mann DG, Meisterfeld R, Mendoza L, Moestrup O, Mozley-Standridge SE, Smirnov AV, Spiegel F. Diversity, nomenclature, and taxonomy of protists. *Syst Biol.* 2007;56:684–9.
- Adl SM, Simpson AGB, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, le Gall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Standridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch CL, Smirnov A, Spiegel FW. The revised classification of eukaryotes. *J Eukaryot Microbiol.* 2012;59:429–93.
- Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *Plos ONE.* 2009;4.
- Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol.* 2010;27:1698–709.
- Comeau AM, Li WKW, Tremblay J-É, Carmack EC, Lovejoy C. Arctic ocean microbial community structure before and after the 2007 record sea ice minimum. *Plos ONE.* 2011;6:e27492.
- Diez B, Pedros-Alio C, Massana R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol.* 2001;67:2932–41.
- Grosillier A, Massana R, Valentin K, Vaulot D, Guillou L. Genetic diversity and habitats of two enigmatic

- marine alveolate lineages. *Aquat Microb Ecol*. 2006;42:277–91.
- Ichinomiya M, Yoshikawa S, Kamiya M, Ohki K, Takaichi S, and Kuwata A. Isolation and characterization of Parmales (Heterokonta/Heterokontophyta/stramenopiles) from the Oyashio region, western North Pacific. *J Phycol*. 2011;47:144–151.
- Keeling PJ. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol*. 2009;56:1–8.
- Kosman CA, Thomsen HA, Ostergaard JB. Parmales (Chrysophyceae) from Mexican, Californian, Baltic, Arctic and Antarctic waters with the description of a new subspecies and several new forms. *Phycologia*. 1993;32:116–28.
- Lovejoy C, Legendre L, Martineau MJ, Bacle J, von Quillfeldt CH. Distribution of phytoplankton and other protists in the North Water. *Deep-Sea Res Part II Top Stud Oceanogr*. 2002;49:5027–47.
- Massana R, Castresana J, Balague V, Guillou L, Romari K, Groisillier A, Valentin K, Pedros-Alio C. Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol*. 2004;70:3528–34.
- Massana R, Unrein F, Rodriguez-Martinez R, Forn I, Lefort T, Pinhassi J, Not F. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J*. 2009;3:588–96.
- Moon-van der Staay SY, De Wachter R, Vault D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*. 2001;409:607–10.
- Not F, Valentin K, Romari K, Lovejoy C, Massana R, Tobe K, Vault D, Medlin LK. Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science*. 2007;315:253–5.
- Not F, Latasa M, Scharek R, Viprey M, Karleskind P, Balague V, Ontoria-Oviedo I, Cumino A, Goetze E, Vault D, Massana R. Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep-Sea Res Part I Oceanogr Res Pap*. 2008;55:1456–73.
- Raven JA, Finkel ZV, Irwin AJ. Picophytoplankton: bottom-up and top-down controls on ecology and evolution. *Vie Et Milieu-Life Environ*. 2005;55:209–15.
- Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, Nakayama T, Ishida K, Bhattacharya D. Differential gene retention in plastids of common recent origin. *Mol Biol Evol*. 2010;27:1530–7.
- Rozanska M, Poulin M, Gosselin M. Protist entrapment in newly formed sea ice in the Coastal Arctic Ocean. *J Mar Syst*. 2008;74:887–901.
- Seenivasan R, Sausen N, Medlin LK, Melkonian M. *Picomonas judraskeda* gen. et sp. nov.: The first identified member of the Picozoa Phylum nov., a widespread group of picoeukaryotes, formerly known as 'picobiliphytes'. *PLoS ONE* 2013;8(3):e59565. doi:10.1371/journal.pone.0059565.
- Skovgaard A, Massana R, Balague V, Saiz E. Phylogenetic position of the copepod-infesting parasite *Syndinium turbo* (Dinoflagellata, Syndinea). *Protist*. 2005;156:413–23.
- Taylor FJR, Hoppenrath M, Saldarriaga JF. Dinoflagellate diversity and distribution. *Biodivers Conserv*. 2008;17:407–18.
- Thaler M, Lovejoy C. Distribution and diversity of a protist predator *Cryothecomonas* (Cercozoa) in Arctic marine waters. *J Eukaryot Microbiol*. 2012;59:291–9.
- Vault D, Eikrem W, Viprey M, Moreau H. The diversity of small eukaryotic phytoplankton ($\leq 3 \mu\text{m}$) in marine ecosystems. *FEMS Microbiol Rev*. 2008;32:795–820.
- Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*. 2011;332:714–7.

DNA Methylation Analysis by Pyrosequencing

Florence Busato and Jörg Tost

Laboratory for Epigenetics and Environment,
Centre National de Génotypage, CEA- Institut de
Génomique, Evry, France

Synonyms

Quantitative sequencing by synthesis

Definition

Pyrosequencing is a sequencing-by-synthesis method that quantitatively monitors the real-time incorporation of nucleotides using an enzymatic conversion of pyrophosphate into a proportional light signal. Quantitative measures are crucial for applications such as the analysis of DNA methylation patterns, which are intensively studied in various developmental and pathological contexts as well as for bacterial identification and determination of allelic imbalance.

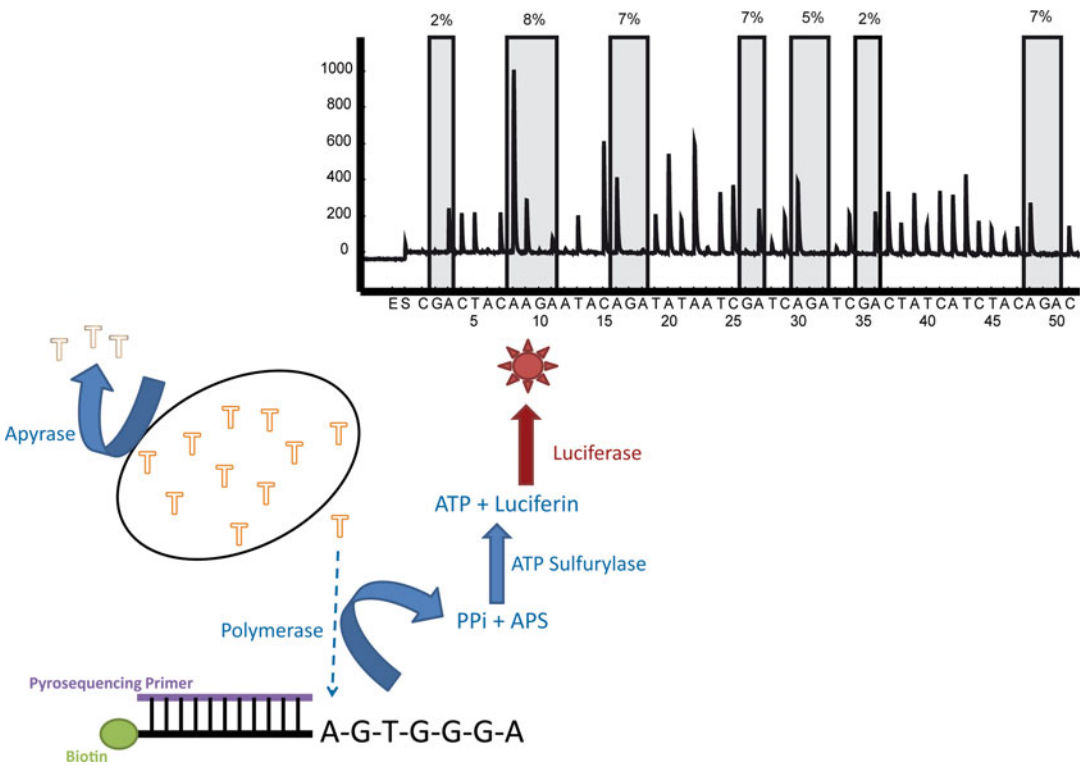
Introduction

While Sanger sequencing has been the “gold standard” for the identification of sequence variants for a long time, pyrosequencing with its improved ability for quantification, decreased limit of detection and accelerated workflow leading to a shorter time to results, has become a valuable alternative notably for many clinical and diagnostic applications. Pyrosequencing is a sequencing-by-synthesis method, where nucleotides are incorporated complementary to a template strand leading to the release of pyrophosphate (PPi) that will – after several enzymatic reactions – produce a light signal proportional to the amount of incorporated nucleotide (Fig. 1).

The experimental procedure of the pyrosequencing assay is simple and relatively robust and results are highly reproducible. Therefore, pyrosequencing has become a widely used analysis platform for various biological and/or diagnostic applications such as routine (multiplex) genotyping of single-nucleotide polymorphisms (SNPs), methylation analysis of bisulfite-treated samples, bacterial typing, mutation detection, and allele quantification (Marsh 2007).

DNA Methylation

DNA methylation is a post-replication modification that occurs in mammals almost exclusively



DNA Methylation Analysis by Pyrosequencing, Fig. 1 Nucleotides added into the pyrosequencing reaction (here exemplified by a thymine) are incorporated by the DNA polymerase extending the pyrosequencing primer when they are complementary to the DNA template sequence. This incorporation releases PPi, which is

used together with APS by an ATP sulfurylase to produce ATP. ATP will be subsequently used by luciferase to oxidate luciferin to oxyluciferin generating a proportional light signal. Unincorporated nucleotides are degraded by apyrase to avoid unspecific background signals. The reactions are detailed in the text

on the 5' position of the pyrimidine ring of cytosines in the context of a dinucleotide CpG (Tost 2009). CpGs represent less than 1 % of all bases and are mostly methylated in the mammalian genome. CpGs are relatively rare because they are easily transformed into TpGs by deamination, and as thymine is a naturally occurring building block of the DNA, these mutations are less well recognized and repaired by the cellular machinery. This elevated mutation rate has led to CpG depletion during evolution.

However, relatively CpG-rich clusters, called CpG islands, are found in the promoter and first exon of approximately two-thirds of all genes. Mostly unmethylated, these CpG islands are distributed throughout the human genome and maintain the chromatin in an open configuration to allow transcription. The absence of DNA methylation is not directly correlated to the transcriptional activity of the corresponding gene, but rather the transcriptional potential. However, a certain number of promoter CpG islands are methylated in a tissue-specific manner, and this DNA methylation helps to maintain transcriptional silence in non-expressed or noncoding regions of the genome. Methylated regions also maintain transcriptional inactivation, as exemplified by the methylation and repression of repetitive and transposable elements. Furthermore, some genes, called imprinted genes, express only one allele depending on their parent of origin (maternal or paternal allele), and the non-expressed allele is associated with a repressed imprinting control region, which is in many cases marked by DNA methylation. Inactivation of one X chromosome in female mammals is another example in which DNA methylation plays an important role in gene dosage and regulation.

During aging and in the context of pathologies, particularly cancer, regions normally unmethylated become methylated, and this hypermethylation can induce or is at least associated with aberrant gene expression patterns. For example, methylation of the DNA repair genes *MLH1* and *MGMT* can lead to their inactivation, resulting respectively in microsatellite instability and increased mutation frequency. Methylation

can also promote spontaneous deamination, enhance DNA binding of carcinogens, or increase ultraviolet absorption by DNA and, as a result, increase the rate of mutations, DNA adduct formation, and subsequent gene inactivation. As DNA methylation has been shown to be influenced by diet and environmental exposure, it has been postulated that DNA methylation might constitute a measurable molecular memory of our lifestyle and environment (Cortessis et al. 2012).

Methylation of cytosines in other sequence contexts (CpNpG, CpA, etc.) has been identified in cultured cells such as mouse embryonic stem cells. In plants, methylation on cytosines is more prevalent and more diverse compared to mammals, and their DNA is highly methylated. The methylcytosines are mainly located in CpG and CpNpG sequences, but they may also occur in other contexts. DNA methylation controls plant growth and development, with a particular involvement in the regulation of gene expression and DNA replication, similar to its function in mammalian cells.

Compared to mammals, bacteria have at least two methylated bases in addition to 5-methylcytosine: N6-methyladenine in the sequence context GpApTpC and GpApNpTpC and N4-methylcytosine (Casadesus and Low 2006). These methylated bases are involved in the protection of bacterial DNA, where they act as a defense mechanism against bacteriophage infection. They play also crucial roles in the control of DNA repair, replication, transposition, and – similar to eukaryotes – gene expression. Particularly, adenine methylation plays an important role in the regulation of gene expression in bacteria, with its absence allowing the binding of specific proteins to the bacterial DNA. Methylation patterns have also been correlated to the virulence of several pathogens.

However, due to their greater diversity, the presence of many “orphan” methyltransferases, i.e., enzymes not part of a restriction enzyme system that methylate bacterial genomes at specific sites and the only recent emergence of appropriate tools to study the DNA modifications, DNA methylation in bacteria has not been

a topic of intensive research. The advent of single molecule sequencing technologies such as the single molecule real-time sequencer from Pacific Biosciences performing sequencing with an immobilized polymerase at the bottom of zero-mode waveguide wells in zeptoliter volumes has revolutionized the possibilities for DNA methylation analysis in bacteria and allowed the direct readout of CpG and other methylation modifications in bacteria (Davis, et al. 2013).

Principles of the DNA Methylation Analysis

As DNA methylation is involved in many biological processes, it is of great importance to analyze DNA methylation patterns and their variability. As DNA methylation is not retained during PCR amplification, it is necessary to make use of procedures that are able to differentiate the epigenetic state. Methods for DNA methylation analysis are based on four main principles: (1) the use of methylation-sensitive restriction endonucleases, i.e., enzymes that are blocked by methylated cytosines in their recognition sequence are widely used for the analysis of methylation patterns in combination with their methylation-insensitive isoschizomers. Although methods based on methylation-sensitive restriction enzymes are simple and cost-effective as they do not require any special instrumentation, they are hampered by the limitation to specific restriction sites as only CpG sites found within these sequences can be analyzed. (2) The methylated fraction of a genome can be enriched by precipitation with a bead-immobilized antibody specific for 5-methylcytosine or (3) affinity purification of methylated DNA with MBD proteins, but these methods do not permit the analysis of DNA methylation patterns at single-nucleotide resolution. (4) The most widely used approach consists of the chemical modification of genomic DNA with sodium bisulfite. This chemical reaction induces the hydrolytic deamination of non-methylated cytosines to uracils, while methylated cytosines are resistant to conversion under the chosen reaction conditions. This method thus

translates the methylation signal into a sequence difference. After PCR amplification the methylation status at a given position is manifested in the ratio C (former methylated cytosine) to T (former non-methylated cytosine) and can be analyzed as a virtual C/T polymorphism spanning the entire allele frequency spectrum from 0 % to 100 % in the bisulfite-treated DNA. The latter principle is commonly used for DNA methylation analysis by pyrosequencing. It should be noted that the reduced complexity of the bisulfite-treated DNA (which essentially consists of a three-letter genome) creating homopolymeric and highly AT-rich sequences provides a challenge for the design of PCR amplification-based assays and induces frequently a preferential amplification of either unmethylated or methylated alleles. This bias has to be monitored and corrected for to ensure accurate quantification of DNA methylation levels of the analyzed CpGs.

Principle of the Pyrosequencing Reaction

Pyrosequencing is a polymerase-based quantitative real-time sequencing method used to analyze multiple sequence variations in a region of interest. In contrast to conventional Sanger sequencing that uses a mixture of the four fluorescently labeled chain-terminating ddNTPs and strand-elongating dNTPs, only one nucleotide is dispensed at a time by an inkjet-type cartridge in pyrosequencing reactions using either a user-defined sequence-specific dispensation order or a repetitive cyclic dispensation order of the four nucleotides for unknown sequences.

This iterative incorporation of unmodified nucleotides by the exonuclease-deficient Klenow fragment of DNA polymerase I will result in the release of inorganic pyrophosphate (PPi), while all unincorporated nucleotides will be degraded prior to addition of the next nucleotide by an apyrase. When the polymerase encounters a noncomplementary nucleotide, it pauses while nucleotide degradation takes place. The pyrophosphate is in the presence of adenosine phosphosulfate (APS) transformed by an ATP

sulfurylase into several products including ATP. The latter will be used in the subsequent step to oxidize luciferin to oxyluciferin by a luciferase resulting in the creation of a proportional amount of photons, which can be monitored by a CCD camera (Fig. 1). The four enzymes are present in a well-balanced mixture allowing the DNA polymerase to extend the newly synthesized DNA strand until it encounters a noncomplementary nucleotide while at the same time avoiding unspecific nucleotide incorporation and out-of-phase sequencing. A key step in the development of applications for pyrosequencing was the addition of a single-stranded DNA binding protein to the reaction mixture (now also included in the commercial kits), which led to a substantial increase in read length and overall greater accuracy through the reduction of the formation of secondary structures and mispriming (Dupont et al. 2004).

Samples of interest are amplified by PCR performed with one of the two amplification primers being biotinylated. This allows the isolation of a single-stranded sequencing template through the capture of the biotinylated amplification product on streptavidin-coated Sepharose beads. After washing steps, the use of a sodium hydroxide solution allows the denaturation of the double-stranded DNA and isolation of the biotinylated single strand used as template in the pyrosequencing procedure. A (pyro)sequencing primer is subsequently annealed to this template, and the sequence is synthesized one nucleotide at a time. The light signals are then generated by the enzymatic cascade by extending the 3' end of the nascent strand described above. It should be noted that the nucleotide dATP acts as a natural substrate for luciferase (although less efficient compared to ATP). Therefore the α -S-dATP analogue is used as nucleotide for primer extension as it is equally well incorporated by the polymerase.

Pyrosequencing can analyze almost any polymorphism in the amplified sequence. As the expected sequence is in most cases known a priori, the sequence to analyze is simply entered into the software creating automatically a dispensation order, and once the sequencing

reaches this polymorphism, both nucleotides of the variable position will be added successively and their proportional luminometric signal quantified by the software.

Since all the enzymatic reactions are quantitative, the intensity of the bioluminometric response is directly proportional to the amount of incorporated nucleotides: the incorporation of two identical consecutive nucleotides will have the double intensity (and therefore peak height in the resulting pyrogram) compared to the signal of a single-nucleotide incorporation. This quantitative nature of the results is the most important characteristic of the pyrosequencing technology because it allows performing quantitative applications such as DNA methylation analysis. Furthermore, as pyrosequencing proceeds at a rate of one dispensation per minute, results on the presence and abundance of variable nucleotides will be available between 10 and 60 min after launching a pyrosequencing reaction. The total time to results starting from the PCR amplification is commonly below 3–4 h and therefore much faster than conventional Sanger sequencing.

Inconveniences

However, there are some inconveniences associated with this technology, mainly concerning the analysis of variation in the close proximity of homopolymers, the size of the amplification product, and the sequencing read length. Pyrosequencing as well as the closely correlated 454 sequencing and semiconductor sequencing (Ion Torrent) suffer from the lack of precision in the analysis of homopolymers. The bioluminometric response is only linear ($R^2 > 0.99$) for the sequential addition of up to five identical nucleotides (C, G, T) or three α -S-dATPs. Sequence variation in close proximity to homopolymer reads might therefore not be easily resolved, and the quantitative accuracy might be limited. Due to the thermal instability of the enzymes, pyrosequencing has to be carried out at 28 °C which limits the size of the amplification product to 350 base pairs as the formation of secondary structures can complicate annealing

of the sequencing primer or increase background signals. The limitation in the read length (less than 100 dispensed nucleotides) is mainly due to dilution effects and increasing background due to frameshifts of subpopulations of sequenced molecules. This drawback can be partly overcome using the below described serial pyrosequencing approach. Lastly, the setup and optimization of robust pyrosequencing assays including the assay design but also the entry of an optimal dispensation order requires a certain degree of experience and expertise, and only few tools are available in the public domain for the assay design.

Serial Pyrosequencing

To overcome the restriction in read length, a solution was found in the “recycling” of the single-stranded template after the pyrosequencing run. As this template is not altered during the pyrosequencing reaction, it can be recovered after the run by the same template preparation protocol used after PCR amplification. Several pyrosequencing primers can therefore be used on the same DNA template to cover the entire amplified sequence with sufficient intensity and good quantitative resolution. This improvement enables the analysis of an entire region amplified in a single PCR. While the approach has initially been devised for DNA methylation analysis (Tost et al. 2006), it could also be used for the analysis of several sequence variations within the same amplification product.

Application: Genotyping and Mutation Detection

Pyrosequencing can be used to genotype single-nucleotide polymorphism (SNP) and detect mutations involved in various diseases (cancer, Alzheimer’s disease, heart diseases, diabetes) or in biological traits such as eye color or lactose intolerance.

Once the sequencing reaches the SNP (entered in the sequence to analyze in the software using the IUPAC single letter code), all possible

nucleotides will be added one after another. Each allele combination will result in a specific pyrosequencing pattern that can easily be read either by the software or by the user. Besides simple qualitative genotyping, pyrosequencing can be used for quantitative applications such as the level of mutation or the potential loss of one allele (loss of heterozygosity (LOH)). LOH can result in a neutral phenotype but can also be involved in cancer as exemplified by the LOH of *BRCA1* or *BRCA2* in breast cancer.

Due to its relatively short read length, pyrosequencing is best suited for the detection and quantification of mutation hotspots such as the codons 12 and 13 of *KRAS* (Ogino et al. 2005), a gene commonly mutated in many cancers including colorectal cancer, where it is the most commonly mutated gene with a prevalence of ~ 40 % of patients, lung, or pancreatic cancer. Similar applications concern the analysis of *BRAF* (V600E) or *JAK2* (V617F) mutations and polymorphisms such as C677T *MTHFR*. Compared to conventional Sanger sequencing, the limit of detection is significantly improved (i.e., 2–7 % for pyrosequencing compared to 10–20 % for conventional Sanger sequencing) which enables the user to call low-level mutations with greater confidence and resolve, e.g., ambiguous Sanger sequencing results. This property of pyrosequencing is also of special importance in situations where, for example, few tumor cells are present among normal cells and/or a subclone of the tumor carries the mutation of interest, which might expand upon a given therapy and induce drug resistance. Pyrosequencing has also been applied to more complex genetic analyses requiring accurate sequencing such as HLA (sub)typing (Ugolotti et al. 2011). A quantitative readout is also of interest for the genotyping of SNPs in polyploidy organisms such as plants where pyrosequencing has proven to be an effective tool.

Application: Transcript Quantification

Just as it can quantify the ratio of mutations in a heterogeneous mixture of DNA, pyrosequencing

can quantify any variation of sequence. In the case of cDNA, it is thereby possible to determine an imbalance in the transcript quantity of different alleles (Yang et al. 2013).

Application: Bacterial Typing

Similar to the analysis of genetic variations, technologies used for bacterial identification have shifted from Sanger sequencing to the more user-friendly pyrosequencing technology enabling a more extensive sampling of microbial diversity with reduced efforts. Pyrosequencing has been used for the identification of microbial species and detection of genetic mutations that confer resistance to antibiotics and antiviral drugs by sequencing well-characterized short hypervariable regions of bacterial genes such as 16S, 23S rRNA, or rnpB. Universal primers are located in the conserved regions amplifying the variable regions, which are subsequently pyrosequenced. The provided sequence (sometimes in addition to biochemical data) gives unambiguous and discriminatory information for microbial identification. It should be noted that due to the limited read length of the pyrosequencing technology, the careful design of the targets and location of the amplification primers are of utmost importance and depend on the biological question.

Pyrosequencing has been successfully used to identify pathogens, which were refractory to biochemical analyses in a hospital setting identifying 78 different genera representing 16 different specimen types. Further it was applied to the identification and subtyping of different strains of, for example, *Helicobacter pylori*, *Mycobacterium*, and *Streptococcus* (Petrosino et al. 2009). It has been used to differentiate between Gram-positive and Gram-negative bacteria using the 16S RNA demonstrating superior results to conventional Gram staining. Pyrosequencing has also been shown to have sufficient discrimination potential to identify highly similar strains of *Yersinia pestis* in a relatively short time and can also be used to identify antimicrobial resistance genes including mutations in the gyrase and other genes in quinolone-resistant *Salmonella* and

ciprofloxacin-resistant *Neisseria gonorrhoeae*. Similar assays have been developed for fungal and viral identification. It should be noted that similar to Sanger sequencing, pyrosequencing requires pure bacterial isolates and thus a culture step prior to the analysis, as mixtures of different bacteria will lead to sequence patterns that will be inconclusive or difficult to interpret. Genome-wide sequencing approaches using the pyrosequencing-based 454 technology for metagenomics which will circumvent this problem through clonal amplification of single DNA molecules (and thus single bacteria) are discussed elsewhere in this encyclopedia in the context of, e.g., the Human Microbiome Project.

Application: DNA Methylation Analysis

Due to the recent interest in epigenetics in general and DNA methylation analysis in particular, DNA methylation analysis by pyrosequencing is probably the prime application of the technology as it allows simultaneous analysis and quantification of the methylation status of several CpG positions in close proximity (Tost and Gut 2007).

This point is of particular interest as successive CpGs might display significantly different levels of methylation particularly in imprinted genes as well as at promoters devoid of a CpG island. Pyrosequencing has been demonstrated to be very reproducible if assays are performed in a quality-controlled and standardized fashion including sufficient amount of input DNA for methylation analysis (Dupont et al. 2004). Furthermore, the possibility to include controls for complete bisulfite conversion (i.e., the measurement at a cytosine outside of a CpG context) avoids a potential pitfall of DNA methylation analysis. Pyrosequencing has a limit of detection of ~ 5 % for the minor unmethylated or methylated allele, respectively, and the technical variability of the pyrosequencing reaction alone is very limited (~ 2 %). Variability increases to about 5 % if independent bisulfite conversion and PCR amplifications are performed (Dupont et al. 2004). Pyrosequencing is therefore much better suited and less complex than standard

Sanger sequencing for DNA methylation analysis. The recording of calibration curves using standards with a known degree of methylation during assay setup or during routine use also allows for correction for potential preferential amplification of methylated or unmethylated alleles, a phenomenon frequently encountered with bisulfite-treated DNA.

The quantitative accuracy can be applied to analyze global or gene-specific DNA methylation patterns of a sample. Pyrosequencing has been widely used to analyze the DNA methylation patterns of genes aberrantly silenced by promoter hypermethylation in cancer and other diseases. It has been used for the distinction between age-related and cancer-associated DNA methylation patterns or the analysis of the epigenetic field defect in cancer. A diagnostic test using pyrosequencing for the detection of aberrant DNA methylation patterns involved in the imprinting disorders Prader-Willi and Angelman syndromes was proposed.

Pyrosequencing can also be used for screening of differential DNA methylation between two sample groups by creating pools stratified for clinical parameters of interest, for example, cancerous versus matched peritumoral tissue (Dejeux et al. 2007). This method helps to concentrate research efforts and available biological material on genes displaying variable methylation patterns.

DNA methylation analysis can be performed in the tissue of interest itself or in biofluids that were in contact with the diseased tissue such as serum, plasma, sputum, or urine (How Kit et al. 2012). However, amounts of DNA that can be isolated are normally very small and thus require an additional step of genome-wide amplification prior to the quantitative DNA methylation detection. As DNA methylation marks are not retained during amplification, the amplification has to be performed on the bisulfite-treated DNA with its lower sequence complexity, decreased integrity due to the harsh conditions of the chemical treatment, and increased potential for secondary structures. The qMAMBA (quantitative methylation analysis of minute DNA amounts after whole bisulfite

amplification) assay has combined this amplification step with a pyrosequencing-based readout starting from as little as 100 pg of DNA (Paliwal et al. 2010). As a significant quantity of DNA is obtained after amplification, the DNA can be analyzed at multiple loci. It should however be noted that the quantitative accuracy of the genome-wide amplification on bisulfite-treated DNA is still controversial.

Nonetheless these approaches have been used for, e.g., forensic trace identification, whereby tissue-specific DNA methylation patterns analyzed by pyrosequencing after bisulfite amplification were used to identify the biofluid of origin (Madi et al. 2012). Another approach to analyze minute amounts of DNA methylation patterns combines the high sensitivity of methylation-specific PCR (MSP) with the specificity of the pyrosequencing-based readout. The replacement of the gel-based detection with the sequencing-based readout avoids some of the problems associated with potentially false-positive results induced by mispriming of the MSP primers (Shaw et al. 2006). Nonetheless, as molecules with a specific DNA methylation patterns are specifically enriched, the resulting pyrosequencing-based analysis is no longer quantitative and/or representative of the methylation patterns present in the analyzed sample.

Application: Global DNA Methylation Analysis

The LUMinometric Methylation Assay (LUMA) is based on a polymerase extension assay using the pyrosequencing platform after digestion by methylation-sensitive and nonsensitive restriction enzymes (Karimi et al. 2006). In this case, the pyrosequencer measures the luminometric signal produced by the nucleotide extension of the resulting number of digested sites. It is a quantitative and highly reproducible method and uses an internal control for DNA input. Besides, no modification of genomic DNA is required.

Furthermore, the global analysis of DNA methylation can be performed using the

pyrosequencing-based analysis of methylation patterns in repetitive elements such as *ALU* or *LINE1* elements (Yang et al. 2004). While *LINE1* elements do have a relatively conserved sequence allowing thus the design of a sequence-specific pyrosequencing assay for DNA methylation analysis, methylation of *ALU* elements is assessed by a cyclic dispensation. These assays have been widely used for the measurement of global DNA methylation changes in response to environmental stimuli (Cortessis et al. 2012).

Application: Allele-Specific DNA Methylation Analysis

Some genes display different methylation patterns on the two alleles either randomly or in a parent of origin-specific manner. Imprinting control regions regulating the expression of imprinted genes are commonly methylated on only one allele (inherited from the mother or the father) so that only one “parental allele” is expressed. Using a heterozygous SNP to differentiate the two alleles, the methylation status of each allele can be interrogated after enrichment of the methylated molecules using the above-described MSP with primers complementary to a specific DNA methylation pattern. The resulting amplification products are subsequently pyrosequenced, and the ratio of the two alleles after methylation enrichment is quantified by genotyping the two alleles of the SNP after methylation enrichment (Kristensen et al. 2013).

To analyze methylation on both alleles separately, it is possible to design two pyrosequencing primers, each specific of one allele using the two alleles of a heterozygous single-nucleotide polymorphism to differentiate the two alleles (Wong et al. 2006). The specificity of the allele-specific enrichment can be further improved by modifying the base complementary to the SNP with an LNA (locked nucleic acid). Locked nucleic acids are RNA monomers with a modified backbone. The sugar phosphate backbone has a 2'-O-4'-C methylene bridge. The bridge increases the monomer's thermal stability, reduces its flexibility, and increases the hybridization interactions

of the base with the template ensuring an amplification of only the exact complementary allele at the chosen temperature.

Summary

Pyrosequencing is a sequencing-by-synthesis, easy-to-use method that can precisely analyze genetic and epigenetic variation in an amplified sequence of up to 350 base pairs. Its applications are wide and various: genotyping, methylation analysis, transcript quantification, bacterial typing, etc. The broad range of applications combined with the above-described advantages has made pyrosequencing a widespread analysis method.

Cross-References

- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes](#)
- ▶ [Extraction Methods, Variability Encountered in](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)
- ▶ [NGS QC Toolkit: A Platform for Quality Control of Next-Generation Sequencing Data](#)

References

- Casadesus J, Low D. Epigenetic gene regulation in the bacterial world. *Microbiol Mol Biol Rev.* 2006;70:830–56.
- Cortessis VK, Thomas DC, Levine AJ, Breton CV, Mack TM, Siegmund KD, et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. *Hum Genet.* 2012;131:1565–89.
- Davis BM, Chao MC, Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr Opin Microbiol.* 2013;16:192–8.
- Dejeux E, Audard V, Cavard C, Gut IG, Terris B, Tost J. Rapid identification of promoter hypermethylation in hepatocellular carcinoma by pyrosequencing of etiologically homogeneous sample pools. *J Mol Diagn.* 2007;9:510–20.

- Dupont JM, Tost J, Jammes H, Gut IG. De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Anal Biochem.* 2004;333:119–27.
- How Kit A, Nielsen HM, Tost J. DNA methylation based biomarkers: practical considerations and applications. *Biochimie.* 2012;94:2314–37.
- Karimi M, Johansson S, Ekström TJ. Using LUMA: a Luminometric-based assay for global DNA-methylation. *Epigenetics.* 2006;1:45–8.
- Kristensen LS, Treppendahl MB, Asmar F, Girkov MS, Nielsen HM, Kjeldsen TE, et al. Investigation of MGMT and DAPK1 methylation patterns in diffuse large B-cell lymphoma using allelic MSP-pyrosequencing. *Sci Rep.* 2013;3.
- Madi T, Balamurugan K, Bombardi R, Duncan G, McCord B. The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. *Electrophoresis.* 2012;33:1736–45.
- Marsh S, editor. *Pyrosequencing protocols, methods in molecular biology vol 373.* Totowa: Humana Press; 2007.
- Ogino S, Kawasaki T, Brahmandam M, Yan L, Cantor M, Namgyal C, Mino-Kenudson M, Lauwers GY, Loda M, Fuchs CS. Sensitive sequencing method for KRAS mutation detection by pyrosequencing. *J Mol Diagn.* 2005;7:413–21.
- Paliwal A, Vaissière T, Hecceg Z. Quantitative detection of DNA methylation states in minute amounts of DNA from body fluids. *Methods.* 2010;52:242–47.
- Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic pyrosequencing and microbial identification. *Clin Chem.* 2009;55:856–66.
- Shaw RJ, Akufo-Tetteh EK, Risk JM, Field JK, Liloglou T. Methylation enrichment pyrosequencing: combining the specificity of MSP with validation by pyrosequencing. *Nucleic Acids Res.* 2006;34:e78.
- Tost J. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Mol Biotechnol.* 2009;44:71–81.
- Tost J, Gut IG. DNA methylation analysis by pyrosequencing. *Nat Protoc.* 2007;2:2265–75.
- Tost J, Elabdalaoui H, Gut IG. Serial pyrosequencing for quantitative DNA methylation analysis. *Biotechniques.* 2006;40:721–6.
- Ugolotti E, Vanni I, Raso A, Benzi F, Malnati M, Biassoni R. Human leukocyte antigen-B (-Bw6/-Bw4 I⁸⁰, T⁸⁰) and human leukocyte antigen-C (-C1/-C2) subgrouping using pyrosequence analysis. *Hum Immunol.* 2011;72:859–68.
- Wong H-L, Byun H-M, Kwan J, Campan M, Ingles S, Laird P, et al. Rapid and quantitative method of allele-specific DNA methylation analysis. *Biotechniques.* 2006;41:734–9.
- Yang AS, Estéicio MRH, Doshi K, Kondo Y, Tajara EH, Issa J-PJ. A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res.* 2004;32:e38.
- Yang B, Wagner J, Yao T, Damaschke N, Jarrard DF. Pyrosequencing for the rapid and efficient quantification of allele-specific expression. *Epigenetics.* 2013;8:1039–42.

E

Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities

Vedran Lucić¹, Masa Roller², Istvan Nagy³ and Kristian Vlahoviček²

¹Molecular Biology Department, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

²Bioinformatics Group, Molecular Biology Department, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

³Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary

Definition

Whole microbial communities exhibit patterns similar to those of single microbial species in terms of synonymous codon usage, regardless of their phyletic composition. Therefore, methods applicable on single microbial genomes to predict for functionally important and lifestyle-relevant genes based on translational optimization of synonymous codons can be applied to the study of the entire metagenomes. Using these predictions opens up a possibility to discover new and functionally unannotated genes relevant for the community metabolism and overall adaptation to a particular environment. This approach presents an integrated approach to the study of microbial community genomic information and

provides an *in silico* functional metagenomic platform to complement metaproteomic studies.

Introduction

Environmental diversity studies have bypassed the common problem where less than 1% of microbes are amenable to cultivation in laboratory conditions (Staley and Konopka 1985) by instead using high-throughput sequencing to extract genomic information directly from the environmental sample, without prior culturing. Various environments and geological sites have been sampled using new-generation sequencing, such as sea (Venter et al. 2004), soil (Tringe et al. 2005a), and various extreme habitats (e.g., acid drainage from a metal mine (Tyson et al. 2004), as well as gastrointestinal tracts of diverse organisms – including human (Gill et al. 2006) and mouse (Turnbaugh et al. 2006)). Most of the analysis of the sampled environments were focused in two main directions. The first one classifies the functions of identified genes (open reading frames) according to annotation available through orthology databases such as COG/KOG (Clusters of Orthologous Groups of genes) (Tatusov et al. 2003) or KEGG-KO (Kyoto Encyclopedia of Genes and Genomes – Orthology) (Kanehisa et al. 2006) and subsequently ranking the relative “importance” of a particular function according to its abundance in the environment. The second direction focuses on estimating the phyletic distribution of microbial species

represented in the environment, based on similarity searches against known microbial species' sequences (Huson et al. 2007).

For a thorough understanding of microbial communities at the systems level, it is necessary to capture the interplay of community constituents and organizational complexity in the community metabolism. Microbes in the same environment live within the same physical and chemical constraints, such as temperature, pH, or ion concentration, probably causing the GC content to be metagenome specific (Foerstner et al. 2005). Furthermore, communities of microbes have been shown to share tRNA pools to facilitate horizontal gene transfer (Tuller et al. 2011), which also implies a limited choice of preferred cognate codons within the shared tRNA pool. It has also been shown that fast growth rates introduce stronger bias in synonymous codon usage at the level of whole metagenomes (Vieira-Silva and Rocha 2010), much like the effect observed in single microbial species (Rocha 2004; Sharp et al. 2005).

Microbial communities living under the same environmental constraints, at the level of genes, can effectively be considered and studied as metagenomes, thereby using approaches and methodology valid for single microbial genome studies. One such approach is the functional characterization by translational optimization through synonymous codon usage bias.

The codon usage (CU) bias within a genome reflects the selection pressure for translational optimization of highly expressed genes – primarily the protein synthesis machinery such as ribosomal genes and elongation factors, but also genes with environmental adaptation functions (Supek et al. 2010). At the level of a single microbial genome, the effect of CU bias is routinely used to predict for functionally relevant and highly expressed genes (Sharp and Li 1987; Karlin and Mrazek 2000; Plotkin and Kudla 2011). The choice of preferred codons in a single genome is most closely correlated with abundance of the cognate tRNA molecules (Ikemura 1985; Kanaya et al. 2001; Tuller et al. 2010) and further influenced by the genome's GC content (Chen et al. 2004).

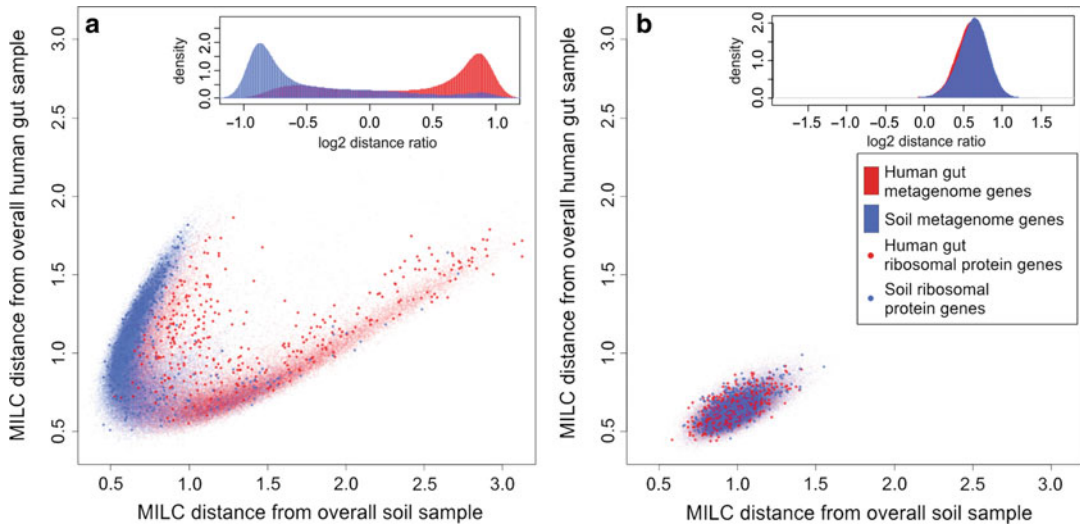
Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Table 1 Metagenomes used to demonstrate the concept of environmental shaping of codon usage

Metagenome	NCBI Project ID	Reference
Global Ocean Sampling Expedition Metagenome, the Sargasso Sea version 1	13694	(Venter et al. 2004)
Waseca County farm soil metagenome	13699	(Tringe et al. 2005b)
Whale fall metagenomes	13700	
5-way (CG) acid mine drainage biofilm metagenome	13696	(Tyson et al. 2004)
Human distal gut biome	16729	(Gill et al. 2006)
Lean mouse 1 gut metagenome	17391	(Turnbaugh et al. 2006)
Obese mouse 1 gut metagenome	17397	
US EBPR sludge metagenome	17657	(Martin et al. 2006)
OZ EBPR sludge metagenome	17659	

Eleven different microbial community sequencing samples (Table 1.) were used to demonstrate that microbes living in the same ecological niche, regardless of their phyletic diversity, share a common preference for codon usage. CU bias is present at the community level and is also different between distinct communities. CU bias also varies within the community, with distributions resembling that of single microbial species, i.e., the intercommunity CU bias can be observed. The effects of intercommunity CU bias and translational optimization concepts are utilized to identify genes with CU close to that of the meta-ribosomal sample. These genes have high predicted expression across the entire microbial community and define its “functional fingerprint.” This approach establishes a functional metagenomic platform that enables functional studies at the level of the entire microbial community samples.

Description

Microbes living in the same ecological niche share a bias in CU. When comparing the distance



Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Fig. 1 Codon usage is metagenome specific. Soil versus human gut metagenome codon usage (CU) frequencies. (a) The distance (MILC) of each gene's CU frequency to overall CU frequencies of two microbial communities. Genes (red in human gut ($N = 33,422$) and blue in Waseca soil ($N = 88,696$) metagenome) are predominantly closer

to their respective metagenome of origin therefore forming two distinct groups (the distribution of \log_2 ratio of the two distances for each gene is shown in the *inset*). If the amino acid composition of metagenomes is kept constant and the codons are randomly chosen, CU bias of each metagenome would be eliminated resulting in uniform distribution of CU distances and overlap of two samples, as shown in **b**

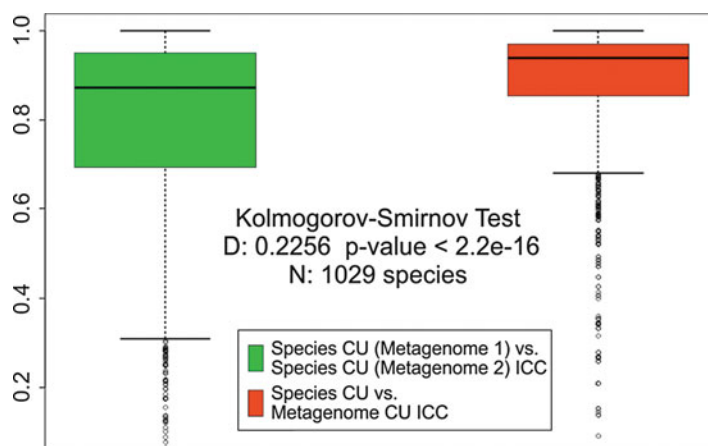
of each gene's CU in a metagenome from overall metagenome CU in the metagenome of origin with all other metagenomes, genes originating from one metagenome form a distinct cluster (as shown in Fig. 1a) and have CU predominantly closer to that of metagenome overall CU than genes from other metagenomes. If the amino acid sequence of each gene is kept constant but the codons randomly chosen (Fig. 1b), the genes' CU becomes equidistant to both metagenomes (i.e., occupy the same portion of the plot) regardless of their metagenome of origin.

The Variability of Single Species' Codon Usage Across Metagenomes

When comparing CU of species present in two distinct metagenomes, they can be compared in terms of CU distance with (i) their respective metagenome overall CU and (ii) CU of genes from the same species in a different metagenome. The resulting distance distributions, quantified with the intraclass correlation coefficient measure (ICC), show a statistically significant difference

in CU patterns of compared phylogenies – the within-species' CU pattern is more variable between metagenomes than in different species within the same metagenome (Fig. 2).

Comparison of CU variability of independently sequenced strains of microbes living in distinct niches is used to demonstrate that CU is a dynamic property that changes with different environmental constraints at the level of single bacterial species. Comparison between 12 strains of *Propionibacterium acnes* (Bruggemann et al. 2004; Hunyadkurti et al. 2011), commensal gram-positive bacteria that live in consistent environmental conditions, with 6 strains of cosmopolitan bacterium *Rhodopseudomonas palustris* (Larimer et al. 2004; Oda et al. 2008), shows that there is less variation in CU per orthologous group in *P. acnes* strains than in the *R. palustris* strains (Fig. 3). Despite the fact that the sampling includes more than twice as many strains from constrained environmental conditions (*P. acnes*) than variable conditions (*R. palustris*), the variability in CU is smaller in



Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Fig. 2 Codon usage variability between same species in different metagenomes is larger than within a metagenome. ORFs from each identified species (using MEGAN) were compared against their originating metagenome (orange, total comparisons $N = 2,058$) and against same-species ORFs in a different metagenome

(green, total comparisons $N = 1,029$ comparisons). ICC measures were calculated, representing how “close” the CU profiles match, with ICC = 1 denoting the perfect match. The orange distribution shows less variability and is shifted toward higher ICC values, denoting the closer overall match of species’ CU to their metagenome of origin

the constrained environmental conditions. *R. palustris* samples show on overall higher variability in CU, suggesting plasticity of codon usage that reflects on translational optimization and adapts to each specific environment. Even though the *R. palustris* strains generally show more variation in CU (Fig. 3), both species, regardless of environmental constraints, show the least relative variation of CU within the COG categories (i.e., orthologous genes) for housekeeping, including ribosomal protein genes.

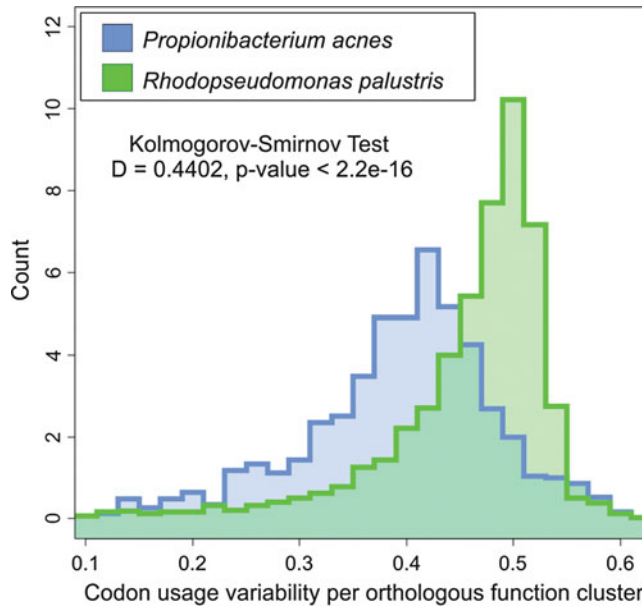
The Variability of Codon Usage in Metagenomes upon Removal of Dominant Phyla

Community-level codon usage bias is not an effect caused by the most abundant species. CU frequencies of the Sargasso Sea metagenome, the largest dataset in this study, were compared to other investigated metagenomes and to itself but with dominant phyla removed. The comparisons between Sargasso Sea CU frequencies and other metagenomes all show ICC < 0.75, while the same Sargasso sample with dominant phyla of the *Alphaproteobacteria* class removed

(~36% of the whole set) and the *Alphaproteobacteria* class itself show virtually no deviation (ICC > 0.98 and 0.95, respectively) from the original metagenome CU.

Codon Usage in Metagenomes Follows Similar Patterns as in Single Microbial Genomes

As has been established at the level of single microbial genomes (Ikemura 1985; Kanaya et al. 2001), the distance of each gene’s CU frequency to the overall CU of the whole genome and to that of a “reference set” of highly expressed genes (ribosomal protein genes) gives a characteristic crescent-shaped plot (Fig. 4a, introduced by (Karlín and Mrazek 2000)). Metagenomes exhibit similar CU distance distributions to those observed in single bacterial genomes, despite the fact that they comprise of genes that originate from diverse phylogenies (i.e., Santa Cruz whale carcass bone in Fig. 4b). If the amino acid composition of genes in a metagenome is kept constant but the codons are randomly chosen, the crescent plot shape analogous to single bacterial genomes and CU bias is lost.



Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Fig. 3 Environmental variability of codon usage. Variability of codon usage per COG category in 6 strains of *Rhodopseudomonas palustris* and in 12 strains of *Propionibacterium acnes*. The codon usage variability (calculated as median CU distance from the ribosomal

set within an orthologous group to its centroid CU) for the strains of *P. acnes* ($N = 15,436$), living in consistent environmental conditions, is shifted to the left, i.e., it shows smaller variation and higher bias than for the *R. palustris* strains ($N = 24,071$) living in diverse environmental conditions

Predicting Metagenomic Expression and Functional Profiles Through Synonymous Codon Usage

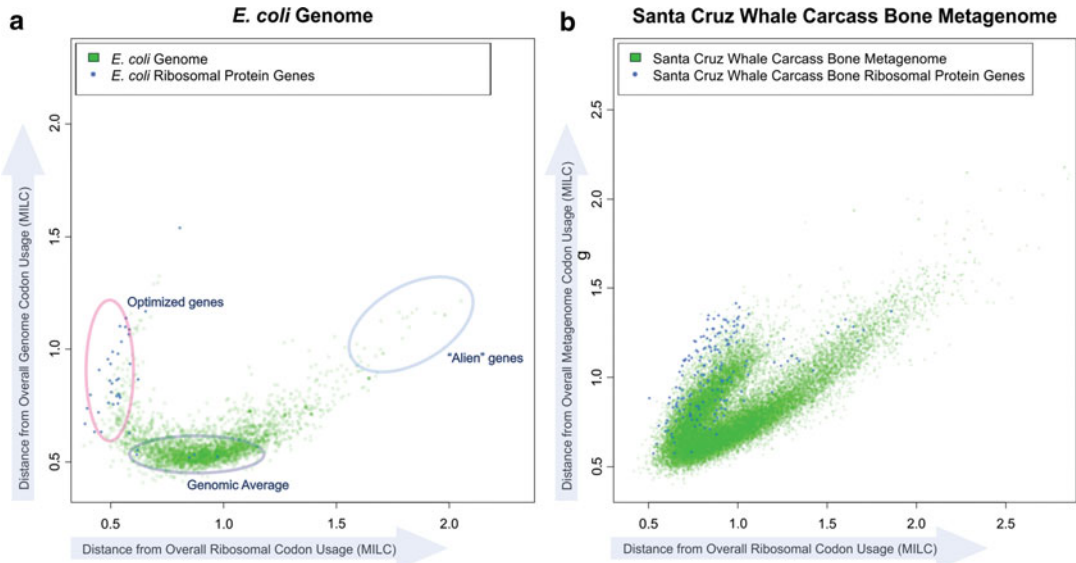
Under different environmental constraints, CU varies in single bacterial species, and metagenomes share synchronized CU as do single bacterial species. CU bias in metagenomes can be used to predict the expression levels of genes in the same manner as is routinely used to predict genes optimized for high levels of expression in single microbial genomes (Sharp and Li 1987; Karlin and Mrazek 2000; Supek and Vlahovicek 2005). Figure 5 depicts the resulting predictions at the level of whole metagenomes using the meta-ribosomal protein reference set. The most significantly enriched functions in the high expression level sets are (i) amino acid transport and metabolism (COG supercategory E) for Sargasso Sea, (ii) energy production and conservation (COG supercategory C) for the Whale fall metagenomes, and (iii) inorganic ion transport and metabolism (COG supercategory P)

for the acid mine biofilm metagenome. The most striking difference between metagenomes was lack of enrichment in energy production and carbohydrate metabolism (COG supercategories C and G) in the obese mice microbiota sample, in contrast to both lean human and mouse microbiota samples, indicating high metabolic activity of lean gut bacteria.

Artificial metagenomes, constructed from randomly selected genes of whole genome bacterial sequences from the NCBI with the same COG composition as their corresponding microbial samples, show loss of environment-specific enrichment of optimization in their expression profiles.

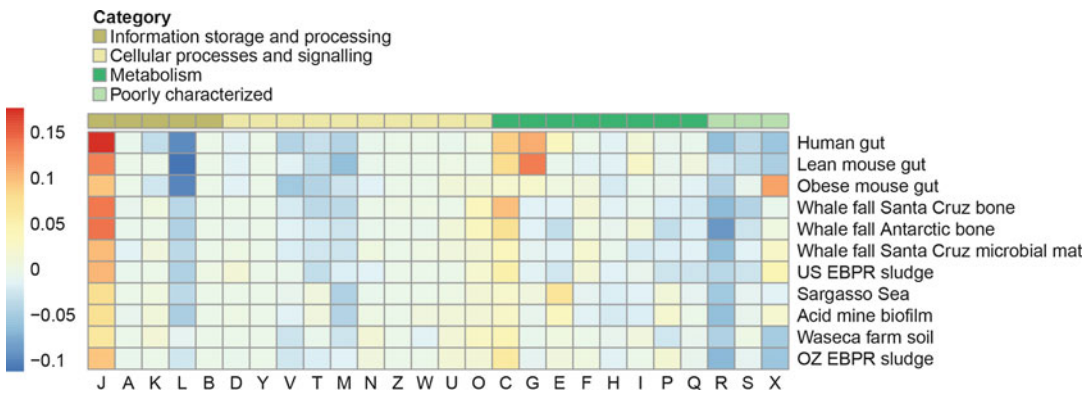
Validation with Metaproteomic Data

Predictions of gene expression for Sargasso Sea metagenome were compared to the Sargasso Sea metaproteomic study (Sowell et al. 2008) and a functionally (COG) classified subset of the human gut metaproteomic study (Verberkmoes et al. 2009). Predicted expression values based



Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Fig. 4 Metagenomes show codon usage distribution similar to single genomes. The distance of each gene's codon usage (CU) frequency forms the overall CU of the (meta) genome and ribosomal reference set, displayed as a Karlin

B-plot for (a) a single microbial genome (*Escherichia coli*, $N = 4,358$) and (b) a metagenome (whale carcass, $N = 33,422$). The metagenome shows the same characteristic distribution as the genome with ribosomal genes closer to the CU of the ribosomal set than the overall CU of the whole (meta)genome



Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities, Fig. 5 Enrichment of functions within highly expressed genes in metagenomes. Enrichment or depletion of functional annotations in the 3% genes with highest predicted expression (highest MELP measure) relative to the abundance of each COG supercategory in the whole metagenome for the OZ EBPR sludge ($N = 29,754$), Waseca farm soil ($N = 88,696$), acid mine biofilm ($N = 79,257$), Sargasso Sea ($N = 688,539$), US EBPR sludge ($N = 20,175$), Whale fall Santa Cruz microbial mat

($N = 40,916$), Whale fall Antarctic bone ($N = 30,503$), Whale fall Santa Cruz bone ($N = 33,422$), obese mouse gut ($N = 4,058$), lean mouse gut ($N = 4,955$), human gut ($N = 47,765$), Santa Cruz whale fall bone ($N = 33,422$), and acid mine ($N = 79,257$). Metagenomes show different functional enrichment patterns that are consistent with environmental requirements (e.g., metabolite transport functions [E] in the Sargasso Sea or energy conversion [C] in the whale carcass metagenome). Letters at the bottom represent COG supercategories

on CU optimization positively correlate with abundance in metaproteomic studies, both for the comparison of each gene with the protein most similar in sequence (Sargasso Sea $\rho=0.34$) and when median values per gene and protein COG are compared (human gut $\rho=0.34$). This opens up for an in silico prediction of overall metagenomic proteome status.

Summary

Analysis of eleven distinct metagenomes shows that microbial communities exhibit codon usage bias similar to that already described for single microbial species. Microbial communities sharing an environment are likely to have similar synonymous codon usage-based translational optimization for expression of environment-specific genes. This effect can be used to identify genes with unknown function and “optimal” codon encoding, indicating their potential for high expression and therefore high relative importance in the community metabolism and lifestyle.

References

- Bruggemann H, Henne A, Hoster F, Liesegang H, Wiezer A, Strittmatter A, et al. The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin. *Science*. 2004;305:671–3.
- Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101:3480–5.
- Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. *EMBO reports*. 2005;6:1208–13.
- Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. *Science*. 2006;312:1355–9.
- Hunyadkurti J, Feltoti Z, Horvath B, Nagymihaly M, Voros A, McDowell A, et al. Complete Genome Sequence of *Propionibacterium acnes* Type IB Strain 6609. *J Bacteriol*. 2011;193:4561–2.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Research*. 2007;17:377–86.
- Ikemura T. Codon Usage and Transfer-RNA Content in Unicellular and Multicellular Organisms. *Molecular Biology and Evolution*. 1985;2:13–34.
- Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution*. 2001;53:290–8.
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. *Nucl Acids Res*. 2006;34:D354–7.
- Karlin S, Mrazek J. Predicted highly expressed genes of diverse prokaryotic genomes. *Journal of Bacteriology*. 2000;182:5238–50.
- Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, et al. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodospseudomonas palustris*. *Nature Biotechnology*. 2004;22:55–61.
- Martin HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*. 2006;24:1263–9.
- Oda Y, Larimer FW, Chain PSG, Malfatti S, Shin MV, Vergez LM, et al. Multiple genome sequences reveal adaptations of a phototrophic bacterium to sediment microenvironments. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105:18543–8.
- Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 2011;12:32–42.
- Rocha EPC. Codon usage bias from tRNA’s point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*. 2004;14:2279–86.
- Sharp P, Li W. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 1987;15(3):1281–95.
- Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*. 2005;33:1141–53.
- Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J*. 2008;3:93–105.
- Staley JT, Konopka A. MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS. *Annual Review of Microbiology*. 1985;39:321–46.
- Supek F, Škunca N, Repar J, Vlahoviček K, Šmuc T. Translational Selection Is Ubiquitous in Prokaryotes. *PLoS Genet*. 2010;6:e1001004.
- Supek F, Vlahoviček K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *Bmc Bioinformatics*. 2005;6:15.

- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. *Bmc Bioinformatics*. 2003;4:14.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. *Science*. 2005a;308:554–7.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative Metagenomics of Microbial Communities. *Science (New York, N Y)*. 2005b;308:554–7.
- Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell*. 2010;141:344–54.
- Tuller T, Girshovich Y, Sella Y, Kreimer A, Freilich S, Kupiec M, et al. Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Research*. 2011;39:4743–55.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*. 2006;444:1027–31.
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 2004;428:37–43.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304:66–74.
- Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, et al. Shotgun metaproteomics of the human distal gut microbiota. *Isme Journal*. 2009;3:179–89.
- Vieira-Silva S, Rocha EPC. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet*. 2010;6:e1000808.

Evaluating Putative Chimeric Sequences from PCR-Amplified Products

Juan M. Gonzalez
Instituto de Recursos Naturales y Agrobiología,
IRNAS-CSIC, Seville, Spain

Introduction

The term chimera has its origins in the Greek mythology defining a creature composed of

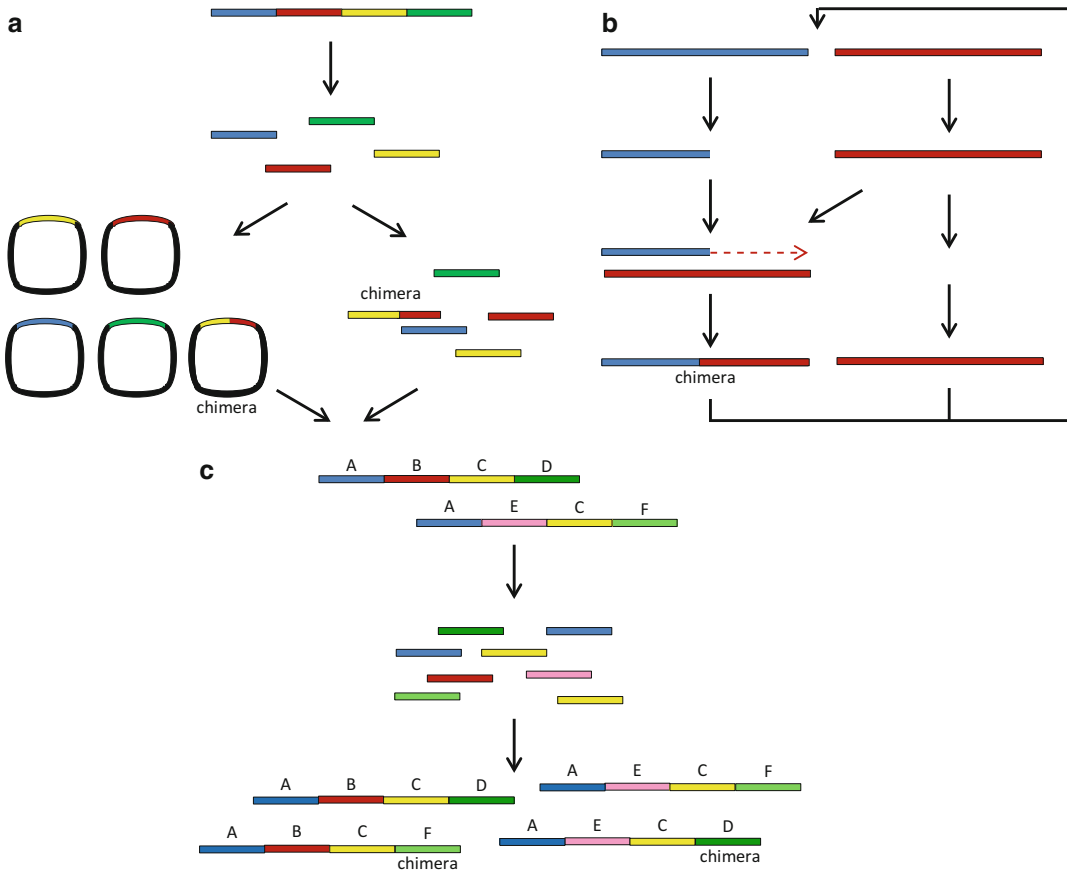
body parts from different living beings. In molecular biology, a chimeric sequence or chimera is a DNA sequence composed of DNA fragments originated from two or more genes or genomes.

Chimeric sequences can be naturally generated during DNA recombination which occurs naturally within a genome or by taking up foreign DNA by an organism. These processes of cross-over recombination are of interest in phylogenetic and evolution studies and need to be identified (Posada and Crandall 2002). Nevertheless, chimeras represent a serious problem to be considered when they are generated as artifacts during DNA manipulation and/or analysis.

Chimeric artifacts can be produced at different stages during experimental DNA studies. Some examples can be described relating to cloning procedures, DNA amplification, and/or DNA assembling during computational analysis (Fig. 1).

During DNA library preparation, genomic DNA is generally broken down into small fragments which will be introduced into cloning vectors or sequenced independently (Sambrook and Russell 2001). These fragments are generated by physical or enzymatic means. The generation of overlapping strand endings can lead to the random fusion of DNA fragments resulting in chimeras which can be detected upon sequencing (Fig. 1a).

By far, DNA amplification procedures represent the most frequently reported processes generating chimeric sequences. Most amplification procedures are prone to generate chimeras. The most studied case is the polymerase-chain reaction (PCR) amplification procedure where multiple sequences of a target DNA region are produced through a cycling amplification reaction. The amplification is exponential and errors during the reaction can be greatly amplified at the end of the PCR (Fig. 1b). Due to a variety of causes, incomplete amplification of the target fragment can behave as a priming sequence in the next cycle potentially originating a DNA fragment from two, or more, different DNA templates. The generation of chimeric sequences during PCR amplification can occur for any gene although the most studied case is that of



Evaluating Putative Chimeric Sequences from PCR-Amplified Products, Fig. 1 Scheme of different possibilities of potential chimera formation during DNA cloning and library preparation (a), PCR amplification (b) and computer processing of assembling DNA fragments (c). Examples are presented on chimeras formed during libraries aimed at both vector cloning (*left* on a) and direct sequencing (*right* on a). During PCR amplification (b),

incomplete synthesis of the target DNA fragment can lead in the next cycle to the annealing to a different target with conserved regions and result in its extension using a different DNA target. The consequence is the generation of a chimera resulting from two different DNAs. During computation of the assembly of small DNA fragments obtained through sequencing (c), different possibilities could be similarly valid and some of them can be chimeras

ribosomal RNA genes (rRNAs) which present both highly conserved and variable regions within their sequences. The rRNAs are present in every organism because the cells require them for protein synthesis. In Microbiology, rRNAs are widely used to detect and classify microorganisms; because most of these microorganisms are often unculturable and cannot be detected otherwise, the rRNAs are, at present, the only mean to survey for these microbes. It is easy to deduce that a chimera would represent a nonexistent microorganism, and so considering chimeras as real sequences can induce serious

overestimations of the microbial diversity in environmental studies (Hugenholtz and Huber 2003; Gonzalez et al. 2005). Thus, it is of most importance to detect and filter out those chimeric DNA sequences.

In addition to the potential to generate chimeras during DNA manipulation, the possibility to produce chimeras during computer processing of DNA sequences should be considered. Small DNA fragments forming DNA libraries are sequenced through a variety of sequencing platforms. These sequences are assembled into larger fragments of gene or genomic DNA. During this

assembly, a potential exists to produce a chimeric final sequence (Fig. 1c). Above all, this can be generated at the extreme of DNA assembled fragments generally induced by the presence of repetitive sequences (which often causes trouble during the assembly process) or by chimeras formed during early DNA manipulation steps or library preparation. As well, these assembling errors can truncate the generation of larger contigs or fragments of genomic DNA during the assembly. The assembly of DNA fragments from different organisms into a single DNA sequence is a risk when working with DNA surveys of complex communities, for instance, on metagenomes, that is, genomic studies of complex microbial communities (Mende et al. 2012).

Independently of the step where chimeric sequences have been generated, they need to be detected and filtered out to clean up these sequence artifacts for further analysis. Numerous strategies and pieces of software have been proposed. Herein, the case of rRNAs will be used as example as most studies on chimera evaluation have been carried out on these genes.

Chimeras and Microbial Diversity

Most surveys of the composition of microbial communities in natural environments are being performed through a PCR amplification step (Gonzalez et al. 2012). Generating a high number of fragments from the rRNAs (rRNA amplicons) represented in a community is a step previous to library preparation and sequencing (Wintzingerode et al. 1997; Roesch et al. 2007).

At present, microbial communities are understood as composed by a highly diverse number of microorganisms most of which remain unculturable (Curtis et al. 2002). If microorganisms cannot be cultured in the laboratory, it implies that the only means to analyze their potential features is through their nucleic acids. Due to the complexity of genomes, accurate taxonomic classification of microorganisms can only be performed with a small number of genes; the most frequently used are the rRNAs. Extensive databases have been built with rRNAs

and today these genes represent the primary way to classify microorganisms which are difficult to differentiate otherwise, either by morphology or physiological traits.

The rRNAs combine highly conserved and variable regions. Thus, partial synthesis of these genes during PCR amplification can lead to a DNA fragment able to anneal to a different rRNA sequence in a complex mixture of DNAs. Annealing of that incomplete DNA fragment to a target DNA from a different organism and extension in the same PCR cycle will result in the formation of hybrid sequences of rRNAs. This rRNA has been originated by portions of sequences from different microorganisms (Fig. 1b). Subsequent PCR cycles will generate multiple copies of that artifact. The result is the generation of chimeras which represent undesired artifacts that need to be detected and eliminated previous to further analysis.

The presence of chimeras in DNA databases have been previously reported (Hugenholtz and Huber 2003; Ashelford et al. 2005; Gonzalez et al. 2005) which affects negatively when users attempt to classify microorganisms by their rRNAs. About 5 % of rRNA gene sequences can represent suspicious or potential chimeras (Ashelford et al. 2005; Haas et al. 2011). The use of curated rRNA-specific databases is recommended. Databases, such as RDP (Ribosomal Database Project; Cole et al. 2009), Greengenes (DeSantis et al. 2006), and SILVA (Quast et al. 2013) (Table 1), have curated entries. These repositories ensure the lack of chimeras and so a realistic approximation to the identification of microorganisms through amplicon sequencing.

In spite of the potential for chimeras in environmental microbial surveys, current understanding of these communities suggests a huge microbial diversity (Curtis et al. 2002). This enormous diversity suggests that chimera detection is more complex than expected. However, the existence of a large set of sequences from microbial rRNAs can be an allied for an increasing accuracy in detecting chimeras. Only by knowing what is real, one can be in situation to discard what is unreal or chimeric (Gonzalez et al. 2005).

Evaluating Putative Chimeric Sequences from PCR-Amplified Products, Table 1 Some resources focused on rRNAs including database and software suites incorporating options and tools for chimera detection

Name	Chimera check procedure	Database/software	Link	Reference
Ribosomal Database Project (RDP)	Pintail	Database and tools	http://rdp.cme.msu.edu	Cole et al. 2003, 2009
SILVA	Pintail	Database and tools	http://www.arb-silva.de	Quast et al. 2013
Greengenes	Bellerophon	Database and tools	http://greengenes.lbl.gov	DeSantis et al. 2006
Mothur	Various ^a	Software suite	http://www.mothur.org	Schloss et al. 2009
QIIME	ChimeraSlayer	Software suite	http://qiime.org	Caporaso et al. 2010
AmpliconNoise	Perseus	Software suite	http://code.google.com/p/ampliconnoise/	Quince et al. 2011

^aVarious options are available: Bellerophon, Ccode, Pintail, ChimeraSlayer, Uchime, Perseus

Evaluating Putative Chimeric Sequences from PCR-Amplified Products, Table 2 Some of the latest software alternatives for chimera detection in sequence data

Program	Link	Reference
Bellerophon	http://comp-bio.anu.edu.au/bellerophon/bellerophon.pl	Hugenholtz and Huber 2003
Ccode	http://www.microextreme.net/downloads.html	Gonzalez et al. 2005
Pintail	http://www.mybiosoftware.com/rna-analysis/1262	Ashelford et al. 2005
WigeoN	http://microbiomeutil.sourceforge.net/#A_WigeoN	Haas et al. 2011
Decipher	http://decipher.cee.wisc.edu/FindChimeras.html	Wright et al. 2011
ChimeraSlayer	http://microbiomeutil.sourceforge.net/#A_CS	Haas et al. 2011
Uchime	http://drive5.com/uchime/uchime_download.html	Edgar et al. 2011
Perseus	http://code.google.com/p/ampliconnoise/	Quince et al. 2011

In fact, the large diversity of microorganisms known so far can provide with a range of variability within specific microbial taxa.

As microbial taxonomy and the sequences of rRNAs become increasingly defined and curated, the detection of chimeric rRNAs is gaining accuracy. Thus, curated and extensive rRNA databases will definitively contribute both to avoid the potential detection of real sequences as chimeras and to improve on the accurate detection of unreal sequences as chimeras.

Chimera Evaluation

Different procedures have been published to check for chimeras in newly generated DNA sequences. There has been a long list of programs

proposed to check or detect chimeras. Table 2 presents some of those alternatives with indication of its original publication and a link to its www homepage. As mentioned above, most of these studies have been carried out to detect chimeras in DNA fragments obtained from PCR amplification and specifically on rRNA genes. Originally, a simple method to intuitively and approximately detect a potential chimera was to search independently for homologues to the initial and final portions of the DNA fragments. This search was usually performed by blast searches (Altschul et al. 1990). If this blast resulted in different organisms for the initial and final portions of the DNA fragment, it was suspicious to be a chimera (Cole et al. 2003). More sophisticated attempts have been designed through the years. A fruitful method was to

analyze potential chimeras by comparison to the sequences obtained from the rRNA gene library being sequenced and analyzed (Hugenholtz and Huber 2003). Similar analysis can be carried out to full DNA databases or repositories (Ashelford et al. 2005; Quast et al. 2013). Further improvements included the analysis of the query sequence in relationship to the known sequences showing highest homology, for instance, within a taxonomic group. These known sequences marked the variability for small portions of the DNA fragment under analysis, and so those sequences showing the highest dispersion than the limited by known sequences were identified as potential chimeras, and these assessments included statistical results of the computational analysis (Ashelford et al. 2005; Gonzalez et al. 2005). Different procedures are periodically proposed to screen for chimeras, mainly performing analyses of portions of the DNA fragment (Wright et al. 2011) by searching if different results are received from DNA database searches. A DNA fragment is proposed as a chimera if it presents different homology results for different portions throughout its length.

As a result of the next-generation sequencing (NGS) platforms, large number of sequences is being generated through whole library sequencing. The screening of such amount of data would not be possible without the latest developments and the recent design of pipelines for the analysis of large data sets of DNA amplicon sequences (Schloss et al. 2009; Caporaso et al. 2010; Quince et al. 2011). The inclusion of chimera checking procedures within these pipelines (Table 1) has greatly facilitated the analysis of massive sequencing data. Nevertheless, the newly introduced algorithms are masked by the advantages presenting the whole pipelines and the easily handling of large sequencing data (Quince et al. 2011). One should confirm that the computational pipeline to process your sequencing data includes a chimera filtering procedure. Besides, some of these pipelines offer the possibility of using different databases. The inclusion in these analyses of curated databases is an important point to be considered.

Amplicon sequencing is still the most used procedure for microbial surveys through rRNAs. The detection of potential chimeras during these studies is a requirement to avoid the false consideration of nonexistent microorganisms and an overestimation of microbial diversity. Current pipelines for the processing of amplicon sequencing data incorporate chimera screening and filtering procedures. Databases must continue their current effort to evaluate newly deposited sequences for potential chimeras. Curated rRNA databases are a required reference for the taxonomically classification of microorganisms through sequencing data. These efforts will result in a more accurate detection of chimeras, a significant decrease in misclassifications due to erroneous sequences included in databases, and an improved knowledge of microbial species, gene, and genomic diversity.

Future Perspectives

As NGS is attracting most research on genomics, metagenomics, transcriptomics, and amplicon sequencing surveys, the massive data they generate and the work needed for processing these results is exponentially increasing. High-throughput procedures are required to cope with this demand. The use of current pipelines, or future improvements, should build a standard for amplicon sequencing. The detection of sequencing errors through algorithms in bioinformatics should also be introduced into these high-throughput pipelines, all aiming to obtain clean and accurate data previous to pursue further analysis. The screening and curation being performed by public repositories must continue in spite of the developments in pipelines and algorithms to ensure that databases remain as clean as possible of chimeric and erroneous sequences. At a time when sequencing analyses are not manually edited anymore, algorithms to automatically filtering off chimeras and the required curation at the scientist and database ends will become of increasing relevance.

Acknowledgments The author acknowledges funding from the Spanish Ministry of Economy and Competitiveness, project CONSOLIDER CSD2009-00006, which includes participation of Feder funds.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol.* 2005;71:7724–36.
- Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7:335–6.
- Cole JR, Chai B, Marsh TL, et al. The Ribosomal Database Project (RDPII): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucl Acids Res.* 2003;31:442–3.
- Cole JR, Wang Q, Cardenas E, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucl Acids Res.* 2009;37:D141–5.
- Curtis TP, Sloan WT, Scannell JW. Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA.* 2002;99:10494–9.
- DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 2006;72:5069–72.
- Edgar RC, Haas BJ, Clemente JC, et al. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics.* 2011;27:2194–200.
- Gonzalez JM, Zimmermann J, Saiz-Jimenez C. Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics.* 2005;21:333–7.
- Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE.* 2012;7(1):e29973.
- Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 2011;21:494–504.
- Hugenholtz P, Huber T. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Intl J Syst Evol Microbiol.* 2003;53:289–93.
- Mende DR, Waller AS, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE.* 2012;7(2):e31386.
- Posada D, Crandall AK. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol.* 2002;54:396–402.
- Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl Acids Res.* 2013;41: D590–6.
- Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinforma.* 2011;12:38.
- Roesch LFW, Fulthorpe RR, Riva A, et al. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 2007;1:283–90.
- Sambrook JJ, Russell DDW. *Molecular cloning. A laboratory manual.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2001.
- Schloss PD, Westcott SL, Ryabin T, et al. Introducing mother: open-source, platform-independent, community supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009;75:7537–41.
- Wintzingerode F, Göbel UB, Stackebrandt E. Determination of microbial diversity in environmental samples: pitfalls of PCR-base rRNA analysis. *FEMS Microbiol Rev.* 1997;21:213–29.
- Wright ES, Yilmaz LS, Noguera DR. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol.* 2011;78:717–25.

Extended Local Similarity Analysis (eLSA) of Biological Data

Fengzhu Sun and Li Charlie Xia

Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Synonyms

Local association analysis; Local similarity analysis

Introduction

The advances in high-throughput low-cost experimental technologies have made possible time series studies of hundreds or thousands biological factors simultaneously. The availability of such

datasets leads to an increased interest in profile similarity analysis techniques that can identify significant association patterns possibly embracing biological insights. In the context of metagenomics, factors of particular interest are operational taxonomic units (OTUs), microbial genomes, and environmental genes. Their association patterns may suggest microbe-environment, symbiotic relationships, and other types of interactions.

Many computational or statistical approaches exist to study the profile similarity at global scale, such as Pearson's correlation coefficients (PCC), Spearman's correlation coefficients (SCC), principal component analysis (PCA), multi-dimensional scaling (MDS), discriminant function analysis (DFA), and canonical correlation analysis (CCA). However, in many biological settings, the interaction may be active within only certain subintervals or the response to regulation may be time lagged. Methods based on the global relationships of profiles may fail to detect these interactions. Extended local similarity analysis (eLSA) method is specifically developed to capture local and potentially time-delayed co-occurrence and association patterns in time series data that cannot otherwise be identified by ordinary correlation analysis.

Description

Local Association with Possible Time Delays

Local association refers to the association that only occurs in a subinterval of the time of interest. Time-delayed association indicates that there is a time lag for the response of one factor to the change in another factor. As an example of local association, in Fig. 1, the top-left panel shows two series X and Y with nonsignificant correlation ($r = 0.26$, $P = 0.273$); however, they are in fact significantly correlated in the time interval from 7 to 16 as shown in the bottom-left panel (eLS = 0.43, $P = 0.028$). As an example of time-delayed local association, in Fig. 1, the top-right panel shows two series X and Y with nonsignificant correlation ($r = -0.26$, $P = 0.272$); however, they are in fact

significantly correlated in time interval from 4 to 17 if X is shifted three units toward origin as shown in the bottom-right panel (eLS = 0.51, $P = 0.006$).

Extended Local Similarity Analysis

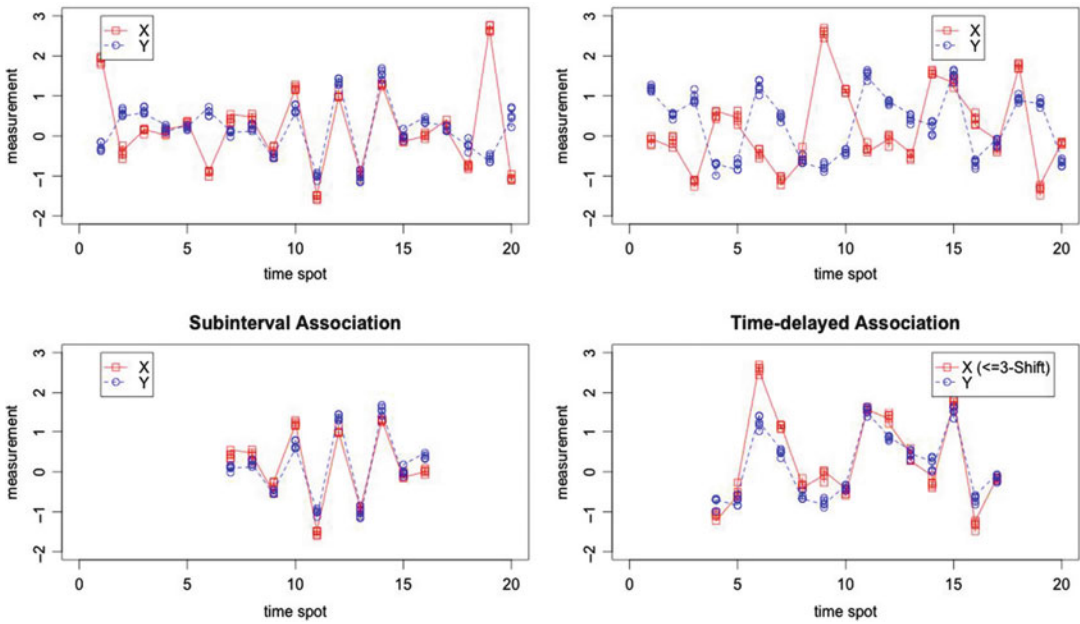
Extended local similarity analysis (eLSA) is an analysis technique designed to capture local associations possibly with time delays. eLSA extends the original local similarity analysis technique (Qian et al. 2001; Ruan et al. 2006) and local shape analysis technique to time series data with replicates (Xia et al. 2011). Improvements in computation efficiency of p -values are also made (Xia et al. 2013). Time series data of a pair of factors X and Y with replicates can be expressed as data matrices $X_{[1:m][1:n]}$ and $Y_{[1:m][1:n]}$, where each column is one sample from the time point and n is the number of time points; each row is a replicate and m is the number of replicates.

Given time series data of two factors and a user-constrained delay limit, eLSA uses dynamic programming algorithm to find the configuration of the data that yields the highest extended local similarity (eLS) score – a similarity metric defined as

$$eLS(X_{[1:m][1:n]}, Y_{[1:m][1:n]}) = \frac{1}{n} \max_{i,j,l \text{ s.t. } |i-j| \leq D} \left| \sum_{k=0}^{l-1} F(X_{[1:m], i+k}) F(Y_{[1:m], j+k}) \right|$$

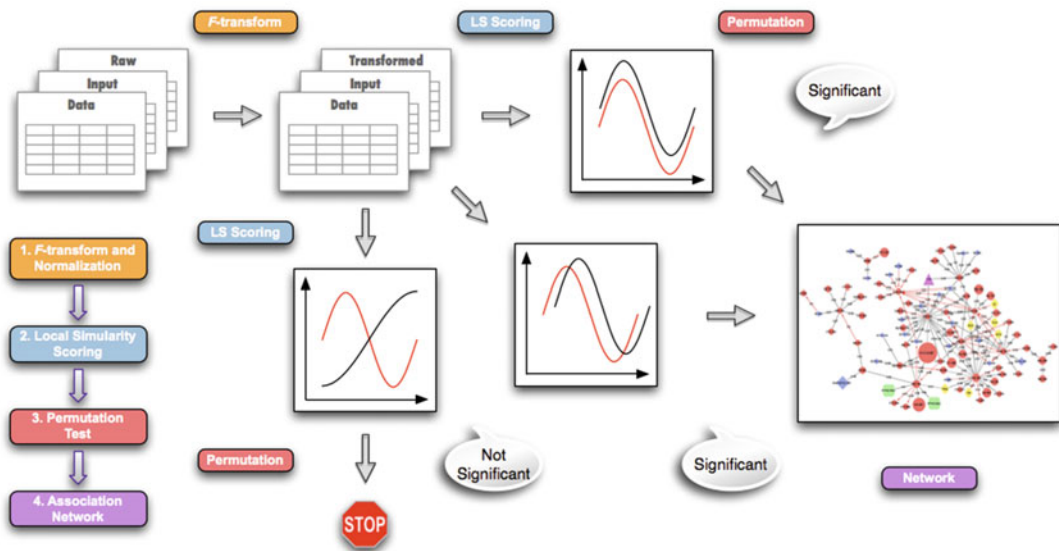
where D is the delay limit and F is the summarizing function for repeated measures (mean, median, etc.). For example, within a delay limit of two units, the first time spot of one series might be aligned to the third time spot of the other series, thus maximizing their eLS.

For a dataset of many factors, eLSA is applied to each pairwise combination of factors in the dataset. Candidate associations are then evaluated statistically by a permutation test, which calculates the p -value – the proportion of scores exceeding the original eLS score after shuffling the first series and reevaluating the eLS score many times – or more efficiently by theoretical approximation. Researchers can use eLSA to detect undirected associations, i.e., association patterns without



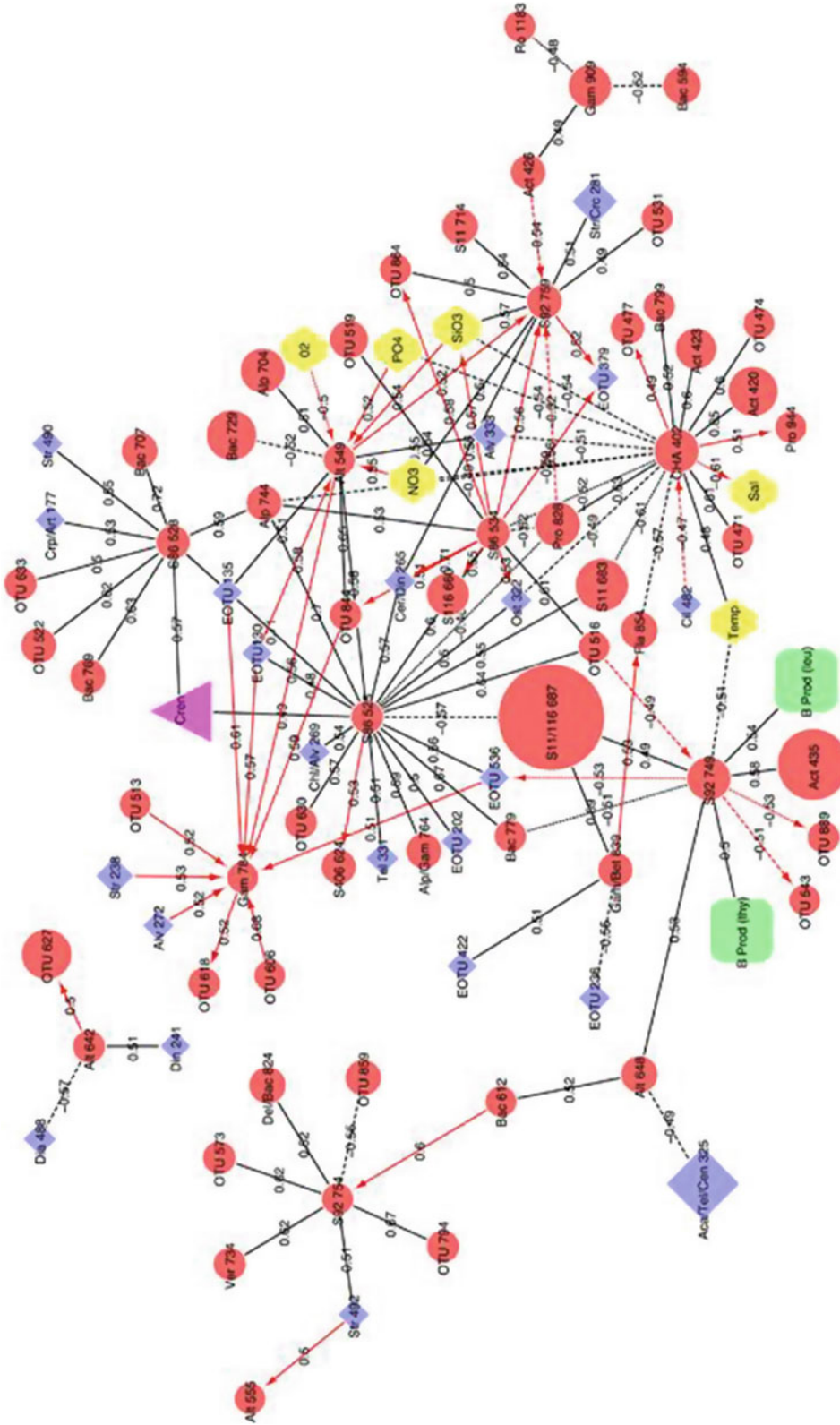
Extended Local Similarity Analysis (eLSA) of Biological Data, Fig. 1 Examples of local and time-delayed associations. *Top left*, two series X and Y with nonsignificant correlation ($r = 0.26, P = 0.273$); *bottom left*, they are in fact significantly correlated in the time interval from

7 to 16 (eLS = 0.43, $P = 0.028$); *top right*, two series X and Y with nonsignificant correlation ($r = -0.26, P = 0.272$); *bottom right*, they are significantly correlated in time interval from 4 to 17 if X is shifted three units toward origin (eLS = 0.51, $P = 0.006$)



Extended Local Similarity Analysis (eLSA) of Biological Data, Fig. 2 The eLSA pipeline. Users start with raw data (matrices of time series) as input and specify their requirements as parameters. The LSA tools subsequently *F*-transform and normalize the raw data and calculate extended local similarity (eLS) scores and Pearson’s

correlation coefficients. The tools then assess the statistical significance (p -values) of these correlation statistics using the permutation test and filter out insignificant results. Finally, the tools construct a partially directed association network from the significant associations



Extended Local Similarity Analysis (eLSA) of Biological Data, Fig. 3 An eLSA subnetwork built around γ -proteobacteria OTUs as central nodes (abbreviated *Alt* alternomas, *CHB* CHABI-7, *Gam* γ -proteobacterium, *S86* SAR86, *S92* SAR92)

time delays, and directed associations, where the change of one factor may temporally lead or follow another factor. Figure 2 shows the analysis pipeline of the eLSA technique.

Inferring Co-occurrence Networks Using eLSA

Studies adopting the local similarity analysis technique have shown interesting and novel discoveries for microbial community network analysis. In one of the studies (Steele et al. 2011), eLSA is used to find associations among relative abundances of bacteria, archaea, protists, total abundance of bacteria and viruses, and physico-chemical parameters. Co-occurrence networks were generated from significant eLSA associations to visualize and identify time-dependent relationship among ecologically important taxa, for example, the SAR11 cluster, stramenopiles, alveolates, cyanobacteria, and ammonia-oxidizing archaea.

A subnetwork from the study is shown in Fig. 3. It is built around γ -proteobacteria OTUs as central nodes (abbreviated Alt, alteromonas; CHB, CHABI-7; Gam, γ -proteobacterium; S86, SAR86; S92, SAR92). This subnetwork identifies 12 γ -proteobacterial OTUs. γ -proteobacteria OTUs correlate with eukaryotes and Crenarchaea (Cren), as well as environmental parameters and bacterial production. γ -proteobacterium SAR92-749 is more likely opportunistic species, as the relative abundance of SAR92-749 positively correlated with bacterial production measured by leucine and thymidine incorporation (eLS = 0.54, $P = 0.003$ and eLS = 0.495, $P = 0.005$, respectively).

Conclusion

eLSA technique uniquely captures local and potentially time-delayed co-occurrence and association patterns in time series data. eLSA technique is also applicable to other types of gradient data, including the response to different levels of treatments, temperature, humidity, or spatial distributions. The analysis pipeline is implemented as a C++ extension to Python, which streamlines data normalization, local similarity scoring,

permutation testing, and network construction. More information about the software is available from eLSA's homepage at <http://meta.usc.edu/softs/lisa>.

Cross-References

- ▶ [Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)

References

- Qian J, Dolled-Filhart M, Lin J, et al. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol.* 2001;314(5):1053–66.
- Ruan Q, Dutta D, Schwalbach MS, et al. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics.* 2006;22(20):2532–8.
- Steele JA, Countway PD, Xia L, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* 2011;5(9):1414–25.
- Xia LC, Steele JA, Cram JA, et al. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol.* 2011;5 Suppl 2:S15.
- Xia LC, Ai D, Cram J, et al. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics.* 2013;29(2):230–7.

Extraction Methods, Variability Encountered in

Paul L. E. Bodelier
Netherlands Institute of Ecology
(NIOO-KNAW), Wageningen, Netherlands

Synonyms

Bias in DNA extractions methods; Variation in DNA extraction methods

Definition

The variability in extraction methods is defined as differences in quality and quantity of DNA observed using various extraction protocols, leading to differences in outcome of microbial community composition assessments using genomic approaches.

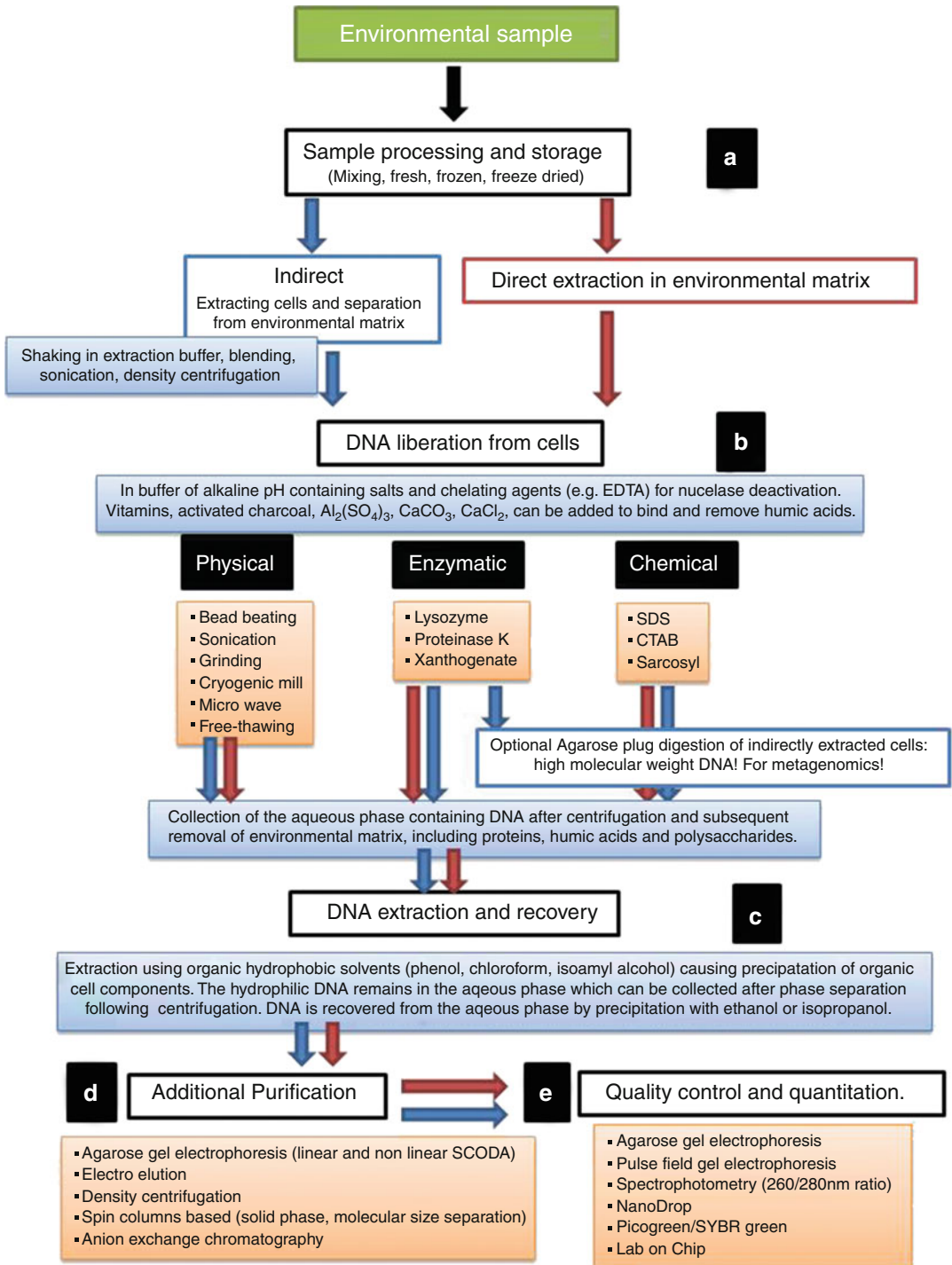
Introduction

Microbial communities are at the very basis of life on Earth, catalyzing biogeochemical reactions and driving global nutrient cycles (Falkowski et al. 2008). As yet, they are not on the global biodiversity conservation agenda, implying that microbial diversity is not under any threat by anthropogenic disturbance or climate change. However, this maybe a misconception caused by the rudimentary knowledge we have concerning microbial communities in their natural habitats as compared to the knowledge we have on plants and animals. The inability to culture the vast majority of microbes present in ecosystems prevents the detailed study of their ecology and physiology. The introduction of culture-independent methods based on DNA and RNA studies has revolutionized our abilities to study microbes and microbial communities in their natural habitat. A vast array of methods has been established going from assessing the complete genomes of single cells to whole communities. A vast number of books, overview articles, as well as reviews have been published on the molecular assessment of ecology, functioning and diversity of environmental microbes, and microbial communities of which the following are recommended: Liu and Jansson (2010), De Bruin (2011), and Kowalchuk et al. (2007). Despite the advances made and insight gained since the genomic revolution, we are still far away from understanding the functioning of microbial communities in situ and especially their individual contributions to biogeochemical reactions. The most challenging task ahead in microbial ecology will be to compare and integrate data from various habitats and

environmental conditions that are collected by different laboratories in order to come to concepts and general principles of community structure, functioning, and regulation. This is a challenge because the most crucial step in any culture-independent molecular microbial study is the extraction of nucleic acids from cells and the recovery from the environment. Within the last decades, countless procedures and protocols have been developed to obtain DNA or RNA from often very complex habitats optimized to yield DNA/RNA amendable to PCR- or non-PCR-based downstream community composition analyses. This was necessary because microbial cells as well as the habitats where they are retrieved from contain compounds which damage the nucleic acids directly, make the DNA/RNA inaccessible, or inhibit downstream applications directly. The major problem microbial ecology research is facing is that the efficiency and outcome of community composition analyses is variable between protocols used and between environmental matrices the protocols are applied to. This entry will give an overview of DNA extraction methods and associated biases and what can be done to improve comparability between different habitats.

DNA Extraction from Environmental Samples and Sources of Variability

When retrieving DNA from complex habitats, there are two main hurdles to take. First, the DNA has to be liberated from the cells. Second, the DNA has to be protected from degradation and precipitation which requires the separation from other cell components and environmental contaminants. As said, countless protocols have been developed which consist mainly of the five key steps which can vary in the way they are executed. An overview of these steps and variants in execution is given in Fig. 1. The overview is a summary of many studies and giving these references goes beyond the goal of this overview. However, most aspects addressed can be found in Kowalchuk et al. (2007), Herrera and Cockell (2007), and Lombard et al. (2011).



Extraction Methods, Variability Encountered in, Fig. 1 Schematic presentation of the steps and procedures to extract and purify DNA from environmental

samples. Step B is the step in all protocols where most biases are introduced

The first step in every DNA-based study is the collection and storage of environmental samples before the DNA is extracted (step A in Fig. 1). Depending on whether the samples are fresh or have been stored cold or frozen or whether they have been freeze dried can already give rise to variations in the extracted DNA quality and quantity depending on the environmental matrix and the community composition. However, recently it has been shown using a pyrosequencing approach that the variation introduced due to sample storage of soil and human-associated samples was insignificant (Lauber et al. 2010). After sample storage two routes can be followed to step B, the liberation of DNA from cells (step B in Fig. 1). Either cells are released from the environmental matrix by shaking or sonication followed by harvesting by, e.g., density centrifugation with subsequent lysis (indirect extraction) or the cells are lysed in the environmental matrix directly (direct extraction). Generally, the direct lysis is preferred because the DNA yield is higher due to no cell loss during cell extraction and purification. However, especially in metagenomic studies where large intact DNA fragments are required to (>20 kb in size) obtain complete genes, operons, and genomes, it has been shown that the indirect method is preferred and does also not lead to a significant difference in overall diversity (Delmont et al. 2011b). The subsequent liberation of DNA from cells is the step in all extraction protocols where most bias is introduced. Cell walls have to be broken. The efficiency is dependent on the cell wall structure (gram + vs. gram -) and the presence of extracellular slime layers composed of polysaccharides and proteins. Also the lyses methods commonly used, physical, enzymatic, and chemical (Fig. 1), differ in their efficiency of lyses, giving rise to variability, strongly depending on the community composition in terms of the presence of difficult to lyse cells. Also at this step, a choice of method can be made on the basis of the downstream application. The physical disruption techniques (e.g., bead beating) yield low molecular weight DNA (<20 kb) not suitable for metagenome studies. In this case the enzymatic lyses methods in

combination with detergents are preferred. The use of agarose plugs to perform enzymatic lysis has shown to be very effective in obtaining high molecular weight DNA (Williamson et al. 2011). Next to the method of lyses the environmental matrix is also a source of variation. The extraction and liberation of DNA always is executed in a "lysis buffer." The buffer normally is of alkaline pH (8–9) which reduces electrostatic interactions between DNA and proteins and which inhibits enzymes degrading DNA (nucleases) and facilitates denaturing of other proteins. Often a chelating agent (e.g., EDTA) is added to the buffer which destabilizes cell walls and membranes as well as proteins by binding cations (Ca^{2+} , Mg^{2+}). Besides protecting DNA from degradation once it is liberated, compounds that bind the DNA should be removed before non-DNA components are removed by centrifugation. Humic acids are derived from plant and animal remains by decomposition and are highly diverse in chemical structure. Due to their variability of functional groups on the molecules that can more or less strongly adhere to DNA and to the fact that the amount and structure depend on the biota and chemical conditions of the environment, the impact of humic acids on DNA extraction is highly variable. Hence, a large number of compounds (step B, Fig. 1) have been tested and used to bind and remove humic acids already at the stage of liberation of DNA. The latest addition was the use of vitamins (Techer et al. 2010). Centrifugation removes cell debris and precipitated components, while the supernatant containing the DNA is taken to step C (Fig. 1) which is the extraction from DNA out of the remaining organic cell and environmental components. This is done by phase separation using hydrophobic solvents (step C, Fig. 1), keeping the DNA in the aqueous which is underneath the hydrophobic phase containing the remaining cell components. Variability in this step can only come from the quality of the chemicals and the pipetting skills of the researcher. Care has to be taken not to collect any of the hydrophobic phase which leads to differences in the amounts of aqueous phase collected. The DNA is recovered by precipitation

using ethanol or isopropanol which destroys the helical structure leading to precipitation. After resuspension in water or buffer, the DNA can be ready for use in various analyses of abundance, diversity, or genomic procedures or has to be additionally cleaned up to remove any remaining impurities as indicated in step D (Fig. 1). The potential additional variation introduced here is that loss of DNA can occur leading to changes in relative abundances of species not reaching the detection limits of the respective downstream method. Hence, when DNA yield from samples is low, additional cleanup is often not an option. Also at this step some procedures are more applicable when HMW DNA is preferred. A procedure where direct current and pulsating nonlinear currents in gel electrophoresis are alternate has been shown to be very effective in purifying HMW DNA from the soil (Engel et al. 2012). The last step before downstream analyses is the quality control and quantification of the DNA concentration. UV spectrophotometry is most often used as an indicator of purity, where the ratio of absorbance at wavelengths 260/280 nm should be 2 when DNA is free from proteins or humic acids. The NanoDrop device is mostly used for this purpose because it only requires a few μ l of the precious extracted DNA. However, the spectrophotometric methods suffer from the fact that co-extracted RNA is also measured and that humic acids also lead to absorbance, eventually overestimating the amount of DNA in the extract. Alternative methods based on fluorescent dyes binding to double-stranded DNA can be used which only detect DNA, but which are also sensitive to interference by humic acids. Bias-free quantification methods are the ones where gel electrophoresis is combined with densitometry, which even is available in a lab-on chip format.

All the procedures described in Fig. 1 have also been combined and offered as commercial ready-to-go DNA extraction kits for various environmental matrices often by machinery for cell lyses. In Table 1 an overview of some commercially available kits and equipment is given.

Variability and Community Composition Assessment

The central question in microbial ecological research is why microbial communities are composed in the way they do and what factors influence community composition. To this end it is essential when comparing one sample with another that differences observed are due to biotic or abiotic factors and not biases introduced by the methods used. It is obvious from the previous section that a bias-free extraction of DNA from all environments is not possible. The matrix as shown in Fig. 1 is a collection of methods developed with the goal to obtain PCR-amplifiable DNA. Hence, the protocols were not designed for bias-free extraction but for obtaining extract enabling downstream applications. Considering the inherent problems specific to various environmental matrices, not a single protocol will suffice to be applied to all environments. The protocols developed were designed and tested to yield the highest quality and quantity of DNA and highest diversity in fingerprinting (denaturing gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (T-RFLP), microarray) methods or highest abundance assessed with quantitative polymerase chain reaction (qPCR) or highest MW DNA in metagenomic studies. Hence, community composition was the criterion for testing performance of protocols, and the amount of protocols available is a good indicator of the biases introduced. However, it was demonstrated that even when applying 1 protocol on exactly the same soil sample, community composition analyses following DNA extraction are not bias-free (Pan et al. 2010). When a single well-homogenized soil sample was extracted in different laboratories using the same protocol, biases were already introduced at the initial extraction. The DNA quantity (Fig. 2a) as well as quality varied significantly between laboratories leading to significant differences in community composition of methane-oxidizing bacteria (Fig. 3) as assessed by PCR-based microarray analyses. Moreover, the same extractions performed by

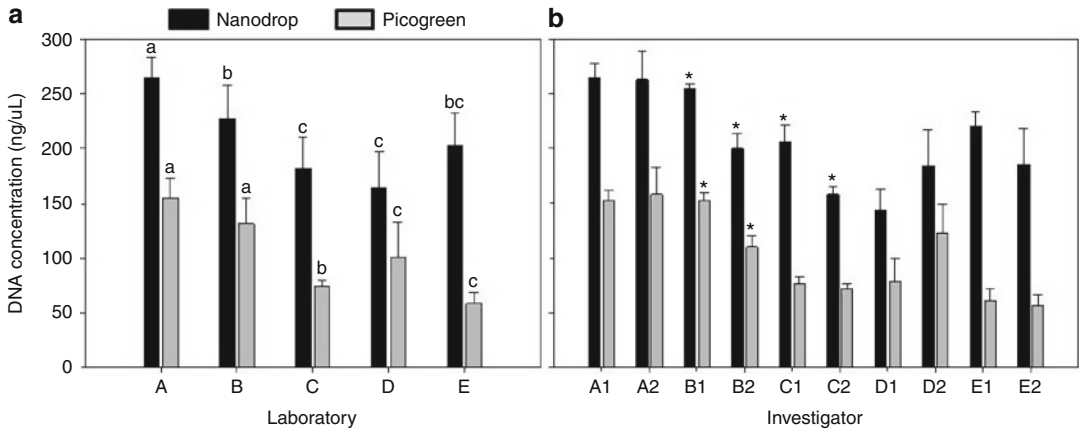
Extraction Methods, Variability Encountered in, Table 1 Overview of a number of commercially available DNA extraction kit, lyses equipment, additional cleanup kits, and DNA quantitation methods

Soil DNA extraction kits	
PowerSoil and PowerMax/Mobio	http://www.mobio.com/soil-dna-isolation/powermax-soil-dna-isolation-kit.html
SoilMaster/Epicentre Technologies	http://www.epibio.com/item.asp?id=388
E.Z.N.A._ Soil DNA Kit/Omega BioTek	http://www.omegabiotek.com/product_detail.php?ID=95
ZR Soil Microbe DNA Kit/Zymo Research	http://www.zymoresearch.com/media/downloads/212/D6001d.pdf
FastDNA_ SPIN kit for Soil/MP Biomedicals	http://www.biocompare.com/11793-DNA-Purification-Kits-Soil/2691724-FastDNA96-Soil-Microbe-DNA-Kit/
Cell disruption equipment	
BioSpec Mini Bead Beater	http://www.biospec.com/product/28/mini_beadbeater/
MP Biomedicals FastPrep [®] -24 or MP Biomedicals FastPrep [®] -96	http://www.mpbio.com/product_info.php?family_key=116004500
Geno/Grinder [®]	http://www.spexsampleprep.com/equipment-and-accessories/equipment_product.aspx?typeid=1
Free/Mill [®]	http://www.spexsampleprep.com/equipment-and-accessories/equipment_product.aspx?typeid=2
Additional cleanup kits	
Wizard [®] SV Gel and PCR Clean-Up System	http://www.promega.com/products/dna-and-rna-purification/dna-fragment-purification/wizard-sv-gel-and-pcr-clean_up-system/
Sepharose 4B [®] columns	http://www.gelifesciences.com/webapp/wcs/stores/servlet/catalog/nl/GELifeSciences-nl/products/AlternativeProductStructure_17546/17075701
Nonlinear electrophoresis (SCODA)	http://www.borealgenomics.com/products/aurora/
DNA quality/quantity	
NanoDrop	http://www.nanodrop.com/
PicoGreen (QuantiT TM)	http://www.invitrogen.com/site/us/en/home/brands/Product-Brand/Quant-iT.html
Microfluidics Agilent Bioanalyzer	http://www.genomics.agilent.com/GenericB.aspx?PageType=Family&SubPageType=FamilyOverview&PageID=183

two investigators simultaneously in the same laboratory using exactly the same chemicals and equipment also yielded significant differences in DNA quantity (Fig. 2b) and quality proving that also the investigator can introduce biases, probably due to pipet handling in step C (Fig. 1) of the protocol. Another source of bias appeared to come from the DNA quantitation method (Fig. 2) leading to significantly different community profiles (Fig. 4) as well as abundance of methane-consuming bacteria. In this case overestimation of DNA quantity by NanoDrop leads to a higher dilution of the DNA to reach the same input amount of target DNA as in the PicoGreen-based PCR reaction. This dilution reduced the remaining inhibition of the PCR by contaminants still present in the DNA with

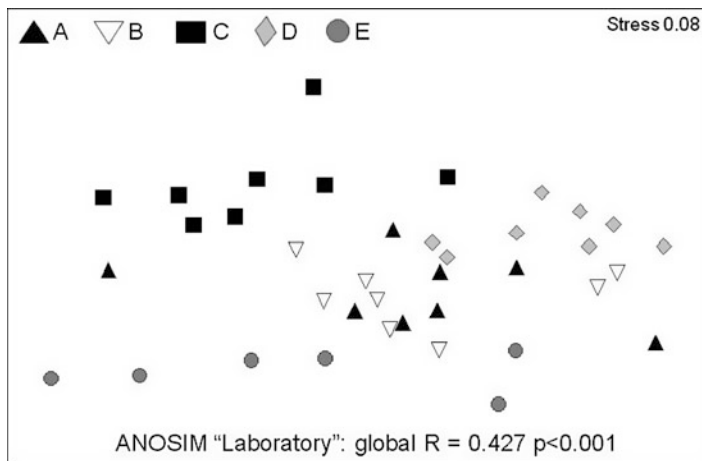
consequences for the subsequent outcome of the downstream analyses.

Important improvements were made to reduce extraction bias by extracting the same sample matrix, remaining in the pellet of step B (Fig. 1) multiple times (Feinstein et al. 2009). After three extractions DNA quantity as well as bacteria abundance reached a plateau which was similar for a number of different lyses protocols. This demonstrates that a single extraction always gives a biased picture of the community composition. Combining multiple extraction protocols has shown to enhance the detected diversity of recovered species by more than 80 % (Delmont et al. 2011a) in soil samples. However, the relative abundance of the various approaches was different, making this approach very important



Extraction Methods, Variability Encountered in, Fig. 2 DNA concentrations (means \pm 1 standard deviation) as analyzed with NanoDrop or PicoGreen, showing the comparisons between laboratories (a) and between investigators in the various laboratories (b). Different letters in panel A indicate significant differences between

countries ($P < 0.05$, unequal honestly significant difference test). In panel B, the asterisk indicates a significant difference between investigators within one laboratory (as assessed using Student's *t* test; $P < 0.01$) (From Pan et al. 2010 with permission)

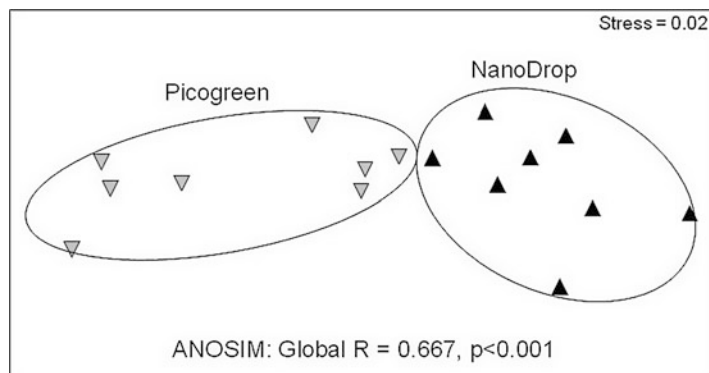


Extraction Methods, Variability Encountered in, Fig. 3 Nonmetric multidimensional scaling plot using log-transformed Bray-Curtis dissimilarity matrices based on signal intensity values of the *pmoA* microarray analyses on DNA extracted in five different laboratories. Distances between symbols represent relative

dissimilarity between MOB communities. Analyses of similarity (ANOSIM) resulted in a significant difference between MOB community structures analyzed in the different laboratories. Only samples from laboratory A and B did not differ from each other ($n = 8$, except for laboratory E [$n = 6$]) (From Pan et al. 2010 with permission)

for complete diversity assessment but not for comparisons between different samples or environments. The first attempt for standardization between samples and environments has been established for soils where an ISO-certified extraction protocol was tested on various soils

and by a number of different laboratories (Petric et al. 2011). The protocol was only standardized up to what is believed to be the step (step B, Fig. 1) causing most variation. Thirteen different laboratories tested a number of soil types. There was variation in DNA quantity and quality and



Extraction Methods, Variability Encountered in, Fig. 4 Nonmetric multidimensional scaling plot using log-transformed Bray-Curtis dissimilarity matrices based on signal intensity values of *pmoA* microarray analyses, performed on the basis of the NanoDrop or PicoGreen DNA quantitation method. Distances between symbols

represent relative dissimilarity between MOB communities. Analyses of similarity (ANOSIM) resulted in a significant difference between MOB community structures when based on different DNA concentration measurements ($n = 8$) (From Pan et al. 2010 with permission)

also in community fingerprinting but acceptable as compared to commonly observed variation. Although the soils did not differ/vary much in their complexity and only one fingerprinting method was used, this standard protocol is a very important step toward comparability of samples. At least for the intensively studied soil habitat, comparisons may be possible and similar standardizations for related habitats may be a way to go.

Conclusions

It is obvious that not one protocol of DNA extraction will be bias-free and that applying a single protocol to a sample will never yield a “true” picture of microbial community composition. The inherent differences in the properties of environmental matrices prevent this. However, important improvements have been made leading to the recommendation to perform multiple extractions on the same matrix and multiple protocols with varying stringency of lyses to maximize diversity assessments of single samples. When different samples have to be compared in time or between treatments or habitats, it is best when extractions are performed in the same

laboratory by the same person using the identical chemicals and machinery, especially the bead-beating apparatus. Of course the latter may not always be feasible, and an extraction robot may be very useful in order to reduce variation caused by pipet handling (e.g., Maxwell-16 system from Promega). However, in order to come to real ecological comparisons of microbial communities, new methods of standardization have to be developed. Internal standardization by spiking samples with a known amount of cells may be an option. The most important, however, will be to assess for every sample matrix what the extent of the bias is and take that into account in the interpretation.

Summary

Microbial communities are the drivers of all ecosystems on Earth but are also the least understood branch on the tree of life. The advent of molecular biological techniques assessing environmental nucleic acids has revolutionized the amount of information on environmental microbial communities. However, in the era of metagenomics and high-throughput sequencing, the critical step in microbial community analyses is still the

extraction of DNA from environmental samples. DNA is extracted by liberation from cells followed by extraction from the matrix using organic solvents and recovered by precipitation with alcohols. The lyses of cells and the removal of contaminants that degrade or adhere to the DNA call for many different approaches varying in effectiveness and leading to substantial bias in downstream genomic or metagenomic applications. Next to this, variation can also be introduced to investigator skills. Improvements have been made for increasing the observed diversity in one single sample, and for soils, an ISO-certified extraction protocol has been established facilitating ecological comparisons for this habitat. For true ecological comparisons, new ways of standardization have to be developed.

References

- De Bruin FJ, editor. Handbook of molecular microbial ecology II: metagenomics in different habitats. Hoboken: Wiley; 2011.
- Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, et al. Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol.* 2011a;77(4):1315–24.
- Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods.* 2011b;86(3):397–400.
- Engel K, Pinnell L, Cheng J, Charles TC, Neufeld JD. Nonlinear electrophoresis for purification of soil DNA for metagenomics. *J Microbiol Methods.* 2012;88(1):35–40.
- Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science.* 2008;320(5879):1034–9.
- Feinstein LM, Sul WJ, Blackwood CB. Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl Environ Microbiol.* 2009;75(16):5428–33.
- Herrera A, Cockell CS. Exploring microbial diversity in volcanic environments: a review of methods in DNA extraction. *J Microbiol Methods.* 2007;70(1):1–12.
- Kowalchuk GA, De Bruin FJ, Head IM, Akkermans AD, Van Elsas JD, editor. Molecular microbial ecology manual, 2nd ed. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2007.
- Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *Fems Microbiol Lett.* 2010;307(1):80–6.
- Liu W-T, Jansson JK, editors. Environmental molecular microbiology. Norfolk: Caister Academic press; 2010.
- Lombard N, Prestat E, van Elsas JD, Simonet P. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *Fems Microbiol Ecol.* 2011;78(1):31–49.
- Pan Y, Bodrossy L, Frenzel P, Hestnes AG, Krause S, Luke C, et al. Impacts of inter- and intralaboratory variations on the reproducibility of microbial community analyses. *Appl Environ Microbiol.* 2010;76(22):7451–8.
- Petric I, Philippot L, Abbate C, Bispo A, Chesnot T, Hallin S, et al. Inter-laboratory evaluation of the ISO standard 11063 "Soil quality – method to directly extract DNA from soil samples". *J Microbiol Methods.* 2011;84(3):454–60.
- Techer D, Martinez-Chois C, D'Innocenzo M, Laval-Gilly P, Bennisroune A, Foucaud L, et al. Novel perspectives to purify genomic DNA from high humic acid content and contaminated soils. *Sep Purif Technol.* 2010;75(1):81–6.
- Williamson KE, Kan J, Polson SW, Williamson SJ. Optimizing the indirect extraction of prokaryotic DNA from soils. *Soil Biol Biochem.* 2011;43(4):736–48.

Extradiol Dioxygenases Retrieved from the Metagenome

Kentaro Miyazaki^{1,2} and Hikaru Suenaga²
¹Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Sapporo, Japan
²Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology, Sapporo, Japan

Synonyms

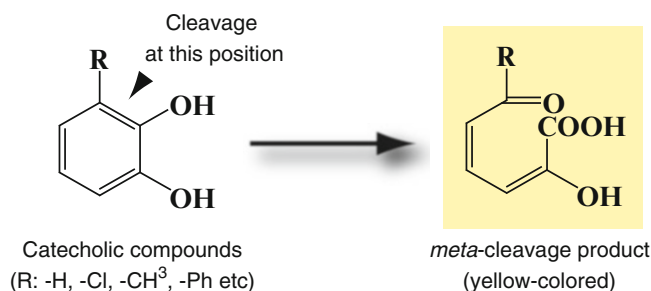
Extradiol Dioxygenases

Definition

Extradiol dioxygenases (EDOs) are mononuclear metalloenzymes that cleave the *meta*-position of the C–C bond of catecholic compounds, yielding yellow-pigmented open-ring products (Fig. 1).

Extradiol Dioxygenases Retrieved from the Metagenome,

Fig. 1 *Meta*-cleavage of catecholic compounds by EDOs



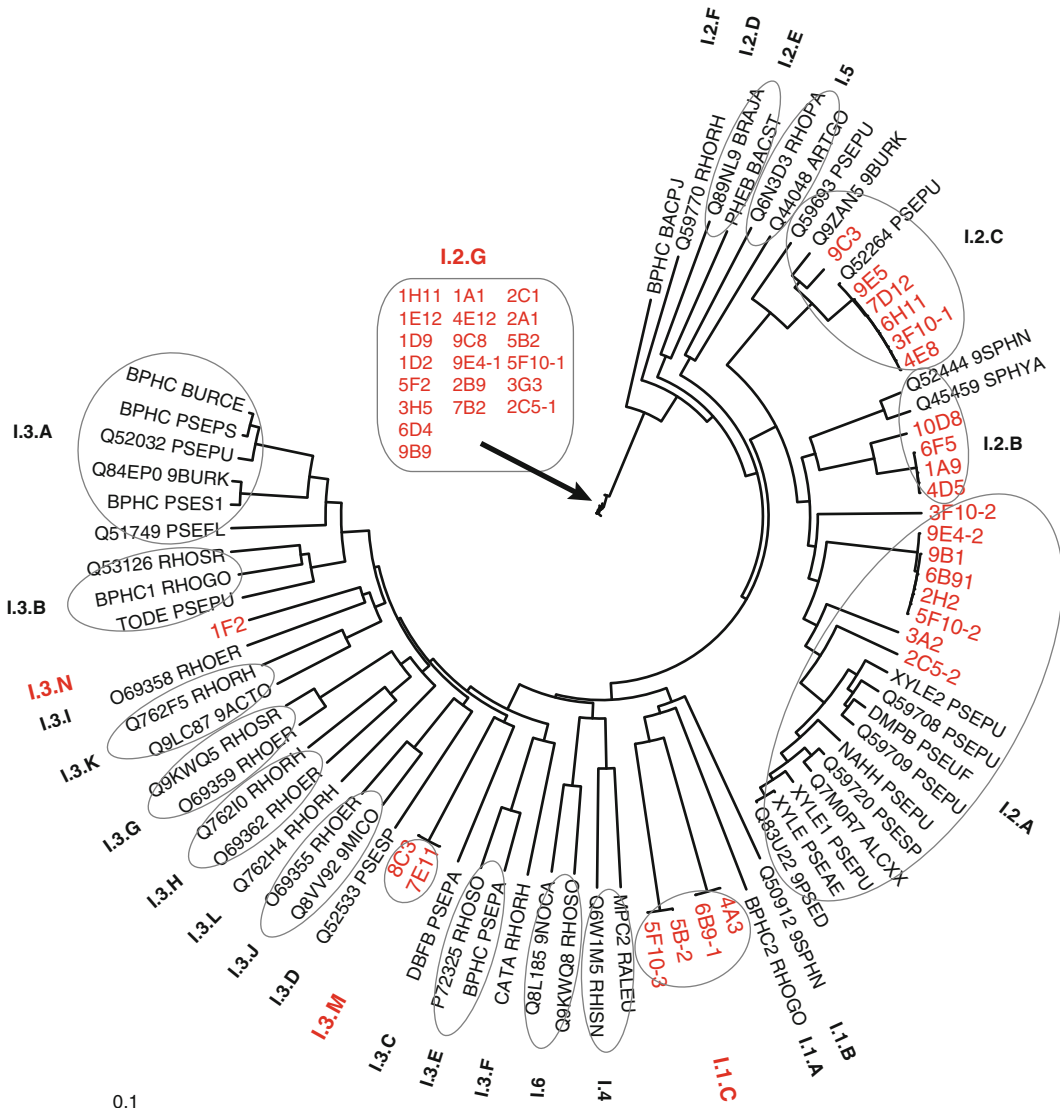
Introduction

Both naturally existing and synthetic aromatic hydrocarbons (e.g., petroleum products and chemical wastes of agricultural and industrial origin) are common contaminants of the environment (US Environmental Protection Agency; <http://www.epa.gov>). Microorganisms, particularly bacteria, play crucial roles in the biodegradation of these compounds and contribute to various biochemical cycles (Abraham et al. 2002; Chakraborty and Coates 2004; Furukawa et al. 2004). Extensive efforts have been directed at surveying and analyzing the pathways and genes responsible for the degradation of aromatic compounds with the aim of reviving polluted environments by using these microorganisms (i.e., bioremediation) (Top and Springael 2003; Janssen et al. 2005; de Lorenzo 2008). These studies have shown that the initial conversion step in the degradation of aromatic compounds is catalyzed by various types of enzymes, depending on the aromatic compound substrate, pathway, or the organism. Despite the variation, however, aromatic compound substrates are converted to a limited number of central intermediates, most commonly the catecholic compounds (Fritsche and Hofrichter 2005). The subsequent cleavage of the aromatic rings of catechol derivatives is catalyzed by extradiol dioxygenases (EDOs); these reactions are considered crucial in the biodegradation of aromatic compounds (Lipscomb 2008). EDOs have thus served as functional markers in the assessment of the biodegradation potential of specific bacterial communities (Vilchez-Vargas et al. 2010).

Most of our knowledge on EDOs has been obtained from activities involving microbial screening. Based on the observation of bacterial colonies that develop yellow pigments attributable to the ring-cleavage products of catecholic substrates, those expressing EDOs were isolated and studied in detail for the past three decades. However, information on the degradation pathways, enzymes, and genes that are harbored by “uncultured” bacteria remain unknown. Screening of those genes using a metagenomic approach should thus shed light on the diversity, evolution, and biochemical properties of novel pathways, enzymes, and genes.

Enzymatic Classification of EDO Family

EDOs can be classified into at least three evolutionarily distinct families (Vilchez-Vargas et al. 2010): type I belongs to the vicinal oxygen chelate superfamily, type II includes enzymes consisting of different subunits, and type III belongs to the cupin superfamily. Type I is considered as a major family and is further divided into subfamilies (e.g., I.2.A) depending on the amino acid sequences of the enzymes (Fig. 2). Enzymes belonging to the same subfamily are defined as those with >54 % sequence identity (Eltis and Bolin 1996). They are roughly classified into two families: those that act on monocyclic aromatics (subfamily I.2) and those that act on bicyclic aromatics (subfamily I.3). Despite differences in substrate specificities, these enzymes share common mechanisms of reaction, occurring at similar catalytic centers that contain a Fe(II) ion in the



Extradiol Dioxygenases Retrieved from the Metagenome, Fig. 2 A phylogenetic tree showing both metagenomic EDOs and previously identified type I EDOs. The metagenomic clones identified from the activated sludge of wastewater from a Coke plant (Suenaga et al. 2007) are shown in red. The accession numbers of the previously identified EDOs are as follows: BPHC BACPJ, Q8945; Q59770 RHORH, Q59770; PHEB BACST, P31003; Q89NL9 BRAJA, Q89NL9; Q59693 PSEPU, Q59693; Q9ZAN5 9BURK, Q9ZAN5; Q52264 PSEPU, Q52264; Q52444 9SPHN, Q52444; Q45459 SPHYA, Q45459; XYLE2 PSEPU, Q04285; Q59708 PSEPU, Q59708; DMPB PSEUF, P17262; Q59709 PSEPU, Q59709; NAHH PSEPU, P08127; Q59720 PSESP, Q59720; Q7M0R7 ALCXX, Q7M0R7; XYLE1 PSEPU, P06622; XYLE PSEAE, P27887; Q83U22 9PSED, Q83U22; Q6N3D3 RHOPA, CGA009; Q44048 ARTGO, Q44048; Q50912

9SPHN, Q50912; MPC2 RALEU, P17296; Q6W1M5 RHISN, Q6W1M5; Q9KWQ8 RHOSR, Q9KWQ8; Q8L185 9NOCA, Q8L185; BPHC2 RHOGO, P47232; DBFB PSEPA, P47243; CATA RHORH, Q53034; BPHC PSEPA, P11122; P72325 RHOSO, P72325; Q52533 PSESP, Q52533; Q8VV92 9MICO, Q8VV92; Q69355 RHOER, Q69355; Q762H4 RHORH, Q762H4; Q69362 RHOER, Q69362; Q76210 RHORH, Q76210; Q69359 RHOER, Q69359; Q9KWQ5 RHOSR, Q9KWQ5; Q9LC87 9ACTO, Q9LC87; Q762F5 RHORH, Q762F5; Q69358 RHOER, Q69358; TODE PSEPU, P13453; BPHC1 RHOGO, P47231; Q53126 RHOSR, Q53126; Q51749 PSEFL, Q51749; BPHC PSES1, P17297; Q84EP0 9BURK, Q84EP0; Q52032 PSEPU, Q52032; BPHC PSEPS, P08695; BPHC BURCE, P47228 (This figure was drawn using the FigTree software (<http://tree.bio.ed.ac.uk/software/figtree/>))

active site and are coordinated by the so-called 2-His-1-carboxylate facial triad motif (Lipscomb 2008).

EDOs Retrieved from the Metagenome

At the time of writing of this report (March 2013), 42,295 “extradiol dioxygenase” sequences have been deposited in the Protein database of NCBI (www.ncbi.nlm.gov/protein), 1,076 of which are derived from “uncultured bacteria.” Of the 1,076 sequences, however, only few contain complete EDO sequences (Vilchez-Vargas et al. 2010; Suenaga 2012).

Based on the yellow coloration of catechol ring-cleavage products, 235 positive clones were identified from the fosmid library constructed from environmental DNA extracted from petrol-contaminated soil (Brennerova et al. 2009). PCR-based classification of the internal sequences of the metagenomic EDO genes showed that only one-fourth of the observed EDOs belong to subfamily I.3.A of I.3.B that would be expected as predominant taking into consideration of the knowledge obtained from isolated bacteria. Genes of subfamily I.2.A, which have frequently been used as DNA markers for assessing the catabolic potential of polluted sites, were also absent (Vilchez-Vargas et al. 2010). Functional analysis of representative proteins indicated that 1 clone, s45, has exceptionally high affinity for different catecholic substrates.

Coke plant wastewater contains various aromatic compounds and activated sludge that is used for decontamination may serve as a rich resource for EDO discovery. Suenaga et al. (2007) created a metagenomic fosmid library using the activated sludge and by functional screening, 91 EDO-positive clones were identified. Based on their substrate specificity for various catecholic compounds, 38 clones were subjected to shotgun DNA sequencing. Some clones contained 2 EDO genes and as a result, a total of 43 EDO genes were identified. Approximately half of these were classified into

known subfamilies, but surprisingly, 23 genes could not be classified into existing subfamilies, and therefore, four new subfamilies, namely, I.1.C, I.2.G, I.3.M, and I.3.N (Fig. 2), were proposed. Among these novel EDOs, the I.2.G subfamily genes were overrepresented among the retrieved metagenomic EDOs and branched at a deep point in the lineage. Enzymatic characterization demonstrated that the I.2.G EDOs have unique properties, including Mn(II) dependence instead of the more common Fe(II) dependence, as well as the highest affinity for catechol reported thus far, and tolerance for thermal and chemical inhibitors (NaCN and H₂O₂) (Suenaga et al. 2009).

EDO Application for Bioremediation

Each polluted site harbors contaminants that carry environment-specific EDO genes. Monitoring these “marker” EDO genes using the metagenomic approach may be a good method in evaluating the bioremediation process (Widada et al. 2002). Furthermore, retrieving novel EDOs, as well as engineering these for higher activity and stability, can enhance the development of bioremediation processes.

Summary

Metagenomic approaches are an effective means of discovering novel enzymes including EDOs, which present specific sequences and enzymatic properties based on their substrate preference, metal dependence, inhibitor tolerance, and various physicochemical properties. Research targeting different environments may help in furthering the knowledge about the diversity of EDOs.

Cross-References

- [Metagenomics Potential for Bioremediation](#)

References

- Abraham WR, Nogales B, Golyshin PN, et al. Polychlorinated biphenyl-degrading microbial communities in soils and sediments. *Curr Opin Microbiol.* 2002;5:246–53.
- Brennerova MV, Josefiova J, Brenner V, et al. Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ Microbiol.* 2009; 11:2216–27.
- Chakraborty R, Coates JD. Anaerobic degradation of monoaromatic hydrocarbons. *Appl Microbiol Biotechnol.* 2004;64:437–46.
- Eltis LD, Bolin JT. Evolutionary relationships among extradiol dioxygenases. *J Bacteriol.* 1996;178:5930–7.
- Fritsche W, Hofrichter M. Aerobic degradation by microorganisms. In: Rehm H-J, Reed G, editors. *Biotechnology: environmental processes II*, vol. 11b. 2nd ed. Weinheim: Wiley-VCH Verlag GmbH; 2008.
- Furukawa K, Suenaga H, Goto M. Biphenyl dioxygenases: functional versatility and directed evolution. *J Bacteriol.* 2004;186:5189–96.
- Janssen DB, Dinkla IJT, Poelarends GJ, et al. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol.* 2005;7:1868–82.
- Lipscomb JD. Mechanism of extradiol aromatic ring-cleaving dioxygenases. *Curr Opin Struct Biol.* 2008;18:644–9.
- De Lorenzo V. Systems biology approaches to bioremediation. *Curr Opin Biotechnol.* 2008;19:579–89.
- Pieper DH, Seeger M. Bacterial metabolism of polychlorinated biphenyls. *J Mol Microbiol Biotechnol.* 2008;15:121–38.
- Suenaga H, Ohnuki T, Miyazaki K. Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environ Microbiol.* 2007;9:2289–97.
- Suenaga H, Mizuta S, Miyazaki K. The molecular basis for adaptive evolution in novel extradiol dioxygenases retrieved from the metagenome. *FEMS Microbiol Ecol.* 2009;69:472–80.
- Suenaga H. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ Microbiol.* 2012;14:13–22.
- Top EM, Springael D. The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol.* 2003;14:262–9.
- Vilchez-Vargas R, Junca H, Pieper DH. Metabolic networks, microbial ecology and “omics” technologies: towards understanding in situ biodegradation processes. *Environ Microbiol.* 2010;12: 3089–104.
- Widada J, Nojiri H, Omori T. Recent developments in molecular techniques for identification and monitoring of xenobiotic-degrading bacteria and their catabolic genes in bioremediation. *Appl Microbiol Biotechnol.* 2002;60:45–59.

F

Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences

Weizhong Li
J. Craig Venter Institute, La Jolla, CA, USA

Synonyms

CD-HIT is a fast program for clustering large amount of protein and nucleotide sequences

Definition

Sequence clustering is a process to group sequences into groups (clusters) such that similar sequences are clustered together and can be potentially represented by a single representative sequence. CD-HIT uses a greedy incremental clustering algorithm enhanced by an efficient word filtering heuristics and an effective parallelization technique to do clustering on big sequence datasets efficiently.

Introduction

Since the development of high-throughput sequencing technologies, the amount of available biological sequences has increased dramatically and continues to increase rapidly. Efficient handling and effective analysis of such massive

amount of data has become one of the major issues and challenges in many sequencing-based research. Such challenges are typically dominated by two factors: huge data size and high sequence redundancy. Sequence clustering is a key technique that can address these two issues at once, by clustering the sequences and reducing them to a smaller subset of representative sequences.

Sequence clustering is a technique to group sequences into groups (clusters), such that similar sequences are clustered together and can be potentially represented by a single representative sequence. A sequence similarity between two sequences is normally defined based on an optimal alignment between them. Such optimal alignment is usually found by dynamic programming techniques, which are computationally expensive. Traditional clustering algorithms that require many pairwise sequence comparisons are impractical for clustering very large sequence datasets. Reducing the number of sequence comparisons is the key to efficient sequence clustering that can cope with the massive amount of sequencing data.

Greedy incremental clustering has been employed in sequence clustering to reduce the number of sequence comparisons since the implementation of a tool by Holm and Sander (1998) to create *nrd90* for protein sequences with a decapeptide filter to further reduce the number of comparisons. To overcome some limitations of that tool and further improve the clustering efficiency, CD-HIT was developed to use

the same greedy incremental algorithm, but with a much more efficient filtering heuristics (Li et al. 2001, 2002). CD-HIT was then extended to support clustering of nucleotide sequences (Li and Godzik 2006) and became one of the most widely used programs for sequence clustering due to its efficiency to handle large datasets.

The rapid increasing amount of sequence data demand even more efficient clustering programs and have lead to the development an enhanced version of CD-HIT (Fu et al. 2012), which has been reengineered to support clustering of very large sequence datasets. In this new CD-HIT, a parallelization technique was developed to safely and efficiently parallelize the greedy incremental clustering algorithm. This parallel CD-HIT can achieve very good speedup (quasilinear speedup for up to eight cores) on multicore computers for sequence clustering.

CD-HIT and its derived programs such as CD-HIT-454, CD-HIT-DUP, CD-HIT-LAP, and CD-HIT-OTU have extensive applications in metagenomics field. A summary of these applications is available from a recent review paper (Li et al. 2012).

Methods

CD-HIT uses a greedy incremental clustering algorithm with filtering heuristics based on shared word counting for efficient clustering. It is further enhanced by an effective parallelization technique that can achieve very good speedup on multicore computers.

Greedy Incremental Clustering

A greedy incremental clustering essentially works in the following way. Given a list of DNA or protein sequences, sort them from long to short, and take the first sequence as a cluster representative sequence. Then, for each (query sequence) of the remaining sequences, check if it is similar to any of the existing representative (reference) sequences, if yes, mark the sequence as a redundant sequence, otherwise, add it to the representative sequence list.

Filtering Based on Shared Words

Checking a query sequence against each of the representative sequences is very inefficient, because such checking involves sequence comparison based on sequence alignment using dynamic programming, which is computationally expensive. To reduce such comparisons, a word (k-mer or q-gram) indexing table can be used to filter out unnecessary comparisons based on the number of words shared between the query sequence and each of the representative sequences.

The idea is that, for two sequences to have identity above an identity cutoff, they must share a minimum number of common words given the sequence lengths. It is easy to see that, given two sequences with an alignment length L and an identity cutoff C , the maximum number of mismatches and gaps that are allowed between two aligned sequences is $E = L(1 - C)$, so the minimum number of shared words of length W should be $L + 1 - (E + 1) * W$. This is also the minimum number of shared words between a query sequence of length L and any other longer reference sequences. In CD-HIT, this threshold is adjusted according to the presence of unknown letters such as “N” and “X,” etc., and to the command line options.

To speed up the counting of shared words, an indexing table is built for the representative sequences to record for each word the indices of the representative sequences and the number of occurrences the word appears. This will allow efficient counting of shared words between a query and each of the representative sequences.

Banded Alignment and Sequence Identity

In CD-HIT, sequence identity is computed based on an optimal alignment between two sequences. To reduce the computational time of dynamic programming, CD-HIT uses heuristics based on short words (shorter than the words for filtering) to estimate an optimal band and does banded alignment. Sequence identity is then calculated as the percentage of matched bases among the aligned bases within the whole or best alignment region.

CD-HIT Core Procedures: Checking and Clustering

In order to simplify CD-HIT and make an efficient implementation possible, the key steps of CD-HIT are abstracted into two core procedures: *checking* and *clustering*. The distinction between *checking* procedure and *clustering* procedure is also the key to an efficient parallelization.

Given a word indexing table, the *checking* procedure will check a query sequence against this table and its associated representatives, using the filtering heuristics and sequence comparison techniques described above. If the query is similar to one of the representatives, the query sequence will be marked as redundant and be skipped in all future clustering steps.

The *clustering* procedure is identical to the *checking* procedure except that, if the query is not marked as redundant, it will be added to the representative sequence list of the table, and the table is updated to index and incorporate the words of the new representative sequence.

The Sequential CD-HIT Algorithm

The sequential CD-HIT algorithm is formed by combining the greedy incremental clustering algorithm and the above described heuristics and techniques, with proper dividing of the input sequences. Basically, the steps are the following (Fig. 1):

1. Given a list of DNA or protein sequences (say S), sort them from long to short.
2. Take a sub-list of the longest sequences from S (and remove them from S) and do the *clustering* procedure on them starting from an empty word indexing table.

3. Use the word indexing table from step 2, do the *checking* procedure on the remaining sequences of S , and remove the sequences that are marked as redundant from S .
4. Repeat steps 2 and 3, until S becomes empty.

The Parallel CD-HIT Algorithm

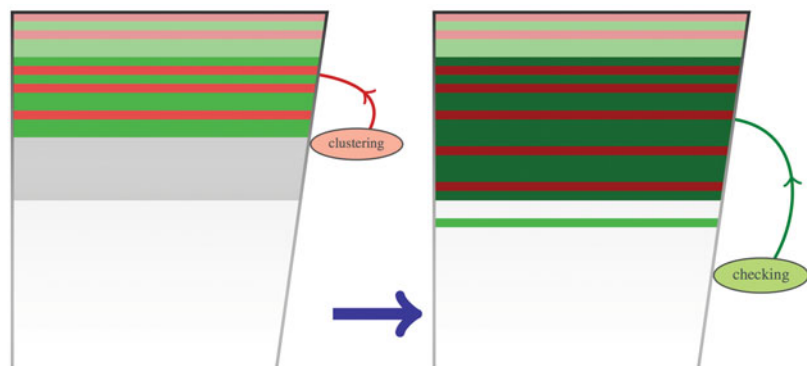
The parallel CD-HIT algorithm uses two word indexing tables to do sequence clustering. Since an efficient parallelization cannot be achieved within each single clustering cycle as described in the sequential algorithm section, the idea of the parallelization technique developed in the parallel CD-HIT is to properly interweave the step 2 and step 3 between two consecutive clustering cycles, as the following (Fig. 2):

1. Given a list of DNA or protein sequences (say S), sort them from long to short.
2. Take a sub-list (say G) of the longest sequences from S (and remove them from S).
3. Use *all* threads to do the *checking* procedure concurrently on G using the word indexing table built by the *clustering* procedure from the previous cycle.
4. Use *all-but-one* threads to do the *checking* procedure on the sequences in S and simultaneously using the remaining *one* thread do the *clustering* procedure on the sequences in G starting from an empty word indexing table.
5. Repeat steps 2, 3, and 4, until S becomes empty.

Here, if the *clustering* procedure finishes processing G before the *checking* procedures finish processing the S , the thread for the *clustering* procedure will switch to do the *checking* procedure on S as well. But if the *checking* procedures

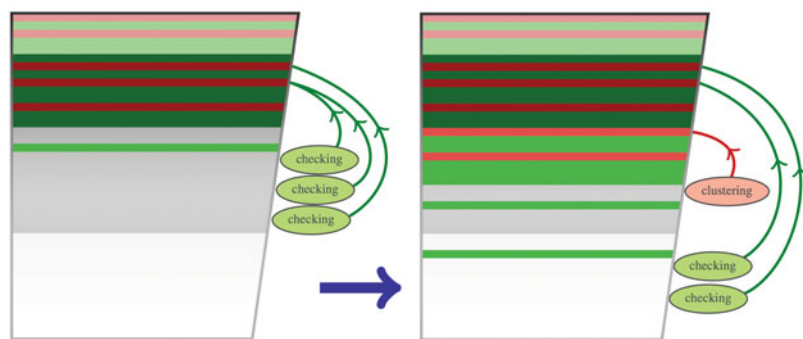
Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences,

Fig. 1 Diagram for the sequential CD-HIT algorithm



Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences,

Fig. 2 Diagram for the parallel CD-HIT algorithm



finish before the *clustering* procedure, the *clustering* procedure will be terminated in order to start a new *clustering* cycle, and the unfinished sequences in G will be put back in S .

In this parallel version of the algorithm, the first and last clustering cycle will effectively use a single thread to do the *clustering* procedure.

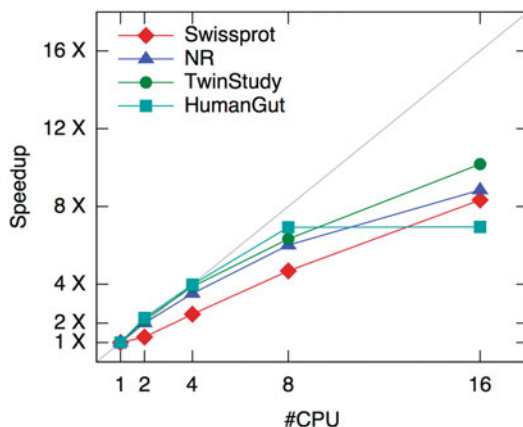
Efficiency of the Parallel CD-HIT

The described parallelization technique is very effective for CD-HIT on large datasets. The main reason is that, in the parallel CD-HIT, all threads are guaranteed to be active simultaneously and do effective computation, and only the first and the last clustering cycle cannot use multithreading. But for large sequence datasets, the time spent on single threaded computation for the first and the last cycle is negligible. So in theory, the speedup should approach linear for large datasets.

Figure 3 shows a benchmarking result on two protein sequence datasets *Swissprot* (437,168 sequences) and *NR* (12,954,819 sequences) and two nucleotide sequence datasets *Twin Study* (8,294,694 sequences) and *Human Gut* (23,285,083 sequences). This test was done on a Debian Linux server with four 12-core *AMD Opteron 6172* processors. As it demonstrated, the parallel CD-HIT can achieve quasilinear speedup for up to eight cores, with good speedup for up to 16 cores.

Summary

CD-HIT is a very fast sequence clustering program that can cluster very big sequence datasets



Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences, Fig. 3 Evaluation of CD-HIT parallelization: computational time speedup with respect to the number of used CPU cores

efficiently. The parallelized version of CD-HIT can further speed up the clustering process by using multiple CPU cores. With the high-throughput sequencing technologies becoming more and more widely used, CD-HIT could play an essential role to facilitate the analysis of the massive amount of sequencing data.

References

- Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*. 2012;28(23):3150–2.
- Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*. 1998;14:423–9.

- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–9.
- Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*. 2001;17:282–3.
- Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 2002;18:77–82.
- Li W, Fu L, Niu B, Wu S, Wooley J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform*. 2012;13:656–68.

Fosmid System

Francisco Rodriguez-Valera
Microbiologia, Universidad Miguel Hernandez,
Campus San Juan, San Juan, Alicante, Spain

Synonyms

BAC; Cosmids; Large insert vectors

Definition

Original molecular cloning vectors were plasmids such as the pBR232, were meant to clone single genes, and were based in multicopy plasmids that have low stringency control of the copy number. Later on with the development of genomics, larger insert vectors were required for the assembly of repeated regions and in general to pair-end the individual shotgun reads. Bacterial artificial chromosomes (BAC) were developed based in the large single-copy plasmids of the F group (Shizuya et al. 1992). These can be propagated in *Escherichia coli* with inserts larger than 300 Kbp. BACs were used by Beja and coworkers in one of the first and more influential papers of the early development of metagenomics in which the existence of an energy-generating rhodopsin was found in a proteobacterial BAC clone (Beja et al. 2000). However, BACs are laborious to generate and do not work well with the limited amount of DNA that normally is available for metagenomics.

In the meanwhile, fosmid (F-based cosmid) was developed. Basically, they contain the replication origin of the *E. coli* F plasmid and can be packaged in a lambda capsid to be transfected rather than transformed. Based loosely on the cosmid vector but adding the F origin of replication, fosmids combine the advantages of BAC vectors (stability and single-copy maintenance) and the easiness of transfection using a cosmid-based vector (Kim et al. 1992). Cosmids have cosN of phage lambda on the vector and use a phage terminase to generate cohesive ends at the cosN. This way, a fosmid insert of 40-kb average size can be cloned very efficiently after packaging in a lambda phage capsid and infected as in conventional cosmid cloning. Extensive libraries of fosmid clones are readily constructed and offer increased insert stability. They can be propagated by standard *E. coli* cultures and the clones isolated as colonies can be collected by an automated colony picking robot. They can be stored as phage suspensions and transferred to the host very efficiently. Also the insert size is very even and can be estimated in the range of between 30 and 40 Kbp in most cases. The *E. coli* F-factor single-copy origin of replication guarantees that there will be only one copy per genome during the cloning phase, avoiding problem with chimera formation during this critical step. However, the inducible high-copy oriV can be used to amplify to up to 50 copies per cell which, while maintaining the stability of the plasmid, increases the DNA yield and the possibilities to be expressed in *E. coli*. For specific protocols of fosmid cloning, see for example: <http://www.epibio.com/item.asp?ID=385>.

Fosmids in Metagenomics

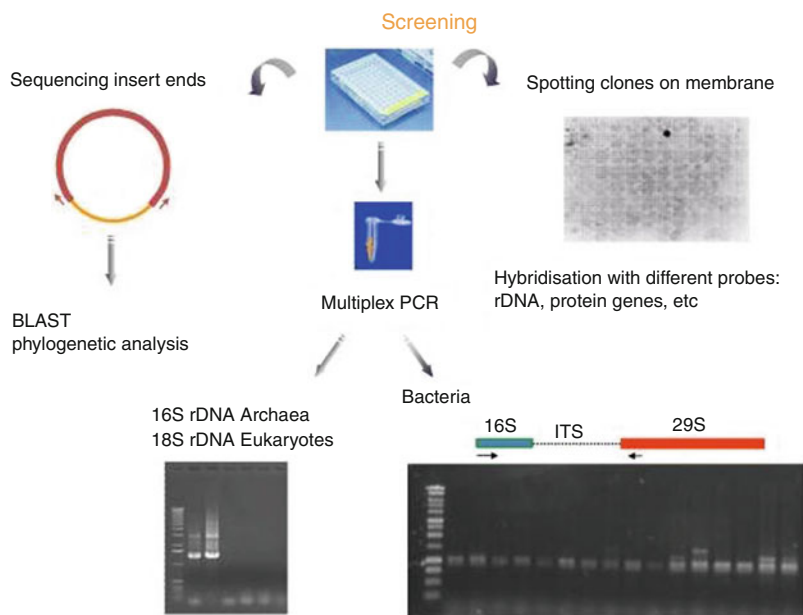
Before the development of second-generation sequencing such as 454 pyrosequencing or Illumina, all metagenomic studies were dependent on cloning of environmental DNA to sequence by Sanger using the vector primers. Small insert vectors have been widespread for the easiness to generate very large libraries and also because the insert that can be sequenced by

Sanger using primer vectors is smaller than the size of most inserts of this size (Venter et al. 2004).

However, large insert and particularly fosmids have been very popular for metagenomic workers (DeLong et al. 2006; Martin-Cuadrado et al. 2007). The main reason is that the insert in a fosmid is a sizeable natural contig that contains typically 30–40 genes. This size is very appropriate for annotation since bacterial and archaeal gene clusters are arranged functionally, i.e., genes with related function, such as different enzymes of a metabolic pathway, are located next to each other, often organized in operons. Therefore, function can be inferred with much more reliability from a large contig. A common approach taken for analysis of fosmid libraries is the fosmid-end sequencing by using the vector primers. This generates datasets that are similar to the short insert (also known sometimes as shotgun) libraries but with the important

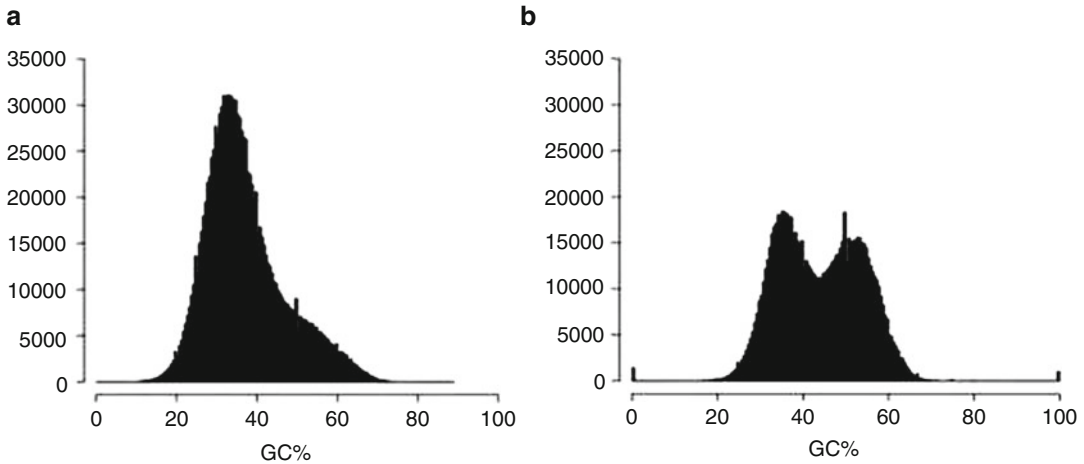
difference that the ends sequenced are separated by a much larger distance. Also, the fosmids that show promise of revealing some interesting activity, or corresponding to an interesting microbe, can be fully sequenced (Fig. 1), traditionally by Sanger dideoxy but now also by high-throughput approaches (Martin-Cuadrado et al. 2009).

Fosmids can also be screened by PCR to select those belonging to selected groups of microbes, largely by using 16S rRNA primers (Martin-Cuadrado et al. 2008). This way, the fosmids containing ribosomal operons can be identified and those containing the target rRNA gene fully sequenced. This approach is a bit tricky when the target group are bacteria because fosmid preparations are always contaminated with *E. coli* DNA and PCR of 16S rRNA gene gives always that amplicon. As an alternative methodology to select bacterial fosmids containing ribosomal operons, primers for 16–23S gene spacer or ITS



Fosmid System, Fig. 1 Methods for selecting fosmids for full sequencing. End sequences can provide clues as to the kind of genes present in the fosmid and allow for selecting those involved in interesting processes or microbes (Martin-Cuadrado et al. 2009). Alternatively, fosmid clones can be screened by PCR or hybridization to select those that contain taxonomically informative

genes such as rRNAs. In the case of bacteria, a strategy to select those containing other rRNAs different from *E. coli* that is present in all the clones is shown. The amplicon includes the internal transcribed spacer (ITS), and the size of this hypervariable region shows the clones containing rRNA genes different from those of *E. coli*



Fosmid System, Fig. 2 Frequency distribution of GC% for the two metagenomic sequence datasets from the Mediterranean water column at the deep chlorophyll maximum (50 m deep). (a) All reads obtained in the DCM

direct 454 pyrosequencing dataset. (b) All reads of the DCM fosmids dataset after removing the vector pCC1fos sequences. GC% of vector pCC1fos = 48 %. For details see Ghai et al. (2010)

were used. The amplicons were run in an agarose gel, and only those with a significantly different size from that of *E. coli* were selected (Quaiser et al. 2008).

With the advent of high-throughput sequencing (HTS), the applications of fosmids are still significant. First of all, they provide a way to assemble much larger contigs, the Achilles' heel of the HTS. Ghai et al. (Ghai et al. 2010) sequenced 1,000 pooled fosmids by 454 pyrosequencing and compared the results with the direct 454 pyrosequencing of the same DNA before cloning. The results indicated a strong bias in the fosmid clones against some specific groups of microbes such as *Candidatus Pelagibacter ubique* and *Prochlorococcus* that happen to be the most abundant microbes in this environment. Besides, the GC distribution plot indicated that high GC of ca. 50 % was enriched versus the reads of the directly sequenced DNA (Fig. 2). The reasons for these biases are obscure, and a similar bias was found for environmental BAC libraries (Feingersch and Beja 2009). However, fosmid cloning provided a complementarity to direct pyrosequencing, providing a way to access microbes that were relatively less abundant in the sample such as marine *Euryarchaea* or

cyanophages in the case of marine samples from the photic zone. Also it provided much larger contigs (up to 44 Kbp and close to 200 contigs over 10 Kbp). The importance of long contigs for interpreting metagenomic datasets cannot be stressed enough since annotation of large clusters of genes is much more reliable (see above). For example, Ghai et al. assemble large fragments of the genomes of marine *Euryarchaea* of group II that later on were instrumental in assembling the complete genome of one of their members from a natural environment (Iverson et al. 2012).

A recent application described for fosmid vectors has been their use for metavirome studies. Metaviromes have a major problem when sequenced by HTS. Viral genes are even more difficult to annotate, and to infer information from their sequence is close to impossible unless large fragments of the viral genome are available. This problem has been solved by fosmid cloning in a pilot study carried out by Garcia-Heredia et al. (2012). These authors have retrieved viral DNA from a natural extreme environment and could reconstruct complete to near-complete viral genomes that prey on microbes which pure culture is very fastidious and hence not adequate for classical phage isolation in pure culture.

Besides, the chances of screening for biological activity are better when using larger inserts, among other things because the complete metabolic pathway might be present, in case more than one gene is needed, and also the genomic context facilitates expression (e.g., better chances of the required promoters and control machinery being present). Many recent examples have used fosmid clones for expression of activities such as enzymes (Selvin et al. 2012) or bioactive compounds (Riaz et al. 2008; Huang et al. 2009; Parsley et al. 2011).

The third generation of high-throughput single-molecule nucleic acid sequencing such as Nanopore or Helicos might generate long reads that, provided they have enough reliability, might make fosmid cloning and sequencing obsolete (Munroe and Harris 2010; Manrao et al. 2012).

Summary

Many authors used the fosmid vectors to describe metagenomes. They allow to generate large libraries with relatively small investment of time and money, and they can be used for multiple purposes. For example, fosmid-end sequencing provides data similar to shotgun libraries (in small insert vectors) but can be screened for sequences of interest for full fosmid sequencing. There are many examples of studies carried out that way. They can be screened by PCR for genes of interest such as 16S rRNA or others. Fosmids are also better vectors for expression screening by biological activity. The advent of high-throughput sequencing technologies provides new opportunities for sequencing and screening fosmids. However, long read single-molecule sequencing might replace the need for fosmid cloning and render this metagenomic approach obsolete.

References

- Beja O, Suzuki MT, et al. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol.* 2000;2(5): 516–29.
- DeLong EF, Preston CM, et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science.* 2006;311(5760):496–503.
- Feingersch R, Beja O. Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries? *ISME J.* 2009;J3(10):1117–9.
- Garcia-Heredia I, Martin-Cuadrado AB, et al. Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. *PLoS One.* 2012;7(3):30.
- Ghai R, Martin-Cuadrado A, et al. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* 2010;9:1154–1166.
- Huang Y, Lai X, et al. Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in *Escherichia coli*. *Mar Biotechnol.* 2009;11(1):124–31.
- Iverson V, Morris RM, et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science.* 2012;335(6068):587–90.
- Kim UJ, Shizuya H, et al. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* 1992;20(5):1083–5.
- Manrao EA, Derrington IM, et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol.* 2012;30(4):349–53.
- Martin-Cuadrado AB, Lopez-Garcia P, et al. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One.* 2007;2(9):e914.
- Martin-Cuadrado AB, Rodriguez-Valera F, et al. Hind-sight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.* 2008;2(8):865–86.
- Martin-Cuadrado AB, Ghai R, et al. CO dehydrogenase genes found in metagenomic fosmid clones from the deep Mediterranean sea. *Appl Environ Microbiol.* 2009;75(23):7436–44.
- Munroe DJ, Harris TJ. Third-generation sequencing fireworks at Marco Island. *Nat Biotechnol.* 2010;28(5): 426–8.
- Parsley LC, Linneman J, et al. Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. *FEMS Microbiol Ecol.* 2011;78(1):176–87.
- Quaiser A, Lopez-Garcia P, et al. Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton. *Environ Microbiol.* 2008;10(10): 2704–17.
- Riaz K, Elmerich C, et al. A metagenomic analysis of soil bacteria extends the diversity of quorum-quenching lactonases. *Environ Microbiol.* 2008;10(3):560–70.
- Selvin J, Kennedy J, et al. Isolation identification and biochemical characterization of a novel halo-tolerant lipase from the metagenome of the marine sponge *Haliclona simulans*. *Microb Cell Fact.* 2012;11(1):72.

- Shizuya H, Birren B, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA*. 1992;89(18):8794–7.
- Venter JC, Remington K, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304(5667):66–74.

FragGeneScan: Predicting Genes in Short and Error-Prone Reads

Yuzhen Ye

Indiana University, School of Informatics and Computing, Bloomington, IN, USA

Definition

Protein-coding genes are functional units in genomes that encode for proteins.

FragGeneScan is a hidden Markov model (HMM)-based predictor of incomplete and complete genes from short reads or complete genomes of prokaryotes.

Introduction

Identification of genes is one of the most important and challenging problems in whole microbial genome sequencing projects (Davidsen et al. 2001; Aziz et al. 2008; Stewart et al. 2009). In metagenomics, gene finding can provide the opportunity to elucidate the activities and interactions of genes within an environmental sample, from which the metabolic and signaling pathways specific to the environment can be reconstructed and identified (Turnbaugh et al. 2009; HMP consortium 2012). Most commonly, genes encoded by metagenomes have been identified by using homology-based methods such as BLASTX (Altschul et al. 1990; Meyer et al. 2008), which however is facing a challenge due to the large amount of sequencing data even with recent developments of faster tools including RAPSearch (Ye et al. 2011; Zhao et al. 2012). Homology searches against known protein databases, however, cannot be used to predict novel genes,

although discovering new genes is one of the most important aspects in metagenomics research. Alternatively, sequence conservation information can be utilized for prediction of novel protein-coding genes (Krause et al. 2006; Yooseph et al. 2008); for example, a Ka/Ks value of ~ 1 for a group of similar sequences indicates that these sequences are under no selective pressure and hence unlikely to code for proteins. This way, novel families that have multiple members in a metagenomic dataset can be identified (Yooseph et al. 2008). The other straightforward solution to novel gene prediction in metagenomics is to use feature-based approaches such as probabilistic models to evaluate the probabilities of open reading frames (ORFs) being protein-coding regions (Noguchi et al. 2006, 2008; Hoff et al. 2009), in a manner similar to conventional gene-finding methods such as Glimmer and GeneMark (Lukashin and Borodovsky 1998; Salzberg et al. 1998; Delcher et al. 1999).

Short read length and sequencing errors are two major issues that pose significant challenges to gene prediction: incomplete genes (gene fragments) are difficult to predict, and sequencing errors may cause frameshifts that further complicate gene prediction. The average length of genes in microorganisms is about 950 bps (Noguchi et al. 2006), which is much longer than the sequencing reads generated by most NGS (Morozova et al. 2009; Metzker 2010; Quail et al. 2012). Different NGS methods now produce sequencing reads of various lengths ranging from 100 bps (from Illumina sequencers) to thousands of base pairs (PacBio sequencing) and have different error profiles (Morozova et al. 2009). Sanger sequencers produce reads with an error rate of up to 1 %, whereas 454 sequencers produce reads with an error rate of up to 3 % (Richter et al. 2008; Hoff 2009). Illumina sequencing technology may produce reads that have high mismatch rates, especially when relatively long reads are acquired (e.g., G is mistaken as T, and in later cycles A, C, and G are mistaken as T) (Kircher et al. 2009). In 454 sequencing reads, sequencing errors tend to occur in the homopolymer regions, resulting in frequent insertions and deletions. Most of the sequencing errors in

PacBio reads are also indels (Carneiro et al. 2012). It has been shown that ORF-based gene prediction methods are more substantially affected by sequencing errors (indels) that cause frameshifts (Hoff 2009; Tang et al. 2013). As a consequence, programs that are currently available for gene prediction from short reads show a significant decrease in their performance as the sequencing error rate increases. For example, a low sensitivity of 26–43 % was observed with sequencing error rate of 2.8 % (Hoff 2009).

FragGeneScan Algorithm

The core of FragGeneScan (Rho et al. 2010) is a hidden Markov model (HMM) (Rabiner 1989), which incorporates codon usage bias, sequencing error models, and start/stop codon patterns in a unified model. FragGeneScan HMM consists of two-level representations based on data abstraction. FragGeneScan considers separate states representing the gene regions in the forward strand and the reverse strand of a nucleotide sequence, such that it can predict genes simultaneously from both strands. The model has seven superstates, representing gene regions, start codons and stop codons in both the forward (three states) and backward strands (three states), and noncoding regions (one state), respectively. The states for gene regions consist of six consecutive sets of a match state, an insertion state, and a deletion state, which collectively correspond to a six-periodic inhomogeneous HMM. Each match state in the gene regions uses a second-order Markov chain to model the codon usage. The state for noncoding regions is based on a first-order Markov chain. FragGeneScan also incorporates the sequence patterns for each start codon (ATG, GTG, and TTG) and stop codon (TAA, TAG, and TGA) in the start and stop state, respectively.

FragGeneScan HMM has a unique feature. By allowing transitions between the insertion/deletion states and the match states, this model effectively detects frameshifts that are caused by indel errors in sequencing. Considering that complete genomic sequences are unlikely to contain indel errors, the

transition probabilities to insertion and deletion states are set to 0 when applying FragGeneScan to gene prediction in complete genomic sequences.

Given a short read (or a complete genome), the gene prediction problem is to find the best path of hidden states (see below) that is most likely to generate the observed nucleotide sequence, which can be solved by the Viterbi algorithm. FragGeneScan reports genes if they meet the following three conditions: (1) the length of the genes is longer than 60 bps, (2) the genes start in a start state (start codon) or in a match state (internal region of genes), and (3) the genes end in a stop state (stop codon) or in a match state (internal region of genes). Therefore, FragGeneScan can predict complete genes as well as partial (fragmented) genes without start and/or stop codons. Since the probability of gene regions and noncoding regions is calculated solely based on the composition of sequences (which is consistent regardless of the read length and gene length), FragGeneScan is more robust when input sequences are of different lengths.

Applications of FragGeneScan

FragGeneScan software is available as open source on <http://omics.informatics.indiana.edu/FragGeneScan>. It has been incorporated into several metagenomic analysis pipelines, including MG-RAST (<http://press.igsb.anl.gov/mgrdev/under-the-hood/mg-rast-tools/fraggenescan/>), IMG/M (Markowitz et al. 2012), WebMGA (Wu et al. 2011), and EBI metagenomics service (Wu et al. 2011).

Summary

Gene prediction in short reads (and assemblies) will remain a challenging problem, even with recent advances in the field (Tang et al. 2013). Proteins predicted from environmental sequences have already greatly expanded the universe of protein sequences. Not surprisingly, an increasingly large number of these proteins we are getting are hypothetical proteins. Functional

prediction of these hypothetical proteins will play a key role in elucidating their functions, which however, will be an even more daunting task.

References

- Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- Aziz R, Bartels D, Best A, et al. The RAST server: rapid annotations using subsystems technology. *BMC Genomics.* 2008;9(1):75.
- Carneiro MO, Russ C, Ross MG, et al. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics.* 2012; 13:375.
- Davidson T, Beck E, Ganapathy A, et al. The comprehensive microbial resource. *Nucleic Acids Res.* 2001;38 Suppl 1:D340–5.
- Delcher AL, Harmon D, Kasif S, et al. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 1999;27:4636–41.
- HMP consortium. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402): 207–14.
- Hoff K. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics.* 2009;10(1):520.
- Hoff KJ, Lingner T, Meinicke P, et al. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 2009;37:W101–5.
- Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* 2009;10(8):R83.
- Krause L, Diaz NN, Bartels D, et al. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics.* 2006;22:e281–9.
- Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998;26: 1107–15.
- Markowitz VM, Chen IM, Chu K, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* 2012;40-(Database issue):D123–9.
- Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet.* 2010;11(1):31–46.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9(1):386.
- Morozova O, Hirst M, Marra M. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet.* 2009;10:135–51.
- Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 2006;34(19):5623–30.
- Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 2008;15: 387–96.
- Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 2012;13:341.
- Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE.* 1989;77:257–86.
- Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010;38(20):e191.
- Richter DC, Ott F, Auch AF, et al. MetaSim – a sequencing simulator for genomics and metagenomics. *PLoS ONE.* 2008;3:e3373.
- Salzberg SL, Delcher AL, Kasif S, et al. Microbial gene identification using interpolated Markov models. *Nucleic Acid Res.* 1998;26:544–8.
- Stewart AC, Osborne B, Read TD. DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics.* 2009;25(7):962–3.
- Tang S, Antonov I, Borodovsky M. MetaGeneTack: ab initio detection of frameshifts in metagenomic sequences. *Bioinformatics.* 2013;29(1):114–6.
- Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature.* 2009;457(7228):480–4.
- Wu S, Zhu Z, Fu L, et al. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics.* 2011;12:444.
- Ye Y, Choi JH, Tang H. RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinforma.* 2011;12:159.
- Yooseph S, Li W, Sutton G. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinforma.* 2008;9:182.
- Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics.* 2012; 28(1):125–6.

FR-HIT Overview

Beifang Niu, Zhengwei Zhu, Limin Fu and Sitao Wu
Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, CA, USA

Definition

A crucial step in metagenomic data analysis is fragment recruitment, a process of aligning

sequencing reads to reference genomes. FR-HIT offers high speed and high sensitivity in recruiting large-scale metagenomic reads.

Introduction

Microbiome data are directly obtained from various environments and contain genomics information of many known and novel microorganisms. An important step to study these organisms' identity and abundance is to align the sequencing reads against the available reference genomes. This process was called fragment recruitment in the Global Ocean Sampling (GOS) project that surveyed the world's oceans (Rusch et al. 2007).

A metagenomic dataset may have many novel species without available reference genomes. Even if references are available, the microbial species may undergo large variations. So a fragment recruitment method needs to find all significant alignments with arbitrary number of mismatches and gaps.

There are many available alignment programs that can be considered for fragment recruitment. In terms of accuracy, BLAST is the best tool because it can identify very remote homology so it was used in earlier studies such as GOS. But it is too slow for computing reads from the next-generation sequencing (NGS) platforms. The new generation of mapping programs, such as SOAP (Li et al. 2008), Bowtie (Langmead et al. 2009), BWA (Li and Durbin 2009), and many others, are orders of magnitude faster than BLAST. However, these mapping programs only tolerate a few mismatches so they are not suitable for recruiting metagenomic reads.

FR-HIT is a very fast program to recruit metagenomic reads to homologous reference genomes (Niu et al. 2011). It offers both high speed and high sensitivity in recruiting NGS reads. A C++ implementation of FR-HIT and more details of this method are available at <http://weizhongli-lab.org/frhit>.

Methods

FR-HIT adopts a seeding strategy with overlapping q -gram hashing to locate candidate matching blocks on the reference sequences and then applies an effective filtering within the candidate blocks to filter out blocks that do not meet the minimum criteria for containing an alignment with specified parameters. For each candidate block that passed the filter, the best matching subregions between a candidate block and a read are determined and used subsequently by the banded Smith-Waterman algorithm to carry out the actual alignment efficiently, which will finally verify if this can be a valid recruitment hit.

Constructing Overlapping Q-Gram

Hash Table

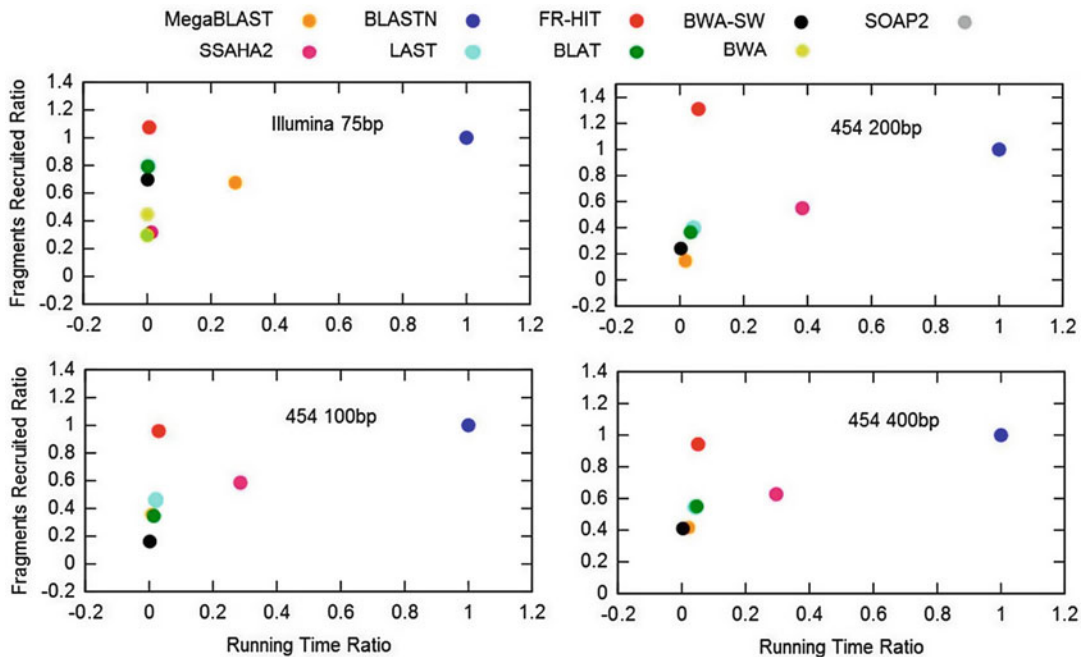
All reference sequences are stored together with a hash lookup table to rapidly locate q character overlapping q -gram. The overlapping q -grams are sampled at equidistant steps along the reference sequences. A reference of length m contains $(m - q)/(q - p) + 1$ q -grams with an overlap of p bases. Here q and p are user-adjustable parameters.

Identifying Candidate Matching Block

The candidate blocks are fragments on reference sequences that will be further considered for alignment with the query. For each query, all its overlapping q -grams are used to scan the q -gram hash table and collect the q -grams shared by reference sequences. Candidate blocks are derived from clusters of pieces on reference genomes marked by the shared q -grams.

Q-Gram Filtering and Banded Alignment

Q -gram filtering strategy was used before in QUASAR (Burkhardt et al. 1999) based on the q -gram lemma (Jokinen and Ukkonen 1991; Owolabi and McGregor 1988), which states two sequences of length n with Hamming distance ϵ share at least $n + 1 - (\epsilon + 1)q$ common q -grams. FR-HIT calculates the maximal number of mismatches according to user-specified alignment cutoff value and rejects the candidate blocks



FR-HIT Overview, Fig. 1 Recruitment rate and speed of FR-HIT and other programs for four datasets. The x-axis is the ratio of CPU time relative to BLASTN; y-axis is the ratio of number of recruited reads relative to BLASTN

that do not have enough common q -grams. In this step, the length of q -gram is 4. After filtering, banded alignments between the query and the candidate blocks that passed the filter are performed.

On average, FR-HIT is ~ 2 orders of magnitude faster than BLASTN with similar recruitment rate. FR-HIT is slower than the mapping programs SOAP2, BWA, and BWA-SW, but it recruits several times more reads.

Performance of FR-HIT

The fragment recruitment performance of FR-HIT was compared to some widely used short-read mapping and sequence alignment tools including BLASTN, MegaBLAST, SOAP2, BWA, BWA-SW, SSAHA2, BLAT, and LAST using four metagenomic datasets of up to one million reads covering 454 GS20, 454 GSFLX, 454 Titanium, and Illumina platforms. Reads are aligned to available microbial reference genomes and considered recruited if the alignments are at least 30 bp and at least 80 % identity.

The overall comparison of CPU time and the number of recruited reads are shown in Fig. 1.

Fragment Recruitment Viewer

The results of alignments from FR-HIT can be interactively visualized using Fragment Recruitment Viewer, a tool that plots the alignments on a 2D map where the x-axis is the genome coordinate and y-axis is the alignment identity (Fig. 2). The map can be operated like a Google Map so that users can explore the recruitment alignments from one or multiple samples to many reference genomes. Fragment Recruitment Viewer is available from <http://weizhongli-lab.org/mgaviewer>. Some pre-calculated recruitment results using FR-HIT are available from the CAMERA project (<http://camera.calit2.net>).



FR-HIT Overview, Fig. 2 Screenshots of the Fragment Recruitment Viewer. The initial view of plot shows all hits to the full reference genome. X-axis is the genome coordinate, and y-axis is the alignment identity. Hits are colored by samples. The *bottom* of the plot shows genes of the reference genome colored by gene type (protein, rRNA,

and tRNA). At *right bottom corner*, there are a few icons to zoom in, zoom out, increase and decrease plot size, and reset to the default view. Mouse wheel can be used to zoom the plot. Plot can be panned using mouse. Information of an alignment or a gene is displayed when the pointer is over it

Summary

FR-HIT is an important tool to perform fragment recruitment analysis for metagenomic sequences. The recruitment results can be visualized using the fragment recruitment reviewer. They can also be analyzed to provide taxonomy and function annotations. As a fast alignment tool, FR-HIT can also be used for many applications such as filtering out human contaminations for human microbiome samples.

References

- Burkhardt S, Cramer A, Ferragina P. q-gram based database searching using a suffix array (QUASAR). RECOMB '99; 1999 Apr 11–14; Lyon; 1999, pp. 77–83.
- Jokinen P, Ukkonen E. 2 algorithms for approximate string matching in static texts. In: Tarlecki A, editor. Mathematical foundations of computer science. Lecture notes in computer science, vol 520. Berlin: Springer; 1991, pp. 240–248.
- Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
- Li R, Li Y, Kristiansen K, et al. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24:713–4.
- Niu B, Zhu Z, Fu L, et al. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*. 2011;27:1704–5.
- Owolabi O, Mcgregor DR. Fast approximate string matching. *Softw Pract Exp*. 1988;18:387–93.
- Rusch DB, Halpern AL, Sutton G, et al. The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*. 2007;5:e77.

Functional Metagenomics of Bacterial-Cell Crosstalk

Tomas de Wouters^{1,3}, Nicolas Lapaque¹,
Emmanuelle Maguin¹, Joël Doré^{1,2,3} and
Hervé M. Blottière^{1,2}

¹INRA, AgroParisTech, Jouy en Josas, France

²US 1367 MetaGenoPolis, INRA, Jouy en Josas, France

³UMR Micalis, AgroParisTech, Jouy en Josas, France

Synonyms

Host-microbiota interactions

Definition

Functional analysis of a metagenome (combined genomes of a defined system) with the aim to understand and/or identify single components of the interaction of a microbe with specific cells.

Introduction

Complex ecosystems often exert several niche-specific functions. Dependent on their entanglement and the accessibility of the ecosystem, the identification and analysis of these single functions can be challenging.

Cultivability and metabolic interdependence of microbes in their ecosystems have confronted microbial ecologists with “the great plate-count anomaly” (Staley and Konopka 1985) since the beginning of their studies. The term summarizes the great discrepancy between the loads of microscopically observed bacteria in an environmental sample and the lower numbers obtained using culture-dependent counting techniques, indicating the lack of representativeness of culture-dependent techniques in the study of most complex bacterial ecosystem.

The development of molecular cloning approaches led microbial ecologists to explore the enzymatic potential of their ecosystems by heterologous expression. They developed techniques to extract total genomic DNA of bacterial origin from complex environmental samples. These metagenomes can subsequently be expressed in a well-known and cultivable host using fosmids, cosmids, or bacterial artificial chromosomes (BACs). The first application of this technique allowed the identification of formerly unknown fibrolytic enzymes from, among others, anaerobic and Gram-positive bacteria (Healy et al. 1995) using *E. coli* (a Gram-negative bacterium) as a host. The use of heterologous expression of the metagenome of an ecosystem to identify functionalities of uncultivable bacteria was later coined “functional metagenomics” as opposed to the use of molecular techniques for phylogenetic characterization and *in silico* functional predictions of microbial ecosystems called metagenomics.

Human Metagenomics

With the discovery of the importance of the human microbiota for human health, the study of the different ecological niches of the human body gained a lot of attention in the late 1990s. All over the human body to date, five principal niches were addressed: the skin, nasal, oral, urogenital, and gastrointestinal microbiota (Huttenhower et al. 2012). Based on the rapid development of the next-generation sequencing technologies, these complex ecosystems have

been explored mainly through metagenomic studies of their phylogenetic composition and their metabolic repertoire as far as in silico prediction is possible.

Most attention has been focused on the intestinal microbiota. Not only because of its unique bacterial density but also because of the large mucosal interface that exposes the human body to this bacterial load. The study of germ-free animals and large human cohorts revealed correlations between the composition of the intestinal microbiota and physiological conditions of the host, such as the proper development of immunity, a balanced metabolism, and the systemic inflammatory status (Cerf-Bensussan and Gaboriau-Routhiau 2010). This systemic impact indicates an interaction between the intestinal microbiota and the host that has since been subject to intensive scientific research.

Functional Studies of the Intestinal Microbiota

The human intestinal microbiota harbors a genetic repertoire >25 times larger than that of each human host (Qin et al. 2010) encoding a multitude of functions that contribute directly or indirectly to host's physiology. Cultivation efforts as compared to molecular techniques revealed that 70–80 % of the dominant bacteria are not yet cultured. Therefore up to 80 % of the intestinal microbes have no representative in any bacterial strain collection for potential functional studies (Suau et al. 1999; Hayashi et al. 2002).

Functional studies of intestinal bacteria have therefore long been limited to the study of

cultivable bacteria or the study of monoxenic and gnotobiotic animal models. In order to circumvent this limitation, culture-independent methods such as functional metagenomics have been adapted and used to study functions of the human intestinal microbiota (Table 1). Initially the approach was used to search for enzymatic activities specific for intestinal metabolic functions.

Using a BAC library prepared in an *E. coli* host, Walter and colleagues screened a mouse intestinal metagenome for β -glucanase activity identifying 3 out of a total 5,760 clones (containing a total of 320 Mb of genomic DNA, each clone bearing on average 55 Kb) encoding enzymes of interest (Walter et al. 2005). Similarly, by screening a small fragment metagenomic library (14,000 clones, representing 77 Mb of genomic DNA, cloned DNA fragments had sizes of up to 8 kb) derived from a cow rumen content, Ferrer and colleagues identified and characterized 22 clones with distinct hydrolytic activities (Ferrer et al. 2005). In these two studies, the screening process only allowed a very limited coverage of the actual metagenome due to the size of the library. Although several studies have identified hydrolytic enzymes using plasmid libraries, one of the key issues of the functional approach is to obtain libraries bearing large fragments of DNA to have access to full operons and operational gene clusters, i.e., from 10 to 50 Kb.

Indeed, Jones and colleagues developed a more promising approach by screening about 90,000 metagenomic fosmid clones derived from a human fecal sample (representing a total of about 3.6 Gb bacterial DNA which is about one

Functional Metagenomics of Bacterial-Cell Crosstalk, Table 1 Reported functional metagenomic screenings of the human intestinal microbiota

	Target	n° of clones tested	Hit rate (%)	Reference
Enzymatic activity	Bile salt hydrolases	89,856	1×10^{-3}	(Jones et al. 2008)
	Carbohydrate-active enzymes	156,000	2×10^{-3}	(Tasse et al. 2010)
	β -Glucuronidase	4,608	1.79	(Gloux et al. 2011)
Host –Microbe interaction	Cell proliferation	20,725	4×10^{-2}	(Gloux et al. 2007)
	NF- κ B activation	2,640	6×10^{-2}	(Lakhdari et al. 2010)

equivalent of the dominant intestinal metagenome) for bile salt hydrolase activity (Jones et al. 2008). They observed that these functions were present and enriched in all major gut bacterial divisions including Archaea, demonstrating the powerful capacities for discovery of the functional metagenomic approach. In the same way, Tasse and colleagues applied high throughput functional screenings to search human gut-derived metagenomics clones (156,000 clones representing 5.5 Gb of DNA) for their capacity to hydrolyze different polysaccharides (Tasse et al. 2010). This exhaustive analysis of carbohydrate-active enzymes allowed the identification of highly prevalent genes encoding enzymes that are involved in the catabolism of dietary fibers in the human intestinal tract, demonstrating again the strategic interest of the functional metagenomic approach.

Functional Metagenomics and Host-Microbiota Interaction

The intestinal microbiota had successfully been screened for its enzymatic activities with the help of metagenomic libraries using fosmids, cosmids, or bacterial artificial chromosomes (BAC) with single, low copy, or copy control vectors. Thus Gloux and colleagues set out to test if these metagenomic libraries were suited for the study of bacteria-host cell interactions at the intestinal interface, targeting the intestinal epithelial cells. They therefore screened a library of over 20,000 clones for their influence on proliferation of HT-29 human intestinal epithelial cells and CV1 kidney fibroblast showing that indeed this approach could reveal genes of interest in the dialogue between the host and its microbiota (Gloux et al. 2007).

The same group further developed this approach performing the screening of over 2,500 clones on human colorectal carcinoma cell lines, namely, Caco-2 and HT-29, which were stably transfected with NF- κ B-dependent

reporter genes (Lakhdari et al. 2010). NF- κ B is a key transcription factor in intestinal epithelial cells controlling, among others, the inflammatory response. This unique combination of reporter cell technology and functional metagenomics established a new approach to identify specific regulatory elements of the intestinal microbiota in the complex interactions between the intestinal microbiota and its host. They further identified the genes implicated in the observed effect using random transposition on the bioactive clones, showing that this approach can be used to identify genes involved in bacteria-cell crosstalk at the level of intestinal epithelium.

In order to reach a reasonable level of coverage of the metagenomic samples, this approach has been automated and in parallel screens have been developed for other transcription factors (AP1, PPAR γ ...) or target genes (*ANGPTL4*, *TSLP*, *TGF β* ...) in order to allow the high throughput application necessary to identify bioactive compounds of the intestinal microbiota (www.mgpps.eu). The identification of these bioactive clones and the corresponding genes, molecules, and mode of action will help to untangle the complex interactions of the intestinal microbiota with its host.

Functional metagenomics can also be applied to identify indirect interactions of the intestinal microbiota with its host. Gloux and colleagues identified β -glucuronidases using a functional metagenomic screen on libraries derived from intestinal samples from healthy individuals and Crohn's disease patients (Gloux et al. 2011). The study revealed the presence of a new class of β -glucuronidase that seems to be gut-specific and is hypothesized to play a role in the metabolism of xenobiotics. On this background, functional metagenomics in the human intestine could be a powerful tool to identify specific biodegradation or conversions observed in the intestine that can have physiological effects and for which the dominant causal agent is often unknown.

Up to now all reported functional metagenomic studies of host microbe interactions published were performed using *E. coli* as a host

strain. Since the Gram + bacteria represent a large part of the intestinal microbiota and most of the probiotic bacteria described to have beneficial effects on human health are Gram+, great efforts have been made to develop easy cloning tools for such studies in Gram + hosts. Since the expression of heterologous genes in *E. coli* gave access to around 40 % of the genes for both Gram + and Gram- bacteria (Gabor et al. 2004), it makes it a suitable but not universal host. The utility of a Gram + bacterial host is based on eventual potential preference for RBSs and hence increased transcription but also on secretion of proteins through Gram + specific signal peptides or eventual surface exposure of bioactive proteins through cell wall anchoring motifs. Screenings of metagenomic libraries in *Streptomyces spp.* (Wang et al. 2000) and even Archaea (Albers et al. 2006) have successfully been performed for other ecosystems. Efforts for targeted expression of candidate proteins of the human intestinal microbiota have been made by developing prediction tools for surface-exposed and secreted proteins in Gram + hosts in order to mine the abundantly available metagenomic data (Barinov et al. 2009). The expression of the identified candidate genes in a Gram + host such as *Bacillus subtilis* or *Lactococcus lactis* will allow functional screening in cell-based assays. Though

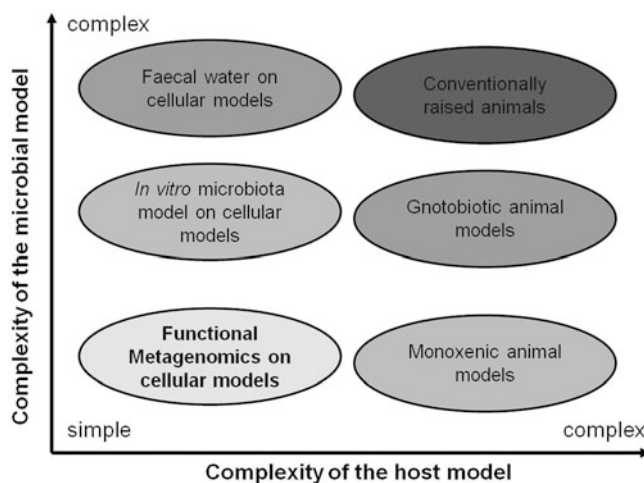
such tools have been used for functional screens of pathogen-cell interaction, a functional metagenomic study of interactions between commensal bacteria and host cells using a Gram + host has not been published yet.

Summary

Metagenomic studies are applied to complex systems. Functional metagenomics is no exception. If we study a complex system, simplification can bring clarity. This is the case if we search for specific enzymatic activities in a complex ecosystem. Simultaneously, simplification harbors the danger of oversimplification and therefore error or deception.

The authors consider functional metagenomics as a very useful and powerful tool to screen complex ecosystems for specific functions and believe it can be extended to the study of host-microbiota interactions as performed in the studies mentioned above. For a full understanding of the complex interaction of a microbiome with its cellular counterpart, this is however only an exploratory tool that will always require validation in a more holistic and thus more complex model (Fig. 1).

Functional Metagenomics of Bacterial-Cell Crosstalk, Fig. 1 Possible models to study host-microbiota interactions ordered by complexity of the microbial (*ordinate*) and cellular model (*abscise*) toward the understanding of human intestinal physiology



Cross-References

- ▶ [Functional Metagenomics of Human Intestinal Microbiome \$\beta\$ -Glucuronidase Activity](#)
- ▶ [Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing](#)
- ▶ [Use of Bacterial Artificial Chromosomes in Metagenomics Studies, Overview](#)

References

- Albers S-V, Jonuscheit M, Dinkelaker S, Ulrich T, Kletzin A, Tampé R, et al. Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Appl Environ Microbiol* [Internet]. 2006 [cited 2011 Aug 21];72(1):102–11. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1352248&tool=pmcentrez&rendertype=abstract>
- Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, et al. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics* [Internet]. 2009 [cited 2012 Sep 5];9(1):61–73. Available from <http://www.ncbi.nlm.nih.gov/pubmed/19053137>
- Cerf-Bensussan N, Gaboriau-Routhiau V. The immune system and the gut microbiota: friends or foes? *Nature Rev Immunol* [Internet]. Nature Publishing Group; 2010 [cited 2011 Jul 20];10(10):735–44. Available from <http://www.ncbi.nlm.nih.gov/pubmed/20865020>
- Ferrer M, Golyshina OV, Chernikova TN, Khachane AN, Reyes-Duarte D, Santos V a PM Dos, et al. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ Microbiol* [Internet]. 2005 [cited 2013 Jan 28];7(12):1996–2010. Available from <http://www.ncbi.nlm.nih.gov/pubmed/16309396>
- Gabor EM, Alkema WBL, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* [Internet]. 2004 [cited 2011 Jun 22];6(9):879–86. Available from <http://www.ncbi.nlm.nih.gov/pubmed/15305913>
- Gloux K, Berteau O, El Oumami H, Béguet F, Leclerc M, Doré J. A metagenomic β -glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc Natl Acad Sci U S A* [Internet]. 2011 [cited 2011 Jul 29];108(Suppl):4539–46. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3063586&tool=pmcentrez&rendertype=abstract>
- Gloux K, Leclerc M, Iliozier H, L'Haridon R, Manichanh C, Corthier G, et al. Development of high-throughput phenotyping of metagenomic clones from the human gut microbiome for modulation of eukaryotic cell growth. *Appl Environ Microbiol* [Internet]. 2007 [cited 2011 Jun 22];73(11):3734–7. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1932692&tool=pmcentrez&rendertype=abstract>
- Hayashi H, Sakamoto M, Benno Y. Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiol Immunol* [Internet]. 2002 [cited 2011 Apr 21];46(8):535–48. Available from <http://www.ncbi.nlm.nih.gov/pubmed/12363017>
- Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, Shanmugam KT. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl Microbiol Biotechnol* [Internet]. 1995 [cited 2011 Aug 17];43(4):667–74. Available from <http://www.ncbi.nlm.nih.gov/pubmed/7546604>
- Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. *Nature* [Internet]. Nature Publishing Group; 2012 [cited 2012 Jun 13];486(7402):207–14. Available from <http://www.nature.com/doi/10.1038/nature11234>
- Jones BV, Begley M, Hill C, Gahan CGM, Marchesi JR. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc Natl Acad Sci U S Am* [Internet]. 2008 [cited 2011 Aug 20];105(36):13580–5. Available from <http://www.pnas.org/cgi/content/abstract/105/36/13580>
- Lakhdari O, Cultrone A, Tap J, Gloux K, Bernard F, Ehrlich SD, et al. Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF- κ B modulation in the human gut. Sturtevant J, editor. *PLoS ONE* [Internet]. 2010 [cited 2010 Oct 1];5(9):e13092. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2948039&tool=pmcentrez&rendertype=abstract>
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* [Internet]. 2010;464(7285):59–65. Available from <http://www.ncbi.nlm.nih.gov/pubmed/20203603>
- Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Ann Rev Microbiol* [Internet]. 1985 [cited 2011 Aug 13];39:321–46. Available from <http://www.ncbi.nlm.nih.gov/pubmed/3904603>
- Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, et al. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* [Internet]. 1999;65(11):4799–807. Available from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=91647&tool=pmcentrez&rendertype=abstract>

- Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, et al. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res* [Internet]. 2010 [cited 2010 Sep 18];20(11):1605–12. Available from <http://www.ncbi.nlm.nih.gov/pubmed/20841432>
- Walter J, Mangold M, Tannock GW, Icrobiol APPLN-M. Construction, analysis, and beta-glucanase screening of a bacterial artificial chromosome library from the large-bowel microbiota of mice. *Appl Environ Microbiol*. 2005;71(5):2347–54.
- Wang GY, Graziani E, Waters B, Pan W, Li X, McDermott J, et al. Novel natural products from soil DNA libraries in a streptomycete host. *Organ Lett* [Internet]. 2000 [cited 2011 Aug 21];2(16):2401–4. Available from <http://www.ncbi.nlm.nih.gov/pubmed/10956506>

Functional Metagenomics of Human Intestinal Microbiome β -Glucuronidase Activity

Petra Louis¹ and Joël Doré^{2,3,4}

¹Rowett Institute of Nutrition and Health, Microbiology Group, Gut Health Programme, University of Aberdeen, Aberdeen, UK

²INRA, AgroParisTech, Jouy en Josas, France

³US 1367 MetaGenoPolis, INRA, Jouy en Josas, France

⁴UMR Micalis, AgroParisTech, Jouy en Josas, France

Definitions

β -glucuronidases: Enzymes belonging to glycoside hydrolase family 2 that catalyze the cleavage of β -D-glucuronic acid residues from a range of different compounds.

Functional metagenomics: Screening of metagenomic DNA cloned into heterologous hosts for the expression of specific functions.

Sequence-based metagenomics/metagenomic sequence mining: In silico analysis of metagenomic sequence libraries for the presence of genes with sequence similarity to known genes.

Degenerate PCR: Usage of a mixture of similar PCR primers designed to amplify the same gene from different organisms, by targeting highly conserved gene regions.

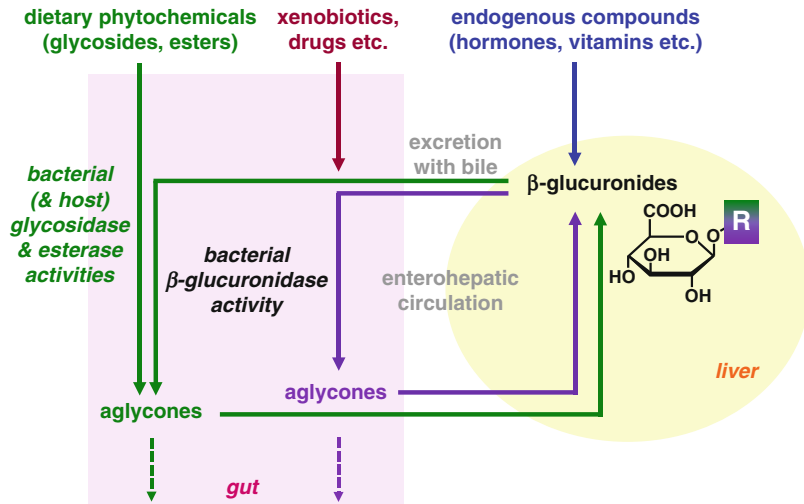
Introduction

Intestinal β -glucuronidases (EC 3.2.1.31) are among the major enzyme families associated with chemical detoxification (Fig. 1). They catalyze the hydrolysis of β -glucuronides naturally present in the human diet, in drugs, or those produced in the liver by glucuronidation via UDP-glucuronosyltransferases (EC 2.4.1.17), which is the major conjugation process in mammals (Tukey and Strassburg 2000; Haiser and Turnbaugh 2013). Numerous lipophilic compounds including metabolic wastes, vitamins, steroid hormones, plant- and animal-derived secondary metabolites, xenobiotics, and pharmaceuticals are thus converted to water-soluble compounds, allowing excretion via the bile and the digestive tract. The β -glucuronidase activity on glucuronide compounds in the gut lumen is primarily due to intestinal bacteria (Rod et al. 1977). This activity regenerates aglycone insoluble forms that are frequently reabsorbed by the host through the enterohepatic circulation, thus increasing circulating aglycone concentrations and extending body exposure. The presence of circulating hormones and xenobiotics is substantially due to this phenomenon and linked to bacterial β -glucuronidase activity. With regard to toxic aglycones, the bacterial activity is largely used as a marker of the potentially harmful effects of commensal bacteria, particularly in studies relating to colorectal cancer (McBain and Macfarlane 1998). β -glucuronidase activity can also lead to increased toxicity of chemotherapeutics; such toxic effects were reduced by coadministration of a β -glucuronidase inhibitor in an animal model (Haiser and Turnbaugh 2013). In contrast, it is also involved in the beneficial bioconversion of dietary compounds including lignans, flavonoids, sphingolipids, glycyrrhizin, or baicalein (Kim et al. 1998, 2000; Schmelz et al. 1999).

β -glucuronidase activity is phylogenetically widely distributed among the microbiota and is present in numerous genera including *Bacteroides*, *Clostridium*, *Eubacterium*, *Lactobacillus*, *Ruminococcus*, *Faecalibacterium*, *Roseburia*, *Streptococcus*, *Peptostreptococcus*,

Functional Metagenomics of Human Intestinal Microbiome β -Glucuronidase Activity, Fig. 1

Role of gut bacterial β -glucuronidase activity in the detoxification of dietary glucuronides, xenobiotics, drugs, and endogenous compounds



Enterococcus, *Bacillus*, *Staphylococcus*, *Corynebacterium*, *Acinetobacter*, *Catenabacterium*, and *Propionibacterium* (Beaud et al. 2005; Dabek et al. 2008; Russell and Klaenhammer 2001; McBain and Macfarlane 1998). However, because of the difficulty differentiating β -glucuronidase from β -galactosidase genes, very few corresponding protein or gene sequences have been clearly annotated as β -glucuronidase in the databases. The β -glucuronidase genes annotated in NCBI primarily are associated with the four major bacterial phyla present in the digestive tract: Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria.

Taking into account the great diversity of glucuronides likely present in the digestive tract, the question of the diversity and specificities of β -glucuronidases is crucial to discriminate beneficial from harmful intestinal bacteria with regard to this activity. A few studies have suggested a diversity of enzyme action. Some bacterial groups are thought to exert activity toward *para*-nitrophenyl glucuronide or phenolphthalein glucuronide (Nanno et al. 1986), while others activate 1-nytropyrene (Morotomi et al. 1985). β -glucuronidases from *Escherichia coli* strains were more strongly induced by methyl- β -D-glucuronide than were those from other bacterial species (Tryland and Fiksdal 1998). In the case of glycyrrhizin metabolism,

highly specialized β -glucuronidase activities were involved in cleavage of the two glucuronic acid residues carried by the molecule (Kim et al. 2000).

Genetic diversity has been demonstrated within the *gusA* genes of *E. coli* (Ram et al. 2004) and *Ruminococcus gnavus* species (Beaud et al. 2005) and for the genetic environment of different *gusA* genes of *Ruminococcus gnavus* strains (Beaud et al. 2005). We herein summarize the most recent results of metagenomic investigations of β -glucuronidase diversity within the human intestinal microbiota, derived from function-based and sequence-based approaches. This provides key elements toward a better understanding of the “ambiguous” roles of these enzymes in handling the large diversity of glucuronides reaching the colon.

Functional Screening-Based Identification of Human Fecal β -Glucuronidases in Metagenomic Clone Libraries

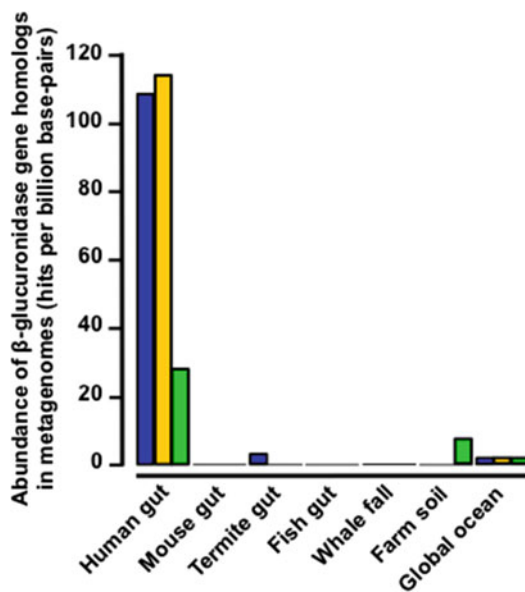
The genetic information from complex microbial ecosystems can be cloned as large fragments of genomes (Handelsman 2004) into libraries that can be used to detect gene clusters or operons allowing functional investigation. This approach has recently been applied to the intestinal

ecosystems and offers the potential to identify new genes from the microbiota, including its uncultured fraction. It is expected that about 40 % of enzymatic activities should be recoverable in *E. coli* (Gabor et al. 2004) and this host can express a significant number of genes (Handelsman et al. 1998; Rondon et al. 2000). The metagenomic approach has revealed new enzymes (Hayashi et al. 2005; Humblot et al. 2007; Yun et al. 2004; Kim et al. 2006, Tasse et al. 2010, Cecchini et al 2013), anticancer products (Piel et al. 2005), and compounds important for industrial, biotechnological, or therapeutic applications (Streit and Schmitz 2004), all having no homolog in the host bacterium (*E. coli*). β -glucuronidase represents an important function of interaction between the intestinal microbiota and the host and a relevant intestinal activity for human health.

Metagenomic libraries from microbiota obtained from human ileum or feces were constructed in *E. coli* and their phylogenetic diversity analyzed (Manichanh et al. 2006). The first functional approach using these libraries argued in favor of an efficiency of functional expression from the four dominant phyla of the digestive tract (Gloux et al. 2007). Despite the presence of β -glucuronidase genes in the host bacterium (*E. coli*), we designed a screening strategy that allowed the identification of numerous bioactive clones. Following primary screening for metagenomic clones overexpressing β -glucuronidase activity, we subcloned the inserts in a *uidA* *E. coli* strain (Gloux et al. 2011). Overall, 19 out of 6,144 metagenomic clones tested had fosmids able to express a β -glucuronidase activity based on *para*-nitro-phenyl- β -D-glucuronide bioconversion (Bardonnet and Blanco 1992), with levels ranging from 0.02 to 0.88 units. Phylogenetic, genetic, and functional characteristics of β -glucuronidase-positive inserts were investigated. A novel BG gene encoding a β -glucuronidase was identified in both Firmicutes and Bacteroidetes genetic backgrounds. The protein encoded by the gene has two conserved glutamate residues required for catalysis (Salleh et al. 2006) and the conserved predicted TIM barrel domain structure

of glycosyl hydrolase family 2 enzymes (Marchler-Bauer and Bryant 2004). The BG protein also had unique features, including an additional C-terminal domain compared to known β -glucuronidases and primary sequence specificities that led to the proposal of novel consensus motifs for the Firmicutes-borne BG and for glycosyl hydrolase family 2 (Gloux et al. 2011).

On the basis of sequence specificities, the frequency of the novel Firmicutes or Bacteroidetes BGs within the human gut metagenomes could be assessed. It was such that at least one homolog could be found within approximately 10^4 bacterial genes, making it by far the most dominant BG gene in human gut metagenomes. It was absent from other environmental metagenomes, including animal guts, making it specific to the human gut metagenome (Fig. 2). It was present in the genomes of numerous human intestinal commensals belonging to the phylogenetic and



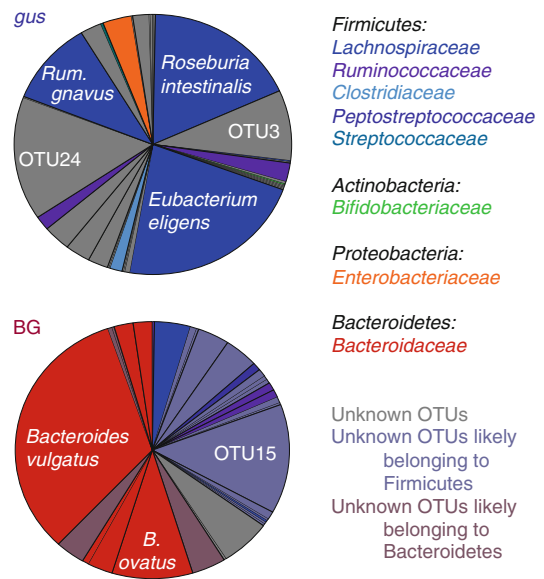
Functional Metagenomics of Human Intestinal Microbiome β -Glucuronidase Activity, Fig. 2 Abundance of Firmicutes BG (blue), Bacteroidetes BG (orange), and *uidA* homologs (green) in different environments. Abundance was assessed as hits per billion base pairs to correct for size difference of metagenomic datasets. The hit threshold was set as at least 50 % identity with 50 % sequence coverage. For full details of the different genomic hits, see Gloux et al. (2011)

metagenomic cores described for the human microbiome (Tap et al. 2009; Qin et al. 2010). Finally, gene duplications and its spread across diverse phylogenetic lineages suggested an ecological drive to ensure the presence of the activity, via functional redundancy, in spite of population variability between individuals.

In conclusion, a novel class of BG was revealed by our functional metagenomic approach that may be part of a functional core specifically evolved to adapt to the human gut environment and potentially important in maintaining health.

Sequence-Based Analysis of Human Fecal β -Glucuronidases

As described above, two different β -glucuronidase genes have been described, the *gus* (also referred to as *gusA* or *uidA*) gene, which is present in many bacteria as well as higher organisms, and the BG gene, which was identified by a functional metagenomic approach (Gloux et al. 2011). Human fecal metagenomic sequences can be searched to establish the distribution of β -glucuronidases within the human gut community. Degenerate primers targeting highly conserved regions were designed to amplify both types of β -glucuronidase gene from human fecal DNA (McIntosh et al. 2012). This revealed that the *gus* gene is present in many different phylogenetic groups, whereas the BG gene appears only to be present in bacteria related to Bacteroidetes and Firmicutes. Over 30 different sequence types (operational taxonomic units, OTUs) were found for both genes, with only a few of those being highly abundant (Fig. 3). The majority of OTUs, including some of the abundant ones, corresponded to sequences that currently cannot be assigned to specific bacteria, as either they remain uncultured or their genomes have not yet been sequenced. Three Lachnospiraceae species appear to be among the main carriers of *gus* (Fig. 3). A recent study that compared levels of human fecal β -glucuronidase activity with overall community diversity also concluded that β -glucuronidase activity was



Functional Metagenomics of Human Intestinal Microbiome β -Glucuronidase Activity, Fig. 3 Distribution of different types of β -glucuronidase gene in feces of human volunteers (*gus* gene: 685 sequences from ten volunteers, BG gene: 400 sequences from six volunteers). For full details of the different OTUs detected, see McIntosh et al. (2012)

mostly due to particular taxa rather than the wider community (Flores et al. 2012). Interestingly, the phylogenetic relationship of *gus* genes in known bacteria often did not agree with their relatedness based on the 16S rRNA gene sequence (McIntosh et al. 2012), which is commonly used to classify bacteria phylogenetically. Thus, it appears that the *gus* gene has been obtained by horizontal gene transfer in several bacteria. For the BG gene, there was a clear phylogenetic distinction between Firmicutes and Bacteroidetes, but many bacteria, especially among the Bacteroidetes, carried several copies of the gene with slightly divergent sequences.

The degenerate PCR approach described here to investigate specific functions within microbial communities may miss sequences with slight variations in the primer regions that may nevertheless encode the same function. The data were therefore compared to a large metagenomic sequence library of 85 healthy human volunteers (Qin et al. 2010), which showed that the vast

majority of sequences had indeed been captured by the degenerate PCR approach (McIntosh et al. 2012). There were slight differences in relative abundance, which is not surprising considering the difference in technical approach as well as volunteer numbers, but overall both approaches correlated significantly in terms of relative abundance as well as prevalence of different OTUs. Thus, a targeted approach based on degenerate primers appears to provide a good coverage of β -glucuronidase genes. It currently also allows for a more in-depth analysis per volunteer, as the actual metagenomic sequence coverage per volunteer in the pioneer metagenomic sequencing studies is relatively low and many genes are only partially covered. With the vast advances in sequencing technology, however, direct metagenomic mining for specific functional genes will become increasingly attractive.

Sequence-based analysis of functional genes poses the risk of assigning functions to genes that may in fact carry out a different activity, and the actual enzyme activity will ultimately have to be established for representatives of gene variants less closely related to biochemically characterized ones. Especially for glycoside hydrolases, it is often difficult to infer function from sequence alone (► [Carbohydrate-Active Enzymes Database, Metagenomic Expert Resource](#)). Both β -glucuronidase genes are remotely related to each other based on protein sequence identity and belong to glycoside hydrolase family 2, which also includes enzymes with other specificities, including β -galactosidases and β -mannosidases (<http://www.cazy.org/GH2.html>). The *gus* gene has been characterized biochemically in bacteria from different phylogenetic backgrounds (Beaud et al. 2005; Russell and Klaenhammer 2001), and the presence of the gene in a panel of human gut isolates correlated relatively well with the detection of β -glucuronidase activity (Dabek et al. 2008). On the other hand, it was shown that different strains of the same species can show differences in enzyme activity levels when grown under the same conditions and that the level to which β -glucuronidase activity is induced varies in dependence of the growth substrate and the

presence of β -glucuronides (Dabek et al. 2008; McIntosh et al. 2012). The BG gene was identified by screening for β -glucuronidase activity (Gloux et al. 2011), thus confirming its function. An investigation of several bacteria that harbor a BG gene but no *gus* gene revealed only low levels of β -glucuronidase activity, in both the absence and presence of β -glucuronide as inducer (McIntosh et al. 2012). Thus, BG genes may only be expressed under specific conditions that are yet to be identified. Alternatively, some variants of this diverse gene family may actually encode enzymes with different substrate specificities.

In conclusion, sequence-based analysis of genes encoding β -glucuronidases can be used to reveal the diversity of the β -glucuronidase-positive community and forms a solid basis for further functional investigation of this activity in representative organisms.

Summary

The metabolic activities of the microbial community present in the human gut are closely linked to the physiological status and overall health of its host. Bacterial β -glucuronidase activity directly interferes with one of the major host detoxification systems for a wide range of lipophilic compounds that enter the body via the diet, drugs, or exposure to environmental pollutants, as well as endogenous molecules. Glucuronidation of those compounds renders them more hydrophilic and facilitates their excretion, but β -glucuronidase activities within the gut microbiota convert them back to their respective aglycones, which leads to an extended retention time in the body. Many of those compounds are toxic or carcinogenic, but potentially health-promoting compounds, such as plant phenolics ingested with the diet, may also be glucuronidated. Metagenomics can be utilized to enhance our understanding of which bacteria in the human gut carry β -glucuronidase activity. A functional metagenomic approach, whereby genes from environmental communities are expressed in a heterologous host, has led to the identification of a novel type of β -glucuronidase gene, which was found to be prevalent within the

human gut microbiota but not commonly found in other environments. Metagenomic sequence mining for this novel gene, as well as a previously known β -glucuronidase gene, revealed the distribution of these genes in different phylogenetic lineages. These results provide a valuable foundation for further functional characterization of this important microbial activity.

Cross-References

- ▶ Carbohydrate-Active Enzymes Database, Metagenomic Expert Resource
- ▶ Fosmid System

References

- Bardonnet N, Blanco C. *uidA*-antibiotic-resistance cassettes for insertion mutagenesis, gene fusions and genetic constructions. *FEMS Microbiol Lett.* 1992;72:243–7.
- Beaud D, Tailliez P, Anba-Mondoloni J. Genetic characterization of the beta-glucuronidase enzyme from a human intestinal bacterium, *Ruminococcus gnavus*. *Microbiology.* 2005;151:2323–30.
- Cecchini DA, Laville E, Laguette S, Patrick Robe P, Leclerc M, Doré J, Henriissat B, Remaud-Siméon M, Pierre Monsans P, Potocki-Véronèse G. Functional metagenomics reveals novel pathways of prebiotic metabolism by human gut bacteria. *PLoS ONE.* 2013;8:1–9.
- Dabek M, McCrae SI, Stevens VJ, Duncan SH, Louis P. Distribution of β -glucosidase and β -glucuronidase activity and of β -glucuronidase gene *gus* in human colonic bacteria. *FEMS Microbiol Ecol.* 2008;66:487–95.
- Flores R, Shi J, Gail MH, Gajer P, Ravel J, Goedert JJ. Association of fecal microbial diversity and taxonomy with selected enzymatic functions. *PLoS ONE.* 2012;7:e39745.
- Gabor EM, Alkema WB, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol.* 2004;6:879–86.
- Gloux K, Leclerc M, Iliozier H, L'haridon R, Manichanh C, Corthier G, Nalin R, Blottière HM, Doré J. Development of high-throughput phenotyping of metagenomic clones from the human gut microbiome for modulation of eukaryotic cell growth. *Appl Environ Microbiol.* 2007;73:3734–7.
- Gloux K, Berteau O, El Oumami H, Béguet F, Leclerc M, Doré J. A metagenomic β -glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc Natl Acad Sci U S A.* 2011;108:4539–46.
- Haiser HJ, Turnbaugh PJ. Developing a metagenomic view of xenobiotic metabolism. *Pharmacol Res.* 2013;69:21–31.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev.* 2004;68:669–85.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:R245–9.
- Hayashi H, Abe T, Sakamoto M, Ohara H, Ikemura T, Sakka K, Benno Y. Direct cloning of genes encoding novel xylanases from the human gut. *Can J Microbiol.* 2005;51:251–9.
- Henriissat B, Cantarel B, Coutinho P. Carbohydrate-active enzymes database, metagenomic expert resource. <http://www.springerreference.com/index.chapterbid/303280>
- Humblot C, Murkovic M, Rigottier-Gois L, Bensaada M, Bouclet A, Andrieux C, Anba J, Rabot S. Beta-glucuronidase in human intestinal microbiota is necessary for the colonic genotoxicity of the food-borne carcinogen 2-amino-3-methylimidazo[4,5-f]quinoline in rats. *Carcinogenesis.* 2007;28:2419–25.
- Kim DH, Jung EA, Sohng IS, Han JA, Kim TH, Han MJ. Intestinal bacterial metabolism of flavonoids and its relation to some biological activities. *Arch Pharm Res.* 1998;21:17–23.
- Kim DH, Hong SW, Kim BT, Bae EA, Park HY, Han MJ. Biotransformation of glycyrrhizin by human intestinal bacteria and its relation to biological activities. *Arch Pharm Res.* 2000;23:172–7.
- Kim YJ, Choi GS, Kim SB, Yoon GS, Kim YS, Ryu YW. Screening and characterization of a novel esterase from a metagenomic library. *Protein Expr Purif.* 2006;45:315–23.
- Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut.* 2006;55:205–11.
- Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 2004;32:327–31.
- McBain AJ, Macfarlane GT. Ecological and physiological studies on large intestinal bacteria in relation to production of hydrolytic and reductive enzymes involved in formation of genotoxic metabolites. *J Med Microbiol.* 1998;47:407–16.
- McIntosh FM, Maison N, Holtrop G, Young P, Stevens VJ, Ince J, Johnstone A, Lobley G, Flint HJ, Louis P. Phylogenetic distribution of genes encoding β -glucuronidase activity in human colonic bacteria and the impact of diet on faecal glycosidase activities. *Environ Microbiol.* 2012;14:1876–87.
- Morotomi M, Nanno M, Watanabe T, Sakurai T, Mutai M. Mutagenic activation of biliary metabolites of

- 1-nitropyrene by intestinal microflora. *Mutat Res.* 1985;149:171–8.
- Nanno M, Morotomi M, Takayama H, Kuroshima T, Tanaka R, Mutai M. Mutagenic activation of biliary metabolites of benzo(a)pyrene by beta-glucuronidase-positive bacteria in human faeces. *J Med Microbiol.* 1986;22:351–5.
- Piel J, Butzke D, Fusetani N, Hui D, Platzer M, Wen G, Matsunaga S. Exploring the chemistry of uncultivated bacterial symbionts: antitumor polyketides of the pederin family. *J Nat Prod.* 2005;68:472–9.
- Qin J, Ruiqiang L, Raes J, Arumugam M, Solvsten K, Burgdorf, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende D, Li J, Xu J, LI S, Li D, Cao J, Wang B, Liang H, Zheng H, Yie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Zhang X, Li S, Yang H, Wang J, Brunak S, Brunak J, Dore J, Guraner F, Kristiansen K, Pedersen O, Parkhill J, Wessenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalog established by deep metagenomic sequencing. *Nature.* 2010;464:59–65.
- Ram JL, Ritchie RP, Fang J, Gonzales FS, Selegean JP. Sequence-based source tracking of *Escherichia coli* based on genetic diversity of beta-glucuronidase. *J Environ Qual.* 2004;33:1024–32.
- Rod TO, Midtvedt T. Origin of intestinal beta-glucuronidase in germfree, monocontaminated and conventional rats. *Acta Pathol Microbiol Scand.* 1977;85(B):271–6.
- Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osborne MS, Clardy J, Handelsman J, Goodman RM. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol.* 2000;66:2541–7.
- Russell WM, Klaenhammer TR. Identification and cloning of *gusA*, encoding a new beta-glucuronidase from *Lactobacillus gasseri* ADH. *Appl Environ Microbiol.* 2001;67:1253–61.
- Salleh HM, Müllegger J, Reid SP, Chan WY, Hwang J, Warren RA, Withers SG. Cloning and characterization of *Thermotoga maritima* beta-glucuronidase. *Carbohydr Res.* 2006;341:49–59.
- Schmelz EM, Bushnev AS, Dillehay DL, Sullards MC, Liotta DC, Merrill Jr AH. Ceramide-beta-D-glucuronide: synthesis, digestion, and suppression of early markers of colon carcinogenesis. *Cancer Res.* 1999;59:5768–72.
- Streit WR, Schmitz RA. Metagenomics-the key to the uncultured microbes. *Curr Opin Microbiol.* 2004;7:492–8.
- Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Munoz-Tamayo R, Paslier DL, Nalin R, Dore J, Leclerc M. Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol.* 2009;11:2574–84.
- Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL, Coutinho PM, Henrissat B, Leclerc M, Doré J, Monsan M, Remaud-Simeon M, Potocki-Veronese G. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* 2010;20:1605–12.
- Tryland I, Fiksdal L. Enzyme characteristics of beta-D-galactosidase- and beta-D-glucuronidase-positive bacteria and their interference in rapid methods for detection of waterborne coliforms and *Escherichia coli*. *Appl Environ Microbiol.* 1998;64:1018–23.
- Tukey RH, Strassburg CP. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu Rev Pharmacol Toxicol.* 2000;40:581–616.
- Yun J, Kang S, Park S, Yoon H, Kim MJ, Heu S, Ryu S. Characterization of a novel amylolytic enzyme encoded by a gene from a soil-derived metagenomic library. *Appl Environ Microbiol.* 2004;70:7229–35.

Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing

Thomas W. Schoenfeld, Michael J. Moser and David Mead
Lucigen Corporation, Middleton, WI, USA

Introduction

The enzymes of phages and other viruses were vital to the early development of molecular biology and are still essential tools. However, the available viral enzymes represent a tiny sample of the potential diversity found in the global virosphere. Viral metagenomics has revealed a vast diversity of novel genes and its virtually limitless potential to provide new enzymes for use in molecular analysis. An important challenge to both the understanding of viral ecology and development of new viral enzymes is functional characterization of metagenomic sequences, which has lagged far behind the ability to collect sequence data. Described is

a program to identify and characterize replication operons of viral metapopulations isolated from natural thermal environments and develop the gene products as thermostable enzymes for nucleic acid amplification and sequencing. Approaches to functionally characterize viral replicases include (1) expression and biochemical analysis of gene products identified by sequence similarity, (2) functional screens to discover new families of genes, and (3) assembly of operons to predict function based on gene position. These approaches have uncovered at least two diverse families of replication operons including dozens of genes for thermostable DNA polymerases and reverse transcriptases, as well as likely replicase subunits. In addition, functional screens have uncovered one viral Pol unrelated to any known protein. These enzymes are being engineered as improved PCR, RT PCR, and DNA sequencing reagents. Diversity in the viral metagenomes is also being explored to optimize the activity of the genes discovered in the libraries and make them more suitable for the targeted applications.

Gene products of phages and other viruses (collectively referred to here as viruses) have historically provided many of the enzymatic tools for molecular biology. However, most of the commonly used viral enzymes are derived from a very limited number of cultivated viruses, primarily phages T4, T7, lambda, SP6, and phi29, and retroviruses Moloney murine leukemia virus (Mo-MLV) and avian myeloblastosis virus (AMV). The program to study hot spring virology in Yellowstone National Park (YNP), California, and Nevada has provided insight into viral ecology (Otto et al. 1998; Breitbart et al. 2004; Schoenfeld et al. 2008) and has revealed a nearly unlimited source of diversity for the search for new enzymes (Beechem et al. 1998; Moser et al. 2012; Perez et al. 2012). However, current approaches to functional analysis of viral metagenomes, while informative, are limited by their reliance on sequence similarity to infer gene function. Improvements in the ability to functionally characterize viral metagenomes are necessary to advance the field.

Thermostable DNA polymerases (Pols) have been a major research focus due mainly to their wide use in molecular detection and analysis. DNA polymerases are essential for PCR (Staley and Konopka 1985) and other target-specific (Petruska et al. 1998; Notomi et al. 2000) and whole genome amplification methods (Goodman and Fygenon 1998) and are also essential components of all the major DNA sequencing platforms. Sanger (dideoxy chain termination) DNA sequencing was the first major sequencing method to use DNA polymerases and was advanced by thermostable Pols (Tang et al. 2008). All of the leading next-generation sequencing-by-synthesis platforms (e.g., Roche/454 FLX, Illumina Genome Analyzer, Helicos Heliscope, Pacific BioSystems SMRT, ABI SOLiD) (Mardis 2008b; Shendure and Ji 2008) use at least one DNA polymerase for base discrimination and/or template preparation. DNA polymerase-based methods are driving discovery in research labs and, increasingly, in the clinic (Bhui-Kaur et al. 1998) as methods for nucleic-acid-based detection of infectious agents, cancer and genetic variation advance next-generation diagnostics, and personalized medicine. Progress in improving all these methods depends in part on more suitable DNA polymerases.

Viruses are rich sources of diverse new DNA polymerases. Compared to their cellular hosts, viruses use a wide array of strategies to replicate their genomes, and their genomes adopt nearly every conceivable form, including double-stranded and both positive and negative single-stranded RNA and DNA forms, with linear, circular, and multipartite topologies ranging in size from 1.2 Mb (mimivirus) down to 3.2 kb (hepatitis B virus) (Blanco et al. 1989; Detter et al. 2002). While many of these replicative strategies rely on host enzymes, a substantial subset of viral families supplies its own replication proteins. There is speculation that viruses may have played a key role in the evolution of replication strategies used by cellular life (Koonin 2006).

As replicases, viral polymerases are functionally distinct from the bacterial and archaeal

enzymes currently used in molecular biology. During prokaryotic cellular replication, processive leading-strand synthesis depends on a multisubunit complex including Pol III holoenzyme, helicases, and primases. *E. coli* Pol III holoenzyme is a 791 kD protein comprised of nine subunits (reviewed in Xiang et al. (2008)). Due to their complexity, no Pol III derivative has been developed as a molecular biology reagent. Cell-derived reagent Pols, e.g., *Taq*, *Pfu*, or *E. coli* DNA polymerases, are all bacterial Pol I or archaeal Pol II derivatives that are mainly responsible in vivo for lagging strand and repair synthesis, neither of which requires strand separation or processive synthesis of long sequences. Viral Pols are functionally more like the leading-strand replicases and, accordingly, exhibit higher fidelity, rates of synthesis, and processivity (Ley et al. 2008). Phage T7 Pol, for example, incorporates 300 nt per second, six times faster than *Escherichia coli* Pol I; T4 phage replicates DNA ten times faster than its *E. coli* host (Heckler et al. 1984). Phi29 Pol has a processivity of >70,000 nucleotides (Blanco et al. 1989) (i.e., it incorporates over 70,000 nucleotides before dissociating), far greater than that of *Taq* Pol I, which has a processivity of between 50 and 80 (Merkens et al. 1995). Phi29 also has a strong strand displacement capability that, together with its processivity, makes it the polymerase of choice for whole genome amplification by multiple displacement amplification (MDA) (Dean et al. 2001). T7 phage Pol holoenzyme has a processivity of 1,000 nucleotides (Tabor et al. 1987) and efficiently incorporates chain-terminating nucleotide analogs, which facilitated Sanger sequencing until it was displaced by Thermo Sequenase, a *Taq* Pol derivative that was engineered based on the nucleotide variation in T7 DNA Pol that conferred efficient incorporation of dideoxynucleotides (Tabor and Richardson 1995). T5 Pol has both high processivity and a potent strand displacement activity that are independent of additional host or viral proteins (Andraos et al. 2004). T4 DNA Pol has a high proofreading activity that is commonly exploited for generating blunt ends, especially in physically sheared DNA (Karam and Konigsberg 2000).

Retroviral replicases (i.e., reverse transcriptases), especially Mo-MLV and AMV, are indispensable for detection, analysis, and cloning of transcripts and RNA viruses (Morin et al. 2008; Wang et al. 2008). Together, these qualities make viral Pols attractive targets for development as reagents.

While the emphasis has been DNA polymerases, viruses encode other useful enzymes. RNA polymerases, for example, are key components of a number of in vitro and in vivo transcription and translation systems, as well as several transcription-mediated amplification methods (Guatelli et al. 1990; Compton 1991). Virtually all ligation methods used for cloning and linker attachment depend on T4 DNA ligase due to its high activity on 5'- and 3'-extended and blunt DNA. The integrases and recombinases of various phages (e.g., lambda *red* and P1 *cre/lox*) have been used to integrate genes into bacterial and eukaryotic genomes. Resolvases (e.g., T4 endonuclease VII and T7 endonuclease I) have been used to detect single nucleotide polymorphisms (SNPs) (Babon et al. 2003). It is likely that these and many other methods that rely on viral enzymes can be further improved by novel enzyme activities. Functional metagenomic-based enzyme discovery and development should benefit a wide range of applications.

The enzymes that have been isolated by cultivation over the years demonstrate the potential of viruses as a source of new enzymes, but greatly underrepresent the richness of this resource. The extreme global abundance and diversity of viruses is well documented (Breitbart et al. 2002; Angly et al. 2006; Dinsdale et al. 2008; McDaniel et al. 2008; Schoenfeld et al. 2008). A liter of ocean water contains as many viruses as there are humans on the planet and much more genetic diversity (Wang et al. 2007). In fact, the bulk of the world's genetic diversity is probably encoded in viral genomes. Despite the richness of the global virosphere as a source of diverse replicative proteins, standard approaches to discovering new enzymes by cultivating the viruses have proven extremely inefficient and few new viral enzymes have been commercialized in the past decades.

Notably, despite their widespread potential applications and notwithstanding substantial effort, thermostable viral Pols have completely eluded discovery by cultivation. There are now 34 fully sequenced genomes from thermophilic viruses in the NCBI database (February 2010): 27 archaeal viruses and 7 bacteriophages. None of these genomes or broad screens of hundreds of cultivated *Thermus* phage (Lopatto et al. 2008) has produced a thermostable DNA polymerase. Extensive analysis of cultivated crenarchaeal viral genomes from high-temperature environments reveals few recognizable features other than a small number of methylases, helicases, glycosyltransferases, and several unknown but shared genes (Rehrauer et al. 1998). At least one presumptive DNA polymerase has been identified in an archaeal viral genome (Baklanov et al. 1984), but not expressed in the lab. At least five Pols have been expressed from thermophilic bacteriophage genomes (Wang et al. 2006; Schmidt et al. 2008; T. Schoenfeld, unpublished); however, for unknown reasons, these enzymes are only moderately thermostable and incapable of surviving thermocycling in PCR or sequencing, despite the thermostability of their host Pols. In order to identify useful thermostable Pols, more efficient approaches are needed.

One of the main barriers to discovery of new viral enzymes is technical challenges associated with cultivation. It is widely noted that cultivation in the lab selects against the great majority of Bacteria and Archaea. Cultivation of new viruses introduces another extreme level of selection against the vast majority of natural populations because cultivation requires the investigator to choose a host that can be grown in the lab, which severely limits the comprehensiveness of the screens. When examining extreme environments like thermal springs, which are dominated by autotrophic microbes, this host selection is even more limiting. Most of these cultivation efforts have focused on viruses that infect heterotrophic Bacteria, especially *Thermus* (Rehkrantz et al. 1998; Karam and Konigsberg 2000; Pavlov and Karam 2000; Bebenek et al. 2001; Blondal et al. 2003;

Ding et al. 2008; Lopatto et al. 2008; Schmidt et al. 2008) or a small number of thermoacidophilic Archaea, particularly *Sulfolobus* and *Acidianus* (reviewed in Rehrauer et al. (1998)), due to the relative ease of cultivating these hosts. Metagenomics promises to overcome these barriers and provide a largely unbiased sampling of viral populations.

In some respects viral metagenomes are especially well suited for discovery of enzymes for use in molecular analysis. Viral genomes are highly diverse and dense with genes associated with nucleic acid metabolism (Paulsen and Wintermeyer 1984). For example, a typical bacterial genome of 2 Mb contains three to five DNA polymerase genes, only one of which, *polA*, encodes enzymes that have been used as reagents. In contrast, a comparable 2 Mb of viral metagenome can yield up to 40 *pol* genes (Schoenfeld et al. 2008). However, the promise of using this diversity to advance the understanding of global ecology and in developing useful enzymes from viral metagenomes is tempered by the challenge in assigning function to the genes. The gigabases of viral metagenomic sequence data that have been generated over the past decade have provided only inferential insight into function or biochemistry of the viral genes and, consequently, few new molecular tools. Efforts to glean insight from metagenomes are hampered by the nearly complete reliance on sequence similarity coupled with the extreme viral genomic diversity and the dearth of annotated sequences. Depending on the environment, 40–90 % of viral metagenomic sequences are unknown, novel sequences (Angly et al. 2006; Dinsdale et al. 2008; Bench 2007; Srinivasiah 2008; Schoenfeld 2008). All the next-generation platforms generate shorter reads that are even more difficult to assemble or align to sequences in GenBank, resulting in artificially low BLASTx homologies or, conversely, artificially high numbers of “unique” sequence (Wommack 2008). The VIROME database (virome.dbi.udel.edu) has cataloged 201 Mb of predicted open reading frames (ORFs) from long read sequence data (Feb 2010), the vast majority of which are novel and functionally uncharacterized.

Functional characterization of viral metagenomes has lagged far behind the ability to collect sequence data. Essentially none of the millions of gene functions inferred by sequence similarity has been proven biochemically by expression and analysis of the gene products. More importantly, the mere description of sequence similarity does little to further the understanding of viral biology or to identify useful new enzymes. Furthermore, sequence-similarity screens only identify genes with an annotated counterpart in a database. The relative scarcity of functionally annotated viral genes in GenBank has likely prevented discovery of truly novel enzyme families, which should be the strength of viral metagenomics.

Finally, a conceptual barrier associated with the definition of related viral types has prevented assembly of viral genomes, and, consequently, inferences into function that are based on gene position. Phage genes of related function, especially replication-related genes, often occur in proximity within operons (El Omari et al. 2006). Assembly of sequence reads should allow reconstruction of operons; however, standard approaches relying on nucleotide identities of greater than 95 % are ineffective in assembly of viral metagenomes and only a few very small, abundant phage genomes have been reconstructed from metagenomic data (Angly et al. 2006). Because even the relatively long Sanger reads are almost always too short to include more than one complete gene, these associations are generally missed. Since traditional shotgun sequencing, used in some of the work described below, involved the construction of clone libraries, success in identifying adjacent genes by sequencing entire inserts from archived clones was achieved, but even this approach is limited by the sizes of inserts in the libraries, generally less than 5 kb. Since none of the next-generation sequencing methods uses clone libraries, this approach is impossible for most of the ongoing viral metagenomic projects. The fundamental problem is that viral populations are too molecularly diverse to accommodate this criterion. Among cultivated viruses, closely related phages are up to 50 % divergent at the nucleotide

level (Truncaite et al. 2006; Wang and Silverman 2006). When assembly criteria are reduced to as low as 50 %, much larger assembled contigs are generated (Schoenfeld et al. 2008). This approach has proven effective in generating contigs that contain identifiable operons that not only allow isolation of genes of related function but allow mapping of diversity onto the protein structure. These sequence variations correspond to biochemical differences in the gene products and provide a guide to enzyme engineering. In the work described below, a tripartite approach was used for functional analysis of viral metagenomes including (1) expression and biochemical characterization of the “BLASTx hits,” (2) functional screens to identify enzymes too dissimilar to known genes to be detected by sequence similarity, and (3) assembly of operons to infer gene function based on position in the genome.

Methods

Sampling, Library Construction, and Sequencing

Sampling, library construction, and sequencing of the YNP samples have been described (Schoenfeld et al. 2008). The Great Boiling Spring samples were collected as described and amplified using the Repli-g kit (GE Healthcare). DNA was sheared and inserted into pETite vector (Lucigen) and the library used to transform *E. coli* HI-Control BL21(DE3) cells (Lucigen). Individual clones from both libraries were sequenced in their entirety using standard chemistry (Life Technologies).

Bioinformatics

Sequence assemblies were performed using Sequencher (Gene Codes) or SeqMan (DNASTAR). ClustalW analysis was performed as described (Nandakumar and Shuman 2005).

Functional Screens

The clones from the Great Boiling Spring samples were grown on Luria broth, pelleted, and resuspended in buffer containing lysozyme. Lysates were incubated for 10 min at 70 °C and

centrifuged, and the supernatants were tested for DNA polymerase activity using the standard assay. Positive clones were cultivated at 50 ml in LB and retested. The inserts of clones with activity were sequenced in their entirety.

Cloning, Expression, Purification, and Mutagenesis

DNA polymerase genes that were further characterized were expressed at higher levels by insertion into pET28 vector and expression in *E. coli* EXPRESS BL21(DE3) (Lucigen). DNA polymerase was purified by heat treatment and standard chromatography methods. Mutagenesis was performed using the QuikChange II Site-Directed Mutagenesis Kits (Agilent).

Biochemical Analysis and Applications Development

Biochemical assays were performed using standard methods (Mardis 2008; Marks et al. 2008).

Results and Discussion

Sequence-Based and Functional Discovery of New DNA Polymerases

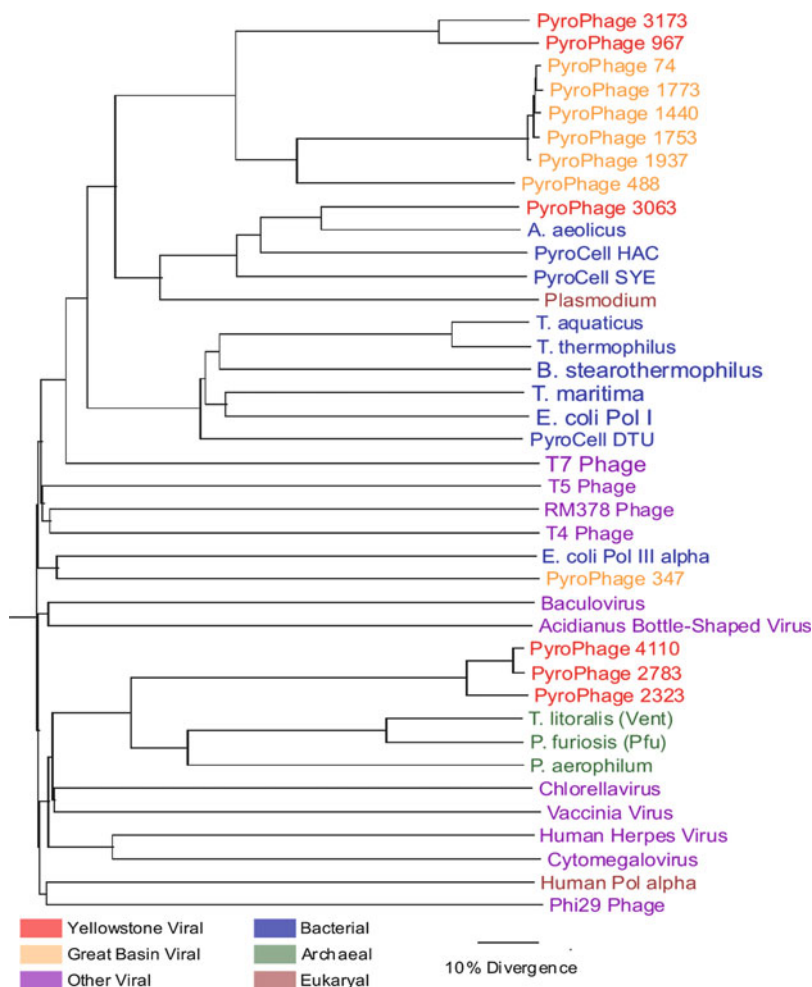
In a recent study of viral metagenomes from Yellowstone hot springs, more than 28,000 Sanger-based long sequence reads (nearly 30 Mb of sequence) were determined (Schoenfeld et al. 2008). BLASTx alignment to the nonredundant protein database indicated that 156 ORFs had similarity to known *pol* genes. Fifty-nine appeared to be complete genes and were tested for DNA polymerase activity. Ten showed activity and seven of these were sequenced in their entirety. Although highly divergent from known viral and cellular genes, four were loosely grouped with family A polymerases and three grouped with family B polymerases. These *pol* genes are referred to as “PyroPhage” followed by an identifying number. The family A *pol*s detected by this screen were too divergent to be grouped, but the family B *Pol*s are referred to below as “PyroPhage 4110-like *Pol*s” in reference to the first one discovered.

The degree of sequence conservation among *pol* genes in these libraries, while relatively low, was higher than most sequences found in viral metagenomes. The discovery of 156 partial genes among roughly 600 viral genome equivalents suggests that sequence-based screens were relatively efficient in identifying *pol* genes. Nonetheless, there are important disadvantages to this approach. One is that the diversity of viral *pol* genes is likely to be high enough that interesting new enzymes are missed. Another problem is that a gene must be situated in the random clone so that an identifiable portion of it is within the read length of the sequencing method (>1,000 nucleotides by Sanger, much less by newer sequencing approaches) and the gene must not extend beyond the boundaries of the random insert so that it is incomplete. It is unknown how many genes failed to fulfill the first criterion and were within the insert, but not within the sequence range. Of the 156 identified candidate *pol* genes, only 38 % fulfilled the second criterion and appeared complete. Finally, the identification of a gene does not mean that the gene will express efficiently in *E. coli*. For unknown reasons, among the 59 likely complete genes, 83 % failed to express at detectable levels.

Functional screens address many limitations of sequence-similarity screens and can often detect completely novel activities regardless of divergence from known genes or position in the insert, as long as the complete gene is present. By their nature, functional screens only detect complete, expression-competent genes. Viral metagenomic DNA from the Great Boiling Spring, Gerlach, NV, kindly provided by Brian Hedlund and Jeremy Dodsworth (University of Nevada-Las Vegas), was used to construct a library that was screened for expression of thermostable *pol* activity. Screening of 2,800 clones resulted in the discovery of 12 that were positive for primer extension activity. Eleven of these were more than 97 % identical to each other and are referred to as the “PyroPhage 74-like polymerases” in reference to the first member discovered. These *pol* genes share up to 45 % identity with the other *polA*-type genes from Yellowstone (PyroPhage 3173 and 967) and 56 %

Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing,

Fig. 1 Polymerase phylogenetic tree. Full-length viral metagenomic DNA polymerase amino acid sequences were compared by ClustalW to representative viral, microbial, and eukaryotic Pols and displayed in a neighbor-joining tree



identity to PyroPhage 488, a *pol* gene isolated 8 years earlier in a sequence-based screen of a metagenome from Little Hot Creek, Long Valley, CA, which is 400 km from Gerlach, NV, but still in the Great Basin. The final clone identified in the functional screen, PyroPhage 347, had no significant similarity to any known *pol* gene. In fact the strongest E value to any known gene had a barely significant 0.750 score to an open reading frame of unknown function in a crenarchaeal virus. Due to this lack of similarity to genes of known function, this gene would never have been identified by sequence similarity.

The *pol* genes discovered by both screens were aligned by ClustalW to each other and to representative cellular and viral *pol* genes to construct a neighbor-joining tree (Fig. 1). Viral genes

from these screens, as well as those retrieved from GenBank, were noticeably more diverse than cellular genes. Most PyroPhage *pol* genes are highly divergent from known cellular or viral *pol* genes. The exception is PyroPhage 3063, which is related to several *polA* genes of Aquificales family, which are known to be quite divergent from other bacterial *polA* genes (Griffiths and Gupta 2004).

Since the libraries were constructed from different hot spring populations, direct comparisons are difficult. However, while the overall rate of discovery of apparent DNA polymerase genes was comparable for the sequence-based and functional screens (156 *pol* genes from 28,000 clones compared to 12 from 2,800 clones, respectively), the rate of discovery of *functional* thermostable

enzymes was much lower for the sequence screens than the functional screens (10 of 28,000 vs. 12 of 2,800). The diversity of the enzymes in the GBS library was much lower than those from Yellowstone springs, presumably reflecting a lower overall population diversity.

Biochemical Characteristics and Directed Engineering Improve Use of PyroPhage Pols in PCR and Sanger Sequencing

PyroPhage 3173 and 347 Pols proved to be the most thermostable of the newly discovered polymerases. In fact, these are the first viral Pols with adequate thermostability for PCR. PyroPhage 3173 Pol, which has been studied in greatest detail (Table 1), has adequate thermostability for thermocycling, inherent reverse transcriptase activity, and high fidelity that enable a number of applications for this enzyme. The proofreading activity proved highly beneficial for high-fidelity PCR amplification (Fig. 2). However, many applications benefit from the absence of proofreading activity. Alignment of the PyroPhage 3173 *pol* gene to *E. coli polA* (Beese and Steitz 1991) identified codons for two acidic residues, either of which could be mutated to eliminate exonuclease activity. This reduced fidelity to very close to that of Taq Pol, but simplifies its use in PCR and other amplification methods. Like most family A Pols, 3173 has a strong discrimination against dideoxynucleotides that made it less effective in Sanger sequencing. Based on

Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Table 1 Biochemical characteristics of PyroPhage 3173 DNAP

3'-5' exonuclease	Strong
5'-3' exonuclease	None
Strand displacement	Strong
Extension from nicks	Strong
T _{1/2} at 95°	10 min
K _m dNTPs	40 μM
K _m DNA	5.3 nM
Processivity	42
Fidelity	8 × 10 ⁴
3' ends of amplicons	Blunt
Template	DNA or RNA

alignment to known proteins (Tabor and Richardson 1995), mutation F418Y (Fig. 3a) reduced discrimination against chain terminators to nearly zero, making the enzyme very effective for dye terminator cycle sequencing (Fig. 3b).

Single-Enzyme RT PCR with 3173 DNA Polymerase

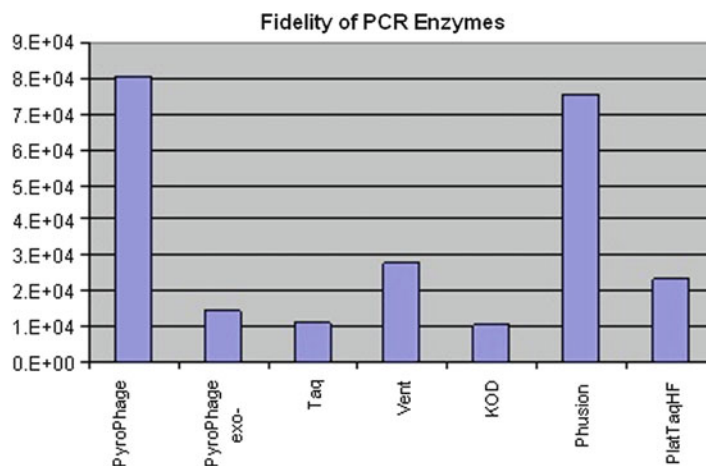
The thermostability and reverse transcriptase activities seen in PyroPhage 3173 Pol allow efficient RT PCR amplification of mRNA and viral RNA genomic targets with improved performance compared to alternative single-enzyme solutions (Fig. 4). Quantitative detection of viral targets is linear over at least seven logs of dilution (Fig. 5). These benefits have significantly improved detection of transcripts and RNA viruses (Moser et al. 2012).

Currently almost all RT PCR depends on retroviral RTs, i.e., M-MLV and AMV RTs, which, despite wide use, have well-documented deficiencies that compromise RT PCR. Side activities in retroviral reverse transcriptases, including RNase H and terminal transferase, lead to mismatch extension artifacts (Blumenthal 1980; Blumenthal and Hill 1980; Harrison and Zimmerman 1984; Pulsinelli and Temin 1991; Shah et al. 1995; Vratskikh et al. 1995; Ho and Shuman 2002; van Dijk et al. 2004). Primer-dependent bias in extension efficiency (Yin et al. 2003) and fidelity (Cheng et al. 2005) likely account for documented inaccuracy of RT PCR quantification (Loeffler et al. 2003), poor correlation between tests (Nelson et al. 2001), and/or complete amplification failure depending on the RT and the abundance of transcript (Damasko et al. 2005). Inherently low synthesis fidelity (up to one error per 500 nt, 20X higher than Taq Pol) results in misincorporations, frameshifts, and deletions (Kerr and Sadowski 1972; Little 1981; Heaphy et al. 1987). Strand-switching (Strauch et al. 2003) probably causes the inter- and intramolecular rearrangement artifacts (Cherepanov and de Vries 2001) that can be preferentially extended (Sharp et al. 1994) and result in recombination or insertion/deletion (indel) artifacts in cDNA synthesis (Evans et al. 1989; Snyder et al. 1992). A consequence of

Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing,

Fig. 2 Fidelity of PyroPhage 3173 Pol and its exo- derivative.

Fidelities of PCR amplification of PyroPhage 3173 wt and exonuclease minus Pols were compared to commercial sources of thermostable Pols in the lacI forward mutation assay (Lundberg et al. 1991)

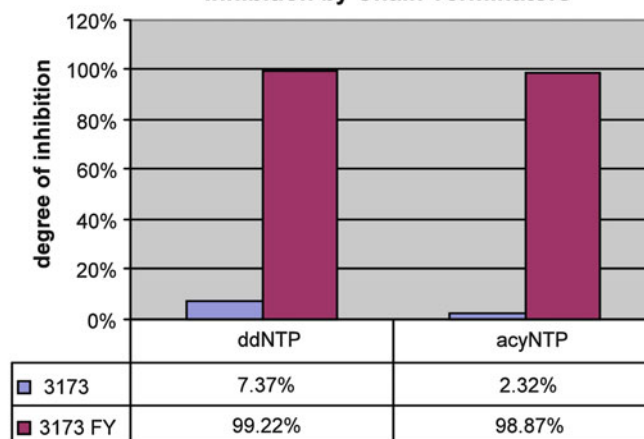


Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing,

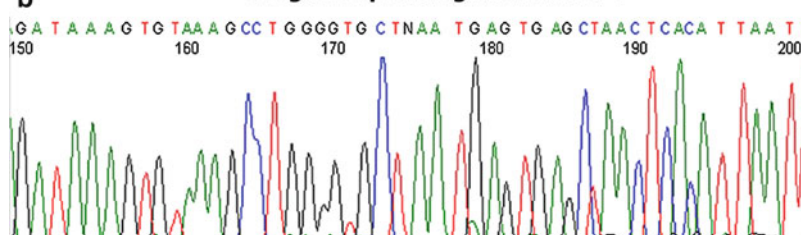
Fig. 3 Directed engineering of 3173 Pol to improve Sanger sequencing.

(a) Shown is the increased incorporation of dideoxy- and acyclo-nucleotides by the F418Y mutant of PyroPhage 3173 Pol, as indicated by increased inhibition of Pol activity by chain-terminating nucleotides. (b) The F418Y mutant was used as a direct substitute for Thermo Sequenase in a BigDye® (ABI) sequencing reaction

a Improved Incorporation of Chain Terminators
Inhibition by Chain Terminators



b Sanger Sequencing with 3173 FY



two-enzyme RT PCR is that the RT step can interfere with subsequent PCR (Harnett et al. 1985; McLaughlin et al. 1985; Evans et al. 1989; Petric et al. 1991; Snyder et al. 1992; Sharp et al. 1994), which compromises quantification of low abundance targets. Efforts to ameliorate

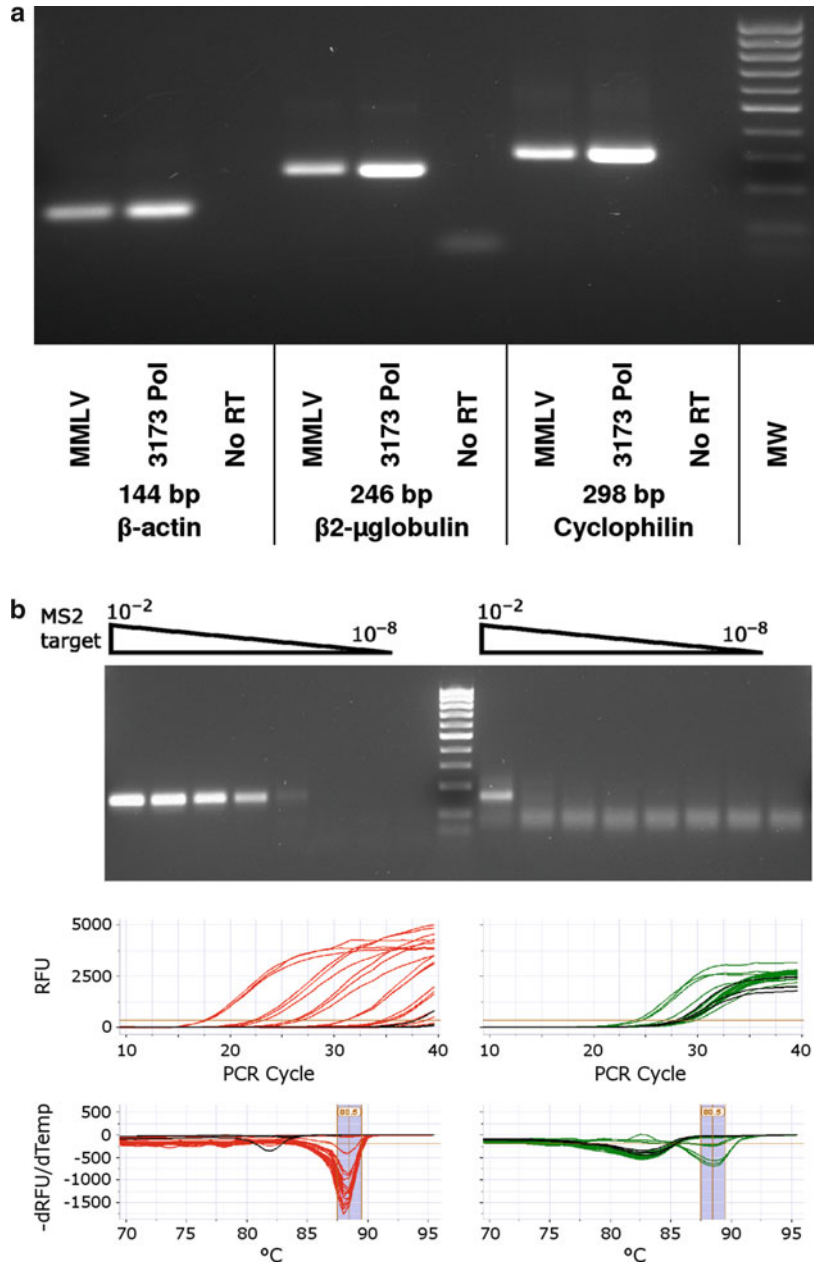
these deficiencies include mutagenesis to disable or remove the RNase H domain (Downie et al. 2004). These mutations reduce rearrangements, but lead to increased substitution errors and bias (Blumenthal and Hill 1980; Middleton et al. 1985; Vratskikh et al. 1995).

Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing,

Fig. 4 Reverse transcription PCR using PyroPhage 3173 Pol.

(a) Total human liver RNA (1 µg, Promega) was reverse transcribed by M-MLV RT or PyroPhage 3173 Pol and then PCR amplified using Lucigen EconoTaq® PLUS Master Mix. Shown are targets of 144, 246, and 298 bp.

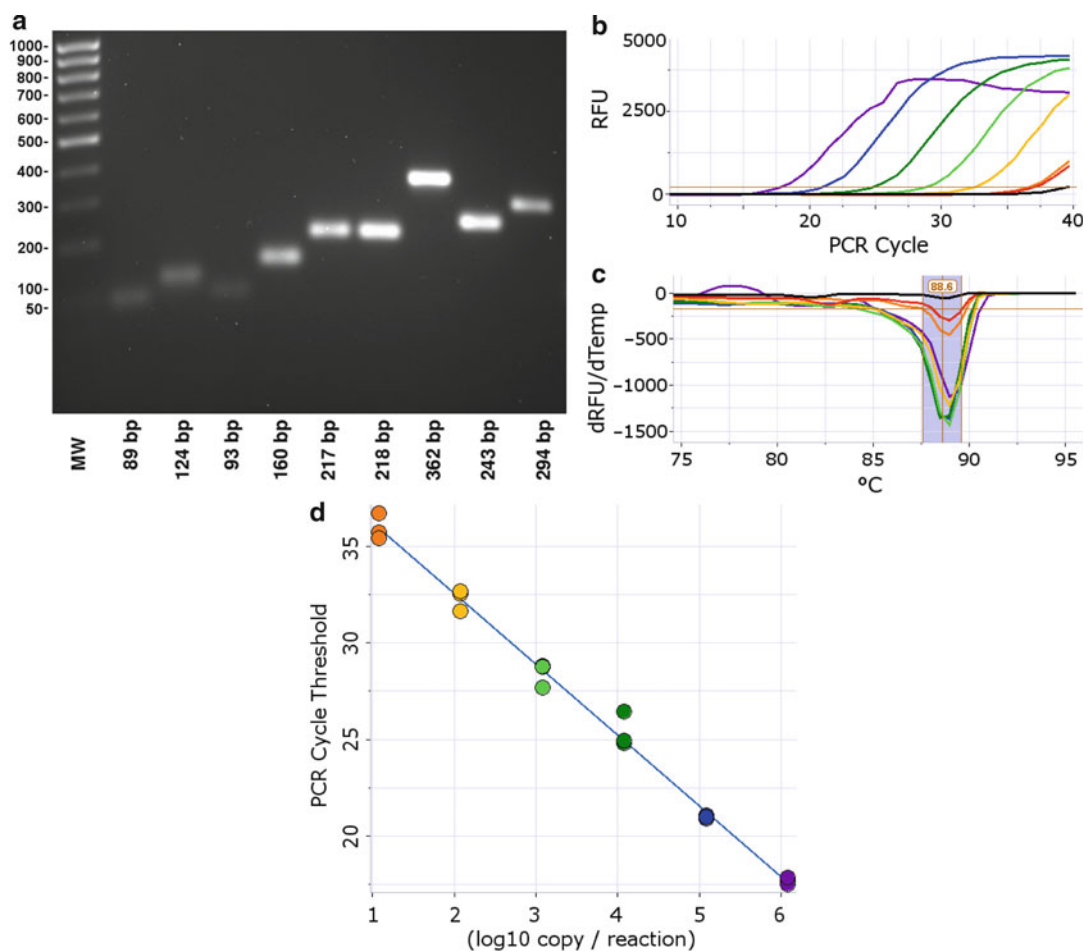
(b) Single-enzyme RT PCR amplifications by PyroPhage 3173 Pol and Tth (Epicentre) were compared using a 160 bp MS2 phage RNA target over a 10²- to 10⁸-fold dilution series. Shown are real-time and post reaction melt data (top) and corresponding end point RT PCR agarose gel (bottom). Tth polymerase was used with Mn2+ as directed. Arrows show correct melt T_m (top) and amplicon (bottom)



Other enzymes have been explored as alternatives to retroviral RTs (e.g., Tth Pol (Rand and Gait 1984)), but none has proven a satisfactory replacement for most methods that rely on reverse transcription of RNA. PyroPhage 3173 is the most efficient Pol for single-enzyme RT PCR and, as such, an alternative to the retroviral RT-dependent methods.

Assembly of Composite Contigs from Viral Metagenomes

One anticipated drawback of using metagenomics as an enzyme discovery tools was the fragmentary nature of the reads, which was expected to hamper efforts to associate subunits of multisubunit enzymes. Many proteins, replicases in particular, function as multiple



Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Fig. 5 Single-enzyme, one-step RT PCR amplification of MS2 phage RNA using 3173 Pol. MS2 RNA was amplified by 40 cycles of RT PCR using the primers shown in Table 1 and 3173 Pol. (a) Products from 89 to 362 bp in length were amplified using one-step single-enzyme RT PCR cycling conditions: 15 s at 94 °C (10 s at 94 °C, 30 s at 72 °C)*40. Products were resolved by 2 % agarose gel

electrophoresis. (b) The MS2 RNA was diluted from 10^1 - to 10^7 -fold and amplified using a primer pair corresponding to the 160 bp fragment in Panel A. Real-time PCR fluorescence in RFU (relative fluorescence units) vs. PCR cycles. (c) Post-amplification thermal melt in $-dRFU/dTemperature$ vs. Temperature (°C). Light blue region indicates melt curves for specific products. (d) Standard curve PCR cycle threshold vs. \log_{10} RNA copy number in triplicate with linear least squares best fit line

subunits. Indeed, the replicases of phages T4, T7, and Phi29 and viruses Mo-MLV, vaccinia, and herpes all function in vivo as multigene replication complexes encoding a number of subunits, e.g., helicases, primases, processivity factors, and clamp loaders (Blanco et al. 1994; Bertram et al. 1998; Goodman 1998; Reha-Krantz et al. 1998; Tang et al. 1998). While, in most cases, the polymerase subunits function

independently in vitro, the utility may be improved by additional subunits. For example, T7 Pol apoenzyme, by itself, has low processivity and was not very effective in Sanger sequencing without its host-derived processivity factor, thioredoxin (Tabor et al. 1987; Tabor and Richardson 1987). Because proteins in replication complexes often have highly specific contacts with one another (Goodman 1998), it is important

that subunits are derived from the same viral genome and not from unrelated viruses.

Because these functionally related genes are often adjacent in operons, it is theoretically possible to identify them given long enough contiguous sequence. Experience shows that operons are almost always too large to be found in the relatively small insert clones seen in typical metagenomic libraries and, without modified assembly rules, are missed. With deep sequencing, these fragments could theoretically be assembled to recover complete viral genomes. In practice, the high degree of sequence polymorphism that characterizes viral metapopulations confounds assembly of related genes and only very limited assembly has been possible by standard protocols.

To accommodate this natural population diversity, assembly stringency was lowered experimentally from the standard 95 % identity to as low as 50 %. Assembly of the YNP Bear Paw (74 °C) and Octopus (93 °C) metagenomes at 50 % identity allowed recovery of composite contigs as large as 35 kb. Fully 7.04 Mb (33 %) of the Octopus reads assembled at this identity into 17 contigs of greater than 10 kb (Schoenfeld et al. 2008). These assemblies appear very reliable in associating orthologous sequences. Particularly in the Octopus library, the sequence reads are evenly distributed throughout the contigs with minimal stacking or other anomalies that would suggest amplification or cloning artifacts. The high numbers of reads on both strands, evenly distributed throughout the contigs, suggest these contigs represent independent clones of closely related genomes. Using the lower stringency assemblies, SNPs can be identified and mapped to the coding sequences. As additional biochemical and structural data become available, molecular diversity may be correlated with variations in function and structure.

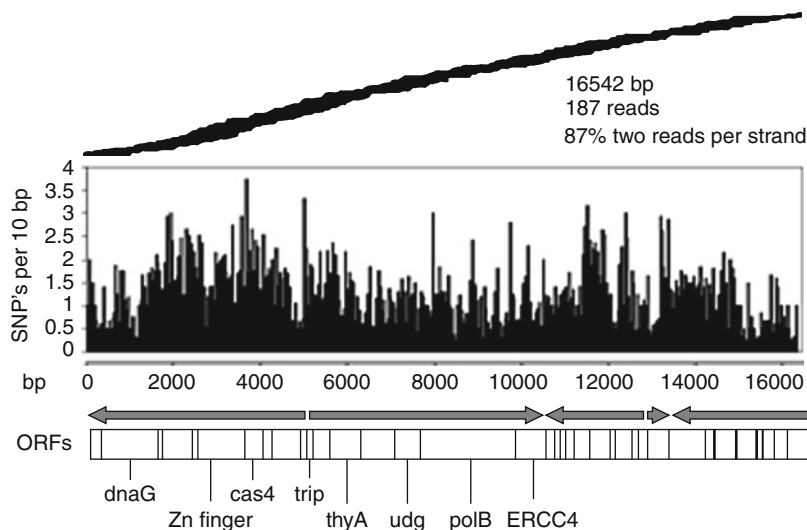
Assembly of a Replication Operon from a Viral Metagenome

One of these contigs provided a unique opportunity to identify potential replicase subunits and associate population diversity of an assembled

metagenome with the biochemistry of the gene products (Fig. 6).

This 16.5 kb contig, assembled at 50 % identity, includes 187 reads (average coverage of 11 reads per nucleotide position). GeneMark (Besemer and Borodovsky 2005) predicted 26 ORFs of greater than 100 nucleotides, which, when translated and annotated by BLASTp, appears to include at least a partial replication operon. The genes with the strongest similarity to four of these ORFs encode two primase subunits, uracil DNA glycosylase, a family B DNA polymerase, and nucleotide excision repair nuclease (*dna G*, *udg*, *pol B*, and ERCC4 genes, respectively). Homologs of these ORFs belong to crenarchaeal DNA replication/repair complexes (Roberts et al. 2003; Dionne and Bell 2005; Barry and Bell 2006). The predicted *pol B* gene showed 28 % identity to *Pyrobaculum islandicus* polB2 (Kahler and Antranikian 2000). Three of the discreet clones that include the *pol B* gene in this contig (PyroPhage 4110, 2783, and 2323 Pols; Fig. 1) have been expressed in *E. coli* to produce a functional thermostable DNA polymerase (data not shown). This contig also contains apparent homologs to a zinc fingerlike protein and a transposon-like integrase/resolvase (*tnp*). Another ORF with highest similarity to the CRISPR-associated sequence *cas4* (Haft et al. 2005) is more likely a separate member of the *cas4* COG, presumably a *recB*-like exonuclease gene.

To correlate the level of sequence divergence with predicted gene function, SNP frequency was calculated and overlaid onto the 50 % assembly consensus sequence of the contig (Fig. 6). Overall distribution of SNPs in the contig was 0.705 per 10 bp. Replication-associated genes showed noticeably lower molecular diversity than the other ORFs. SNP distribution in the *dna G*, *udg*, *pol B*, and ERCC homologs was 0.565, 0.617, 0.569, and 0.548 per 10 bp, respectively, while the distribution in the Zn finger, *cas4*, and *thy A* homologs was 0.979, 1.31, and 0.728, respectively. Finer mapping of this diversity is being used to understand the functional differences in the enzymes encoded by the constituent clones of this contig.



Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Fig. 6 Assembly of a 16.5 kb viral metagenome consensus contig from Octopus hot spring showing single nucleotide polymorphism heterogeneity. (a) 16.5 kb contig was assembled at 50 % identity from the NYP Octopus hot spring library. Sequence coverage is shown on the top, with each line representing a separate read. Single nucleotide

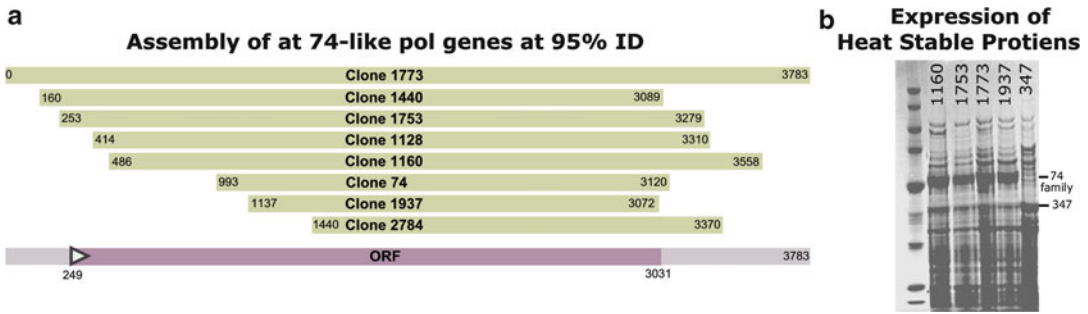
polymorphisms per 10 base pairs were normalized to the number of reads covering the respective nucleotide and are aligned with predicted open reading frames from the consensus sequence in the contig and the gene name of the strongest BLASTx similarity. Direction of transcription is shown by the arrows. Similarities to known genes were identified by BLASTp (Reprinted with permission (Schoenfeld et al. 2008))

Identification of a Replicase Polyprotein from the Great Boiling Spring Metagenome

Based on the large number of highly similar isolates (<3 % amino acid divergence), the PyroPhage 74-like family of *pol A-like* genes from the Great Boiling Spring in Nevada (Fig. 1) appears to be derived from abundant viruses with limited molecular diversity. Unlike the previously described *pol* genes, these were identified by functional screening, precluding the assembly of large contigs. However, this group of *pol* genes proved particularly useful for dissecting the molecular biology of a different replicase. The various polymerase positive clones contain the carboxy terminal half of an apparent polyprotein, but vary in the amount of coding sequence for the amino terminal half (Fig. 7a), implying that the carboxy terminal half of the polyprotein is sufficient for polymerase activity. The polymerase gene appears to be part of an open reading frame that would encode a polyprotein of at least 100 kD. After expression

in *E. coli*, this polyprotein is processed, either in vivo or in vitro, to produce a protein of about 55 kD (Fig. 7b). The amino terminal half of this apparent polyprotein has no known function and no significant sequence similarity to known proteins, but is likely to be associated with replication and, therefore, the target of ongoing investigation.

Polyproteins are common elements used by RNA viruses (Nandakumar et al. 2004). The retroviral reverse transcriptases, for example, are all expressed as polyproteins that are proteolytically processed (Clepet et al. 2004). Heterologous viral polyproteins from hepatitis C have been shown to be active and properly processed in *E. coli* (Yin et al. 2004). However, replicases expressed as polyproteins are much rarer in DNA viruses. PyroPhage 74-family Pol described here and the PyroPhage 3173 Pol described below are the first documented examples of thermophilic phage polyproteins that are actively processed in *E. coli*.



Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Fig. 7 Putative polyprotein from Great Boiling Spring viral metagenome. The PyroPhage 74-like pol genes are aligned to the consensus sequence (Panel A). All of the clones contain the

C-terminal half of a 100 kD ORF, but vary in the amount of N-terminal sequence. Despite differences in the sizes of open reading frames of the inserts, all PyroPhage 74-like clones express a thermostable protein of about 55 kD (Panel B). The 347 clone, in contrast, produces a 35 kD thermostable protein

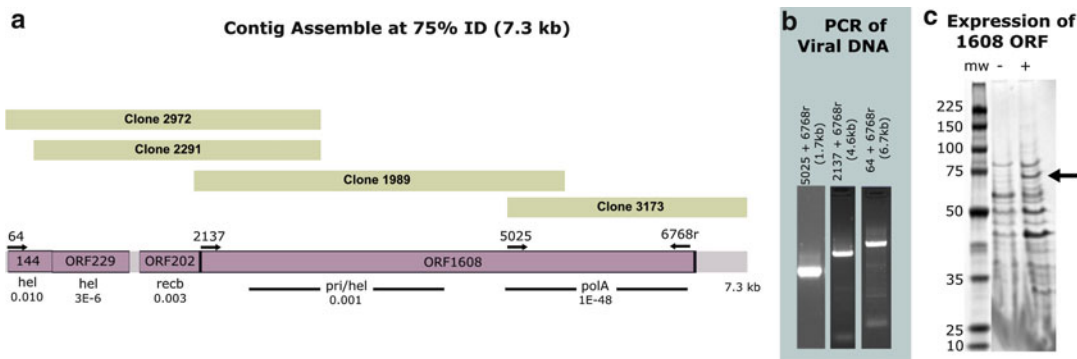
Molecular Biology of the PyroPhage 3173 Replicase Operon

Expression of PyroPhage 3173 Pol, described above, illustrates another challenge in metagenomic-based enzyme discovery. Since, as with all metagenomes, the intact virus has never been cultivated and the sequence data is fragmentary, delineation of the open reading frame of the *pol* gene was unclear. For production and study of the 3173 Pol, expression was initiated at an ATG codon that appeared to be the most probable start site based on alignment to bacterial *pol* genes. Despite the success in using this 55 kD expression product in RT PCR and other applications (see above), anomalies were apparent in the open reading frame that was used for expression of this enzyme. First, there was no obvious adjacent ribosome binding site or transcriptional promoter. Second, there was no homologous ATG codon in the related 488 and 967 clones (Fig. 1), despite overall alignment with the 3173 gene. Finally, an open reading frame extended upstream from the putative start codon to the insertion site of the viral sequence in the cloning vector.

Low identity assembly of the 3173 clone proved useful in dissecting the molecular biology of this gene and allowed production of the complete enzyme corresponding to the likely in vivo product. In contrast to the 4110-like and 74-like polymerase families, the 3173 clone was derived

from a highly divergent, less abundant virus, since reads from this clone failed to assemble at 95 % identity with any other read in the library. Assembly at 75 % identity resulted in a 7,299 nt contig (Fig. 8a), comprised of four reads. This assembly was confirmed by PCR amplification of nearly the entire contig from viral DNA isolated from the same hot spring 4 years later to produce a product of the predicted size (Fig. 8b). This amplification also suggests the 3173-encoding virus is more persistent in the environment than other viral families, none of which was detectable in the later samples. This contig encodes four open reading frames of greater than 100 nt. The largest of these encodes a protein of 1608 amino acids (170 kD), the carboxy terminal portion of which includes the 55 kD PyroPhage 3173 DNA polymerase. The amino terminal portion contains a coding sequence with only weak similarity to known genes. The other open reading frames encode putative helicases and a *cas4/recB* endonuclease protein.

The amplification product of the entire 1608 amino acid ORF expressed in *E. coli* produced an 80 kD protein (Fig. 8c) that co-purified with thermostable DNA polymerase activity. The simplest explanation is that the 1608 amino acid protein (expected MW of 170 kD) is processed in vivo or in vitro to generate the 80 kD product and that the original 55 kD PyroPhage 3173 Pol was a cloning anomaly. Supporting this



Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Fig. 8 Analysis and PCR amplification of a 7.3 kb contig from 75 % NIAID assembly. A 7.3 kb contig was assembled from four clones in the hot springs viral metagenome. GeneMark identified four open reading frames of greater than 100 amino acids, the sizes of which (144, 229, 202, and 1,608 amino acids) are indicated (Panel A). These genes had BLASTx similarity to helicases, cas4 (recB), and DNA polymerases, with the

indicated E values. Primers derived from the assembly are indicated by *arrows* and their positions on the contig are indicated by the associated numbers. These primers were used to amplify viral DNA isolated 4 years after the original collection (Panel B). An amplicon covering the 1608 amino acid ORF (Panel B, lane 2) was used; inserted into an expression system and used to produce an apparent truncation product of ~80 kD, indicated by the *arrow* (Panel C); that co-purified with the Pol activity

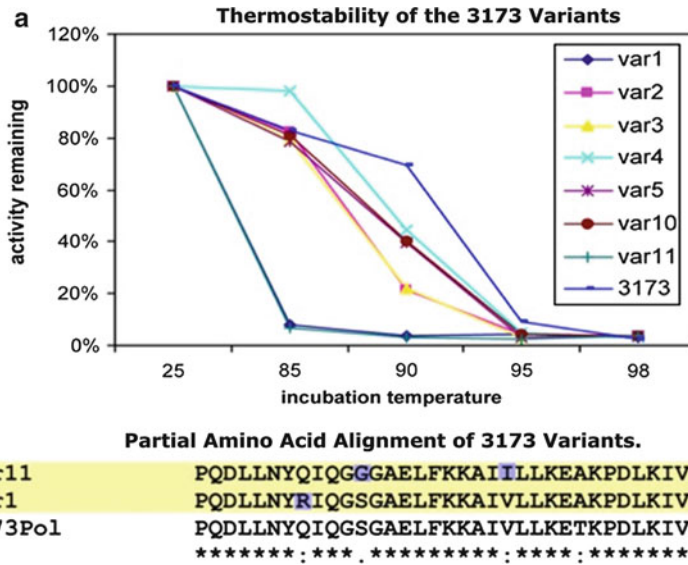
interpretation, amino acids 884 to 894 form the motif AYYILGSIFVE, which was predicted by cleavage site analysis to be both labile to autolytic cleavage and accessible on the surface (Cosstick et al. 1984). Cleavage between G and S would result in a 704 amino acid (80 kD) protein. The amino terminal amino acids from the 80 kD protein aligns with the 5'–3' exonuclease domains of *T. aquaticus* and *E. coli*. The amino acids involved in nucleotide binding are conserved, but not the amino acids required for hydrolysis. Although the 55 kD protein has shown great utility, it is possible that addition of this 25 kD amino terminal sequence, or a portion thereof, would improve its function for certain applications. In addition to the 80 kD Pol protein, the other ORFs are being expressed to reconstitute the presumptive replicase holoenzyme.

This work highlights an important caveat of enzyme discovery by metagenomics. The fragmentary sequences can result in the recovery of partial genes. Assembly of sequences can be the only means of verifying ORFs. In this case, the partial gene proved highly useful, but in many cases, a functional protein could easily be missed by recovery of partial sequences.

Sequence Variants of PyroPhage 3173 DNA Polymerase Isolated from the Viral Metagenome

Metagenomics has proven quite useful for new enzyme discovery. The utility of viral metagenomes is greatly expanded when it is used to guide engineering. One approach to improving DNA polymerases is directed evolution (Ghadessy et al. 2001) based on random mutagenesis. While effective, quite daunting is the sheer number of mutants that must be screened to approach saturation. For an enzyme of the size of *Taq* Pol (832 amino acids), this would require 20^{832} clones to completely saturate the entire gene with mutagenized codons and test all the possible amino acids at each positions. Even a fraction of this number overwhelms any current or conceivable screening capability. To limit the search, algorithms have been developed to target mutagenesis to specific domains (Voigt et al. 2001).

Metagenomic libraries are an alternative to random degenerate libraries as a source of molecular diversity. Since, in native populations, nature selects for active proteins, activities of variants in the libraries may differ, but they should all retain function. To study sequence variants, the 55 kD



Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing, Fig. 9 Thermostability of PyroPhage 3173 Pol variants. The amplification product from Fig. 7b, lane 2, was cloned and expressed to produce thermostable protein. The clones grouped into at least four families that were 97 % identical to one another and 93 % identical to the original clone. The expressed Pol activity

was purified and tested for thermostability by incubating for 10 min at the indicated temperature and assaying using the standard DNA polymerase assay (Panel A). Shown are amino acid alignments of a portion of the Q-helix from the prototype PyroPhage 3173 and the two least thermostable sequence variants (variants 1 and 11) (Panel B). These thermolabile variants had one or two unique amino acids, respectively, that mapped to this region

version of PyroPhage 3173 amplified from viral DNA collected at Octopus hot spring (Fig. 8b) was cloned in an expression vector. Eleven clones were used to express DNA polymerase activity and the inserts were sequenced. The variants were 93 % identical to the original 3173 isolate and at least 97 % identical to one another. When the polymerases were partially purified and tested, they had a significant range of thermostability (Fig. 9). The two most labile enzymes had only one or two unique nucleotide polymorphism each. Two of these independent sequence polymorphisms map within four codons of each other. No three-dimensional structure is available for PyroPhage 3173 Pol, but, based on sequence alignment to *Taq* DNA polymerase and its known protein structure (Kim et al. 1995), the polymorphisms associated with reduced thermostability likely map to the same alpha helix (the Q-helix) within one of the “fingers” of the Pol structure. If so, the two affected amino acids are at the proper spacing to be adjacent on the alpha

helix (four amino acids apart) and likely interact to stabilize or destabilize the alpha helix and thereby alter thermostability.

While a goal of screening hot spring viromes was to find the most thermostable enzymes possible, the lower thermostability variants have value. Isothermal amplification methods such as LAMP (Notomi et al. 2000) use intermediate temperature (i.e., 50–70 °C) and do not require extreme thermostability. Less thermostable enzymes will likely have higher activity at these intermediate temperatures (Giver et al. 1998). Equally important, amino acids that reduce thermostability map to regions that can be targeted to increase thermostability (Bae and Phillips 2004) and are attractive targets for mutagenesis.

Prospects

The focus of the efforts has been discovering and improving thermostable DNA polymerases.

Metagenomics is playing a role in both the discovery and development phases of this project. Viral metagenomics has revealed new replicase operons, thermophilic polyproteins, and entirely new classes of Pols with novel and useful activities for a number of methods of DNA and RNA detection and analysis. In the near future, it may be possible to assemble complete genomes from uncultivated viruses from thermal environments and recover intact replicase operons using the appropriate combination of sequencing strategy, assembly paradigm, and genome walking techniques. The information encoded in the viral metagenomes is being used to direct an enzyme improvement program. Additional applications can likely be improved by the discovery of enzymes other than Pols. In many cases, viral metagenomes are excellent sources of diversity for these discovery programs and presumably any biochemical characteristic that can be measured can be further improved by application of the knowledge gained through metagenomics.

References

- Andraos N, Tabor S, Richardson CC. The highly processive DNA polymerase of bacteriophage T5. Role of the unique N and C termini. *J Biol Chem.* 2004;279(48):50609–18.
- Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The marine viromes of four oceanic regions. *PLoS Biol.* 2006;4(11):e368.
- Babon JJ, McKenzie M, Cotton RG. The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Mol Biotechnol.* 2003;23(1):73–81.
- Bae E, Phillips Jr GN. Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *J Biol Chem.* 2004;279(27):28202–8.
- Baklanov MM, Riazankin IA, Butorin AS, Nechaev Iu S, Iamkovo VI. Purification and characteristics of an RNA-ligase preparation from bacteriophage T4. *Prikl Biokhim Mikrobiol.* 1984;20(2):191–9.
- Barry ER, Bell SD. DNA replication in the archaea. *Microbiol Mol Biol Rev.* 2006;70(4):876–87.
- Bebek A, Dressman HK, Carver GT, Ng S, Petrov V, Yang G, Konigsberg WH, Karam JD, Drake JW. Interacting fidelity defects in the replicative DNA polymerase of bacteriophage RB69. *J Biol Chem.* 2001;276(13):10387–97.
- Beechem JM, Otto MR, Bloom LB, Eritja R, Reha-Krantz LJ, Goodman MF. Exonuclease-polymerase active site partitioning of primer-template DNA strands and equilibrium Mg^{2+} binding properties of bacteriophage T4 DNA polymerase. *Biochemistry.* 1998;37(28):10144–55.
- Beese LS, Steitz TA. Structural basis for the 3'-5' exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. *EMBO J.* 1991;10(1):25–33.
- Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE. Metagenomic characterization of Chesapeake Bay viroplankton. *Appl Environ Microbiol.* 2007;73(23):7629–41.
- Bertram JG, Bloom LB, Turner J, O'Donnell M, Beechem JM, Goodman MF. Pre-steady state analysis of the assembly of wild type and mutant circular clamps of *Escherichia coli* DNA polymerase III onto DNA. *J Biol Chem.* 1998;273(38):24564–74.
- Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 2005;33:W451–4. Web Server issue.
- Bhui-Kaur A, Goodman MF, Tower J. DNA mismatch repair catalyzed by extracts of mitotic, postmitotic, and senescent *Drosophila* tissues and involvement of mei-9 gene function for full activity. *Mol Cell Biol.* 1998;18(3):1436–43.
- Blanco L, Bernad A, Lazaro JM, Martin G, Garmendia C, Salas M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem.* 1989;264(15):8935–40.
- Blanco L, Lazaro JM, de Vega M, Bonnin A, Salas M. Terminal protein-primed DNA amplification. *Proc Natl Acad Sci U S A.* 1994;91(25):12198–202.
- Blondal T, Hjorleifsdottir SH, Fridjonsson OF, Aevarsson A, Skirmisdottir S, Hermannsdottir AG, Hreggvidsson GO, Smith AV, Kristjansson JK. Discovery and characterization of a thermostable bacteriophage RNA ligase homologous to T4 RNA ligase I. *Nucleic Acids Res.* 2003;31(24):7247–54.
- Blumenthal T. Interaction of host-coded and virus-coded polypeptides in RNA phage replication. *Proc R Soc Lond B Biol Sci.* 1980;210(1180):321–35.
- Blumenthal T, Hill D. Roles of the host polypeptides in Q beta RNA replication. Host factor and ribosomal protein S1 allow initiation at reduced GTP concentration. *J Biol Chem.* 1980;255(24):11713–6.
- Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A.* 2002;99(22):14250–5.
- Breitbart M, Wegley L, Leeds S, Schoenfeld T, Rohwer F. Phage community dynamics in hot springs. *Appl Environ Microbiol.* 2004;70(3):1633–40.

- Cheng Q, Nelson D, Zhu S, Fischetti VA. Removal of group B streptococci colonizing the vagina and oropharynx of mice with a bacteriophage lytic enzyme. *Antimicrob Agents Chemother*. 2005;49(1):111–7.
- Cherepanov AV, de Vries S. Binding of nucleotides by T4 DNA ligase and T4 RNA ligase: optical absorbance and fluorescence studies. *Biophys J*. 2001;81(6):3545–59.
- Clepet C, Le Clainche I, Caboche M. Improved full-length cDNA production based on RNA tagging by T4 DNA ligase. *Nucleic Acids Res*. 2004;32(1):e6.
- Compton J. Nucleic acid sequence-based amplification. *Nature*. 1991;350(6313):91–2.
- Cosstick R, McLaughlin LW, Eckstein F. Fluorescent labelling of tRNA and oligodeoxynucleotides using T4 RNA ligase. *Nucleic Acids Res*. 1984;12(4):1791–810.
- Damasko C, Konietzny A, Kaspar H, Appel B, Dersch P, Strauch E. Studies of the efficacy of Enterocolitacin, a phage-tail like bacteriocin, as antimicrobial agent against *Yersinia enterocolitica* serotype O3 in a cell culture system and in mice. *J Vet Med B Infect Dis Vet Public Health*. 2005;52(4):171–9.
- Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*. 2001;11(6):1095–9.
- Detter JC, Jett JM, Lucas SM, Dalin E, Arellano AR, Wang M, Nelson JR, Chapman J, Lou Y, Rokhsar D, Hawkins TL, Richardson PM. Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics*. 2002;80(6):691–8.
- Ding L, Getz G, Wheeler DA, Mardis ER, McLellan DM, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chiriac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstein GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 2008;455(7216):1069–75.
- Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. Functional metagenomic profiling of nine biomes. *Nature*. 2008;452(7187):629–32.
- Dionne I, Bell SD. Characterization of an archaeal family 4 uracil DNA glycosylase and its interaction with PCNA and chromatin proteins. *Biochem J*. 2005;387(Pt 3):859–63.
- Downie AB, Dirk LM, Xu Q, Drake J, Zhang D, Dutt M, Butterfield A, Geneve RR, Corum 3rd JW, Lindstrom KG, Snyder JC. A physical, enzymatic, and genetic characterization of perturbations in the seeds of the brownseed tomato mutants. *J Exp Bot*. 2004;55(399):961–73.
- El Omari K, Ren J, Bird LE, Bona MK, Klarmann G, LeGrice SF, Stammers DK. Molecular architecture and ligand recognition determinants for T4 RNA ligase. *J Biol Chem*. 2006;281(3):1573–9.
- Evans GF, Snyder YM, Butler LD, Zuckerman SH. Differential expression of interleukin-1 and tumor necrosis factor in murine septic shock models. *Circ Shock*. 1989;29(4):279–90.
- Ghadessy FJ, Ong JL, Holliger P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A*. 2001;98(8):4552–7.
- Giver L, Gershenson A, Freskgard PO, Arnold FH. Directed evolution of a thermostable esterase. *Proc Natl Acad Sci U S A*. 1998;95(22):12809–13.
- Goodman MF. Purposeful mutations. *Nature*. 1998;395(6699):221–3.
- Goodman MF, Fyngenson KD. DNA polymerase fidelity: from genetics toward a biochemical understanding. *Genetics*. 1998;148(4):1475–82.
- Griffiths E, Gupta RS. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. *Int Microbiol*. 2004;7(1):41–52.
- Guatelli JC, Whitfield KM, Kwok DY, Barringer KJ, Richman DD, Gingeras TR. Isothermal, in vitro amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. *Proc Natl Acad Sci U S A*. 1990;87(19):7797.
- Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*. 2005;1(6):e60.
- Harrett SP, Lowe G, Tansley G. A stereochemical study of the mechanism of activation of donor oligonucleotides by RNA ligase from bacteriophage T4 infected *Escherichia coli*. *Biochemistry*. 1985;24(25):7446–9.
- Harrison B, Zimmerman SB. Polymer-stimulated ligation: enhanced ligation of oligo- and polynucleotides by T4 RNA ligase in polymer solutions. *Nucleic Acids Res*. 1984;12(21):8235–51.
- Heaphy S, Singh M, Gait MJ. Effect of single amino acid changes in the region of the adenylation site of T4 RNA ligase. *Biochemistry*. 1987;26(6):1688–96.
- Heckler TG, Chang LH, Zama Y, Naka T, Chorghade MS, Hecht SM. T4 RNA ligase mediated preparation of

- novel “chemically misacylated” tRNAPheS. *Biochemistry*. 1984;23(7):1468–73.
- Ho CK, Shuman S. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. *Proc Natl Acad Sci U S A*. 2002;99(20):12709–14.
- Kahler M, Antranikian G. Cloning and characterization of a family B DNA polymerase from the hyperthermophilic crenarchaeon *Pyrobaculum islandicum*. *J Bacteriol*. 2000;182(3):655–63.
- Karam JD, Konigsberg WH. DNA polymerase of the T4-related bacteriophages. *Prog Nucleic Acid Res Mol Biol*. 2000;64:65–96.
- Kerr C, Sadowski PD. Gene 6 exonuclease of bacteriophage T7. I. Purification and properties of the enzyme. *J Biol Chem*. 1972;247(1):305–10.
- Kim Y, Eom SH, Wang J, Lee DS, Suh SW, Steitz TA. Crystal structure of *Thermus aquaticus* DNA polymerase. *Nature*. 1995;376(6541):612–6.
- Koonin EV. Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases. *Biol Direct*. 2006;1:39.
- Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66–72.
- Little JW. Lambda exonuclease. *Gene Amplif Anal*. 1981;2:135–45.
- Loeffler JM, Djurkovic S, Fischetti VA. Phage lytic enzyme Cpl-1 as a novel antimicrobial for pneumococcal bacteremia. *Infect Immun*. 2003;71(11):6199–204.
- Lopatto D, Alvarez C, Barnard D, Chandrasekaran C, Chung HM, Du C, Eckdahl T, Goodman AL, Hauser C, Jones CJ, Kopp OR, Kuleck GA, McNeil G, Morris R, Myka JL, Nagengast A, Overvoorde PJ, Poet JL, Reed K, Regisford G, Revie D, Rosenwald A, Saville K, Shaw M, Skuse GR, Smith C, Smith M, Spratt M, Stamm J, Thompson JS, Wilson BA, Witkowski C, Youngblom J, Leung W, Shaffer CD, Buhler J, Mardis E, Elgin SC. Undergraduate research. Genomics education partnership. *Science*. 2008;322(5902):684–5.
- Lundberg KS, Shoemaker DD, Adams MW, Short JM, Sorge JA, Mathur EJ. High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*. 1991;108(1):1–6.
- Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 2008a;24(3):133–41.
- Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008b;9:387–402.
- Marks JL, Gong Y, Chitale D, Golas B, McLellan MD, Kasai Y, Ding L, Mardis ER, Wilson RK, Solit D, Levine R, Michel K, Thomas RK, Rusch VW, Ladanyi M, Pao W. Novel MEK1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. *Cancer Res*. 2008;68(14):5524–8.
- McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F, Paul JH. Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS One*. 2008;3(9):e3263.
- McLaughlin LW, Piel N, Graeser E. Donor activation in the T4 RNA ligase reaction. *Biochemistry*. 1985;24(2):267–73.
- Merkens LS, Bryan SK, Moses RE. Inactivation of the 5'-3' exonuclease of *Thermus aquaticus* DNA polymerase. *Biochim Biophys Acta*. 1995;1264(2):243–8.
- Middleton T, Herlihy WC, Schimmel PR, Munro HN. Synthesis and purification of oligoribonucleotides using T4 RNA ligase and reverse-phase chromatography. *Anal Biochem*. 1985;144(1):110–7.
- Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V, Mardis ER, Sahinalp SC, Unrau PJ. Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Res*. 2008;18(4):571–84.
- Moser MJ, Difrancesco RA, Gowda K, Klingele AJ, Sugar DR, Stocki S, Mead DA, Schoenfeld TW. Thermostable DNA polymerase from a viral metagenome is a potent rt-PCR enzyme. *PLoS One*. 2012;7(6):e38371.
- Nandakumar J, Shuman S. Dual mechanisms whereby a broken RNA end assists the catalysis of its repair by T4 RNA ligase 2. *J Biol Chem*. 2005;280(25):23484–9.
- Nandakumar J, Ho CK, Lima CD, Shuman S. RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2. *J Biol Chem*. 2004;279(30):31337–47.
- Nelson D, Loomis L, Fischetti VA. Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. *Proc Natl Acad Sci U S A*. 2001;98(7):4107–12.
- Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, Hase T. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res*. 2000;28(12):E63.
- Otto MR, Bloom LB, Goodman MF, Beechem JM. Stopped-flow fluorescence study of precatalytic primer strand base-unstacking transitions in the exonuclease cleft of bacteriophage T4 DNA polymerase. *Biochemistry*. 1998;37(28):10156–63.
- Paulsen H, Wintermeyer W. Incorporation of 1, N6-ethenoadenosine into the 3' terminus of tRNA using T4 RNA ligase. 2. Preparation and ribosome

- interaction of fluorescent *Escherichia coli* tRNAMetf. Eur J Biochem. 1984;138(1):125–30.
- Pavlov AR, Karam JD. Nucleotide-sequence-specific and non-specific interactions of T4 DNA polymerase with its own mRNA. Nucleic Acids Res. 2000;28(23):4657–64.
- Perez LE, Merrill GA, Delorenzo RA, Schoenfeld TW, Vats A, Moser MJ. Evaluation of the specificity and sensitivity of a potential rapid influenza screening system. Diagn Microbiol Infect Dis. 2012;75(1):77–80.
- Petric A, Bhat B, Leonard NJ, Gumport RI. Ligation with T4 RNA ligase of an oligodeoxyribonucleotide to covalently-linked cross-sectional base-pair analogues of short, normal, and long dimensions. Nucleic Acids Res. 1991;19(3):585–90.
- Petruska J, Hartenstine MJ, Goodman MF. Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. J Biol Chem. 1998;273(9):5204–10.
- Pulsinelli GA, Temin HM. Characterization of large deletions occurring during a single round of retrovirus vector replication: novel deletion mechanism involving errors in strand transfer. J Virol. 1991;65(9):4786–97.
- Rand KN, Gait MJ. Sequencing and cloning of bacteriophage T4 gene 63 encoding RNA ligase and tail fibre attachment activities. EMBO J. 1984;3(2):397–402.
- Reha-Krantz LJ, Marquez LA, Elisseeva E, Baker RP, Bloom LB, Dunford HB, Goodman MF. The proof-reading pathway of bacteriophage T4 DNA polymerase. J Biol Chem. 1998;273(36):22969–76.
- Rehrauer WM, Bruck I, Woodgate R, Goodman MF, Kowalczykowski SC. Modulation of RecA nucleoprotein function by the mutagenic UmuD' C protein complex. J Biol Chem. 1998;273(49):32384–7.
- Roberts JA, Bell SD, White MF. An archaeal XPF repair endonuclease dependent on a heterotrimeric PCNA. Mol Microbiol. 2003;48(2):361–71.
- Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, McGrath S, Mardis E, Stein LD. Gallus GBrowse: a unified genomic database for the chicken. Nucleic Acids Res. 2008;36(Database issue):D719–23.
- Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol. 2008;74(13):4164–74.
- Shah JS, Liu J, Buxton D, Hendricks A, Robinson L, Radcliffe G, King W, Lane D, Olive DM, Klinger JD. Q-beta replicase-amplified assay for detection of *Mycobacterium tuberculosis* directly from clinical specimens. J Clin Microbiol. 1995;33(6):1435–41.
- Sharp RL, May PC, Mayne NG, Snyder YM, Burnett JP. Cyclothiazide potentiates agonist responses at human AMPA/kainate receptors expressed in oocytes. Eur J Pharmacol. 1994;266(1):R1–2.
- Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135–45.
- Snyder YM, Guthrie L, Evans GF, Zuckerman SH. Transcriptional inhibition of endotoxin-induced monokine synthesis following heat shock in murine peritoneal macrophages. J Leukoc Biol. 1992;51(2):181–7.
- Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T, Wommack KE. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Res Microbiol. 2008 Jun;159(5):349–57.
- Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol. 1985;39:321–46.
- Strauch E, Kaspar H, Schaudinn C, Damasko C, Konietzny A, Dersch P, Skurnik M, Appel B. Analysis of enterocolitacin, a phage tail-like bacteriocin. Adv Exp Med Biol. 2003;529:249–51.
- Tabor S, Richardson CC. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc Natl Acad Sci U S A. 1987;84(14):4767–71.
- Tabor S, Richardson CC. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. Proc Natl Acad Sci U S A. 1995;92(14):6339–43.
- Tabor S, Huber HE, Richardson CC. *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. J Biol Chem. 1987;262(33):16212–23.
- Tang M, Bruck I, Eritja R, Turner J, Frank EG, Woodgate R, O'Donnell M, Goodman MF. Biochemical basis of SOS-induced mutagenesis in *Escherichia coli*: reconstitution of in vitro lesion bypass dependent on the UmuD'2C mutagenic complex and RecA protein. Proc Natl Acad Sci U S A. 1998;95(17):9755–60.
- Tang H, Yang X, Wang K, Tan W, Li H, He L, Liu B. RNA-templated single-base mutation detection based on T4 DNA ligase and reverse molecular beacon. Talanta. 2008;75(5):1388–93.
- Truncaite L, Zajanckauskaite A, Arlauskas A, Nivinskas R. Transcription and RNA processing during expression of genes preceding DNA ligase gene 30 in T4-related bacteriophages. Virology. 2006;344(2):378–90.
- van Dijk AA, Makeyev EV, Bamford DH. Initiation of viral RNA-dependent RNA polymerization. J Gen Virol. 2004;85(Pt 5):1077–93.
- Voigt CA, Mayo SL, Arnold FH, Wang ZG. Computationally focusing the directed evolution of proteins. J Cell Biochem Suppl. 2001;37:58–63.
- Vratskikh LV, Komarova NI, Yamkovoy VI. Solid-phase synthesis of oligoribonucleotides using T4 RNA ligase and T4 polynucleotide kinase. Biochimie. 1995;77(4):227–32.
- Wang Y, Silverman SK. Efficient RNA 5'-adenylation by T4 DNA ligase to facilitate practical applications. RNA. 2006;12(6):1142–6.

- Wang LK, Schwer B, Shuman S. Structure-guided mutational analysis of T4 RNA ligase 1. *RNA*. 2006;12(12):2126–34.
- Wang LK, Nandakumar J, Schwer B, Shuman S. The C-terminal domain of T4 RNA ligase 1 confers specificity for tRNA repair. *RNA*. 2007;13(8):1235–44.
- Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*. 2008;3(12):e3839.
- Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. *Appl Environ Microbiol*. 2008 Mar;74(5):1453–63.
- Xiang Z, Zhao Y, Mitaksov V, Fremont DH, Kasai Y, Molitoris A, Ries RE, Miner TL, McLellan MD, DiPersio JF, Link DC, Payton JE, Graubert TA, Watson M, Shannon W, Heath SE, Nagarajan R, Mardis ER, Wilson RK, Ley TJ, Tomasson MH. Identification of somatic JAK1 mutations in patients with acute myeloid leukemia. *Blood*. 2008;111(9):4809–12.
- Yin S, Ho CK, Shuman S. Structure-function analysis of T4 RNA ligase 2. *J Biol Chem*. 2003;278(20):17601–8.
- Yin S, Kiong Ho C, Miller ES, Shuman S. Characterization of bacteriophage KVP40 and T4 RNA ligase 2. *Virology*. 2004;319(1):141–51.

G

Genome Atlases, Potential Applications in Study of Metagenomes

Asli Ismihan Ozen¹ and David Wayne Ussery²

¹The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark

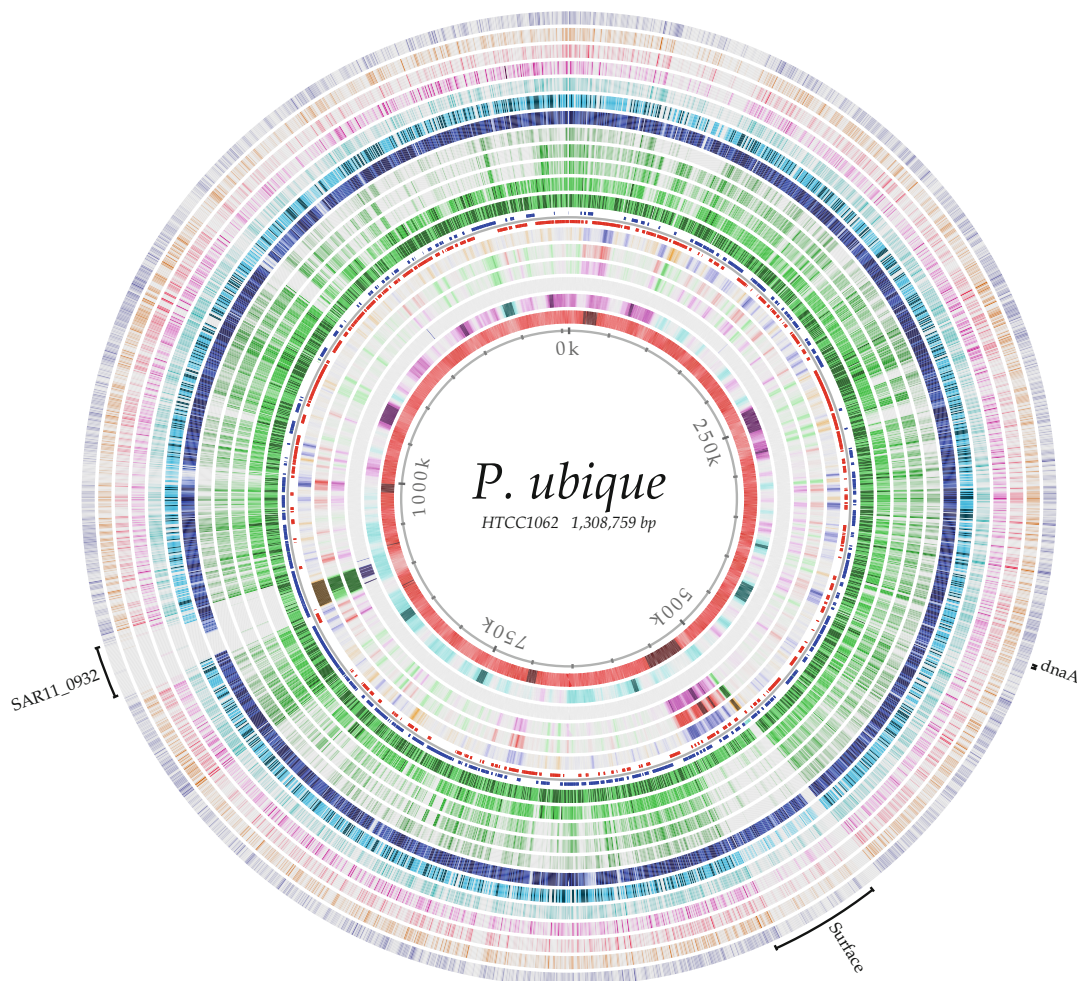
²Bioscience Division of Oak Ridge National Laboratory, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Traditional microbiology has used a single species approach, as in Koch's postulates, where a bacterium is shown to be pathogenic by first isolation from infected organisms, then grown in monoculture, and finally reintroduced into healthy individuals and causing the disease. In contrast, microbial ecology studies multispecies and community structures. Both of these areas have been very successful, and these two different approaches can be seen in comparative genomics, with the traditional analysis of single genomes versus many genomes or metagenomes isolated from an environment. It is possible to relate microbial ecology to reductionist, monoculture microbiology by comparing the two different data types. In this case, the reference is the single genome of an organism, the other being the metagenome samples where most of the DNA in the environment is sampled. Surely, the comparisons are most reliable when

the environmental DNA is preferably in chunks containing at least several genes – from fosmids, longer read lengths, or assembled short reads.

In recent years, there has been many metagenomic data available on public databases such as CAMERA (Sun et al. 2011) or IMG/M (Markowitz et al. 2012). Some of these databases also provide analysis tools as Web servers, e.g., a BLAST (Altschul et al. 1990), or other fast alignment tool is implemented. This allows quick comparison of any sequence data against the metagenomes provided. However, if one is not looking for a single sequence but rather chromosome-wide comparisons, then the interpretation of results might become difficult and complicated. Therefore, a visualization tool such as BLAST Atlas (Hallin et al. 2008) is a very useful way of looking at conservation of proteins in various metagenomic samples, along a given reference chromosome.

Figure 1 is a BLAST Atlas, as an example to illustrate this. An abundant ocean bacterium, *Candidatus Pelagibacter ubique* strain HTCC1062 (Giovannoni et al. 2005) has been chosen as a reference genome, to compare against several genomes and metagenome samples that are found on CAMERA Projects. *P. ubique* is a member of *Alphaproteobacteria*, found in the SAR11 cluster, and known to be a very common inhabitant of marine environments (García-Martínez and Rodríguez-Valera 2000; Brown et al. 2012). It is a free-living cell with a relatively small



Genome Atlases, Potential Applications in Study of Metagenomes, Fig. 1 A BLAST Atlas representing the comparison of marine bacterium *Pelagibacter ubique* to the other four *Pelagibacter* genomes and seven metagenome samples. The six innermost lanes show the DNA properties of the reference genome *P. ubique*

HTCC1062 followed by the genome's annotation lane. Then comes the BLAST lanes, where the BLAST result for the query genome against the reference is shown. The BLAST hit significance is indicated with the color intensity, where higher intensity corresponds to a more significant hit

genome of 1.3 Mbp, first isolated from Saragossa Sea (Giovannoni et al. 1990), and requires added reduced sulfur for growth (Tripp et al. 2008).

The genome comparisons in this study include other *Pelagibacter* species and *Pelagibacterium halotolerans* B2 (Huo et al. 2012). Note the darker green colors for the *P. ubique* lane and for other closely related *Pelagibacter* species. However, apart from the reference strain, there are some regions of missing genes (gaps) that can be seen.

The Metagenome projects that are chosen are Moore Marine Microbial Sequencing (Sun et al. 2011), Global Ocean Sampling (GOS) (Yooshep et al. 2007), Whale Fall (Tringe et al. 2005), Acid Mine Drainage (Tyson et al. 2004), Microbial Community Genomics at the HOT/ALOHA (DeLong et al. 2006), Waseca County Farm Soil (Tringe et al. 2005), and Washington Lake (Kalyuzhnaya et al. 2008). In all comparisons, the *P. ubique* proteins were compared against the metagenomes using the BLAST

tool of the database itself with default parameters, and the results are then visualized with BLAST Atlas. Moore Marine Microbes, GOS, and HOT/ALOHA samples have protein annotations; therefore, a BLASTP search was used. The other metagenomes are assembled but not annotated, so TBLASTN comparison was made. Metagenomes that are not assembled were not used in this study, because protein comparison against metagenome reads was not very reliable.

In the BLAST Atlas, the six innermost lanes show some of the DNA properties (Jensen et al. 1999; Pedersen et al. 2000) of the reference chromosome, *P. ubique* HTCC1062; these are, from innermost to outermost: the average AT percentage (over a 10,000 bp average), GC Skew (10,000 bp average), Global Direct Repeats, Nucleosome Position Preference (green regions represent chromatin-free areas; Satchwell et al. 1986; Baldi et al. 1996), DNA helix stacking energy (on this scale, red regions will melt more readily, and green regions are more stable; Ornstein et al. 1978), and intrinsic curvature (blue means highly curved areas, and yellow indicates low levels of curvature; Bolshoy et al. 1991; Shpigelman et al. 1993). The next outer lane is the annotations, coding sequences on plus and minus strand. After the annotations the BLAST lanes start, which show the BLAST hits on each position. The color intensity indicates how good a BLAST hit is, with darker colors representing regions of conserved proteins and grey areas contain poor or no matches. The first BLAST lane is *P. ubique* itself as a control. The next few lanes are other *Pelagibacter sp.*, and they show high resemblance to the reference *P. ubique*. The 5th lane is a *Pelagibacterium* which should not be mixed because it is classified as a completely different clade in *Alphaproteobacteria*, as can be seen from the low protein similarity. However its BLAST hit profile still resembles the other *Pelagibacter sp.*

According to this figure, we can see that almost all the coding genes of *P. ubique* are found in the CAMERA Marine Microbes samples, and most are also found in the GOS data, which means that the bacterium is present in these environments, as expected. One of the gap

regions around 510–564 kb contains the genes that are related to amino sugar metabolism (*rfaD, rfaE*), pentose phosphate pathway (*tktC*), lipopolysaccharide synthesis (*gmhA* and *gmhB*), streptomycin biosynthesis (*rpbB*), and transferase activity (*spsA, rfaG, rfaK*). This gap region and a few bases downstream is marked as “surface” because this area contains proteins related to surface features (*ompS*, LPS biosynthesis, etc.). Another gap includes a “giant protein” (Strom et al. 2012), annotated as “hypothetical protein SAR11_0932,” and is 7,317 amino acid residues long. The reason why this protein seems to be partially found in Marine Microbes and GOS metagenomes (dark blue lane) is due to the many repeat regions in the protein, which might look like other regions in the proteins of the other genomes. But the whole protein itself is not found because it varies even within the same species; these “giant proteins” are known to be variable and thought to be involved in protection against viral attacks, as well as predation by protists (Strom et al. 2012). Some of the other gaps are due to tRNA or rRNAs, because the BLAST lanes only compare protein sequences. When looked at the other metagenome BLAST lanes, the BLAST hits are seen very weak meaning that *P. ubique* genes that are compared here are not present in those metagenome samples.

In summary, BLAST Atlas is a way to visualize the mapping of bacterial genomes against metagenomes, and this can be used to compare many different environments. If a certain protein, a set of proteins, or a genomic region is being investigated, this tool will guide in finding the presence or absence of those proteins. It is also possible to zoom in to desired ranges of the genome to see local differences (Hallin et al. 2008).

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
- Baldi P, Brunak S, Chauvin Y, Krogh A. Naturally occurring nucleosome positioning signals in human exons and introns. *J Mol Biol.* 1996;263(4):503–10.

- Bolshoy A, McNamara P, Harrington RE, Trifonov EN. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc Natl Acad Sci USA*. 1991;88:2312–6.
- Brown MV, Lauro FM, DeMaere MZ, et al. Global biogeography of SAR11 marine bacteria. *Mol Syst Biol*. 2012;8:595.
- DeLong EF, Preston CM, Mincer T, et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*. 2006;311(5760):496–503.
- García-Martínez J, Rodríguez-Valera F. Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of group I. *Mol Ecol*. 2000;9(7):935–48.
- Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*. 1990;345(6270):60–3.
- Giovannoni SJ, Tripp HJ, Givan S, et al. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*. 2005;309(5738):1242–5.
- Hallin PF, Binnewies TT, Ussery DW. The genome BLAST atlas – a GeneWiz extension for visualization of whole-genome homology. *Mol Biosyst*. 2008;4(5):363–71.
- Huo Y-Y, Cheng H, Han X-F, et al. Complete genome sequence of *Pelagibacterium halotolerans* B2(T). *J Bacteriol*. 2012;194(1):197–8.
- Jensen LJ, Friis C, Ussery DW. Three views of microbial genomes. *Res Microbiol*. 1999;150(9–10):773–7.
- Kalyuzhnaya MG, Lapidus A, Ivanova N, et al. High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol*. 2008;26(9):1029–34.
- Markowitz VM, Chen I-MA, Chu K, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*. 2012;40-(Database issue):D123–9.
- Ornstein RL, Rein R, Breen DL, MacElroy R. An optimised potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers*. 1978;17:2341–60.
- Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW. A DNA structural atlas for *Escherichia coli*. *J Mol Biol*. 2000;299(4):907–30.
- Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol*. 1986;191(4):659–75.
- Shpigelman ES, Trifonov EN, Boishoy A. Curvature: software for the analysis of curved DNA. *Comput Appl Biosci*. 1993;9:435–40.
- Strom SL, Brahamsha B, Fredrickson KA, Apple JK, Rodríguez AG. A giant cell surface protein in *Synechococcus* WH8102 inhibits feeding by a dinoflagellate predator. *Environ Microbiol*. 2012;14(3):807–16.
- Sun S, Chen J, Li W, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res*. 2011;39(Database issue):D546–51.
- Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. *Science*. 2005;308(5721):554–7.
- Tripp HJ, Kitner JB, Schwalbach MS, et al. SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature*. 2008;452(7188):741–4.
- Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 2004;428(6978):37–43.
- Yooseph S, Sutton G, Rusch DB, et al. The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol*. 2007;5(3):e16.

Genome Portal, Joint Genome Institute

Igor V. Grigoriev, Susannah Tringe and Inna Dubchak
US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

Synonyms

Comparative genomics; Data integration; Genome analysis; Genome projects; Metagenomics

Definition

The US Department of Energy (DOE) Joint Genome Institute (JGI) is a national user facility with massive-scale DNA sequencing and analysis capabilities dedicated to advancing genomics for bioenergy and environmental applications.

The JGI Genome Portal is an integrated genomic resource, which provides for the research community around the world access to the large collection of genomic data for plants, fungi, microbes, and metagenomes and to web-based interactive tools for their analysis.

Introduction

The Department of Energy (DOE) Joint Genome Institute (JGI) was established for the Human

Genome Project (Lander et al. 2001) and later was transformed into a national user facility for genome research in the DOE mission areas of bioenergy, carbon cycling, and biogeochemistry. JGI provides expertise and resources in DNA sequencing, technology development, and bioinformatics to the broader scientific community. Scientists around the world can make proposals to the JGI Community Sequencing Program (CSP; e.g., Martin et al. 2011) to sequence genomes, transcriptomes, and metagenomes and address important scientific questions of DOE mission relevance. Massive amounts of genomic data are assembled, annotated, and delivered to users by means of integrated databases and interactive analytical tools interconnected within the JGI Genome Portal (<http://genome.jgi.doe.gov>; Grigoriev et al. 2012).

Leading the world in the number of sequenced plants, fungi, microbes, and metagenomes (according to the Genomes Online Database (GOLD; Pagani et al. 2012)), JGI has dramatically increased its sequencing capabilities using new sequencing technologies. JGI projects evolved from sequencing three of the human chromosomes (Lander et al. 2001) to the large-scale “Grand Challenge” projects such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA; Wu et al. 2009), the 1,000 Fungal Genome Project (Grigoriev et al. 2011), and the metagenomic projects targeting soil and rhizosphere. Since tracking individual organisms and samples at such a scale becomes critical, genomes and metagenomes sequenced or selected for sequencing are carefully catalogued and made available to the public along with their status and links to the produced data and available tools.

The sequenced data are assembled, annotated, and analyzed using various computational pipelines developed for each of the products delivered by JGI to its users. The resulting annotations are available for download and also can be interactively viewed using the JGI Genome Portal offering a wide array of databases and analytical systems to interpret the data. Some systems work across multiple JGI databases, while others allow users to specifically manage datasets on

plants (Phytozome; Goodstein et al. 2012), fungi (MycoCosm; Grigoriev et al. 2012), microbes (Integrated Microbial Genomes or IMG; Markowitz et al. 2012b), and metagenomes (IMG/M; Markowitz et al. 2012a).

The JGI Genome Portal provides a unified access point to all JGI genomic databases and analytical tools, as well as worldwide statistics on the usage of the JGI resources and the information about the latest genome releases and new tool development. A user can find all DOE JGI sequencing projects and their status, search for and download raw data, assemblies and annotations of sequenced genomes and metagenomes, and interactively explore those datasets and compare them with other sequenced microbes, fungi, plants, or metagenomes using specialized systems tailored to each particular class of organisms. All these can serve as building blocks in comprehensive analyses of individual organisms or systems of interacting organisms.

A Catalogue of Genome Sequencing Projects

Metagenomic analysis requires reference genomes for better interpretation of sequence data derived from complex microbial communities. The democratization of sequencing allows many scientists to sequence appropriate genome references in their own labs prior to approaching metagenomes. Consolidation of genomic data sequenced in different places around the world is an important step in both genomics and metagenomics.

JGI’s collection of genomic projects includes thousands of projects of different types and is publicly available and searchable. Product types include standard or improved genome drafts, finished genomes, gene expression profiling, resequencing, metagenome projects, and others. The *Project List* (<http://genome.jgi.doe.gov/genome-projects>) is available from most of the Portal pages as a menu item and includes a detailed description of each project including its scope and current status, taxon, the JGI program, and the project lead. The Resources

column lists tools available for this project. Some of these tools, e.g., *download*, are available for all genomes, while others are taxon, project type, or stage dependent. For example, a plant or fungal genome will be linked to Phytozome or MycoCosm, respectively.

All JGI projects are also registered in the GOLD database, which includes a larger collection of projects sequenced around the world (Pagani et al. 2012). Currently it contains a list of about 16,000 genomes including over 3,000 that are complete and over 2,000 metagenomes. Besides utility for metagenomics, having a comprehensive list of sequencing projects from all laboratories around the world also helps to avoid redundancy when sequencing targets are selected for the large-scale projects like GEBA or 1,000 Fungal Genomes.

Annotated Genomes and Metagenomes

Finding genes in metagenomes is challenging, especially for eukaryotes with their complex intron-exon gene structure and often relies on gene prediction based on similarity to proteins from other organisms. This requires a comprehensive collection of genes from different organisms across all domains of life. Besides the human genome (Lander et al. 2001), JGI sequenced and annotated genomes of the first poplar tree (Tuskan et al. 2006) and its ectomycorrhizal symbiont (Martin et al. 2008); lignocellulose degrading fungi (Berka et al. 2011; Eastwood et al. 2011) and microbial communities (Hess et al. 2011); diverse eukaryotes, often the first representatives of the Tree of Life branches (Tyler et al. 2006; Bowler et al. 2008; King et al. 2008; Fritz-Laylin et al. 2010; Colbourne et al. 2011); and prokaryotes (Wu et al. 2009) as well as soil (Tringe et al. 2005) and ocean metagenomes (Walsh et al. 2009). There are over 3,000 annotated reference genomes in the JGI database and three ways to find a particular genome of interest: using an interactive *Tree of Life*, *search*, and *select* functions.

The Tree of Life organizes the annotated genomes by the domains of life and links to

Organism home pages. Clicking on a branch name produces a menu displaying available genomes in this kingdom, phylum, class, or order (Fig. 1), each connected to pages in different analytical resources. The same pages can be reached in a step-by-step genome selection from a hierarchical selection menu on the top of the page or searching for genomes by keyword (e.g., plants, Eukaryota), name, taxonID, or projectID.

Each of the genomic datasets can be analyzed with a collection of tools linked directly to their genome databases. Each organism's home page contains a description of the project, BLAST, download, and links to specialized resources as described in the next section.

Comparative Databases and Tools

Comparative genomics is a more powerful approach for functional annotation and evolutionary studies of genomes than analysis of individual genome sequences. It is also a primary method for annotation and analysis of metagenomes. The JGI Genome Portal includes a set of efficient comparative tools, such as gene clustering, whole-genome alignment, and building phylogenetic trees that are used across different genomic resources at JGI. *VISTA Point* (<http://genome.lbl.gov/vista>) is an example of such tools. It was designed for visualization and analysis of pairwise and multiple DNA alignments (Frazer et al. 2004) at different levels of resolution in three visualization modes: (a) *VISTA Browser*, for visual comparative analysis of complete genome assemblies using pairwise and multiple large-scale alignments; (b) *VISTA Synteny Viewer*, a multi-tiered graphical display of pairwise alignments at three different levels of resolution; and (c) *VistaDot*, an interactive two-dimensional dot-plot genome synteny viewer across multiple chromosomes/scaffolds. Several specialized domain-specific computational systems for comparative genome analysis built at JGI include Phytozome, a comparative hub for plant genome and gene family data and analysis; MycoCosm to enable users to navigate across sequenced fungal

[Home](#) | [Project List](#) | [Login](#)

all JGI Organisms and Acetohalobium arabaticum Z-7288, DSM 5501 Search Genomes

Archaea
 Crenarchaeota
 Korarchaeota
 Euryarchaeota
 unclassified Archaea

Eukaryota
 Fungi
 Metazoa
 Choanozoa
 Viridiplantae
 Heterobosha
 Heterokonta
 Rhizaria
 Cryptophyta
 Haptophyta

Bacteria
 Bacteroidetes
 Acidobacteria
 Aquificae
 Deinococcus-Thermus
 Thermodesulfobacteria
 Alphaproteobacteria
 Betaproteobacteria
 Gammaproteobacteria
 Deltaproteobacteria
 Epsilonproteobacteria
 Fibrobacteres
 Chlamydiae
 Verrucomicrobia
 unclassified Bacteria

Metagenomes
 Engineered
 Marine
 Fresh Water
 Thermal Springs

Tree of Life drawing by Lella H. [SITE](#) [BLAST](#) [Download](#)
 Anaerobic methane oxidation (AOX) community from Eel River Basin sediment, California
 Coastal water and sediment microbial communities from Arctic, Canada
 Oceanic planktonic microbial community from high methane PC12-225-485cm (High methane PC12-225-485cm Dec. 2010 assembly)

Integrated Microbial Genomes (IMG) and Metagenomes (IMG/M) - resources for comparative analysis and annotation of all publicly available genomes. [Credits](#) [Disclaimer](#) [Comments/Questions](#) [Site Map](#)

What's New? [Download](#) [Help](#)
 The JGI does not distribute genomic DNA, strains or clones used in the sequencing projects. DNA, eukaryotic strains, bacterial strains should be obtained by contacting the collaborator for a specific organism listed on the [Info](#) page of the organism.

Genome Releases
 • [Fungal Releases](#)
 • [Metagenomics Releases](#)
 • [Microbial Releases](#)
 • [Plant Releases](#)

From this site you can get details about our current and upcoming projects. We also provide a [list of projects](#) for your convenience.

Individual genomes sites can be reached using the [Tree of Life](#). These sites provide download access to sequence files and tools for annotation and BLAST alignment. Eukaryotic genome sites include additional resources such as genome browsing, search, and functional (KEGG/GENCO) annotations.

Get started using the genome portal with our [genome tutorial](#) and [genome help](#) for Eukaryotes.

Fungal Genomics Program
Metagenomics Program
Microbial Genomics Program
Plant Genomics Program

Who's visiting the portal?
 Click on the map to see how the portal community is spread across the globe.

Phytozome - a comparative hub for green plant genomes and gene family data and analysis.

Mycocosm - the Fungal Genomics Resource. Provides access to the annotated fungal genomes and tools for their analysis.

DOE Joint Genome Institute | [Credits](#) | [Disclaimer](#) | [Comments/Questions](#) | [Site Map](#)
 © 1997-2012 The Regents of the University of California.
 Genome Portal version 0.2.205.10511 content:10511 content:10511 content:3428 system: JGI | [jgi-joint-genome-institute.org](#) | [caltech.edu](#) | [lbl.gov](#) | [pcr.gov](#)

Genome Portal, Joint Genome Institute, Fig. 1 The JGI Genome Portal. A pull-down menu for the “Marine” category of Metagenomes is shown. *BLAST* and *Download* functions are available for the entire selected group. Each genome is linked to the associated resources. “Project list” on the top leads users to the list of all sequencing projects at the DOE JGI. The bottom portion of the page connects to the specialized databases in microbes (IMG) and metagenomes (IMG/M), fungi (MycocoCosm), and plants (Phytozome)

genomes and to conduct comparative and genome-centric analyses and community annotation; and the IMG family of tools for large-scale comparative analysis of microbial genomes and metagenomes.

Phytozome (<http://phytozome.net>; Goodstein et al. 2012) gives access to the sequences and functional annotations of a growing number of complete plant genomes (31 in release v8.0), including land plants and selected algae. Phytozome provides both organism-centric and gene family-centric views as well as access to the BLAST, BLAT, and Search capabilities.

Phytozome provides a view of the evolutionary history of every plant and every plant gene at the level of sequence, gene structure, gene family, and genome organization. The Phytozome project organizes the proteomes of green plants into gene families defined at the nodes on the green plant evolutionary tree. Genes have been annotated with PFAM, KOG, KEGG, and PANTHER assignments, and publicly available annotations from RefSeq, UniProt, TAIR, and JGI are hyperlinked and searchable. The gene family view gives access to the information on each family and its members, organized to highlight shared attributes.

GBrowse provides genome-centric views for all genomes included in Phytozome. Each organism browser displays a number of tracks including a gene prediction track, a track of homologous sequences from related species aligned against the genome, supporting EST and VISTA tracks identifying regions of this genome that are syntenic with other plant genomes.

MycCosm (<http://jgi.doe.gov/fungi>; Grigoriev et al. 2012) brings together genomic data and analytical tools for diverse fungi that are important for energy and environment. Genomic data from the JGI and its users are integrated and curated via user community participation in data submission, curation, annotation, and analysis. Over 150 newly sequenced and annotated fungal genomes are available to the public through MycoCosm for *genome-centric* and *comparative* analyses. Visual navigation across the MycoCosm tree (Fig. 2b), where each node represents a group of phylogenetically related fungi and is linked to analysis tools, allows users to redefine the search

and analysis space from a single organism to the entire list of fungal genomes.

The Genome browser with configurable selection of tracks displays predicted gene models and annotations along with different lines of evidence in support of these predictions, such as gene and protein expression profiles. Gene models and annotations are linked to community annotation tools to revise them if needed. Functional profiles of each genome summarize gene annotations according to the GO, KEGG, and KOG classifications and can be compared with each other to study gene family expansions or contractions at different levels of granularity. Clustering using BLAST alignments of all proteins and MCL can expand these analyses to gene families even without annotation and enable side-by-side comparison of each of the cluster members for pattern of protein domains, intron-exon structure, and synteny.

MycoCosm comparative views combine the abovementioned tools to study entire groups of genomes corresponding to MycoCosm nodes. Unlike the genome-centric view, there is no reference genome in this analysis, and, for example, a keyword or BLAST search for protein kinases in *Basidiomycota* or *Ascomycota* will show differences in the number of found genes or BLAST hits across different members of these phyla.

IMG, the Integrated Microbial Genomes database (<http://img.jgi.doe.gov>; Markowitz et al. 2012a, b), is a system designed for flexible comparative analyses of microbial genomic data, which incorporates all complete public microbial genomes as well as those sequenced at JGI. IMG with microbiome samples (IMG/M) is an expanded database that includes metagenome data from diverse environments, both sequenced at JGI and submitted by external users.

In addition to importing all public genomes and their annotations from NCBI's RefSeq, IMG curates the data by adding features missed by many annotation pipelines, such as small RNAs; assigning proteins and domains to all major protein family databases (e.g., COG, TIGRFam); and linking to organism metadata stored in GOLD, such as oxygen requirements or environment of origin. Annotations can be viewed in detailed gene pages or summarized in genome pages that

a

Species > Tools > Info > Help > Contact Us

phytozome

Nitrogen-fixing node selected

2. Choose a tool:

Keyword search expand

BLAST search submit

Target Nitrogen-fixing gene families

BLAST program: BLASTP - protein query to protein db

Query name: (optional)

Query sequence: enter manually upload file

Show results in browser (default) Notify by email when job completes (long jobs)

Algorithm parameters

BLAT search

Genome browser

Info page

Bulk data

Select all species

JGI Joint Genome Institute ©2011 University of California Regents. All rights reserved. Center for Integrative Genomics

b

by keyword Search Genome

MycoCosm the fungal genomics resource

Video Tutorials

Fungi

Basidiomycota

Dikarya

Ascomycota

Pezizomycotina

Pucciniomycotina

Ustilaginomycotina

Agaricomycotina

Pezizomycetes

Eurotiomycetes

Dothideomycetes

Lecanoromycetes

Leotiomycetes

Sordariomycetes

Saccharomycotina

Taphrinomycotina

Glomeromycotina

Mucoromycotina

Mucoromycetes

Zoopagomycotina

Entomophthoromycotina

Kickxellomycotina

Blastocladiomycotina

Chytridiomycotina

Neocallimastigomycotina

Microsporidia

Mucoromycotina

Search

BLAST

Clusters

Mucor circinelloides CBS277.49 v2.0

Phycomyces blakesleeanus NRRL1555 v2.0

Rhizopus oryzae 99-880 from Broad

Genome Portal, Joint Genome Institute, Fig. 2 (continued)

c

img/m INTEGRATED MICROBIAL GENOMES with MICROBIOME SAMPLES

IMG/M Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart My IMG Companion Systems Using IMG/M

IMG/M Genomes

	Total
Bacteria	2901
Archaea	119
Eukarya	121
Plasmids	1182
Viruses	2697
Metagenomes	353
All Genomes	7378
GEBA	235

Metagenome Projects Map
What's New
Using IMG
IMG/M Addendum
About IMG
FAQ

Hands on training available at the
Microbial Genomics & Metagenomics Workshop

Database updated: 2012-04-24
Next IMG release: June 2012

IMG/M (Nucleic Acids Research, Vol 40, 2012) provides tools for analyzing the functional capability of microbial communities based on their metagenome sequence, in the context of reference isolate genomes included from the Integrated Microbial Genomes (IMG) system. This version of IMG/M contains all the isolate genomes from IMG 3.4 (Nucleic Acids Research, Vol 40, 2012) as well as genomes from the GEBA project, which provide a comprehensive context for metagenome data analysis.

IMG/M contains metagenome data generated from

Metagenome samples (unique sample names)	351
Metagenome studies (unique proposal names)	94
New samples since 3.5 (Mar 2012)	138
New studies since 3.5 (Mar 2012)	0

IMG/M Statistics IMG Family Systems

Metagenome Environment Category

- Engineered
- Environmental
- Host-associated

Environmental

Version 3.5 January 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer imo-edoe1 2012-04-27-10.18.35

JGI U.S. DEPARTMENT OF ENERGY Office of Science GOLD

Genome Portal, Joint Genome Institute, Fig. 2 Comparative genomic resources at JGI: (a) Phytozome for plants, (b) MycoCosm for fungi, and (c) IMG family of tool

include organism metadata in addition to statistics on genome size and gene counts within various categories.

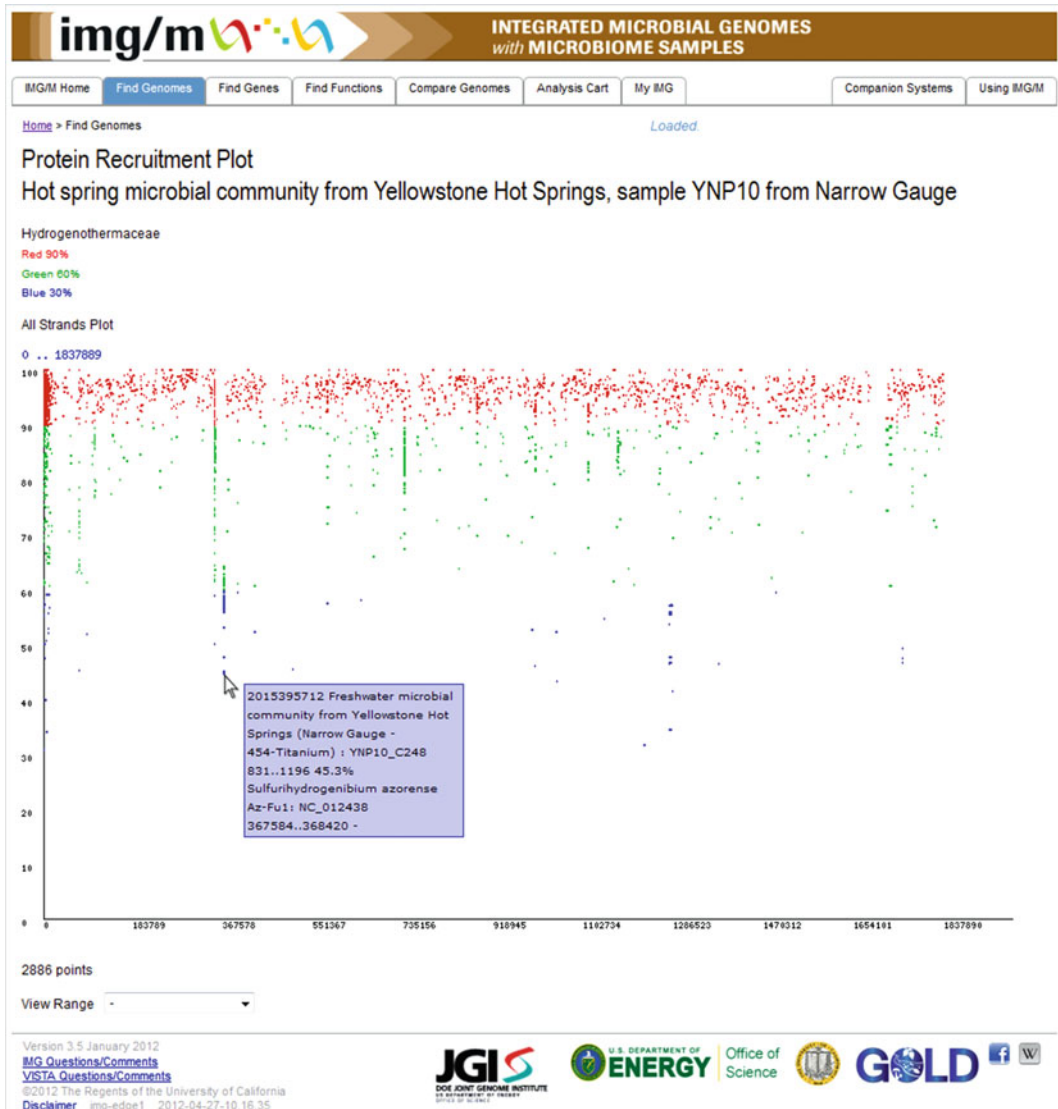
The tools available in IMG allow for analyses at the gene, function, or genome level, using customizable “carts” for each of these data types. Thus, any given analysis can readily be performed on a single (meta)genome or several and can be extended to many individual genes, functions, or pathways. IMG/M includes a number of metagenome-specific functions, including the option to account for different organism abundances by weighting comparative analyses according to estimated gene copies, based on the contig read coverage reported in the assembly rather than simple gene counts.

It also includes a “scaffold cart” for exploring genes within a given set of contigs or scaffolds as well as the option to categorize contigs/scaffolds into population “bins” based on oligonucleotide composition or other features.

Recent developments in IMG and IMG/M include the capacity to add and view (meta) transcriptome and (meta)proteome data in the context of a reference and compare expression profiles across experiments.

Metagenome Analysis

Analysis of metagenome data presents a number of challenges beyond those faced in isolate



Genome Portal, Joint Genome Institute, Fig. 3 Metagenomic analysis. A protein recruitment plot showing alignment of genes from a hot spring sample to genomes from the family *Hydrogenothermaceae*

genome analysis, in particular the wide variation in individual organism abundances and the shallow coverage of low-abundance, but nonetheless biologically important, taxa. Both of these tend to result in highly fragmented assemblies, which are most readily interpreted when high-quality reference genome data are available.

Most metagenome analyses approach the data from either a phylogenetic perspective (i.e., who is

there?) or a functional one (i.e., what are they doing?). Each of these uses a specific suite of tools, though nearly all rely on a well-curated database of genes with known phylogenies and functions. For phylogenetic analysis, genes or gene fragments are assigned to phylogenetic lineages based on homology to genes of known phylogenetic origin. This can be done for all genes from a metagenome dataset, for example, using

MEGAN (Huson and Mitra 2012), or for a set of conserved phylogenetic markers which can be placed onto a tree of known sequences from isolate genomes and/or amplified from uncultivated organisms, for example, using pplacer (Matsen et al. 2012). IMG/M allows for both approaches – an overall perspective of all the genes in a dataset or on a specific set of contigs is provided through the “Phylogenetic Distribution of Genes” option on the main metagenome page or in the scaffold cart, and genes with homology to particular phyla, families, genera, or species can be retrieved. When there are good reference genomes available, alignments of protein-coding genes to those genomes can be viewed in a recruitment plot (Fig. 3). Phylogenetic marker genes can also be extracted and incorporated into trees using the “Phylogenetic Marker COGs” option under the “Find Functions” tab.

Functional or “gene-centric” approaches enable the comparison of metagenome datasets at the functional level to both assess their relative similarity and identify genes or functions that are over- or underrepresented in a given dataset. This type of approach is utilized by metagenome analysis systems like MG-RAST (Meyer et al. 2008). IMG/M provides several options for whole metagenome comparisons. Metagenomes can be clustered (under the “Compare Genomes” tab) according to gene content, using either functional (e.g., COG, Pfam) or phylogenetic criteria, and the results visualized via hierarchical clustering, principal components analysis (PCA), or a correlation matrix. Relative abundances of specific gene families can be viewed via the abundance profile function also under the “Compare Genomes” tab. As mentioned above, these comparisons can be made between partly assembled genomes by taking contig read depth into account when calculating gene abundance.

Summary

Technological innovations leading to the democratization of genome sequencing have resulted in

large amounts of genomic data being produced in different parts of the world. Effective analysis of genomic and metagenomic data depends on the availability of comprehensive catalogues of reference genome data for annotation and comparative genomics as well as computational tools able to process the large amounts of sequence data. The JGI Genome Portal (<http://genome.jgi.doe.gov>) provides a unified access point to all JGI genomic databases and analytical tools including list of sequencing projects at JGI and around the world, a comprehensive collection of annotated genomes in all domains of life, and specialized databases for comparative analysis of plant, fungal, and microbial genomes and metagenomes. The latter is still in early stages of development, and data generated at unprecedented scale and complexity for metagenomes will require new approaches to data processing, analysis, and visualization.

References

- Berka RM, Grigoriev IV, Otiillar R, et al. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. *Nat Biotechnol.* 2011;29:922–7.
- Bowler C, Allen AE, Badger JH, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature.* 2008;456:239–44.
- Colbourne JK, Pfrender ME, Gilbert D, et al. The ecoresponsive genome of *Daphnia pulex*. *Science.* 2011;331:555–61.
- Eastwood DC, Floudas D, Binder M, et al. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science.* 2011;333:762–5.
- Frazer KA, Pachter L, Poliakov A, et al. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 2004;32:W273–9.
- Fritz-Laylin LK, Prochnik SE, Ginger ML, et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell.* 2010;140:631–42.
- Goodstein DM, Shu S, Howson R, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 2012;40:D1178–86.
- Grigoriev IV, Cullen D, Goodwin SB, et al. Fueling the future with fungal genomics. *Mycology.* 2011;2:192–209.

- Grigoriev IV, Nordberg H, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* 2012;40: D26–32.
- Hess M, Sczyrba A, Egan R, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science.* 2011;331:463–7.
- Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol.* 2012;856:415–29.
- King N, Westbrook MJ, Young SL, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature.* 2008;451:783–8.
- Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860–921.
- Markowitz VM, Chen IM, Chu K, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* 2012a;40: D123–9.
- Markowitz VM, Chen IM, Palaniappan K, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* 2012b;40: D115–22.
- Martin F, Aerts A, Ahren D, et al. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature.* 2008;452:88–92.
- Martin F, Cullen D, Hibbett D, et al. Sequencing the fungal tree of life. *New Phytol.* 2011;190:818–21.
- Matsen FA, Hoffman NG, Gallagher A, et al. A format for phylogenetic placements. *PLoS One.* 2012;7: e31009.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9:386.
- Pagani I, Liolios K, Jansson J, et al. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 2012;40:D571–9.
- Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. *Science.* 2005;308:554–7.
- Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science.* 2006;313:1596–604.
- Tyler BM, Tripathy S, Zhang X, et al. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science.* 2006;313: 1261–6.
- Walsh DA, Zaikova E, Howes CG, et al. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science.* 2009;326: 578–82.
- Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature.* 2009;462:1056–60.

Genome-Based Studies of Marine Microorganisms

Xinqing Zhao¹, Chao Chen², Liangyu Chen², Yumei Wang² and Xiang Geng²

¹School of Life Science and Biotechnology, Dalian University of Technology, Dalian, People's Republic of China

²Dalian University of Technology, Dalian, China

Synonyms

Genome mining of marine microorganisms

Definition

Genome-based studies of marine microorganisms mean utilizing genetic information retrieved from genomic sequences of marine microorganisms to guide the discovery of useful enzymes and natural products from marine microorganisms. Chemical structures of natural products potentially synthesized by marine microorganisms can be predicted by aligning the biosynthetic genes with known gene sequences that are responsible for the biosynthesis of natural products, and the physicochemical properties (UV spectrum, molecular weight, polarity, etc.) obtained from the prediction can be used to guide further purification and structure elucidation of the compounds. In case that the interested genes or gene clusters are not expressed or express in low level, various methods can be employed to activate the expression of biosynthetic genes. Identification of target natural products can be achieved by comparative metabolic profiling, heterologous expression, and other genome-mining strategies. For unculturable or yet-uncultured marine microbes in given environments, metagenomic, metatranscriptomic, and metaproteomic sequences can be employed. Function-based or

sequence-based screening of metagenomic libraries is subsequently performed to identify novel enzymes and natural products.

Introduction

Marine microorganisms are important sources for novel natural products and industrial enzymes, and many unique small molecules and proteins produced by marine microorganisms have been reported in the recent years, which facilitate novel drug discovery, agricultural biocontrol, as well as industrial applications. In case of marine natural products, it has been clear that vast diversity of chemistry can be explored from marine microorganisms, mainly including marine bacteria and marine fungi (Imhoff et al. 2011). However, bioassay-guided screening of natural products has limitations in identification of compounds with novel functions that are not readily assayed, as well as in the discovery of novel compounds which exist in low amount, or even not be produced under normal culture conditions. In addition, some marine microbes may grow very slowly under laboratory conditions or unculturable using currently available methods. Therefore, it is important to develop new strategies to fully explore the biosynthetic potential of marine microorganisms.

The development of high-throughput sequencing technologies has facilitated the exploration of the full biosynthetic potential of marine microorganisms. It has become increasingly evident through the analysis of abundantly available genomic sequences and metagenomic sequences that microorganisms have much greater potential than we expected to produce various metabolites. It was estimated by comparing the known secondary metabolite and the analysis of the genomic sequences of several actinobacteria that as much as 90 % of the biosynthetic potential of actinomycetes remains undiscovered (Wilkinson and Micklefield 2007). The available genomic sequences of marine microorganisms enable us to rapidly identify useful enzymes and natural products by genome mining.

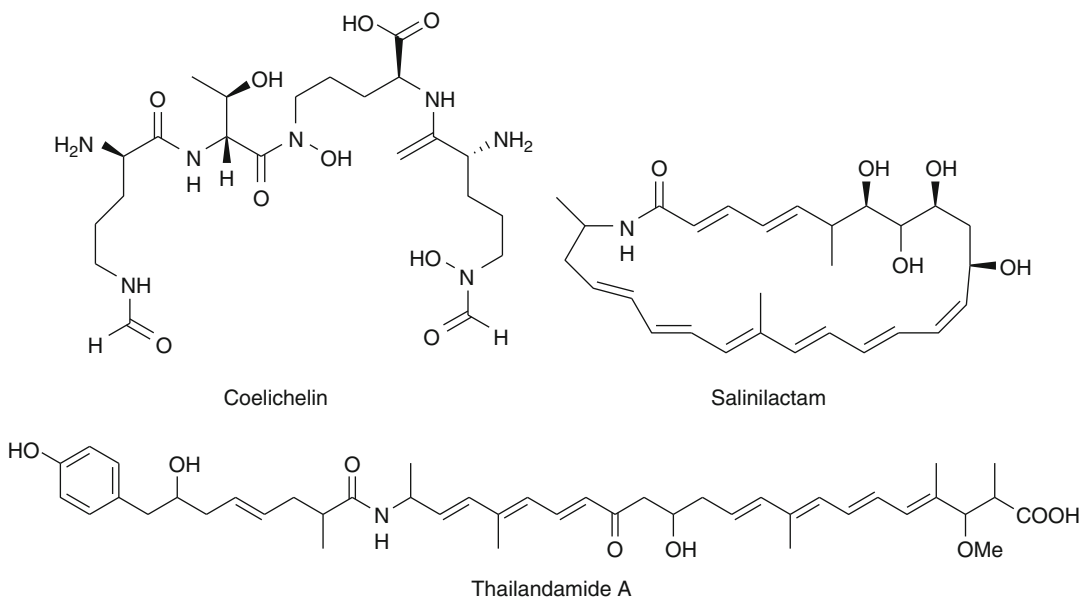
Genome Mining for Natural Product Discovery

Typically, biosynthetic genes of small molecules in microorganisms are clustered together in the genome to form gene clusters, and bioinformatic analysis allows the rapid identification of gene clusters similar to the known ones, thus speeding up the discovery of natural products. Genome mining involves prediction of biosynthetic potential of organisms by analyzing their genomic sequences, followed by screening or activation of enzymes and natural product biosynthesis by process optimization and/or genetic manipulations (Scheffler et al. 2013). Two types of small molecules encoded by multimodular polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) have been extensively focused. The biosynthesis of many polyketides and non-ribosomal peptides follows a colinearity rule and is assembled based on the number and type of domains within the enzymes, which makes it possible to predict the molecule structures (Winter et al. 2011; Nikolouli and Mossialos 2012).

Similar to genome scanning method (Zazopoulos et al. 2003), genome mining has the limitation that only the genes with similar functions to those of known ones are focused; new or unusual pathways are poorly explored. However, the presence of PKS and NRPS genes is good indication of natural products with possible broad spectrum of activities (Nikolouli and Mossialos 2012).

Genome mining of microorganisms was first started in 2000, with the identification of coelichelin as one of the first examples (Challis and Ravel 2000), while the first compound identified in marine actinobacteria by genome mining is the polyene macrolactam salinilactam A from *Salinispora tropica* (Udwary et al. 2007). The structures of coelichelin and salinilactam A were shown in Fig. 1.

Various genome-mining techniques have been reviewed elsewhere (Scheffler et al. 2013). Prediction of gene functions and chemical structures can be achieved using computer programs such as BLAST and THREADER, as well as other useful



Genome-Based Studies of Marine Microorganisms, Fig. 1 Structures of compounds discovered by genome mining

bioinformatic tools such as antiSMASH and NP-searcher (Nikolouli and Mossialos 2012). Due to the limited knowledge on enzymatic functions and metabolic cross talks, the prediction of chemical structures is not always correct, and accurate annotation of gene functions and prediction of chemical structures requires more advanced bioinformatic tools.

In case that the biosynthetic genes are actively expressed under lab conditions, information on the physicochemical properties of the target molecules such as UV spectrum, molecular weight, and polarity obtained from the bioinformatic prediction can be used to guide the further purification of the compounds. Thailandamide A was discovered by genome mining of *Burkholderia thailandensis* (Nguyen et al. 2008). Being temperature and light sensitive and also being produced in the early growth stage, thailandamide A may not have been identified using classical methods without the genomic-guided isolation (Nguyen et al. 2008). The structure of thailandamide A was shown in Fig. 1.

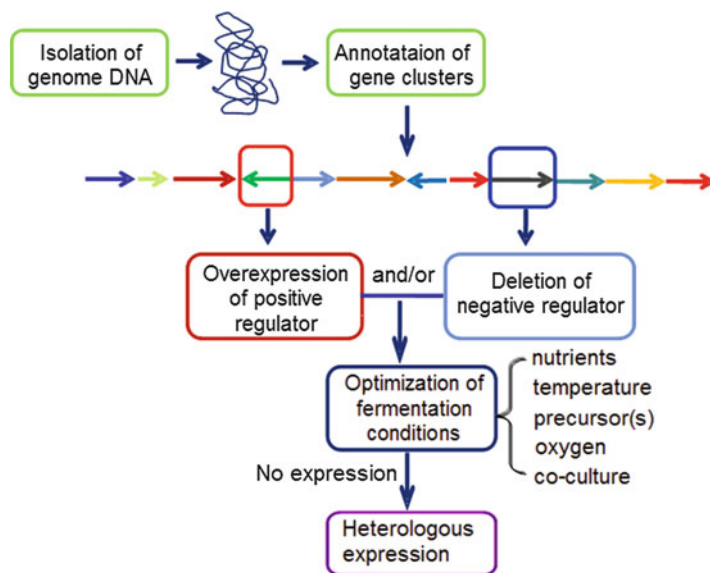
Genom isotopic approach was first described with the discovery of orfamides from *Pseudomonas fluorescens* (Gross et al. 2007), which stable

isotope amino acid precursors feeding into the culture broth and subsequent detection of the labeled molecule to identify NRPS or mixed PKS/NRPS compounds.

Although low-level production of target molecules can be identified by genom isotopic method, some metabolites are only produced under special circumstances; activation of production of these molecules requires mimicking specific nutritional, environmental, and biological conditions, such as special carbon and nitrogen source, high temperature, UV irradiation, osmotic stress treatments, and coculture with another microbial strain (Scherlach and Hertweck 2009). In addition, genetic methods can also be employed to activate production of certain metabolites identified by genome mining, including overexpression of activation regulators and deletion of repressive regulators (Scheffler et al. 2013). Heterologous expression of the entire gene cluster in well-defined host strains, including *E. coli*, *Streptomyces*, *Bacillus*, and *Saccharomyces cerevisiae*, has also been employed in genome mining (Zhang et al. 2011). Selection of suitable host strains and expression vectors are critical to achieve

Genome-Based Studies of Marine Microorganisms,

Fig. 2 Genome mining for identification of natural products



heterologous production of target active molecules. Scheme of genome mining was depicted in Fig. 2.

Metatranscriptomic and Metaproteomic Studies for Discovery of Novel Enzymes and Small Molecules

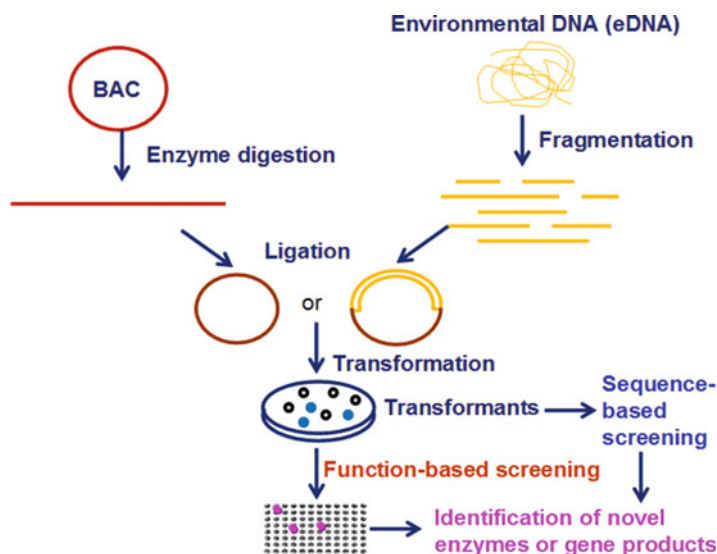
In addition to culture-dependent genome-mining studies, genome-based discovery of novel enzymes and natural products from environmental samples can also be achieved using culture-independent tools. It has been estimated that less than 1 % of the bacteria in most environmental samples are culturable (reviewed by Brady et al. 2009), and it is thus important to study the yet-uncultured microorganisms in marine environment. Metagenome stands for a collection of genetic materials (genomic DNA) of a mixed community of organisms recovered directly from given environmental samples. Environmental DNA (eDNA) extracted from marine sediments, seawater, or marine sponges, plants, or animals can serve as starting point for metagenomic studies. Metagenomic DNA is cloned into various host cells, the most popular host being *E. coli*. Phenotypic-based screening and DNA sequencing-based screening of

metagenomic libraries yield positive clones with aimed sequences (reviewed by Brady et al. 2009). Novel enzymes such as laccase, aromatic hydrocarbon dioxygenase, and halogenase have been isolated from marine metagenomic studies (Fang et al. 2011; Marcos et al. 2012; Bayer et al. 2013, reviewed by Kennedy et al. 2011), which have great potential for industrial applications and environmental bioremediation. In addition, novel natural products were also identified in metagenomic libraries (reviewed by Brady et al. 2009), and *Streptomyces* and *Ralstonia metallidurans* were used as hosts for heterologous expression of metagenomic library. Metagenome mining of symbiotic bacteria of marine sponge *Theonella swinhoei* resulted in the discovery of polytheonamides which are extensively posttranslationally modified ribosomal peptides (Freeman et al. 2012). Metagenomic workflow was illustrated in Fig. 3.

Metatranscriptomic and metaproteomic studies focus on the expression of certain genes in a given environment at a given time (Schweder et al. 2008; Stewart et al. 2012) and have been used to characterize metabolic behavior of microbial community. Such techniques have not been employed to study the isolation of novel enzymes

Genome-Based Studies of Marine Microorganisms,

Fig. 3 Metagenomic method to discover novel natural products or enzymes



and small molecules from marine environment. In comparison to metagenomic studies, metatranscriptomics and metaproteomics overlook genes that are not expressed in certain time and thus have limitation to fully explore the biosynthetic potential of marine microorganisms. However, same problems of silent gene expression can also be encountered when the metagenomic libraries are propagated in certain host cells; therefore, choosing diverse host cells and testing various conditions for expression of metagenomic libraries are important to identify novel enzymes and small molecules in marine environment.

Summary

Genome mining has speeded up the discovery of natural products and novel enzymes from microorganisms by exploring their full biosynthetic potentials. Metagenomic studies combined with genome mining promote the advancement of studies of yet-uncultured marine microorganisms. The discovery of marine natural products and novel enzymes using genome-based methods is still in its early stage; however, development of genome mining and metagenomic approaches

will facilitate discovery of more novel marine enzymes and natural products for biotechnological applications.

Acknowledgments The authors are regretful for not being able to cite more references due to space limitation.

References

- Bayer K, Scheuermayer M, Fieseler L, Hentschel U. Genomic mining for novel FADH(2)-dependent halogenases in marine sponge-associated microbial consortia. *Mar Biotechnol* (NY). 2013;15(1):63–72.
- Brady SF, Simmons L, Kim JH, Schmidt EW. Metagenomic approaches to natural products from free-living and symbiotic organisms. *Nat Prod Rep*. 2009;26(11):1488–503.
- Challis GL, Ravel J. Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett*. 2000;187(2):111–4.
- Fang Z, Li T, Wang Q, Zhang X, Peng H, Fang W, Hong Y, Ge H, Xiao Y. A bacterial laccase from marine microbial metagenome exhibiting chloride tolerance and dye decolorization ability. *Appl Microbiol Biotechnol*. 2011;89:1103–10.
- Freeman MF, Gurgui C, Helf MJ, Morinaka BI, Uria AR, Oldham NJ, Sahl HG, Matsunaga S, Piel J. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science*. 2012;338(6105):387–90.

- Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. *Chem Biol.* 2007;14(1):53–63.
- Imhoff JF, Labes A, Wiese J. Bio-mining the microbial treasures of the ocean: new natural products. *Biotechnol Adv.* 2011;29(5):468–82.
- Kennedy J, O’Leary ND, Kiran GS, Morrissey JP, O’Gara F, Selvin J, Dobson ADW. Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *J Appl Microbiol.* 2011;111(3):787–99.
- Marcos MS, Lozada M, Di Marzio WD, Dionisi HM. Abundance, dynamics, and biogeographic distribution of seven polycyclic aromatic hydrocarbon dioxygenase gene variants in coastal sediments of Patagonia. *Appl Environ Microbiol.* 2012;78(5):1589–92.
- Nguyen TA, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T, Taudien S, Platzer M, Hertweck C, Piel J. Exploiting the mosaic structure of *trans*-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol.* 2008;26(2):225–33.
- Nikolouli K, Mossialos D. Bioactive compounds synthesized by non-ribosomal peptide synthetases and type-I polyketide synthases discovered through genome-mining and metagenomics. *Biotechnol Lett.* 2012;34:1393–403.
- Scheffler RJ, Colmer S, Tynan H, Demain AL, Gullo VP. Antimicrobials, drug discovery, and genome mining. *Appl Microbiol Biotechnol.* 2013;97(3):969–78.
- Scherlach K, Hertweck C. Triggering cryptic natural product biosynthesis in microorganisms. *Org Biomol Chem.* 2009;7:1753–60.
- Schweder T, Markert S, Hecker M. Proteomics of marine bacteria. *Electrophoresis.* 2008;29:2603–16.
- Stewart FJ, Ulloa O, DeLong EF. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol.* 2012;14(1):23–40.
- Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc Natl Acad Sci.* 2007;104(25):10376–81.
- Wilkinson B, Micklefield J. Mining and engineering natural-product biosynthetic pathways. *Nat Chem Biol.* 2007;3(7):379–86.
- Winter JM, Behnken S, Hertweck C. Genomics-inspired discovery of natural products. *Curr Opin Chem Biol.* 2011;15(1):22–31.
- Zazopoulos E, Huang K, Staffa A, Liu W, Bachmann BO, Nonaka K, Ahlert J, Thorson JS, Shen B, Farnet CM. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat Biotechnol.* 2003;21(2):187–90.
- Zhang H, Boghigian BA, Armando J, Pfeifer BA. Methods and options for the heterologous production of complex natural products. *Nat Prod Rep.* 2011;28:125–51.

GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities

Zhili He¹, Joy D. Van Nostrand¹ and Jizhong (Joe) Zhou^{1,2,3}

¹Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

²Department of Environmental Science and Engineering, Tsinghua University, Beijing, China

³Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Synonyms

Functional gene array; Metagenomic technology

Definition

Functional gene arrays (FGAs) are a special type of microarray containing probes for key genes involved in microbial functional processes, such as biogeochemical cycling of carbon, nitrogen, sulfur, phosphorus, and metals, biodegradation of environmental contaminants, antibiotic resistance, energy processing, and stress response. GeoChips are considered to be the most comprehensive FGAs and an important metagenomic tool for microbial community analysis.

Introduction

Microorganisms are the most diverse group of organisms known in terms of phylogeny and functionality. However, they do not live alone but form distinct communities and play integrated and unique roles in ecosystems, such as biogeochemical cycling of carbon (C), nitrogen (N), sulfur (S), phosphorus (P), and metals (e.g., iron, copper, zinc), biodegradation or stabilization of environmental contaminants, and

interaction with hosts. Therefore, one of the most important goals of microbial ecology is to understand the diversity, composition, structure, function, dynamics, and evolution of microbial communities and their relationships with environmental factors and ecosystem functioning. Toward this goal, several challenges remain. First, microorganisms are generally too small to see or characterize with most approaches used for plant or animal studies. Second, microbial communities are extremely diverse. It is estimated that 1 g of soil contains 2,000–50,000 microbial species (Torsvik et al. 2002) and even up to millions of species (Gans et al. 2005). Third, a vast majority of microorganisms (>99 %) are uncultured (Whitman et al. 1998), making it difficult to study their functional ability and molecular mechanisms. Finally, establishing mechanistic linkages between microbial diversity and ecosystem functioning is even more difficult. To address these challenges, culture-independent, high-throughput technologies for analysis of microbial communities are necessary.

Indeed, many culture-independent approaches are available including PCR-based cloning analysis, denaturing gradient gel electrophoresis (DGGE), terminal-restriction fragment length polymorphism (T-RFLP), quantitative PCR, and in situ hybridization. However, these methods only provide snapshots of a microbial community but fail to provide a comprehensive view. Therefore, high-throughput metagenomic technologies are necessary for providing a rapid, specific, sensitive, and quantitative analysis of microbial communities and their relationships with environmental factors and ecosystem functioning.

Microarray-based technology can examine thousands of genes at one time, providing a much more comprehensive analysis of microbial communities. This technology, like GeoChip, has been developed and adopted to analyze microbial communities (He et al. 2007, 2010a; Hazen et al. 2010) and has been used to profile the functional diversity, composition, structure, and dynamics of microbial communities from different habitats (He et al. 2011, 2012a, b). A variety of studies demonstrate that microarrays can provide phylogenetic and functional

information on a microbial community in a rapid, high-throughput, and parallel manner.

This overview is focused on the analysis of functional diversity, structure, and activity of microbial communities using GeoChip-based metagenomic technologies but also includes a brief introduction of GeoChips, GeoChip development, and GeoChip hybridization and data analysis.

GeoChips as the Most Comprehensive Functional Gene Arrays

Functional gene arrays (FGAs) are special microarrays containing probes for key genes involved in microbial functional processes, such as biogeochemical cycling of carbon (C), nitrogen (N), phosphorus (P), sulfur (S), and metals, antibiotic resistance, biodegradation of environmental contaminants, energy processing, and stress response. Since the exact functions of selected genes on FGAs are known, this type of array is especially useful for examining the functional diversity, composition, and structure of microbial communities across different times and scales. Several FGAs have been reported and evaluated, and they generally target specific functional processes, populations, or environments, including *nodC* and *nifH* arrays, a methanotroph gene (*pmoA*) array, a virulence marker gene (VMG) array, pathogen detection/diagnosis arrays, and a bioleaching array (He et al. 2012b). However, GeoChips are the most comprehensive FGAs to date, especially the later versions (GeoChips 2.0, 3.0, and 4.0), which target a variety of key microbial functional processes, such as C, N, P, and S cycling, contaminant bioremediation, and antibiotic resistance (He et al. 2012a).

GeoChips, constructed with 50-mer oligonucleotide probes, have evolved over several generations. The prototype GeoChip contained 89 PCR-amplicon probes for N-cycling genes (*nirS*, *nirK*, *amoA*, and *pmoA*) derived from pure-culture isolates and marine sediment clone libraries (Wu et al. 2001). The first-generation GeoChip (GeoChip 1.0) was constructed with

763 gene variants involved in nitrogen cycling (*nirS*, *nirK*, *nifH*, *amoA*), methane oxidation (*pmoA*), and sulfite reduction (*dsrAB*). Then, an expanded array was developed with 2,402 genes involved in organic contaminant biodegradation and metal resistance to monitor microbial populations and functional genes involved in biodegradation and biotransformation (Rhee et al. 2004). Specificity evaluation with representative pure cultures indicated that the designed probes appeared to be specific to their corresponding target genes. The detection limit was 5–10 ng of genomic DNA in the absence of background DNA and 50–100 ng of pure-culture genomic DNA in the presence of background DNA. Real-time PCR analysis was very consistent with the microarray-based quantification (He et al. 2011).

Although the prototype and GeoChip 1.0 arrays were used to probe specific functional groups or activities, they lacked a truly comprehensive probe set covering key microbial functional processes occurring in different environments. Therefore, more comprehensive GeoChips have been developed and evaluated. For example, GeoChip 2.0, containing 24,243 (50-mer) oligonucleotide probes, targeting ~10,000 functional gene variants from 150 gene families involved in the geochemical cycling of C, N, and P, sulfate reduction, metal reduction and resistance, and organic contaminant degradation, was developed as the first comprehensive FGA (He et al. 2007). After 2 years, GeoChip 3.0 was developed, which contained about 28,000 probes and targeted ~57,000 sequences from 292 gene families (He et al. 2010a). GeoChip 3.0 is more comprehensive and has several other distinct features compared to GeoChip 2.0, such as a common oligo reference standard (CORS) for data normalization and comparison, a software package for data management and future updating, the *gyrB* gene for phylogenetic analysis, and additional functional groups including those involved in antibiotic resistance and energy processing (He et al. 2010a). Based on GeoChip 3.0, GeoChip 4.0 was developed, which contains ~84,000 probes and targeting >152,000 genes from 410 functional families.

GeoChip 4.0 not only contains all functional categories from GeoChip 3.0 but also includes additional functional categories, such as genes from bacterial phages and those involved in stress response and virulence (Hazen et al. 2010; He et al. 2012a). All evaluation and studies demonstrate that GeoChip is a powerful tool for specific, sensitive, and quantitative analysis of microbial communities from a variety of habitats (He et al. 2011, 2012a, b).

GeoChip Development

GeoChip development involves several major steps, including selection of target genes, sequence retrieval and verification, oligonucleotide probe design, probe validation, and array construction as well as future automatic update, which are generally implemented by a GeoChip development and data analysis pipeline (<http://ieg.ou.edu/>) (He et al. 2010a).

Selection of Target Genes and Sequence Retrieval

A variety of functional genes can be used as functional markers targeting different processes, such as biogeochemical cycling of C, N, S, P, and metals, contaminant bioremediation, antibiotic resistance, and stress response. For example, 292 functional gene families were selected for GeoChip 3.0 with 41 for C cycling, 16 for N cycling, 3 for P utilization, 4 for S cycling, 173 for biodegradation of a variety of organic contaminants, 41 for metal reduction and resistance, 11 for antibiotic resistance, and 2 for energy processing. In addition, a phylogenetic marker (*gyrB*) was also chosen (He et al. 2010a). More importantly, when sequences for a known functional gene are available, they can be added in an updated GeoChip. For example, when GeoChip was updated to GeoChip 4.0, functional gene families involved in stress responses, bacterial phages, and virulence were added, resulting in 410 functional gene families on GeoChip 4.0 (Hazen et al. 2010; He et al. 2012a). Generally, genes are chosen for key enzymes or proteins with the corresponding

function(s) of interest. If a process involves multiple steps or a protein complex, those genes responsible for catalytic subunits or with the active site(s) will be selected (He et al. 2011).

Sequence retrieval is performed generally with a pipeline with a database integrated for managing all retrieved sequences and subsequently designed probes. For each functional gene, the first step is to submit a query to the GenBank protein database and fetch all candidate amino acid sequences. The key words may include the name of the target gene/enzyme, its abbreviation and enzyme commission number (EC), and affiliated domains of bacteria, archaea, and fungi. Second, retrieved sequences are validated by seed sequences (those sequences that have been experimentally confirmed to produce the protein of interest and that the protein functions as expected) with the HMMER program. Finally, all confirmed protein sequences are searched against GenBank again to obtain their corresponding nucleic acid sequences for probe design (He et al. 2010a).

Oligonucleotide Probe Design

A new version of CommOligo (He et al. 2012a) with group-specific probe design features can be used to design both gene- and group-specific oligonucleotide probes with different degrees of specificity based on the following criteria: (i) a gene-specific probe must have $\leq 90\%$ sequence identity, ≤ 20 -base continuous stretch, and ≥ -35 kcal/mol free energy; (ii) a group-specific probe has to meet the above requirements for nontarget groups, and it also must have $\geq 96\%$ sequence identity, ≥ 35 -base continuous stretch, and ≤ -60 kcal/mol free energy within the group. Computational and experimental evaluation indicates that these designed probes are highly specific to their targets (He et al. 2007, 2010a).

Probe Validation and GeoChip Construction

All designed probes are subsequently verified against the GenBank (NR) nucleic acid database for specificity. Normally, multiple (e.g., 20) probes for each sequence or each group of sequences are designed, but only the best probe set for each sequence or each group of closely

related sequences will be chosen for array construction. GeoChip can be constructed in-house, such as GeoChips 2.0 and 3.0 (He et al. 2007, 2010a), or commercially, like GeoChip 4.0 (Hazen et al. 2010; He et al. 2012a).

GeoChip Operation and Data Analysis

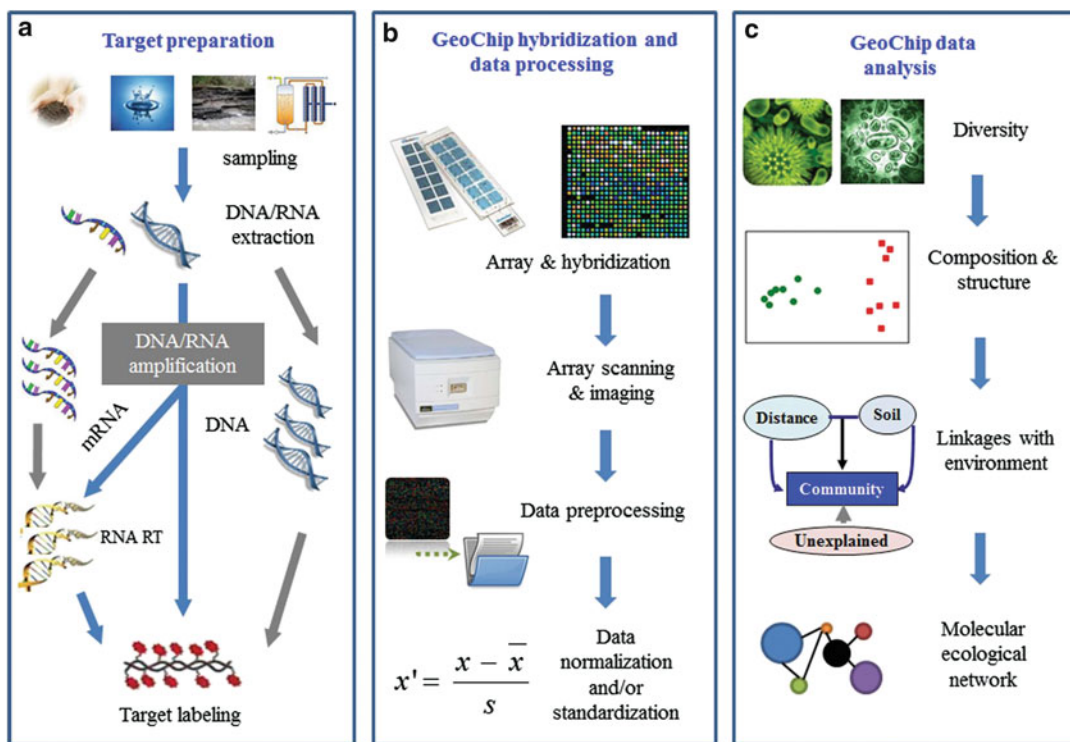
Generally, GeoChip operation and data analysis include target preparation, GeoChip hybridization, image and data preprocessing, and data analysis (Fig. 1).

Target Preparation

Target preparation involves a few steps, including nucleic acid extraction and purification, labeling, and hybridization (Fig. 1a). The most important step for successful GeoChip analysis is nucleic acid extraction and purification from environmental samples generally using a well-established method, which is able to produce large fragments of DNA. High-quality DNA should have ratios of $A_{260}/A_{280} \sim 1.8$ and $A_{260}/A_{230} > 1.7$. Low A_{260}/A_{230} ratios indicate impurities in the DNA sample and can negatively influence subsequent labeling and hybridization. Generally, since 1–5 μg of DNA or 5–20 μg of RNA is required for GeoChip hybridization, whole-community genome amplification (WCGA) for DNA and whole-community RNA amplification (WCRA) for RNA are necessary (He et al. 2012b). Non-amplified or amplified nucleic acids are then labeled with fluorescent dye (e.g., Cy3, Cy5) using random priming with the Klenow fragment of DNA polymerase for DNA and SuperScriptTM II/III RNase H-reverse transcriptase for RNA. The labeled nucleic acids are then purified and dried for hybridization (Fig. 1a).

Hybridization, Imaging, and Data Preprocessing

Labeled nucleic acid target is suspended in a hybridization buffer containing 40–50% formamide and hybridized on GeoChip at 42–50 °C (He et al. 2007, 2010a, 2012b). The hybridization stringency can be adjusted by changing the temperature and/or formamide concentration.



GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities, Fig. 1 A schematic presentation of target preparation, GeoChip operation, and data analysis of

microbial communities from a variety of habitats. (a) Target preparation, (b) GeoChip hybridization and data processing, (c) GeoChip data analysis (This figure is adapted from Fig. 1 by He et al. (2012b))

For every 1 % increase in formamide, the effective temperature increases by 0.6 °C (He et al. 2011).

Hybridized arrays are imaged with a microarray scanner having a resolution of at least 10 μm for homemade arrays and 2 μm for commercially manufactured arrays. Microarray analysis software is then used to quantify the signal intensity (pixel density) of each spot. Spot quality is also evaluated at this point using predetermined criteria, and positive spots are called generally based on signal-to-noise ratio [SNR; $\text{SNR} = (\text{signal mean} - \text{background mean}) / \text{background standard deviation}$] or signal-to-both-standard-deviations ratio [SSDR; $\text{SSDR} = (\text{signal mean} - \text{background mean}) / (\text{signal standard deviation} - \text{background standard deviation})$] (He et al. 2012b).

Raw GeoChip data are further evaluated via the GeoChip data analysis pipeline (He et al. 2010a). The quality of individual

spots, evenness of control spot hybridization signals across the slide surface, and background levels are assessed to determine overall array quality. Spots flagged as poor or low quality are removed along with outliers: positive spots with (signal – mean signal intensity of all replicate spots) greater than three times the replicate spots signal standard deviation (He et al. 2011). The signal intensities are then normalized for further statistical analysis (Fig. 1b).

GeoChip Data Analysis

Data analysis is the most challenging part in the use of GeoChip for microbial community analysis, and a variety of methods have been used to address fundamental microbial ecology questions (Fig. 1c). First, various diversity indices (e.g., richness, evenness, diversity) based on the number of functional genes detected and their abundances are used to examine the functional

diversity of microbial communities. The relative abundance of specific genes or gene groups can be determined based on the total signal intensity of the relevant genes or the number of genes detected. The percentage of genes shared by different samples can also be calculated to compare microbial communities examined. Second, for statistical analysis of the overall microbial community composition and structure with FGA data, ordination techniques can be used such as principal component analysis (PCA), detrended correspondence analysis (DCA), cluster analysis (CA), and nonmetric multidimensional scaling (NMDS). PCA and DCA are multivariate statistical methods, which reduce the number of variables needed to explain the data and highlight the variability between samples. CA groups samples based on the overall similarity of gene patterns. NMDS finds both a nonparametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items and the location of each item in the low-dimensional space. Also, the response ratio can be used to determine changes of specific functional genes between the control and the treatment. In addition, analysis of variation (ANOVA), analysis of similarities (ANOISM), nonparametric multivariate analysis of variance (Adonis), and multi-response permutation procedure (MRPP) can be used to discern dissimilarities of microbial communities over time and space (He et al. 2011, 2012b). Third, if environmental data or other metadata are available, GeoChip data can be used to correlate environmental variables with the functional microbial community structure. These include the Pearson's correlation coefficient (PCC), canonical correspondence analysis (CCA), and Mantel test. PCC measures the strength of linear dependence between two variables, such as functional gene abundances detected by GeoChip, and environmental variables. CCA has been used in many cases in GeoChip-based studies to better understand how environmental factors affect the community structure (He et al. 2011, 2012b). Also, based on the results of the CCA, the relative influence of environmental variables on the microbial community structure can be

determined using variance partitioning analysis (VPA). In addition, further correlations of GeoChip data with environmental parameters can be performed with the Mantel test (He et al. 2007, 2010a, b, 2011, 2012b). Finally, GeoChip data can be used to infer functional molecular ecological networks for revealing interactions of functional genes and their associated populations. A recent study indicated that elevated CO₂ substantially altered the network interaction of soil microbial communities and the shift in network structures is significantly correlated with soil properties (He et al. 2012b; Zhou et al. 2010) (Fig. 1c).

GeoChip Applications

Different versions of GeoChip have been used to analyze microbial communities from different habitats, such as aquatic systems, soils, extreme environments, human microbiomes, and bioreactors for addressing fundamental scientific questions related to global change, bioenergy, bioremediation, agricultural management, land use, and human health and disease as well as ecological theories (He et al. 2011, 2012b). Several recent studies are highlighted, especially with a focus on soil and water microbial communities. A list of representative studies with different GeoChip versions is shown in Table 1.

Soils

Soil may harbor the most complex microbial communities among known habitats, and recently GeoChips have been used to investigate soil microbial communities to address fundamental ecological questions related to global change (e.g., elevated CO₂, elevated O₃, warming), bioremediation of oil-contaminated fields, land use, agricultural management, and livestock grazing.

Three recent studies focused on the response of soil microbial communities to global change, including elevated CO₂, temperature, and O₃. First, GeoChip 3.0 was used to analyze soil microbial communities under elevated CO₂ at a multifactor grassland experiment site, BioCON (biodiversity, CO₂, and nitrogen deposition), in

GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities, Table 1 Summary of representative GeoChip applications. If no references are cited, those studies are described in a previous review (He et al. 2012b)

Habitat or ecosystem	Ecosystem/sample type	GeoChip	Objectives of study/biological questions
Aquatic systems	Marine sediment	GeoChip 1.0	Functional microbial community structure of marine sediments in the Gulf of Mexico
	Ebro and Elbe river sediment	GeoChip 2.0	Pesticide impacts on European rivers
	Coral-associated marine water	GeoChip 2.0	Microbial communities in healthy and yellow-band diseased coral (<i>Montastraea faveolata</i>)
Soils	Antarctic latitudinal transect soil	GeoChip 2.0	Microbial C and N cycling across an Antarctic latitudinal transect
	Deciduous forest soil	GeoChip 2.0	Gene-area relation in microorganisms
	Native grassland soil	GeoChip 2.0	Afforestation impacts soil microbial communities and their functional potential
	Strawberry farmland soil	GeoChip 2.0	Microbial responses to farm management
	Grassland soil	GeoChip 2.0	Microbial responses to plant invasion
	Agricultural soil	GeoChip 2.0	Agricultural practices/land use (Xue et al. 2013)
	Grassland soil	GeoChip 3.0	Global change (elevated CO ₂) (He et al. 2010b)
	Grassland soil	GeoChip 3.0	Global change (warming) (Zhou et al. 2012)
	Wheat rhizosphere soil	GeoChip 3.0	Global change (elevated O ₃) (Li et al. 2013)
	Citrus rhizosphere soil	GeoChip 3.0	Rhizosphere microbial community responses to <i>Candidatus Liberibacter asiaticus</i> -infected citrus trees
	Grassland soil	GeoChip 4.0	The effect of grazing on microbial communities (Yang et al. 2013)
Contaminated sites	U-contaminated underground water (Oak Ridge, TN)	GeoChip 1.0	Bioremediation of U-contaminated groundwater
		GeoChip 2.0	Bioremediation of U-contaminated groundwater (Van Nostrand et al. 2011)
	U-contaminated sediment (Oak Ridge, TN)	GeoChip 2.0	Bioremediation of U-contaminated sediments
	U-contaminated underground water (Rifle, CO)	GeoChip 2.0	Bioremediation of U-contaminated groundwater (Liang et al. 2012)
	PCB-contaminated soil	GeoChip 2.0	Microbial bioremediation of PCB-contaminated soil
	Oil-contaminated soil	GeoChip 2.0	Bioremediation of oil-contaminated soil
	Arsenic-contaminated soil	GeoChip 3.0	Rhizosphere microbial community responses to arsenic contamination and phytoremediation
	Landfill groundwater	GeoChip 3.0	Microbial responses to landfill-derived contaminants in groundwater (Lu et al. 2012)
Oil-spill seawater	GeoChip 4.0	Microbial bioremediation of oil-spill sites (Hazen et al. 2010)	

(continued)

GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities, Table 1 (continued)

Habitat or ecosystem	Ecosystem/sample type	GeoChip	Objectives of study/biological questions
Extreme environments	Deep-sea hydrothermal vent (chimney)	GeoChip 2.0	Functional gene diversity of deep-sea hydrothermal vent microbial communities
	Deep-sea basalt samples	GeoChip 2.0	Functional gene diversity and structure of deep-sea basalt microbial communities
	GSL hypersaline water	GeoChip 2.0	Functional gene diversity and structure of hypersaline water microbial communities
	Acid mine drainage (water)	GeoChip 2.0	Functional gene diversity of microbial communities in acid mine drainage (AMD) systems
Bioreactors	Fluidized bed reactor for bioremediation	GeoChip 2.0	Microbial bioremediation of hydrocarbon-contaminated water
	Microbial electrolysis cell for hydrogen production	GeoChip 3.0	Microbial hydrogen production using wastewater

the Cedar Creek Ecosystem Science Reserve, MN (He et al. 2010b). The results showed that the functional microbial community structure was markedly different between ambient CO₂ and elevated CO₂ as indicated by DCA of GeoChip 3.0 data and 16S rRNA gene-based pyrosequencing data. Also, genes involved in labile C degradation and C and N fixation were significantly increased under elevated CO₂ although the abundance of recalcitrant C degradation genes remained unchanged. In addition, changes in the microbial community structure were significantly correlated with soil C and N contents and plant productivity (He et al. 2010b). Second, GeoChip 3.0 was used to understand the effect of increased temperature on soil microbial communities and their roles in regulating soil carbon dynamics at a tallgrass prairie ecosystem in the US Great Plains of Central Oklahoma. The results suggest soil microorganisms may regulate soil carbon dynamics through three primary feedback mechanisms: (i) shifting microbial community composition, leading to the reduced temperature sensitivity of heterotrophic soil respiration; (ii) differentially stimulating labile C but not recalcitrant C degradation genes to maintain long-term soil carbon stability and storage; and (iii) enhancing nutrient-cycling processes to promote plant growth (Zhou et al. 2012). Third,

GeoChip 3.0 was used to investigate the functional composition, and structure of rhizosphere microbial communities from O₃-sensitive and O₃-relatively-sensitive wheat (*Triticum aestivum* L.) cultivars under elevated O₃ (eO₃). Based on GeoChip hybridization signal intensities, although the overall functional structure of rhizosphere microbial communities did not significantly change by eO₃ or cultivars, the results showed that the abundance of specific functional genes involved in C fixation and degradation, N fixation, and sulfite reduction did significantly alter in response to eO₃ and/or wheat cultivars. Also, the O₃-sensitive cultivar appeared to harbor microbial functional communities in the rhizosphere more sensitive in response to eO₃ than the O₃-relatively sensitive cultivar. In addition, CCA suggested that the functional structure of microbial communities involved in C cycling was largely shaped by soil and plant properties including pH, dissolved organic carbon (DOC), microbial biomass C, C/N ratio, and grain weight (Li et al. 2013). Those studies indicate that global change significantly impacts soil microbial communities, which may in turn regulate ecosystem functioning through different feedback mechanisms.

Various agriculture management practices may have significant influences on soil microbial communities and their ecological functions.

GeoChip 2.0 was used to evaluate the potential functions of soil microbial communities under conventional (CT), low-input (LI), and organic (ORG) management systems at an agricultural research site in Michigan. Compared to CT, a high diversity of functional genes was observed in LI. The functional gene diversity in ORG did not differ significantly from that of either CT or LI. The abundance of genes encoding enzymes involved in C, N, P, and S cycling was generally lower in CT than in LI or ORG, but functional genes involved in lignin degradation, methane generation/oxidation, and assimilatory N reduction remained unchanged. Also, significant correlations were observed between NO_3^- concentration and denitrification gene abundance, NH_4^+ concentration and ammonification gene abundance, and N_2O flux and denitrification gene abundance, indicating a close linkage between soil N availability or utilization and associated functional potential of soil microbial communities (Xue et al. 2013).

Livestock grazing is a type of global land-use activity. However, the effect of free livestock grazing on soil microbial communities at the functional gene level remains unclear. GeoChip 4.0 was used to examine the effects of free livestock grazing on the microbial community at an experimental site in Tibet, a region known to be very sensitive to anthropogenic perturbation and global warming. The results showed that grazing changed the microbial community functional structure, in addition to aboveground vegetation and soil geochemical properties. Further statistical analysis showed that microbial community functional structures were closely correlated with environmental variables and variations in microbial community functional structures were mainly controlled by aboveground vegetation, soil C/N ratio, and NH_4^+ -N. Therefore, these results indicated that soil microbial community functional structure was very sensitive to livestock grazing and revealed the role of soil microbial communities in the regulation of soil N and C cycling, supporting the necessity to include microbial components in evaluating the consequence of land use and/or climate change (Yang et al. 2013).

Groundwater and Aquatic Ecosystems

Due to human activities, groundwater and aquatic ecosystems are often contaminated from various sources (e.g., mining, oil spill, landfill) and with a variety of toxic compounds (e.g., heavy metals, herbicides, antibiotics, pesticides) and conditions (e.g., low pH, high salinity). To understand how such contamination impacts groundwater and aquatic ecosystems, GeoChips were used to investigate those microbial communities to explore the potential of in situ bioremediation of contaminated sites by indigenous microbial communities.

A pilot-scale system was established to examine the feasibility of in situ U(VI) immobilization at a highly contaminated aquifer in Oak Ridge, TN. Ethanol was injected intermittently as an electron donor to stimulate microbial U(VI) reduction, leading to a decrease of U(VI) concentrations below the Environmental Protection Agency drinking water standard. GeoChip 2.0 was used to monitor microbial communities in three wells during active U(VI) reduction and maintenance phases. The results showed that the overall microbial community structure exhibited a considerable shift over the remediation phases examined and functional populations of Fe(III)-reducing bacteria (FeRB), nitrate-reducing bacteria (NRB), and sulfate-reducing bacteria (SRB) reached their highest levels during the active U(VI) reduction phase (days 137–370), in which denitrification, Fe(III) reduction, and sulfate reduction occurred sequentially, suggesting that these functional populations could play an important role in both active U(VI) reduction and maintenance stability of reduced U(IV) (Van Nostrand et al. 2011).

To better understand the microbial functional diversity changes with subsurface redox conditions during in situ U(VI) bioremediation, GeoChip 2.0 was applied to examine groundwater microbial communities at a uranium mill tailings remedial action (UMTRA) site (Rifle, CO). The results indicated that functional microbial communities altered with a shift in the dominant metabolic process and the abundance of *dsrAB* and *mcr* genes increased when redox conditions shifted from Fe-reducing to sulfate-reducing conditions, while cytochrome genes were primarily

detected from *Geobacter* species and decreased with lower subsurface redox conditions. Statistical analysis of environmental parameters and functional genes indicated that acetate, U(VI), and redox potential were the most significant geochemical variables linked to the microbial functional gene structures. This study indicates that microbial functional genes could be very useful for tracking microbial community structure and dynamics during bioremediation (Liang et al. 2012).

In another study, GeoChip 3.0 was used to study the functional gene diversity and structure of groundwater microbial communities in a shallow landfill leachate-contaminated aquifer in Norman, OK. Samples were taken from eight wells at the same aquifer depth immediately below a municipal landfill or along the predominant downgradient groundwater flowpath. The results showed that functional gene richness and diversity immediately below the landfill and the closest well were considerably lower than those in downgradient wells and that landfill leachate impacted the diversity, composition, structure, and functional potential of groundwater microbial communities as a function of groundwater pH and concentrations of sulfate, ammonia, and dissolved organic carbon (Lu et al. 2012).

In 2010, the Deepwater Horizon oil spill occurred in the Gulf of Mexico. GeoChip 4.0 was used to examine the functional composition and structure of water microbial communities from the oil plume and control sites. The results indicated that the water microbial community composition and structure were dramatically altered in deep-sea oil plume samples. A variety of functional genes involved in both aerobic and anaerobic hydrocarbon degradation were highly enriched in the plume compared with outside the plume, indicating a great potential for intrinsic bioremediation or natural attenuation in the deep sea. Various other microbial functional genes that are relevant to C, N, P, S, and iron cycling, metal resistance, and bacteriophage replication were also enriched in the plume. Overall, this study suggests that indigenous microbial communities could have a significant role in biodegradation of oil spills in deep-sea environments (Hazen et al. 2010).

Other Environments

GeoChips were also used to analyze microbial communities from other habitats/ecosystems, including various contaminated sites (e.g., chromate-contaminated water, U-contaminated sediments, polychlorinated biphenyl- and arsenic-contaminated soils), extreme environments (e.g., acid mine drainage, hypersaline lakes, deep-sea basalts, deep-sea hydrothermal vents), bioleaching systems, and bioreactors as well as the human microbiome (He et al. 2011, 2012b).

Summary

Although GeoChip technology has been demonstrated to be specific, sensitive, and quantitative and applied to analyze microbial communities from different habitats, some key issues and challenges still remain, including probe coverage, specificity, sensitivity, quantitative capability, nucleic acid quality, the detection of microbial community activity, and challenges by high-throughput sequencing technologies. It should be noted that probe coverage on GeoChip is relatively low compared to the availability of functional gene sequences in databases, especially for earlier versions of FGAs. One of the reasons is that some sequences do not have specific probes based on the availability of sequence databases and software. Also, GeoChip probe sets need continuous updates to reflect the current status of functional gene sequence information.

Critical issues with GeoChip design and detection are specificity, sensitivity, and quantitative capability, which are especially important since many gene variants within each environmental sample are unknown. Array specificity is controlled by probe design and hybridization conditions. A novel microarray probe design software tool, *CommOligo* (He et al. 2012a), and its improved versions were used to design probes for GeoChip 2.0, GeoChip 3.0, and GeoChip 4.0. Experimental evaluations of GeoChip 2.0 and GeoChip 3.0 indicated that low percentages of false positives (0.002–0.025 %) were observed (He et al. 2007; He et al. 2010a). GeoChip hybridizations are generally performed at 42–50 °C

with 50 % formamide. Sensitivity is another major concern since many gene variants are expected to be low abundant in environmental samples. The current level of sensitivity for oligonucleotide arrays using environmental samples is approximately 50–100 ng or 10^7 cells, or approximately 5 % of the microbial community, providing a coverage of only the most dominant community members. Several strategies have been utilized to increase sensitivity. For example, with WCGA and WCRA approaches, the sensitivity of GeoChip hybridization could increase to 10 fg. Also, array surface modifications, a decrease of hybridization solution, and the use of new labeling techniques could increase GeoChip detection sensitivity (He et al. 2011, 2012a). An important goal in microarray analysis is to provide quantitative information. GeoChip has been shown to have a linear relationship between target DNA or RNA concentrations and hybridization signal intensities. However, this relationship can be affected by sequence divergence (i.e., the more divergent the sequence, the lower the signal intensity). Therefore, two strategies are used to improve quantitative ability: mismatch probes and using relative comparisons across samples rather than absolute comparisons (He et al. 2012a).

The quality and quantification of environmental nucleic acids are one of the most important for successful GeoChip hybridization and reliable data generation. DNA with large fragments and minimal amounts of contaminants are especially important when samples need to be amplified using WCGA. Accurate measurement of DNA yields is also important, so quantification should be based on double-strand DNA (dsDNA) measurement (e.g., PicoGreen) rather than via absorbance. While DNA detection provides information on the presence of functional genes in the environment, it does not provide unconditional evidence for microbial activity. Population changes can be used to infer microbial activity, but this may not be accurate. To monitor microbial activity, mRNA should be used. However, since mRNA is easily degraded with rapid turnover, usually has a low abundance, and has a small proportion of the total RNA, improved

RNA extraction methods are necessary to use environmental RNA for GeoChip analysis. Alternatively, other techniques, such as stable isotope probing (SIP), enzyme activity, metaproteomic analysis, and metabolite assays, may be used to study the functional activity and ecosystem functions of microbial communities.

High-throughput sequencing technologies (e.g., 454, Illumina) are available for microbial community analysis, which challenge GeoChip technologies. However, although these sequencing-based technologies can discover novel sequences, it can be expensive to do in-depth shotgun sequencing of a community. In addition, it suffers from lack of appropriate conserved primers for many target genes. Also, sequencing-based technologies have a disadvantage of random sampling, and/or under-sampling, making it difficult to compare different samples, while microarray-based technologies have a defined probe set, which is good for community comparisons (He et al. 2012b). Therefore, due to the unique features and advantages and disadvantages of both microarray-based and sequencing-based technologies, it is preferable that they be used complementarily for microbial community analysis in order to address fundamental questions in microbial ecology and environmental biology.

Acknowledgments This work conducted by ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies) (<http://enigma.lbl.gov>), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory, was supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under Contract No. DE-AC02-05CH11231 and by the Oklahoma Applied Research Support (OARS), Oklahoma Center for the Advancement of Science and Technology (OCAST), State of Oklahoma, through AR11-035 and AR062-034.

References

- Gans J, Wolinsky M, Dunbar J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*. 2005;309:1387–90.
- Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, et al. Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science*. 2010;330:204–8.

- He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, et al. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J.* 2007;1:67–77.
- He Z, Deng Y, Van Nostrand JD, Tu Q, Xu M, Hemme CL, et al. GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *ISME J.* 2010a;4:1167–79.
- He Z, Xu M, Deng Y, Kang S, Kellogg L, Wu L, et al. Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO₂. *Ecol Lett.* 2010b;13:564–75.
- He Z, Van Nostrand JD, Deng Y, Zhou J. Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. *Front Environ Sci Engin China.* 2011;5:1–20.
- He Z, Deng Y, Zhou J. Development of functional gene microarrays for microbial community analysis. *Curr Opin Biotechnol.* 2012a;23:49–55.
- He Z, Van Nostrand JD, Zhou J. Applications of functional gene microarrays for profiling microbial communities. *Curr Opin Biotechnol.* 2012b;23:460–6.
- Li X, Deng Y, Li Q, Lu C, Wang J, Zhang H, et al. Shifts of functional gene representation in wheat rhizosphere microbial communities under elevated ozone. *ISME J.* 2013;7(3):660–71.
- Liang Y, Van Nostrand JD, N’Guessan LA, Peacock AD, Deng Y, Long PE, et al. Microbial functional gene diversity with a shift of subsurface redox conditions during in situ uranium reduction. *Appl Environ Microbiol.* 2012;78:2966–72.
- Lu Z, He Z, Parisi VA, Kang S, Deng Y, Van Nostrand JD, et al. GeoChip-based analysis of microbial functional gene diversity in a landfill leachate-contaminated aquifer. *Environ Sci Technol.* 2012;46:5824–33.
- Rhee S-K, Liu X, Wu L, Chong SC, Wan X, Zhou J. Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl Environ Microbiol.* 2004;70:4303–17.
- Torsvik V, Ovreas L, Thingstad TF. Prokaryotic diversity – magnitude, dynamics, and controlling factors. *Science.* 2002;296:1064–6.
- Van Nostrand JD, Wu L, Wu W-M, Huang Z, Gentry TJ, Deng Y, et al. Dynamics of microbial community composition and function during in situ bioremediation of a uranium-contaminated aquifer. *Appl Environ Microbiol.* 2011;77:3860–9.
- Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA.* 1998;95:6578–83.
- Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J. Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol.* 2001;67:5780–90.
- Xue K, Wu L, Deng Y, He Z, Van Nostrand J, Robertson PG, et al. Functional gene differences in soil microbial communities from conventional, low-input, and organic farmlands. *Appl Environ Microbiol.* 2013;79:1284–92.
- Yang Y, Wu L, Lin Q, Yuan M, Xu D, Yu H, et al. Responses of the functional structure of soil microbial community to livestock grazing in the Tibetan alpine grassland. *Glob Chang Biol.* 2013;19:637–48.
- Zhou J, Deng Y, Luo F, He Z, Tu Q, Zhi X. Functional molecular ecological networks. *mBio.* 2010;1(4):e00169.
- Zhou J, Xue K, Xie J, Deng Y, Wu L, Cheng X, et al. Microbial mediation of carbon-cycle feedbacks to climate warming. *Nat Clim Chang.* 2012;2:106–10.

GHOSTM

Yutaka Akiyama

Department of Computer Science, Tokyo
Institute of Technology, Meguro-ku,
Tokyo, Japan

Definition

GHOSTM is a homology search tool developed for metagenomics and accelerated by GPU-computing. GHOSTM can be used as the alternative of BLASTX program, which searches protein databases using a translated nucleotide query. The GHOSTM system achieved calculation speeds that were 130 times faster than BLAST with 1 GPU. It also had a calculation speed that was 3.4 times faster than BLAT with higher search sensitivity. GHOSTM is distributed under the MIT license and its source code is available for download at <http://code.google.com/p/ghostm/>.

Introduction

In metagenomic analysis, the DNA sequence fragments obtained from environmental samples frequently include DNA sequences from many different species, and closely related reference

genome sequences are often unavailable. Thus, sensitive approaches are required for the identification of novel genes. Metagenomic DNA fragments are often translated into protein coding sequences and then further assigned to protein families, such as COG and Pfam databases. The BLASTX (Altschul et al. 1990) program has been used for such binning and classification because it can identify homologues that do not have high nucleotide sequence identity, but once these sequences are translated, the homologue can be found in a distantly related member of a protein family (Turnbaugh et al. 2006). The BLAST algorithm is sufficiently sensitive for searching protein families, but its performance is insufficient for analyzing the large quantities of data produced by a next-generation sequencer. In practice, approximately 1,000 CPU days were needed for querying 20 million short reads against the KEGG database using BLASTX program.

To address the issue, the GHOSTM software (Suzuki et al. 2012) had been developed. GHOSTM can efficiently search homologous sequences for a database based on GPU-computing technique. Graphics processing units (GPUs) were originally designed for graphics applications, but new generation GPUs have been transformed into powerful coprocessors for general purpose computing because their computational power supersedes that of CPUs. For example, the peak performance of a GPU, such as the NVIDIA Tesla K20, is approximately 3.5 TFLOPS. This speed is more than tenfold faster than the most recent CPUs. GPUs have already been used for several bioinformatics applications, such as CUDASW (Liu et al. 2010) and CUDA-BLASTP (Liu et al. 2011).

GHOSTM employs a new and efficient homology search algorithm suitable for GPU calculation. The system accepts a large number of short DNA fragment sequences produced by a next-generation sequencer as the input like the BLASTX program and performs DNA sequence homology searches against a protein sequence database. The system demonstrated a calculation speed that was 130 times faster with one GPU than BLAST on a CPU.

Overview of the Algorithm

The GHOSTM is mainly composed of three components, as shown in Fig. 1. The first component searched the candidate alignment positions for a sequence from the database using the indexes. The second component calculated local alignments around the candidate positions using the Smith-Waterman algorithm for calculating the alignment scores. Finally, the third component sorted the alignment scores and output the search results.

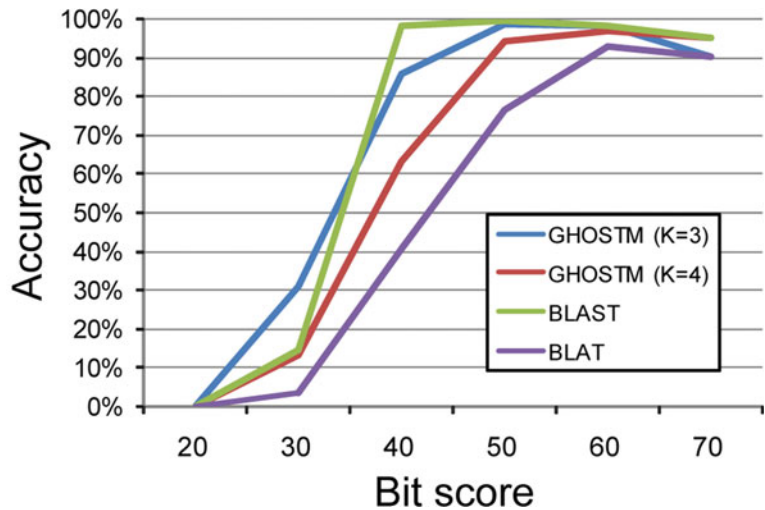
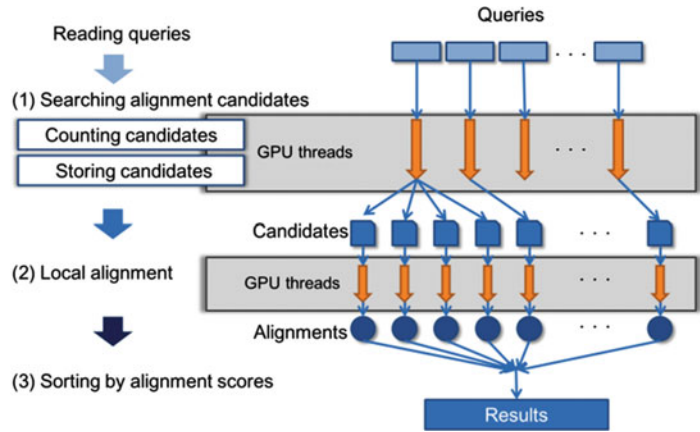
Both the candidate search and local alignment components required a large amount of computing time. Therefore, queries on both components are processed in parallel and they are mapped onto GPUs. Thus, multiple queries were simultaneously processed on different GPU cores. GPUs have many computing cores (the Tesla S1070 has 240 cores per GPU) and this is the reason for the acceleration of GHOSTM in processing time. Importantly, the GHOSTM system requires a sufficient number of queries for maximum efficiency, and in fact, when using only one query sequence, the calculation of GHOSTM becomes much slower than BLAST.

Search Performances

Because metagenomic analyses require highly sensitive searches, it is difficult to use homology search program with high speed but low sensitivity, such as BLAT (Kent 2002). In contrast, GHOSTM has sufficient search sensitivity for metagenomic analysis.

Figure 2 shows the comparison of search sensitivity for each homology search program. To evaluate the search sensitivity, the search results obtained with the Smith-Waterman local alignment method implemented in SSEARCH (Pearson 1991) were assumed to be the correct answers. The performance of a particular method is evaluated in terms of the fraction of its results that corresponded to the correct answers obtained by SSEARCH. The search accuracy of GHOSTM was clearly higher than BLAT. Low-scoring hits (e.g., <50) are generally not used in practice because such hits can occur by chance. With the

GHOSTM, Fig. 1 Data flow and processing within GHOSTM



GHOSTM, Fig. 2 Search accuracy of GHOSTM

exception of the low-score hits, GHOSTM successfully identified more than 90 % of the hits identified by SSEARCH. This result suggests that GHOSTM is sufficiently accurate for general usage.

The computational times of BLAST, BLAT, and GHOSTM for 100 thousand reads are shown in Table 1. Each query read has the length from 60 to 75 bp and the search target is KEGG Genes (“genes.pep”) database (Kanehisa et al. 2010) with approximately 2.5 GB. The GHOSTM program achieved a calculation speed approximately 130 and 400 times faster than the BLAST program using 1 thread and 4 threads, respectively. Moreover, GHOSTM was approximately 3.4 times

GHOSTM, Table 1 Comparison of search speed

Program	#GPUs	Time (s)	Acceleration ratio
GHOSTM (K = 4)	1	2,855	129.5
GHOSTM (K = 4)	4	909	406.7
BLAT		9,898	37.3
BLASTX (1 thread)		369,678	1
BLASTX (4 threads)		102,255	3.6

faster than BLAT despite of its higher search sensitivity. GHOSTM achieves both high search speed and high search sensitivity compared with previous homology search tools.

Installation and Requirements

The source code of GHOSTM is distributed under the MIT license and is available for download at <http://code.google.com/p/ghostm/>. GHOSTM was implemented in C++ and the NVIDIA CUDA library and requires CUDA version 2.2 or higher. Thus, the user has to prepare NVIDIA's GPU card, such as Tesla K20, for executing the GHOSTM program. The user can also execute GHOSTM on a general GeForce graphics card as well as Tesla. The performance of GHOSTM basically depends on the number of CUDA cores and their clocks. Thus, several GeForce GTX cards show better performance than Tesla. However, current GeForce cards do not have Error Check and Correct (ECC) memory, and thus, the search results obtained using such cards are unreliable because of the GPU memory error. Therefore, Tesla GPUs were recommended especially if the user have to process large amount of sequences.

Summary

Currently, sequencing technology continues to improve, and sequencers are increasingly producing larger and larger quantities of data. This explosion of sequence data makes computational analysis with contemporary tools more difficult.

However, GHOSTM is an efficient tool based on GPU-computing techniques and it would be a potential solution to this problem.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
- Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 2010;38(Database issue):D355–60.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12(4):656–64.
- Liu Y, Schmidt B, Maskell DL. CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions. *BMC Res Notes.* 2010;3:93.
- Liu W, Schmidt B, Müller-Wittig W. CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware. *IEEE/ACM Trans Comput Biol Bioinforma/IEEE ACM.* 2011;8(6):1678–84.
- Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics.* 1991;11(3):635–50.
- Suzuki S, Ishida T, Kurokawa K, Akiyama Y. GHOSTM: a GPU-accelerated homology search tool for metagenomics. *PLoS One.* 2012;7(5):e36060.
- Fernandez-Fuentes N, ed.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 2006;444(7122):1027–31.

H

Horizontal Gene Transfer and Bacterial Diversity

Chitra Dutta¹ and Munmun Sarkar²

¹Structural Biology & Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, West Bengal, India

²CSIR-Indian Institute of Chemical Biology, Kolkata, India

Synonyms

Lateral gene transfer

Definition

Horizontal gene transfer (HGT) is the process in which genetic material is transmitted between two organisms that are not parent and offspring. HGT is pervasive among bacteria, even among very distantly related ones. Through transmission of distinct physiological traits from one organism to another, it may cause drastic changes in the ecological and pathogenic character of bacterial species and thereby may catalyze the diversification of bacterial lineages.

Introduction

Bacteria are the most diverse and versatile life forms of our planet. In view of the fact that they

are tiny, unicellular organisms with relatively small genomes, variations observed in their cellular architectures, metabolic properties, and ecological preferences are remarkable. Such enormous diversity may be attributed to the extremely dynamic genomes of bacteria that evolve rapidly through alteration, acquisition, deletion, and rearrangements of relevant genetic information through various molecular mechanisms. These mechanisms include not only the processes of internal modification of genetic materials like mutation or homologous recombination but also exchange of specific set of genes with other species through the process of horizontal transfer (Ochman et al. 2000). Mutations usually lead to slow, subtle, but continuous refinement and alteration of existing genes that may foster diversification and speciation of microorganisms on an evolutionary time scale. HGT, on the contrary, is capable of introducing abrupt large-scale changes in the gene repertoire of an organism that may confer novel physiological traits to the recipient and enable an organism to explore new ecological niches and even can generate new variants of bacterial strains by “genetic quantum leaps” (Groisman and Ochman 1996).

Mechanisms of HGT

HGT is in sharp contrast with the process of vertical transfer that propagates genes from the parental generation to offspring via sexual or asexual reproduction.

There are three principal mechanisms for interspecies transmission of DNA elements in HGT (Ochman et al. 2000):

- (i) Transformation – uptake of naked DNA element from environment
- (ii) Transduction – the bacteriophage-mediated transmission of genetic materials between organisms recognized by the phage
- (iii) Conjugation – transfer of DNA from the donor to the recipient through cell-to-cell contact via sexual pilus

However, mere insertion of the donor DNA element into a recipient cytoplasm does not ensure a successful HGT, unless this foreign DNA sequence becomes stable in the host chromosome. Though the transfer or uptake of a short DNA sequence is usually indiscriminate with respect to the functional or compositional features of the transmitted sequence, stabilization of this foreign DNA element into the host organism depends critically on the compatibility of the transferred genes with the transcriptional and translational machinery of the host (Dutta and Pan 2002). Stable incorporation of the newly acquired DNA into the host genome can be mediated by any of the following processes: (i) homologous recombination, which normally limits the process among closely related organisms; (ii) persistence as an episome, if favored by natural selection; (iii) integration mediated by mobile genetic elements; and (iv) illegitimate incorporation through chance events of double-strand break repair.

Factors Regulating the Events of HGT and Their Outcomes

Depending on the organisms involved and the gene transfer mechanisms that are operational, there are a number of factors that can foster or limit the transfer, uptake, stabilization, and expression of foreign DNA molecules in bacteria.

Factors that may foster an event of HGT (Wiedenbeck and Cohan 2011) include both mechanistic as well as functional aspects. The phylogenetic closeness of the donor and the recipient often facilitate HGT, since close

relatives are likely to have greater sequence identity and hence higher probability of homologous recombination as well as HFIR. Bacteria with the same restriction–modification system can more easily share a phage or a plasmid and exchange their DNA elements. DNAs of short length (carrying one to several genes) usually has a greater probability of undergoing a successful adaptive HGT, even across deeply divergent bacteria, as it may allow an organism to selectively pick up a niche-transcending gene or set of genes without acquiring the niche-specifying genes of the donor. Furthermore, a short DNA may also survive in a host with distinct restriction–modification system, as it is less likely to contain a given recognition sequence and may thereby be more protected from cleavage by the restriction system of the host. And, needless to say, a niche-transcending HGT that provides an important adaptation to a recipient will always have a selective advantage.

Among the mechanistic barriers limiting unregulated uptake of foreign DNA in bacteria are the lack of similarity between the donor and the recipient, which may prohibit the integration of new sequence into a replicating genetic unit, surface exclusion that may create an effective barrier against conjugative transfer into cells, and presence of distinct restriction/modification systems present in the host (Thomas and Nelsen 2005).

A protein's connectivity may be another important factor for the transferability of genes across organisms. The complexity hypothesis (Jain et al. 1999) predicts a low rate of transfer of genes, products of which are involved in many complex interactions. Transfer of only one part of a complex set of coadapted structures is likely to bring about an incompatibility and loss of function. It is thought that bacterial genes may be broadly classified into two categories according to their transferability (Nakamura et al. 2004): (i) less transferable “informational” genes involved in replication translation and transcription and (ii) frequently transferable “operational” genes involved in metabolism. It has also been reported that among operational genes, those involved in cell surface, DNA binding, and

pathogenicity-related functions have higher probability of HGT as compared to the genes related to amino acid biosynthesis, biosynthesis of cofactors, energy metabolism, intermediary metabolism, fatty acid and phospholipid metabolism, and nucleotide biosynthesis.

Any recipient organism would also try to resist an event of HGT that might incur harmful pleiotropic effects. The deleterious side effects of a new acquisition often drive natural selection toward “domesticating” the acquired DNA, i.e., toward ameliorating its negative fitness effects (Wiedenbeck and Cohan 2011). Newly acquired genes may have higher rates of evolution than other genes in the genome. Another mechanism for domesticating a horizontally acquired adaptation involves initial repression of the acquired gene(s) in the host genome by histone-like nucleoid-structuring proteins (H-NS) (Dorman 2004). The compositional differences between a donor segment and the recipient are diminished over time as incorporated genes are subjected to the mutational bias of the host (Lawrence and Ochman 1997).

Bacterial Diversity Incurred by HGT

HGT is thought to be a prime contributor to bacterial evolution. As more and more genome sequences are being determined, it is becoming clear that cross-species transmission of genetic information through HGT is pervasive among bacteria and that it may occur at vast phylogenetic distances and that it may confer novel phenotypes and functions to the host organism by introducing fully functional genes and gene clusters. Unlike point mutations that can only adjust preexisting phenotypes, HGT may result in drastic changes in metabolic, pathological, or ecological character of a microbial species, thereby allowing effective and competitive exploitation of new niches (Lawrence 1999; Hacker and Kaper 2000). In cases where habitat differences suggest ecological differentiation between close relatives, a genome-based analysis often identifies one or more events of HGT as the primary cause of the ecological divergence. Some of the

niche-transcending traits that are commonly introduced in bacterial species through HGT are as follows.

Novel Metabolic Traits and Niche Adaptation

In bacteria, a substantial portion of species-specific functions can be attributed to HGT. Through HGT, divergent bacterial populations may share an adaptation that transcends their differences in cellular architectures, physiological capabilities, and ecological niches. For instance, enterotoxigenic *Escherichia coli* that attacks the epithelial cells of the small intestine shares the class 5 fimbriae with *Burkholderia cepacia* that resides in human lungs of cystic fibrosis patients and attacks the respiratory epithelium. On the other hand, closely related bacteria or even strains of same species may exhibit radically different metabolic, physiological, or pathogenic traits – thanks to HGT. *Bacillus anthracis* (strain *Ames ancestor*), *Bacillus cereus* (ATCC1098), and *Bacillus thuringiensis* (serovar *konkukian str. 97–27*), all are considered as a single species, as they show more than 94 % ANI and have highly syntenic gene repertoire. And yet they are drastically different in their phenotypes – a highly virulent pathogen and potentially lethal bioterror agent, a source of food poisoning, and an eco-friendly organic bio-pesticide, respectively (Doolittle and Papke 2006).

HGT, in many cases, endows the recipient with novel metabolic capabilities that enable it either to invade a new niche or to improve its performance in its current niche (Cohan and Koeppel 2008). For example, acquisition of the lac operon has enabled *Escherichia coli* to utilize the milk sugar lactose as a carbon source and thereby to explore a new niche, the mammalian colon, where it has established a commensal relationship (Ochman et al. 2000). An event of HGT may even allow for conversion of the recipient into a radically different organism that may inhabit niches completely unexplorable by the organisms relying on mutational processes alone. Examples include the aerobic methanotrophs that have gained the ability to synthesize critical cofactors for

H4MPT-mediated methyl-group transfer by acquiring genes from methanogenic archaea, bacteria that exploit halorhodopsin homologues as light-driven proton pumps, and cyanobacteria gaining the capability of oxygenic photosynthesis through acquisition of a second photosystem (Gogarten et al. 2002).

Speciation and Sub-speciation in Bacteria

A substantial part of the speciation and sub-speciation in bacteria can be explained as the result of macroevolution events mediated by HGT (Cruz and Davies 2000). Using *E. coli* and *Salmonella* as a model system, it has been demonstrated that 17 % of their genomes (~800 kb) appear to have been acquired by HGT during the past 100 million years. As the majority of these DNAs seem to be recently recruited, it is apparent that considerable genetic flux may still be occurring across these two species and the 234 detectable HGT events that have persisted are probably “the tip of the iceberg of the thousands of mobile sequences” that have been acquired or shaded off by any particular *E. coli* strain (Cruz and Davies 2000). Comparison of the members of a well-known collection of *E. coli* strains (the ECOR collection) revealed that these strains are quite variable in the size and macro-organization of their chromosomes and plasmids. These observations point toward the fact that a significant proportion of the genome of any strain of a single bacterial species may comprise fragments of functional genetic elements from various origins, which, if properly “nurtured,” can give rise to new bacterial species.

Adoption of Pathogenic/Symbiotic Lifestyle Through Acquisition of Genome Islands

Horizontal acquisition of virulence factors is a common strategy of bacterial organisms for undergoing transformation from the benign form into a pathogen. A pathogenic strain of any bacterial strain is often distinguished from the nonpathogenic variants of the same or related species by the presence of a cluster of virulence factors like toxins, invasion factors, adherence factors, and secretion systems, the G+C composition of which may differ significantly from that

of the core genome of the respective species. Such discrete gene clusters, referred to as “virulence cassettes” or “pathogenicity islands” (Groisman and Ochman 1996; Hacker et al. 1997), usually reside at tRNA and tRNA-like loci, which appear to be common sites for integration of foreign sequences (Hacker et al. 1997; Ochman et al. 2000) and are flanked by 16–20 bp perfect or almost perfect direct repeats. They may also carry insertion elements or transposons. All these observations strongly argue in favor of horizontal acquisition of these islands by their host genomes. Conversion of laboratory strains of *E. coli* from avirulent to virulent forms upon experimental introduction of genes from other species (Isberg and Falkow 1985; McDaniel and Kaper 1997) or presence of large virulence plasmids in pathogenic *Shigella* and *Yersinia* (Gemski et al. 1980; Portnoy et al. 1981; Maurelli et al. 1985; Sasakawa et al. 1988) supported the notion of horizontal transfer of virulence factors in bacteria.

With accumulation of genome sequences of diverse bacterial species, it became clear that pathogenicity islands represent a subclass of a more diverse group of genetic elements, designated as genomic islands (GI). A GI refers to a part of genome – usually 10–200 bp in length – containing a set of horizontally acquired genes that might be beneficial for the host bacterium under specific environmental conditions. GIs may be associated with diverse adaptive functions that enable the respective species to survive or colonize within a specialized niche or to adopt a distinct lifestyle. For instance, nitrogen fixation genes harbored by “symbiosis islands” in various Rhizobiaceae species enable these organisms to develop a symbiotic relationship with legumes, which, in turn, facilitate their survival inside the root nodules of the legumes (Sullivan and Ronson 1998).

Dissemination of the gene clusters (operons) involved in the catabolism of xenobiotics in polluted environment is often attributed to transfer of specific integrative and conjugative elements (ICElands) – a special type of genome islands – across bacterial populations (van der Meer and Sentchilo 2003; Cruz and Davies 2000).

Examples include the ICElands containing the *clc* element for chlorobenzoate and chlorocatechol degradation in *Pseudomonas* sp. strain B13 or in *Ralstonia* spp. strain JS705. It may be mentioned in this context that the xenobiotic degradation pathways usually require complex genetic systems like operons of ten or more genes or even regulons of several operons along with their control circuits. For instance, in *Sphingomonas aromaticivorans*, there are 15 gene clusters – directly associated with the catabolism or transport of aromatic compounds – in a large conjugative plasmid pNL1 that have enabled the host bacteria to degrade compounds such as biphenyl, naphthalene, xylene, and cresol (Romine et al. 1999).

The same or similar GIs may carry out distinct functions in different species, depending upon the genetic background and lifestyle of its hosts (Dutta and Paul 2012). For instance, GIs carrying secretion systems of type III in the pathogenic strains of *Salmonella*, *Shigella*, and *Yersinia* groups or type IV in *Legionella pneumophila* and *Helicobacter pylori* are known to be involved in the infectious process of their respective hosts. But similar GIs encoding the type III system of rhizobia or the type IV system of F plasmids function as symbiotic islands that enhance the fitness of the host organisms in their natural niches. GIs encoding the adherence factors like P-, S-, and F1C-fimbriae in *E. coli* strains of the human gut microbiome act as a saprophytic island that facilitate colonization of these microbes at the gut. But if under special circumstances the P-, S-, or F1C-positive *E. coli* reaches the urinary tract of human, the same island may serve as a pathogenicity island that helps its host microbe to infect the bladder/kidney.

Antibiotic Resistance

A major health concern over past few decades is the emergence of numerous antibiotic-resistant pathogenic strains. Horizontal gene transfer is one of the major reasons for the dissemination of various antibiotic-resistant factors throughout diverse microbial species. The resistant genes located in various mobile DNA elements (such as plasmids) are easily transferred from one

species to another especially in hospital environment among close contaminants and in patients with compromised immunity, thus resulting in nosocomial infections caused by multidrug resistance bacterial strain.

Among different antibiotic-resistant classes of organisms, the cases of two most widely studied phenotypes include resistance to β -lactams and resistance to fluoroquinolones (Barlow 2009). β -lactam antibiotics, one of the major groups of antibiotics used globally, act by inhibiting bacterial cell wall biosynthesis mainly in gram-positive bacteria. They contain a β -lactam ring in their structures and require this ring to be intact in order to be effective. The transfer of β -lactamase (acts by invading the β -lactam ring) gene into many previously sensitive strains, predicted to be transferred from different gram-negative species such as *E. coli*, has resulted in various pathogenic strains resistant to most available antimicrobials. One most cited example is methicillin-resistant *Staphylococcus aureus* (MRSA), one of the most virulent strain of *S. aureus*, resistant to most β -lactams. In addition to the β -lactamase activity, another gene *mecA* is found to be associated with resistance to most β -lactams. This gene acts by producing an altered penicillin-binding protein having lower affinity for β -lactam antibiotics. Another group of antibiotics, the fluoroquinolones (cephalosporin), effective against many gram-negative bacteria, widely used in both human medicine and veterinary practice is also becoming less functional because of the growing incidence of resistant strains.

Different strains of *Enterococci*, a natural commensal in human gut, have shown to contribute in several cases of HGT due to having a large number of plasmids. Cases of vancomycin resistance in *E. faecalis* and *E. faecium* have been shown to be mediated through a type of pheromone-independent plasmids (Palmer et al. 2010). Recent cases showing plasmid-mediated transfer of vancomycin resistance from *Enterococci* to MRSA are producing an alarming rate of last line antibiotic failure, thus leading to combined growth of nosocomial pathogens having no effective antibiotic.

Evolutionary or Ecological Implications of HGT

Discoveries of rampant interspecies gene transfer across the entire microbial world and even beyond have underscored the need for reviewing the basic concepts of biological evolution. As proposed by Doolittle (1999), a single universal phylogenetic tree might not be the best way to depict relationships between all living and extinct species. Instead, a web- or netlike pattern might provide a more appropriate representation (Doolittle 1999). It appears that some genes have flowed “randomly” through the biosphere, almost as if all life forms constituted one global superorganism, divided into subpopulations, within and between which genes are exchanged at varying frequencies (Dutta and Pan 2002).

The microbial niches are also no longer considered as a static domain. A microbial niche may be considered as a dynamic domain, which is continuously being redefined after each gene transfer event. This alternation of niche boundaries then imposes a different filter on the influx of foreign DNA, imparting distinct selective pressures on incoming genes. Recently Martin et al. (2013) proposed a new model of ecological speciation via gradual genetic isolation, instigated by differential niche acclimatization of nascent bacterial populations. The model predicted how microbial populations, despite having ecological cohesion, can display high genomic diversity through employment of selective, local HGT events, by tapping into a gene pool that is adaptive toward continuously changing, local organismic interactions.

Pervasiveness of HGT across the entire living world has also redefined the concept of the “universal ancestor” (Woese 1998). The presence of a gene in all three domains of life – Bacteria, Archaea, and Eukarya – not necessarily ensures its existence in their common ancestor; it could have arisen at a later age in one domain and spread to the others. As stated by Woese (1998): “the universal ancestor is not a discrete entity. It is, rather, a diverse community of cells that survives and evolves as a biological unit.”

Summary

Horizontal gene transfer (HGT) – the process of interspecies transfer of genetic material via mobile genetic elements such as plasmids, phages, genomic islands, and genomic modules – plays an important role in bacterial evolution, speciation, and diversification. HGT is pervasive among bacteria and may occur at vast phylogenetic distances. By introducing fully functional genes and gene clusters, an event of HGT may confer novel phenotypes and functions to the host organism; may result in drastic changes in its metabolic, pathological, or ecological character, thereby allowing effective and competitive exploitation of new niches; and even can generate new variants of bacterial strains by “genetic quantum leaps.” The widespread distribution of various antibiotic-resistant genes throughout diverse microbial species, dissemination of the gene clusters (operons) involved in biodegradative pathways, transformation of various bacterial organism from the benign form into a pathogen, evolution of rhizobia–legume symbiosis or interstrain variations in size, and macro-organization in chromosomal structures of any specific bacterial species all may be attributed to HGT. Genetic elements that can be transferred as a functional unit and provide a niche-transcending adaptation have a greater probability of undergoing a successful adaptive HGT. Informational genes involved in replication translation and transcription are, in general, less transferable than the operational genes involved in metabolism. Stabilization of the transferred material within the host is often limited by the genetic and ecological similarity of the donor and the recipient. Recognition of HGT as a prime factor for bacterial speciation and diversification has revolutionized the basic concepts of biological evolution. It has been proposed that all prokaryotes together might be considered as one “global superorganism” divided into subpopulations within and across which genes are frequently exchanged. It has also been proposed that the bacterial niches and HGT constantly interact with one another, each affecting the other as lineages evolve.

Reference

- Barlow M. What antimicrobial resistance has taught us about horizontal gene transfer. *Methods Mol Biol.* 2009;532:397–411.
- Cohan FM, Koeppel AF. The origins of ecological diversity in prokaryotes. *Curr Biol.* 2008;18:R1024–34.
- de la Cruz F, Davies J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 2000;8:128.
- Doolittle WF. Lateral genomics. *Trends Cell Biol.* 1999;9:M5–8.
- Doolittle WF, Papke RT. Genomics and the bacterial species problem. *Genome Biol.* 2006;7:116.
- Dorman CJ. H-NS: a universal regulator for a dynamic genome. *Nat Rev Microbiol.* 2004;2:391–400.
- Dutta C, Pan A. Horizontal gene transfer and bacterial diversity. *J Biosci.* 2002;27 Suppl 1:27–33.
- Dutta C, Paul S. Microbial lifestyle and genome signatures. *Curr Genomics.* 2012;13:153–62.
- Gemski P, Lazere JR, Casey T, Wohlhieter JA. Presence of a virulence-associated plasmid in *Yersinia pseudotuberculosis*. *Infect Immun.* 1980;28(3):1044–7.
- Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol.* 2002;19:2226–38.
- Groisman EA, Ochman H. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell.* 1996;87:791–4.
- Hacker J, Blum-Oehler G, Mühlendorfer I, Tschäpe H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol.* 1997;23:1089–97.
- Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol.* 2000;54:641–79.
- Isberg RR, Falkow S. A single genetic locus encoded by *Yersinia pseudotuberculosis* permits invasion of cultured animal cells by *Escherichia coli* K-12. *Nature.* 1985;317(6034):262–4.
- Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A.* 1999;96:801–3806.
- Lawrence JG. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol.* 1999;2:519–23.
- Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol.* 1997;44:383–97.
- Martin F, Polz MF, Alm EJ, Hanage WP. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Gen.* 2013;29:170–5.
- Maurelli AT, Baudry B, d’Hauteville H, Hale TL, Sansonetti PJ. Cloning of plasmid DNA sequences involved in invasion of HeLa cells by *Shigella flexneri*. *Infect Immun.* 1985;49(1):164–71.
- McDaniel TK, Kaper JB. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol Microbiol.* 1997;23(2):399–407.
- Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet.* 2004;36:1126.
- Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature.* 2000;405:299–304.
- Palmer KL, Kos VN, Gilmore MS. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Curr Opin Microbiol.* 2010;13(5):632–9.
- Portnoy DA, Falkow S. Virulence-associated plasmids from *Yersinia enterocolitica* and *Yersinia pestis*. *J Bacteriol.* 1981;148(3):877–83.
- Romine MF, Stillwell LC, Wong KK, Thurston SJ, Sisk EC, Sensen C, Gaasterland T, Fredrickson JK, Saffer JD. Complete sequence of a 184-kilobase catabolic plasmid from *Sphingomonas aromaticivorans* F199. *J Bacteriol.* 1999;181(5):1585–602.
- Sasakawa C, Kamata K, Sakai T, Makino S, Yamada M, Okada N, Yoshikawa M. Virulence-associated genetic regions comprising 31 kilobases of the 230-kilobase plasmid in *Shigella flexneri* 2a. *Bacteriol.* 1988;170(6):2480–4.
- Sullivan JT, Ronson CW. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci U S A.* 1998;95:5145–9.
- Thomas CM, Nelsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.* 2005;3:711.
- van der Meer JR, Senthilo V. Genomic islands and the evolution of catabolic pathways in bacteria. *Curr Opin Biotechnol.* 2003;14:248–54.
- Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev.* 2011;35:957–76.
- Woese C. The universal ancestor. *Proc Natl Acad Sci U S A.* 1998;95:6854–9.

Host-Virus Interaction: From Metagenomics to Single-Cell Genomics

Arbel D. Tadmor¹ and Rob Phillips²

¹TRON – Translational Oncology at the University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

²Departments of Applied Physics and Bioengineering California Institute of Technology, California Institute of Technology, Pasadena, CA, USA

Synonyms

DNA packaging gene; Large terminase subunit gene; TerL

Definition

dPCR (digital PCR) is a PCR reaction performed in a nanoliter or subnanoliter volume making it possible to detect single molecules.

MetaCAT (metagenome cluster analysis tool) is a metagenome data mining tool that uses an iterative dynamic clustering approach to identify the most abundant genes in a given metagenome with respect to a reference dataset containing potentially homologous genes.

Bacteriophage is a virus that infects and replicates within bacteria (*phage* for short).

Prophage is a phage genome that is integrated into the bacterial genome or exists in the form of a plasmid within the cell.

Introduction

It is widely appreciated today that viruses are a dominant and critical part of Earth's biosphere. Yet despite the major advances in the study of environmental viruses in most cases, our knowledge of which viruses go with which hosts is meager. In the classic phage isolation technique, known as the plaque assay, a confluent layer of host cells is infected with a low density of viral particles. When a virus infects a cell within this "lawn" of host cells, the cell lyses, and new viral particles infect adjacent cells thereby creating a clearing, or plaque, in the lawn. This technique for isolating viruses requires that the host of the virus be culturable. However, given that >99% of microbes on Earth cannot be cultured at this time, the vast majority of phage-host systems cannot be investigated in the laboratory using these classical phage enrichment techniques. Consequently, little is known about the biology of most viruses and their host specificity in the wild.

Metagenomic studies of environmental viruses circumvent the cultivation limitation and therefore have offered a unique glimpse into the genetic composition of environmental viruses (Kristensen et al. 2010; Mokili et al. 2012). In low complexity environments such as natural acidophilic biofilms, metagenomic analysis can utilize antiviral defense systems known as

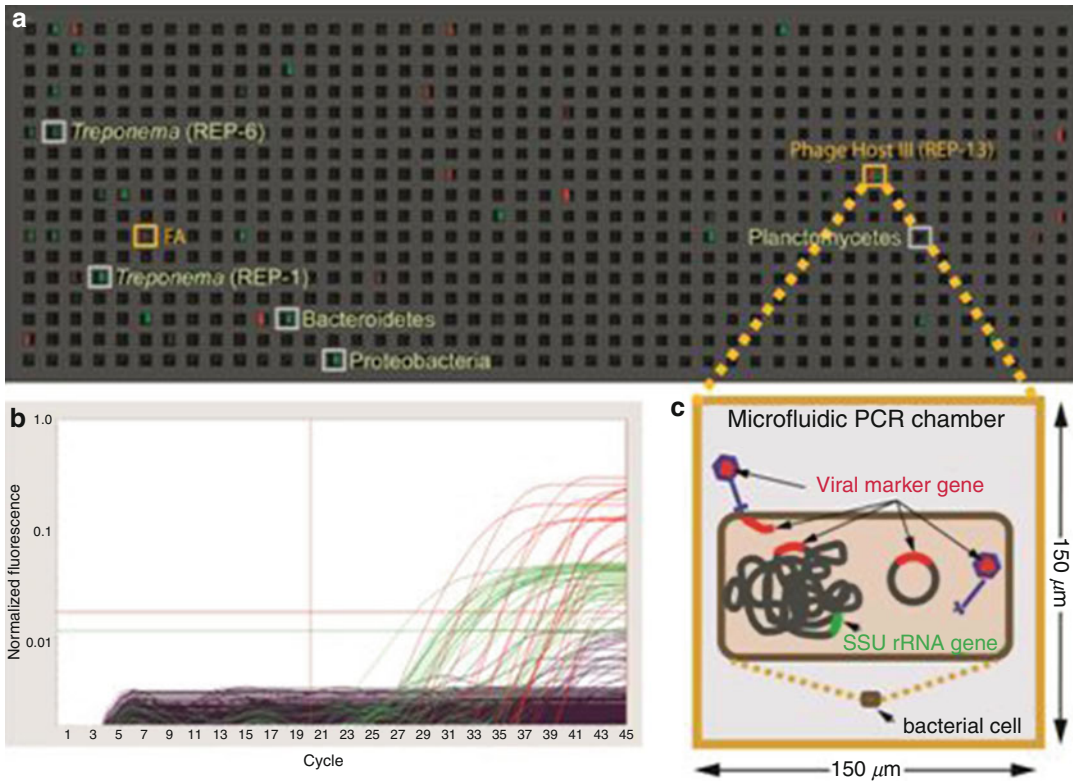
CRISPRs to pair viruses with their hosts by matching spacer sequences that occur both in the genome of the virus and in the genome of the host (Andersson and Banfield 2008). Metagenomes, however, are generally limited in their ability to shed light on the nature of host-virus interaction since most environments are more complex. More importantly, the physical entity of the cell is lost in the process of preparing the metagenome thus destroying the possibility of assigning a given virus to a corresponding host. A way to circumvent this problem is through culture-independent single-cell analysis methods that use microfluidic devices to trap and manipulate single cells.

Single-Cell Genomics

Microfluidic devices are currently routinely used to control and manipulate small volumes of liquid, including trapping and analyzing single cells (Kalisky et al. 2011). Once trapped, individual cells are lysed, and their genetic content can be probed. In the case of host-virus interaction, the genome of the virus forms a unique association with the bacterial cell (Fig. 1). Thus, in an ideal scenario, both the genome of the host and its virus would be sequenced. Although single-cell sequencing has been demonstrated (Kalisky et al. 2011; Kalisky and Quake 2011), for practical reasons, the number of cells that may be interrogated using this method remains at present quite low. As an alternative approach, it is possible to analyze single bacterial cells by PCR using microfluidic digital polymerase chain reaction (dPCR) arrays. This method, which is relatively high throughput, can currently interrogate several thousands of single cells within days.

Microfluidic Digital PCR

In microfluidic digital PCR a sample consisting of either DNA or cells is partitioned uniformly onto an "array" of nanoliter or subnanoliter PCR chambers, with each chamber ideally containing a single DNA molecule or a single cell



Host-Virus Interaction: From Metagenomics to Single-Cell Genomics, Fig. 1 End point fluorescence measured in a panel of a microfluidic digital PCR array. (a) The measured end point fluorescence from the rRNA channel (right half of each chamber) and the terminase channel (left half of each chamber) in a microfluidic array panel. Each panel in this array (one of twelve) consists of 765 $150 \times 150 \times 270 \mu\text{m}^3$ (6 nL) reaction chambers. Retrieved colocalizations are outlined in orange, and positive rRNA chambers randomly selected

for retrieval are outlined in gray. FA indicates false alarm (a probable terminase primer-dimer). (b) Normalized amplification curves of all chambers in (a) after linear derivative baseline correction (red/viral, green/rRNA). (c) Specific physical associations between a bacterial cell and the viral marker gene resulting in colocalization include, for example, an attached or assembling virion, an injected DNA, an integrated prophage, or a plasmid containing the viral marker gene (Tadmor et al. 2011)

(Kalisky et al. 2011; Kalisky and Quake 2011). Each chamber is loaded with a mixture of primers and fluorescent probes that target the loci of interest. The advantage of performing quantitative PCR (qPCR) reactions in such tiny volumes is that the likelihood of contamination is reduced, and the fluorescent signal per PCR chamber is greatly intensified. In a standard benchtop qPCR reaction, for example, the reaction volume is 15 μl compared to a dPCR reaction volume of 6 nL. Therefore dPCR Ct values are reduced by about $\log_2(2,500) = 11.3$ cycles. In addition, the large dilution factor ensures that the vast majority of dPCR chambers are free from

spurious reactions and contaminating molecules such as residual genomic DNA that is intrinsic to some reagents. These factors together provide the sensitivity required to PCR amplify and detect single molecules. Once thermocycling is completed, chambers containing the targets of interest are identified via the fluorescent signal, sampled and post-amplified in the laboratory for sequencing using conventional benchtop PCR machines. An appealing aspect of this technology is that cells may be harvested directly from the environment and loaded onto a microfluidic dPCR array. This method therefore does not require that cells be cultured beforehand and does not depend on

gene expression, the position of the targets in the genome or on the physiological state of the cell at the time of harvest (Ottesen et al. 2006).

The first application of microfluidic digital PCR technology to study environmental bacteria involved colocalization of two genes present in the same individual bacterium (Ottesen et al. 2006). In this study, one marker targeted an important functional gene expressed by certain members of the microbial community resident in the hindgut of termites, and the second marker targeted the small subunit ribosomal RNA (SSU rRNA) gene that was used to phylogenetically identify the bacterium (also known as the 16S marker). By colocalizing and subsequently sequencing both markers from many individual cells, the identity of cells carrying the functional gene was ascertained in cases of repeated colocalizations.

To study host-virus interaction, the dPCR approach described above was extended to colocalize the SSU rRNA gene with a marker targeting a certain viral gene prevalent in the environment of interest, demonstrating proof-of-concept on the termite system (Tadmor et al. 2011). Targeting viruses, however, which are fundamentally different biological entities than bacteria, raised certain questions that needed to be addressed. First and foremost, unlike prokaryotes that have universal markers such as the SSU rRNA gene, viruses do not have a single shared gene that can be used as a universal marker (Rohwer and Edwards 2002). In fact, viral metagenomic studies have shown that viruses are likely the largest reservoirs of unknown genetic diversity with the majority of putative viral sequences exhibiting no significant similarity to currently known genes (Edwards and Rohwer 2005; Kristensen et al. 2010; Mokili et al. 2012). To make matters worse, viruses are notorious for replicating their genetic material with borderline fidelity. Consequently the definition of a viral gene in the environment is relatively fluid. Finally, many genes present in the genome of the virus may be of prokaryote origin making them poor signature markers for the virus. Thus, to utilize digital PCR to study host-virus interaction, an unequivocal viral marker

that is ubiquitous in the environment of interest should be identified.

Requirements from a Viral Marker Gene

Not all viral genes are suitable to be unequivocal markers of a viral entity. As an example, the integrase gene, which codes for an enzyme that is used by the virus to integrate into the genome, is prevalent not only among phages, but also among certain nonviral entities such as plasmids, pathogenicity islands, and integrons (Casjens 2003). Similar arguments apply to viral genes involving lysis, regulation of gene expression, and DNA replication in viruses (Casjens 2003). Casjens therefore argues that ideal “cornerstone” phage genes (or at least prophages genes) are genes involved in the assembly of the virion. Of these, genes that appear to be not only virus specific but also particularly conserved are the large terminase subunit (TerL) and portal protein genes (Casjens 2003).

TerL genes have certain additional features that make them particularly attractive as viral markers. The TerL gene is a component of the DNA packaging and cleaving mechanism present in numerous double-stranded DNA phages (Rao and Feiss 2008). It contains an N-terminal ATPase domain, which is the “engine” of the DNA packaging motor, and a C-terminal nuclease domain (Rao and Feiss 2008). The ATPase domain of the TerL gene is conserved in a wide variety of dsDNA phages, including the eukaryotic herpes virus (Przech et al. 2003), suggesting it is an ancient viral domain (Rao and Feiss 2008). Indeed, Koonin et al., who define “hallmark viral genes” as “genes shared by many diverse groups of viruses with only distant homologs in cellular organisms and with strong indications of monophyly of all viral members of the respective gene families” and thus “can be viewed as distinguishing characters of the virus state” (Koonin et al. 2006), identified the ATPase subunit of the terminase gene as such a hallmark viral gene. Since TerL genes have particularly well-conserved functional residues and motifs (Rao and Feiss 2008), they are well suited for

degenerate primer design. At the same time, across biology TerL genes do not share overall significant sequence similarity (Rao and Feiss 2008), thereby making them sensitive viral markers.

Targeting a “cornerstone” or “hallmark” gene of a virus may, however, be of questionable use if the selected marker tags a defective prophage. Since a necessary condition for the virus to be active is that its cornerstone gene be functional, it is important to verify that the cornerstone gene is under negative selection pressure (Nei and Kumar 2000). Nonfunctional genes may contain errant stop codons, frameshift mutations, or mutations in certain highly conserved residues essential for the proper function of the protein.

Yet demonstrating that a particular family of TerL alleles from the environment of interest is under negative selection pressure does not guarantee that the virus is active in this environment since a viral gene may remain functional while the prophage itself is defective. Such a situation can occur if there was insufficient time for point mutations to have accumulated in the gene of interest after the prophage was inactivated (Casjens 2003). Therefore, viruses carrying the viral marker may have been active only in recent evolutionary history. In an alternative scenario, the putative marker indeed degraded over time upon prophage inactivation, but it was subsequently repaired by a recombination event with another phage that was likely functional (Casjens 2003). In such a case the marker can serve as an indicator for indirect infection.

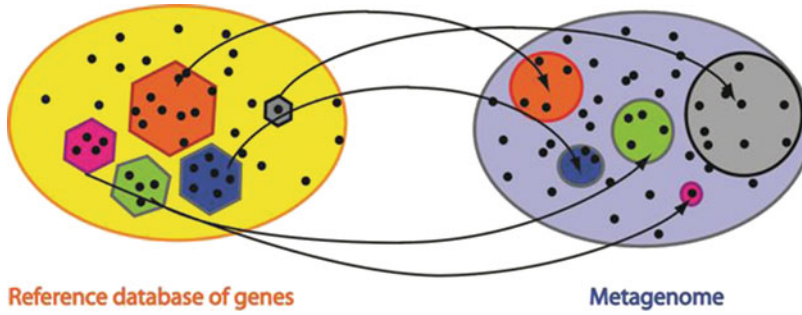
Another possibility is that the putative marker was recruited by the bacterium because it confers on the bacterium a competitive advantage, as is the case with lysogenic conversion genes. In this case the gene would remain under negative selection pressure, while the rest of the prophage degenerated (Boyd and Brüssow 2002; Casjens 2003). It is unlikely, however, that the host will recruit TerL genes since these are highly specialized motors required for virion synthesis. Alternatively, the putative marker could be part of a functional non-phage entity such as a gene transfer agent (GTA) or a bacteriocin (Casjens 2003). In the case of TerL genes, bacteriocins can be ruled out since these tail-like structures do not

contain head-related proteins (Daw and Falkner 1996). GTAs can only be ruled out if the entire genome of the putative viral entity is obtained. Full length viral genomes may be obtained by means of single-cell sequencing techniques.

Identifying Ubiquitous Viral Markers

Although universally shared viral genes do not exist, it is beneficial to select a viral marker that is ubiquitous in the environment of interest. Ubiquitous markers not only have the potential to recover greater genetic diversity from the environment, but can possibly also be found in similar or related environments. Identification of a ubiquitous viral marker in the environment of interest, assuming one exists, is not straightforward and requires sophisticated metagenome data mining approaches.

To address this problem the authors developed a bioinformatic program called MetaCAT (metagenome cluster analysis tool), which employs a heuristic clustering and ranking approach that aims to identify the most abundant genes of a given class (e.g., viral genes) in a metagenome, without relying on superficial features such as gene annotation (Tadmor et al. 2011). The input to MetaCAT is a metagenome (either assembled translated contigs or raw nucleotide reads) and a reference library of known genes (e.g., all known viral genes). The output of MetaCAT is a list of known reference genes (derived from an input reference library) that were found to be present in the metagenome, ranked by their abundance in the metagenome. Abundance of a reported gene is defined as the number of metagenome gene objects or reads that yield significant alignments with respect to this gene. A key feature of MetaCAT is that it uses an iterative dynamic clustering algorithm to group putative homologous reference genes from the input reference library. The clustering is dynamic in the sense that it is performed on the fly based on the matches found in the given metagenome, thereby avoiding loss of information that would occur if the reference library was a priori clustered. The clustering is performed iteratively until all



Host-Virus Interaction: From Metagenomics to Single-Cell Genomics, Fig. 2 Schematic illustration of the MetaCAT algorithm. MetaCAT maps clusters of similar known reference genes to groups of metagenome gene objects or reads. MetaCAT defines two known reference genes as being similar or “related” if their footprint in the metagenome has a significant overlap. The abundance of a given cluster of related known reference genes in the metagenome is defined as the number of metagenome gene objects (or reads) with an E value below a given threshold found when BLASTing a representative from the gene cluster against the metagenome. The key feature of MetaCAT lies in its

ability to cluster the list of known reference genes per metagenome and report a minimally redundant list of known genes that have putative homologs in the metagenome, ranked by their abundance in the metagenome. This list can then be used to generate hypotheses about the given metagenome. In this figure the left oval depicts a reference database of genes (black dots), and the right oval depicts a metagenome, with black dots representing metagenome gene objects. Hexagons in the reference database represent clusters of related reference genes identified by MetaCAT. Each hexagon is linked to a corresponding cluster of metagenome gene objects depicted by a circle of matching color

identified redundancy is removed. In this way the final reported list of ranked genes (or clusters of genes) contains orders of magnitude fewer elements than the reference library and is amenable to manual inspection (Fig. 2).

If gene annotation information is included in the reference database this information will be provided by MetaCAT in the ranked list of genes making it a straightforward task to identify genes of interest. As an example, Table 1 lists all the TerL genes identified by MetaCAT in the metagenome of the hindgut of a higher termite collected from Costa Rica (Tadmor et al. 2011).

Each known reference gene found by MetaCAT to be present in the metagenome can be paired with a metagenome gene object that yielded the lowest E value. This metagenome gene object is referred to as the “representative contig” of the known reference gene. By BLASTing the representative contigs corresponding to the top ranking candidate markers against other metagenomes from similar environments, or against genomes of organisms isolated from similar environments, it is possible to identify ubiquitous viral genes, if present (Fig. 3). Closely related genes found in multiple

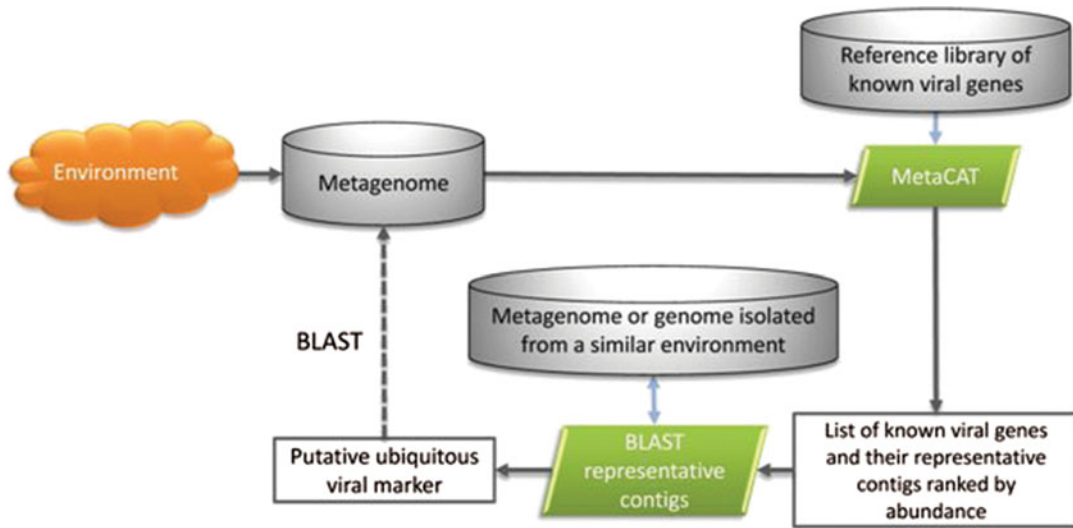
datasets are favorable candidates for putative ubiquitous markers.

In the case of the termite hindgut, the list of representative contigs corresponding to reported genes in Table 1 was BLASTed against the genome of *Treponema primitia*, a spirochete isolated from a lower termite collected from Northern California. Performing this analysis revealed that the representative contig of the top ranking gene found by MetaCAT indeed had significant hits (E value of $\sim 10^{-5}$) in the genome of *T. primitia* and mapped to two prophage-like elements. In this case, BLASTing the TerL gene from the prophage-like element back against the metagenome revealed close homologs with a similarity of 70 to 78% identity (Tadmor et al. 2011). Such a bootstrapping approach enabled the identification of a ubiquitous viral marker in the termite hindgut environment. Indeed, degenerate primers designed against this marker were able to amplify closely related homologs of this marker in other species of termites (as well as a wood-feeding roach) collected from various locations in the United States (Tadmor et al. 2011).

In this context, it is worthwhile to mention that MetaCAT is not restricted to ranking only viral

Host-Virus Interaction: From Metagenomics to Single-Cell Genomics, Table 1 TerL genes identified by MetaCAT in the metagenome of a hindgut of a higher termite collected from Costa Rica. The following table lists TerL genes with minimal E values $\leq 10^{-7}$ and abundances ≥ 5 that were identified by MetaCAT to have putative homologs in the metagenome of the hindgut of a *Nasutitermes* sp. termite. TerL genes are ranked by the number of metagenome gene objects yielding alignments with E value scores below 0.001. Also shown are the E value scores obtained when BLASTing the representative contig of each RefSeqTerL gene cluster against the genome of *Treponema primitia* (ZAS-2), using a cutoff value of 0.01, with values above this threshold marked as not significant (N.S.)

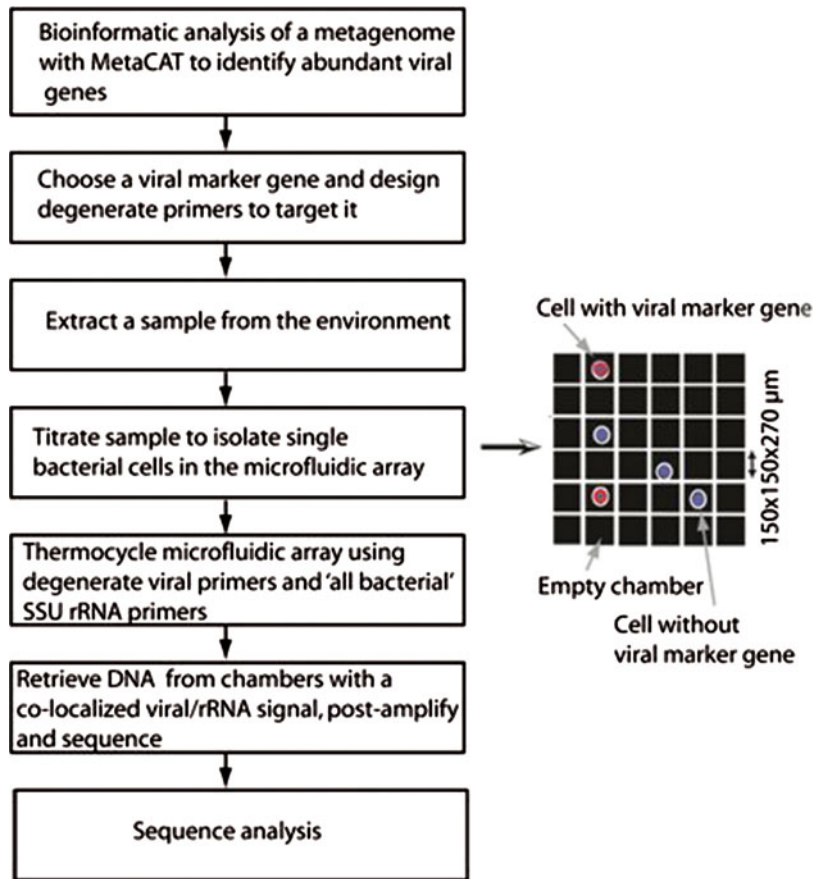
Organism name	Virus classification		No. of hits in metagenome	BLAST RefSeq gene against metagenome (E value)	BLAST representative contig against ZAS-2 (E value)
Clostridium phage phiC2	dsDNA viruses	Caudovirales; Myoviridae	19	4.0E-40	2.0E-05
Streptococcus phage SMP	dsDNA viruses	Caudovirales	11	3.0E-34	N.S.
Pseudomonas phage PaP3	dsDNA viruses	Caudovirales; Podoviridae	7	2.0E-09	N.S.
Enterobacteria phage lambda	dsDNA viruses	Caudovirales; Siphoviridae	6	2.0E-180	N.S.
Enterobacteria phage HK022	dsDNA viruses	Caudovirales; Siphoviridae	6	8.0E-69	N.S.



Host-Virus Interaction: From Metagenomics to Single-Cell Genomics, Fig. 3 Bioinformatic approach to identify ubiquitous viral markers in a given environment. In the proposed approach to identify putative ubiquitous viral markers, a metagenome from a given environment is first analyzed by MetaCAT to produce a list of candidate viral genes abundant in the metagenome. The corresponding representative contig of each candidate viral gene (defined as the contig yielding the lowest E value) is then BLASTed against a second

dataset, such as another metagenome from a similar environment or a genome of an isolate from a similar environment. If the percent identity is sufficiently high allowing for degenerate primer design, this candidate can be considered a putative viral marker and can be further evaluated by experiment. If the percent identity is not high, but the E value is significant, a bootstrap-like approach may be employed where the contig/gene from the new dataset is BLASTed back against the original metagenome, thereby potentially revealing more conserved markers.

Host-Virus Interaction: From Metagenomics to Single-Cell Genomics, Fig. 4 Workflow using the microfluidic digital PCR array for host-virus colocalization in a novel environmental sample (Tadmor et al. 2011)



genes, but it is possible to define other taxonomic groups as input reference libraries. For example, one can use MetaCAT to find the most abundant genes in a given environment involved in a certain metabolic pathway, the most abundant mitochondrion-related genes in a given sample, the most abundant antibiotic genes in a soil sample, and so on. MetaCAT can therefore be thought of as a useful tool for generating hypotheses regarding a given environment. (Requests to obtain a beta version of MetaCAT may be sent to arbel.tadmor@tron-mainz.de.)

Colocalizing Viral-SSU rRNA Genes on Digital PCR Arrays

Once a viral marker has been selected, a diversity of this marker can be retrieved from various

bioinformatic sources (e.g., metagenomes and sequenced genomes), and degenerate primers targeting the marker of interest may be designed. Colocalization of viral genes is, however, complicated by the fact that the low replication fidelity of viruses makes it unlikely to recover the exact same allele twice from the dPCR arrays, contrary to the case of colocalizing two bacterial genes. P-values can, nevertheless, still be assigned to repeated SSU rRNA ribotypes retrieved from a given array, irrespective of the paired gene, using knowledge of the frequency of the given ribotype on the array. It is possible to estimate ribotype frequencies by randomly selecting chambers positive for the SSU rRNA gene and constructing a phylogenetic library of array ribotypes (Tadmor et al. 2011). Host-phage cophylogeny can then be reconstructed from genuine colocalizations, providing a unique glimpse

into the evolutionary dynamics of the system and shedding light on the flow of viral genes between hosts in the given environment. An overview of the workflow using dPCR to colocalize host-virus genes is provided in Fig. 4.

Summary and Outlook

The method outlined in this review provides a general scheme for analyzing host-virus interactions at the single-cell level without having to culture either host or virus. The method first involves a bioinformatic analysis of a metagenomic dataset or datasets from the environment of interest to recover a ubiquitous viral marker. This marker is then colocalized with a universal gene identifying the host by means of dPCR performed on single cells. The methods presented in this review are general and can be applied to other environments.

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution](#)
- ▶ [Viral MetaGenome Annotation Pipeline](#)

References

- Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*. 2008;320(5879):1047–50.
- Boyd EF, Brüssow H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*. 2002;10(11):521–9.
- Casjens S. Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol*. 2003;49(2):277–300.
- Daw MA, Falkiner FR. Bacteriocins: nature, function and structure. *Micron*. 1996;27(6):467–79.
- Edwards R, Rohwer F. Viral metagenomics. *Nat Rev Microbiol*. 2005;3(6):504–10.
- Kalisky T, Quake SR. Single-cell genomics. *Nat Methods*. 2011;8(4):311–4.

- Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet*. 2011;45:431–45.
- Koonin E, Senkevich T, Dolja V. The ancient virus world and evolution of cells. *Biol Direct*. 2006;1(1):29.
- Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol*. 2010;18(1):11–9.
- Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol*. 2012;2(1):63–77.
- Nei M, Kumar S. *Molecular evolution and phylogenetics*. USA: Oxford University Press; 2000.
- Ottesen E, Hong J, Quake S, Leadbetter J. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science*. 2006;314(5804):1464–7.
- Przech AJ, Yu D, Weller SK. Point mutations in exon I of the herpes simplex virus putative terminase subunit, UL15, indicate that the most conserved residues are essential for cleavage and packaging. *J Virol*. 2003;77(17):9613–21.
- Rao VB, Feiss M. The bacteriophage DNA packaging motor. *Annu Rev Genet*. 2008;42:647–81.
- Rohwer F, Edwards R. The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol*. 2002;184(16):4529–35.
- Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science*. 2011;333(6038):58–62.

Human Gut Microbial Genes by Metagenomic Sequencing

Jun Wang
BGI Shenzhen, Shenzhen, China

Synonyms

Genes in the human gut microbial community; Metagenome of the human gut microbiota

Definition

A gene is identified in human distal gut (colon) microbes when reads from high-throughput sequencing of fecal samples are assembled and an open reading frame (ORF) is predicted from the resulting DNA sequence. Such a gene could usually be mapped to a group of bacterial species and linked to certain functions. Metagenomic

studies on other parts of the gastrointestinal tract are often performed invasively using animals and are not discussed here.

Introduction

The human gut has long been known to contain microbial species. Until the advent of high-throughput metagenomic sequencing, however, these mysterious microbes largely eluded interrogations by their human host. Recent advancements described here and in other entries reveal awe-inspiring complexity, dynamics, and significance of the gut microbiota.

Eubacteria dominate the microbial community in the human gut (Scanlan and Marchesi 2008; Marchesi 2010; Parfrey et al. 2011). Both eubacteria and archaeobacteria species are routinely classified to genus level according to their 16S rRNA gene sequences. Unfortunately, taxonomic classification of commensal eukaryotes in the gut has remained a tedious process (Parfrey et al. 2011). As a consequence, our understanding of the eukaryotic minorities in the gut lags far behind that of the bacterial communities. The term “gut microbes” is equivalent to “gut bacteria” hereafter.

Metagenomic sequencing of total DNA extracted from fecal samples constitutes a key step in forging our understanding of gut bacteria beyond taxonomy. The approach allows researchers to obtain complete genome sequences, identify genes, and predict functions. Such metagenomic information is especially precious for those bacteria that are yet to succumb to laboratory culture conditions.

This overview is intended to briefly summarize our current roll call of the various genes present in the human gut flora as well as the functional relevance of these genes to the microorganisms and human beings under normal and perturbed states.

Identification of Gut Microbial Genes

Next-generation, high-throughput, and cost-efficient short-read data (mainly produced by

Illumina sequencing technology) have come of age in metagenomic studies. Considering the nonuniform abundance of gut microbial species and the high level of discordance between individual humans, deep sequencing and wide sampling are critical for a comprehensive understanding of the human gut flora. In 2010, high-throughput short-read sequencing was introduced into human gut microbiome research and showed great potential (Qin et al. 2010).

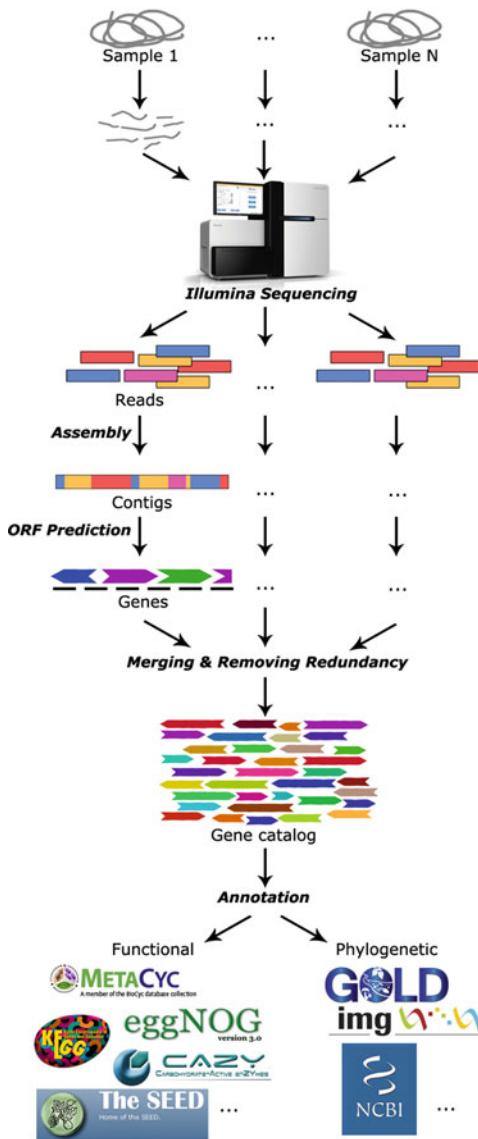
Bacterial DNA obtained from human fecal samples could be readily used for high-throughput sequencing on the Illumina platform. After a few quality control steps, the short reads from each sample were assembled *de novo* (Fig. 1), using software such as SOAPdenovo (Kultima et al. 2012). Protein-coding genes were then predicted from the assembled contigs (Kultima et al. 2012). Genes from multi-samples were pooled together and compared with one another to remove redundancy. Finally, a nonredundant gene catalog was generated and could serve as a basis for functional and phylogenetic analyses (Fig. 1) (Qin et al. 2010).

Alternative to *de novo* assembly, mapping of reads to an existing gene catalog allows convenient identification of genes in a sample. Naturally, such a time-saving approach requires the gene catalog to encompass a complete set of high-quality reference genes.

Total Gene Number and Its Variability

Metagenomic sequencing of 124 Europeans (as part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project) resulted in a gut microbial gene catalog containing 3.3 million nonredundant genes (Qin et al. 2010). Although this gene number might still increase as more samples are sequenced, especially those from patients of a particular disease (e.g., in Qin et al. 2012), this number of known gut microbial genes is already 150-fold greater than the number of genes encoded by the human genome.

Two hundred ninety-four thousand one hundred ten of the gut microbial genes were found in at least 50 % of individuals, which were termed



Human Gut Microbial Genes by Metagenomic Sequencing, Fig. 1 High-throughput metagenomic analysis of the human gut flora. DNA from fecal samples are sequenced using the Illumina platform. The short reads generated are assembled into contigs and open reading frames (ORFs) are predicted. A nonredundant gene catalog is created from the ORFs. The genes are then annotated functionally and phylogenetically according to databases

“common” genes (Qin et al. 2010). The remaining ~90 % genes, although typically seen in multiple samples, were not widely shared. Each individual carried $536,112 \pm 12,167$

nonredundant genes, of which $204,056 \pm 3,603$ (around 38 %) were common genes. Thus, significant interpersonal differences exist in terms of the number, type, and sequence of the genes.

Common Functions Encoded by Gut Bacteria

Functional annotation of the gut metagenome involves aligning the genes to databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, COG (Clusters of Orthologous Groups), and eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) databases (Fig. 1). At present, a significant fraction of genes remain functionally unknown regardless of the database used, although common genes could usually be annotated with greater success. The wealth of information in the gut metagenome awaits exploration in both global and targeted fashion.

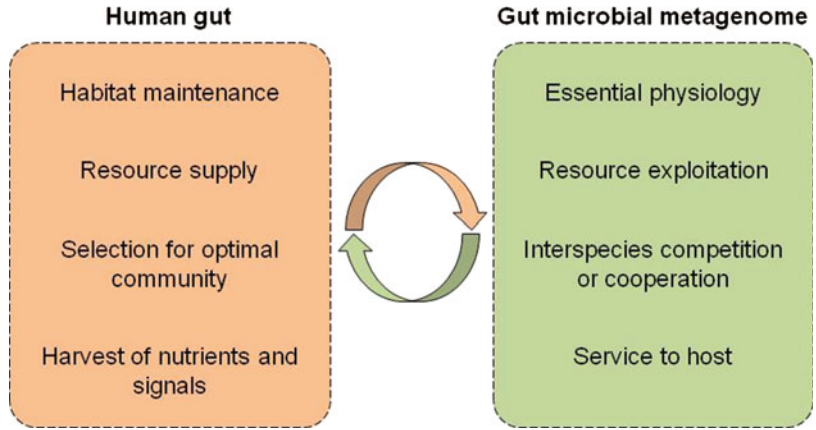
Just as gut microbial genes are to some extent shared between individuals, there are functionalities that are common to the human gut microbiota (Qin et al. 2010; Human Microbiome Project Consortium 2012). Major metabolic pathways such as central carbohydrate metabolism and amino acid synthesis can be seen in all samples. Essential protein complexes, for example, DNA replication machinery, RNA polymerases, ribosome, and secretory apparatus, are also part of the core gut microbiota genes. Moreover, genes not required for all bacteria but are important for life in the gut are expected in the common set. Such genes would presumably reflect adaptation to gut temperature, oxygen level, and nutrients as well as interaction with host cells and other microbes (Fig. 2). The distinction between common and rare functions, however, becomes semantic as one looks into these genes. We find it more convenient to discuss these in the following section.

Genes Influenced by Host Environmental Factors

Traditionally viewed as a place for water and salt resorption, the colon’s integral role in human

Human Gut Microbial Genes by Metagenomic Sequencing,

Fig. 2 Functions encoded by the human gut microbial metagenome in relation to the gut. Gut microbes contain genes important for the survival and success of themselves, at the same time depend on, serve, and manipulate their human host. Diseases follow when the symbiotic relationship goes awry



nutrition has only become realized through studies of the gut (fecal) microbiota. Various digested or indigestible components of the diet arrive at the colon and constitute a major environmental factor shaping the gut microbial ecosystem (Fig. 2). Complex carbohydrates are fermented by bacteria of the phylum *Firmicutes*, producing short-chain fatty acids (SCFAs, including acetate, propionate, and butyrate) for use by the host cells. In contrast, if the host diet relies more on simple sugars, as has become common in the United States, enzymes for metabolizing mono- and disaccharides could be more prominent in the gut flora (Yatsunenko et al. 2012). Similarly, dietary intake of amino acids and vitamins appears to modulate the balance between their catabolism and anabolism by gut bacteria.

Bile acids (BAs) secreted by the host to emulsify dietary fats make a strong impact on the gut microbiota. On one hand, primary BAs are known to be converted to more effective secondary BAs through 7α -dehydroxylation by intestinal bacteria. On the other hand, with their amphipathic properties, BAs show a strong antimicrobial activity. Rats on a diet supplemented with the BA cholic acid recapitulated effects of high-fat diet on the gut flora (reported in mice), namely, an increased ratio of *Firmicutes* to *Bacteroidetes* and a declining microbial diversity (Islam et al. 2011; Ley et al. 2005). Thus, elevated bile secretion stimulated by high-fat diet likely plays a major role in reshaping the gut microbiome (Islam et al. 2011).

Antibiotic administration could lead to profound and long-lasting alterations in the intestinal microbiome (Dethlefsen and Relman 2010; Cho et al. 2012). The distortion is typically manifested as a sharp decrease in microbial diversity accompanied by an overgrowth of *Proteobacteria*, especially in pathogenic *Enterobacteriaceae* populations (Nyberg et al. 2007). Antibiotic intake exerts a strong selective pressure on the intestinal flora and increases transfer of antibiotic-resistant genes (ARGs) among gut microbes, leading to an accumulation of resistance strains (Sullivan et al. 2001; Schj rting and Krogfelt 2011). These antibiotic-resistant pathogens and nonpathogens could persist in the gut well after removal of the selective pressure.

Notably, current evidence suggests that while commensal bacterial species vary between hosts of different genetic background and environmental factors, the individuality is smaller at the functional level, i.e., similar genes in different gut bacteria could serve similar purposes and are selected by similar factors (Spor et al. 2011).

Gut Microbiota and Diseases

A growing body of evidence suggests that the gut microbial flora is central to human health. Although we are very far from a definitive comprehension of healthy versus diseased gut

microbiota, it is fair to say that a productive and well-balanced symbiotic relationship with our little gut residents is of key importance for us human beings. Altered gut microbial composition has been reported in various gut-related diseases such as colorectal cancer and inflammatory bowel diseases (IBDs) and extend to conditions like anorexia, allergies, cardiovascular diseases, and even autism (Clemente et al. 2012; Tremaroli and Bäckhed 2012). These diseases are more or less accompanied by dysbiosis, a state where benign or beneficial gut microbes are overtaken by pathogens and normal processes like fermentation, synthesis of metabolites, barrier function, etc. become disrupted.

On a metagenomic level, the gut microbiome of leptin-deficient obese mice (*ob/ob*) showed an increased capacity for energy harvest from the gut, encoding enzymes that could initiate breakdown of otherwise indigestible polysaccharides (Turnbaugh et al. 2006). However, the end products of bacterial fermentation, SCFAs especially butyrate, appear protective and negatively regulate inflammation in the gut (Maslowski et al. 2009). Butyrate synthesis genes in the gut flora were depleted in diabetes and symptomatic atherosclerosis patients compared to healthy controls (Qin et al. 2012; Karlsson et al. 2012). Together with studies on butyrate and IBDs (Thibault et al. 2010; Scharl and Rogler 2012), current results point to a key role of butyrate metabolism in colon health, with extensive interplays between the gut flora and the host.

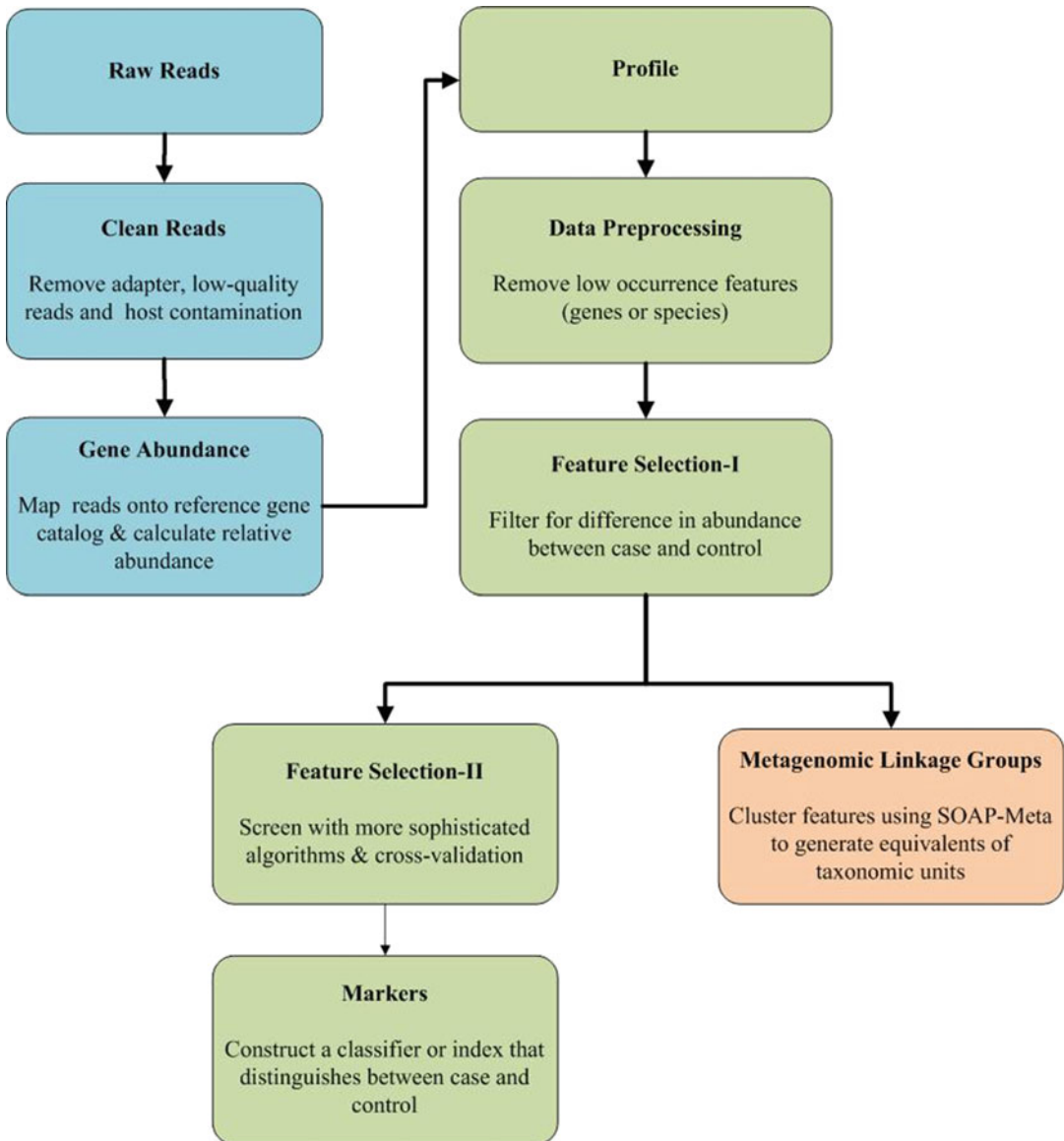
Another common theme in gut microbial homeostasis may be the handling of oxidative stress. The gut metagenome of diabetes patients was enriched for genes involved in sulfate reduction and oxidative stress resistance (Qin et al. 2012). Atherosclerosis was associated with an underrepresentation of phytoene dehydrogenase gene and a matching reduction in the antioxidant β -carotene in patient serum (Karlsson et al. 2012). Oxidative stress is also known to contribute to IBDs such as Crohn's disease (Iborra et al. 2011).

Metagenome-Wide Association Study for Diagnosis

To go beyond a descriptive account of genes present in healthy or unhealthy human gut microbiota, it could be very helpful to perform a metagenome-wide association study (MGWAS) for identification of disease markers and evaluation of disease prospect (Fig. 3). A standard genome-wide association study (GWAS) looks for genetic variants, typically single-nucleotide polymorphisms (SNPs) in a genome, and relates them to a phenotype such as a disease. MGWAS stems from the concept of "metagenome." Accordingly, the relative abundance of a gene in a metagenome, instead of the presence of a SNP, is used to establish correlation with disease.

The proof-of-principle study for MGWAS was performed on type 2 diabetes mellitus (Qin et al. 2012). In a reference gene catalog updated from previous work (Qin et al. 2010), 3,298,811 genes were found in the healthy or diabetic cohorts (total $n = 145$). After filtering for shared genes and clustering based on numerical relationships and phylogeny, the dimensionality was reduced to 1,138,151 genes. The first stage of analysis concluded with 278,168 statistically significant gene markers for diabetes. In Stage II, new samples ($n = 100$ for each cohort) were sequenced and profiled with the markers from Stage I. The analysis further reduced the number of gene markers to 52,484. For lowest error rate, as few as 50 gene markers were found to be optimal and were successfully applied to diabetic/nondiabetic classification of 23 additional samples. Besides gene markers, markers from functional annotations (KEGG orthologous groups, eggNOG orthologous groups) and metagenomic linkage groups (MLG) that represent taxonomic units also present valuable information (Qin et al. 2012).

In addition to diabetes, the same study identified gene markers and orthologous group markers for IBDs and for enterotypes (Qin et al. 2012), raising the stakes for routine application of MGWAS to other microbiota-related diseases. It remains to be seen how factors such as age, gender, and BMI (body mass index) confound MGWAS in various diseases, especially during initial marker



Human Gut Microbial Genes by Metagenomic Sequencing, Fig. 3 Metagenome-wide association study for gut flora-related diseases. For each sample, sequencing reads are mapped to the reference gene catalog (Fig. 1) and relative abundance of genes is computed.

Genes and species that are under- or overrepresented in patients are selected following a rigorous procedure. The analysis results in gene markers and metagenomic linkage groups that can be used for diseased/undiseased classification and potentially for prognosis and diagnosis

selection. Things like sample size, read length, and ecological and genomic diversity all need to be taken into consideration during study design and interpretation. The emergence of an optimum MGWAS workflow and subsequent biological investigations would probably involve effort from researchers across disciplines.

Summary

Metagenomic analyses of the human gut microbiota offer road maps for elucidating the interplay between the gut symbionts and their human host. The information could guide detailed characterization of bacterial species

individually and as a community. The range of metabolites flowing in and out of the microbes and the plasticity of the gut flora are expected to revolutionize nutrition science. Causal relationships between host factors, gut microbiota, and diseases, when established, hold great promise for human health.

References

- Cho I, Yamanishi S, Cox L, Methé BA, Zavadil J, Li K, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*. 2012;488:621–6.
- Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. *Cell*. 2012;148:1258–70.
- Dethlefsen L, Relman DA. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc Natl Acad Sci*. 2010;108:4554–61.
- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486:207–14.
- Iborra M, Moret I, Rausell F, Bastida G, Aguas M, Cerrillo E, et al. Role of oxidative stress and antioxidant enzymes in Crohn's disease. *Biochem Soc Trans*. 2011;39:1102–6.
- Islam KBMS, Fukiya S, Hagio M, Fujii N, Ishizuka S, Ooka T, et al. Bile acid is a host factor that regulates the composition of the cecal microbiota in rats. *Gastroenterology*. 2011;141:1773–81.
- Karlsson FH, Fåk F, Nookaew I, Tremaroli V, Fagerberg B, Petranovic D, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat Commun*. 2012;3:1245.
- Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE*. 2012;7:e47656.
- Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A*. 2005;102:11070–5.
- Marchesi JR. Prokaryotic and eukaryotic diversity of the human gut. *Adv Appl Microbiol*. 2010;72:43–62.
- Maslowski KM, Vieira AT, Ng A, Kranich J, Sierro F, Yu D, et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature*. 2009;461:1282–6.
- Nyberg SD, Osterblad M, Hakanen AJ, Löfmark S, Edlund C, Huovinen P, et al. Long-term antimicrobial resistance in *Escherichia coli* from human intestinal microbiota after administration of clindamycin. *Scand J Infect Dis*. 2007;39:514–20.
- Parfrey LW, Walters WA, Knight R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front Microbiol*. 2011;2:153.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464:59–65.
- Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55–60.
- Scanlan PD, Marchesi JR. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J*. 2008;2:1183–93.
- Scharl M, Rogler G. Inflammatory bowel disease pathogenesis: what is new? *Curr Opin Gastroenterol*. 2012;28:301–9.
- Schjörring S, Krogfelt KA. Assessment of bacterial antibiotic resistance transfer in the gut. *Int J Microbiol*. 2011;2011:312956.
- Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol*. 2011;9:279–90.
- Sullivan A, Edlund C, Nord CE. Effect of antimicrobial agents on the ecological balance of human microflora. *Lancet Infect Dis*. 2001;1:101–14.
- Thibault R, Blachier F, Darcy-Vrillon B, De Coppet P, Bourreille A, Segain J-P. Butyrate utilization by the colonic mucosa in inflammatory bowel diseases: a transport deficiency. *Inflamm Bowel Dis*. 2010;16:684–95.
- Tremaroli V, Bäckhed F. Functional interactions between the gut microbiota and host metabolism. *Nature*. 2012;489:242–9.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*. 2006;444:1027–31.
- Yatsunenkov T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012;486:222–7.

Human Oral Microbiome Database (HOMD)

Tsute Chen¹ and Floyd Dewhirst²

¹Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA

²Department of Molecular Genetics, The Forsyth Institute, Cambridge, MA, USA

Introduction

The human oral cavity is a rich biological site with several microbial niches including teeth, gingival sulcus, tongue, cheek, hard and soft

palates, tonsils, throat, and saliva. The microbiome of the oral cavity (Dewhirst et al. 2010) and its niches have been examined based on 16S rRNA sequencing (Aas et al. 2005; Bik et al. 2010; Human Microbiome Project 2012a, b). The metagenome of the oral cavity has been studied to a limited degree prior to 2012 due to the complexity of the site (Alcaraz et al. 2012; Belda-Ferre et al. 2012; Xie et al. 2010). More than 700 prevalent species comprise the oral microbiome, but many taxa are present at less than 0.1 % of the microbial population (Dewhirst et al. 2010). As oral bacterial reference genomes are becoming available, primarily through the efforts of the Human Microbiome Project (Human Microbiome Project 2012a, b), it is becoming possible to attribute metagenomic sequences to organisms at genus and species level (Martin et al. 2012). The anchoring of metagenome sequence information to specific organisms in a taxonomic framework is key to developing a full description of the bacteria-bacteria and bacteria-host interactions that underlie human oral health and disease.

The Human Oral Microbiome Database (HOMD) was developed in response to the lack of any naming or taxonomic scheme for the thousands of human oral 16S rRNA clone sequences that were being generated in the early 2000s and dumped into GenBank without any taxonomic anchor. Investigators were publishing manuscripts using clone names (such as BU063) as provisional taxonomic names. The only way to phylogenetically place an oral clone was to personally align sequences and generate one's own phylogenetic trees. We recognized that there was a need for a 16S rRNA-based provisional taxonomic scheme to name and provide reference sequences for unnamed taxa known only from clone or isolate 16S rRNA sequences. The naming scheme had to be provisional because formal naming under the bacterial code requires isolation in pure culture and full phenotypic characterization; 16S rRNA sequence by itself is insufficient for formal naming. The taxonomic scheme described more fully below is based on a Human Oral Taxon number which runs currently from 001 to 918.

At about the time we recognized the need to create a taxonomic framework for the oral microbiome, the National Institute of Dental and Craniofacial Research released a request from proposal on "The metagenome of the oral microbiome." We responded with a proposal entitled "A foundation for the oral microbiome and metagenome," which was funded as DE016937. The goals of the grant were to (1) set up the HOMD web-accessible database with a provisional taxonomic scheme and to present all oral genomes in a graphical interface, (2) to complete reference genomes for oral taxa, and (3) to obtain isolates of previously uncultivated taxa and make them available to the research community by placing them in national-type culture collections. We have made steady progress in achieving these goals, and this project is currently in its seventh year of funding.

The HOMD Website

The HOMD contains various types of information on human oral microorganisms including taxonomic, genomic, and bibliographic. The purpose of the HOMD website (<http://www.homd.org>) is to provide an easy-to-use online interface to search, retrieve, and navigate among these different types of information. HOMD also provides web-based bioinformatics software tools for data mining and analyses.

Technically, the HOMD website is constructed using a LAMP system and hosted on the web server computers. The LAMP system provides a Linux operating system, Apache web service, MySQL relational database, and PHP dynamic web page rendering. Textual contents such as the taxonomy and metagenomic information are queried and results dynamically displayed in the web browser by the LAMP system. A dedicated high-performance computer cluster is deployed to handle the computational demanding analysis such as homology sequence searches.

The HOMD has been designed to be compatible with most commonly used web browsers such as Microsoft Internet Explorer, Firefox,

Google Chrome, and Safari. We suggest the use of one of these popular web browsers to ensure the functionalities of HOMD web pages and tools. All the HOMD information and tools are viewable and available to the general public without having to log in or acquiring a user account. The log-in function is mainly for the purpose of maintaining the website and the curation of the database information. If a user has been designated a curator, he or she will see additional administrative submenus.

Detailed functionalities, web interfaces, and tools as well as useful usage tips are presented below. Technical information such as the implementation and design of the HOMD has been published elsewhere (Chen et al. 2010).

Features of the HOMD Web Pages

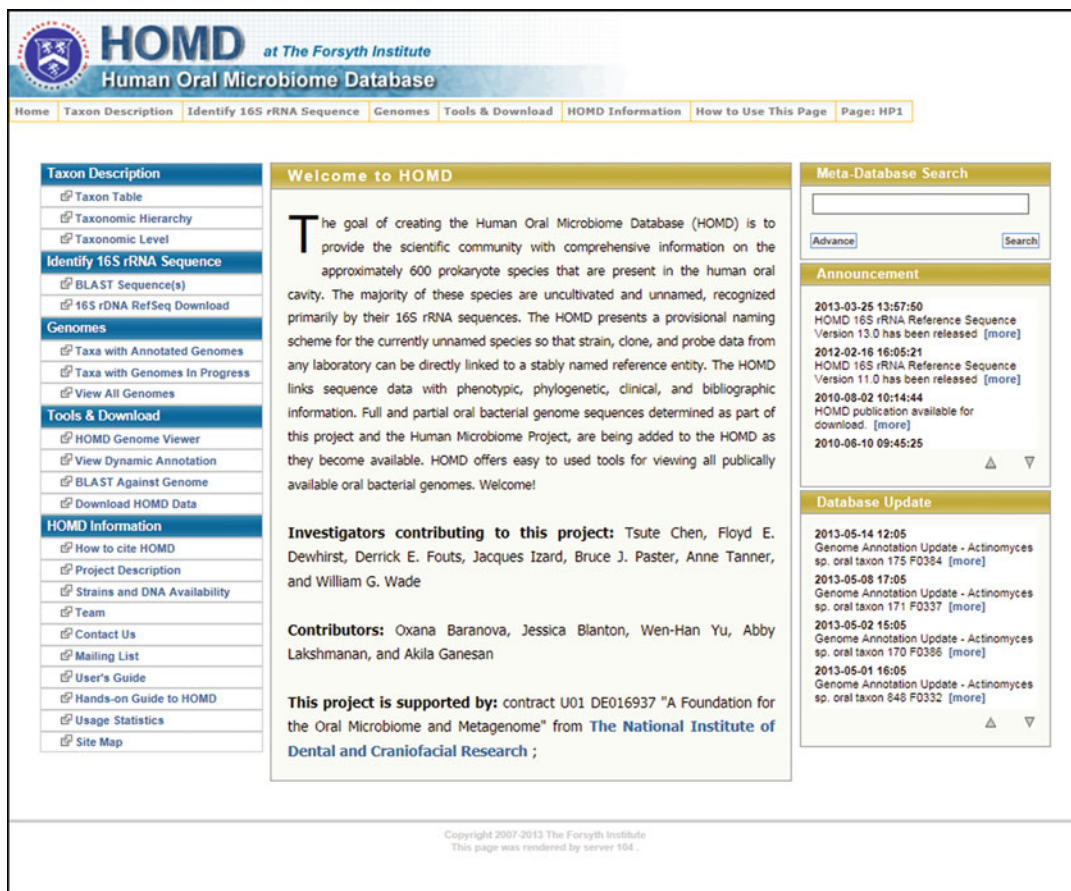
The design of the website was based on the feedback of several researchers in the field of oral microbiology over the past several years. The user interface was designed to be user-friendly, intuitive, and practical. On top of every HOMD page (Fig. 1), there is a top banner for the HOMD logo, which automatically reduces to smaller size (in height) once the user navigates away from the home page so that the banner will not take up too much space from the requested content. Clicking the top banner image also brings the user back to the HOMD home page. Top navigation menu is located right below the top banner and is also accessible throughout all the HOMD pages. The top navigation menu provides access points to all HOMD's tools and information on all the web pages.

Another useful feature of the HOMD web pages is the unique page ID system. The rightmost item displayed on the top navigation menu is the page ID – a unique code that distinctly identifies the current page that a user is viewing. For example, the page ID of the HOMD home page is “HP1” (Fig. 1), and once a user navigates away from home page to, e.g., the Taxon Table page, the page ID automatically changes to “TT1.” This feature allows precise page referencing. This is particularly useful when a user needs to refer to a specific page on HOMD site for discussion, bug reporting, or suggestion.

The HOMD home page also includes a top-down oriented expandable menu on the left side and an introductory paragraph in the center. On the right side are the Meta-Database Search, the Announcement, and the Database Update boxes. The Meta-Database Search is very useful for searching desired information across all the subsets of HOMD databases, including the taxonomy, the metagenomic information, as well as the dynamic genome annotations. The result lists the number of matches to the keyword that provides links, leading to detailed information. The Announcement box displays the important system-wise updates and news for the HOMD. The Database Update box is automatically updated by the HOMD dynamic genome annotation pipeline (see “[Dynamic Annotation of Genomic Sequences](#)” section) to keep track of the status of the genome annotation.

HOMD also provides comprehensive documentation and updates history of data and tools. The HOMD User's Guide (i.e., the help documentation) was designed to help users to use the tools, navigate the information, and interpret the results provided by HOMD. The User's Guide is accessible through the top navigation menu on all pages and is dynamically linked to the relevant guide for each different tool. For example, when users are viewing the Taxon Table page, the “How to Use This Page” menu item shown in the top navigation menu will lead directly to the page that explains the use of the Taxon Table. Alternatively users can also browse the entire user documentation by clicking the “Table of content” tab shown on top of each documentation page as well as the “User's Guide” links on top menu and side menu of home page. Every document of HOMD can be searched either through the search box located at the bottom of the table of contents of the documentation page or through the Meta-Database Search box located at the top-right part of the home page.

The design of the online interfaces of HOMD has been driven by suggestions from HOMD users. HOMD is open to suggestions and feedback from the research community to further improve its interface and content. Currently, HOMD provides several different ways to communicate with the



HOMD at The Forsyth Institute
Human Oral Microbiome Database

Home Taxon Description Identify 16S rRNA Sequence Genomes Tools & Download HOMD Information How to Use This Page Page: HP1

Taxon Description

- Taxon Table
- Taxonomic Hierarchy
- Taxonomic Level

Identify 16S rRNA Sequence

- BLAST Sequence(s)
- 16S rDNA RefSeq Download

Genomes

- Taxa with Annotated Genomes
- Taxa with Genomes In Progress
- View All Genomes

Tools & Download

- HOMD Genome Viewer
- View Dynamic Annotation
- BLAST Against Genome
- Download HOMD Data

HOMD Information

- How to cite HOMD
- Project Description
- Strains and DNA Availability
- Team
- Contact Us
- Mailing List
- User's Guide
- Hands-on Guide to HOMD
- Usage Statistics
- Site Map

Welcome to HOMD

The goal of creating the Human Oral Microbiome Database (HOMD) is to provide the scientific community with comprehensive information on the approximately 600 prokaryote species that are present in the human oral cavity. The majority of these species are uncultivated and unnamed, recognized primarily by their 16S rRNA sequences. The HOMD presents a provisional naming scheme for the currently unnamed species so that strain, clone, and probe data from any laboratory can be directly linked to a stably named reference entity. The HOMD links sequence data with phenotypic, phylogenetic, clinical, and bibliographic information. Full and partial oral bacterial genome sequences determined as part of this project and the Human Microbiome Project, are being added to the HOMD as they become available. HOMD offers easy to used tools for viewing all publically available oral bacterial genomes. Welcome!

Investigators contributing to this project: Tsute Chen, Floyd E. Dewhirst, Derrick E. Fouts, Jacques Izard, Bruce J. Paster, Anne Tanner, and William G. Wade

Contributors: Oxana Baranova, Jessica Blanton, Wen-Han Yu, Abby Lakshmanan, and Akila Ganesan

This project is supported by: contract U01 DE016937 "A Foundation for the Oral Microbiome and Metagenome" from **The National Institute of Dental and Craniofacial Research** ;

Meta-Database Search

Advance Search

Announcement

2013-03-25 13:57:50
HOMD 16S rRNA Reference Sequence
Version 13.0 has been released [more]

2012-02-16 16:05:21
HOMD 16S rRNA Reference Sequence
Version 11.0 has been released [more]

2010-08-02 10:14:44
HOMD publication available for
download. [more]

2010-06-10 09:45:25

Database Update

2013-05-14 12:05
Genome Annotation Update - Actinomyces
sp. oral taxon 175 F0394 [more]

2013-05-08 17:05
Genome Annotation Update - Actinomyces
sp. oral taxon 171 F0337 [more]

2013-05-02 15:05
Genome Annotation Update - Actinomyces
sp. oral taxon 170 F0396 [more]

2013-05-01 16:05
Genome Annotation Update - Actinomyces
sp. oral taxon 848 F0332 [more]

Copyright 2007-2013 The Forsyth Institute
This page was rendered by server 154.

Human Oral Microbiome Database (HOMD), Fig. 1 Screenshot of the HOMD home page

research team and research community. The contact information provides e-mail addresses for direct communication with the HOMD research team. There is also a mailing list for important updates and announcement. Users can use their own e-mail address to subscribe to the **HOMD Mailing List** (<https://groups.google.com/forum/#!forum/homd-mail>) by sending an empty e-mail to the e-mail address: **homd-mail+subscribe@googlegroups.com**. An automatic e-mail will be sent to the subscriber for confirmation. HOMD also provides a discussion platform for the research community (<https://groups.google.com/forum/#!forum/homd-forum>). Note that these web links may change over time. In any case, current or updated web links provided here will be available on the HOMD website.

The HOMD Database Schema

The information and data provided by HOMD are stored in several databases. The Oral Taxon IDs and the genome IDs serve as the keys to cross-link these databases. The database table structures and the contents can be downloaded from the HOMD FTP (file transfer protocol) site at <ftp://ftp.homd.org> to allow users to reconstruct the databases and perform advance queries on their own computers.

Download Data from HOMD

Most of the data recorded in HOMD, including taxonomy, genomics, and 16S rRNA reference sequences, can be downloaded from the HOMD FTP site (<ftp://ftp.homd.org>). The FTP site provides both current and archived versions of the

data for comparison. The FTP site can be accessed directly in the web browser. Each folder comes with a “readme” text file explaining the data, data format, and potential usage. Selected data such as the aligned reference sequence dataset, aligned 16S rRNA datasets for each taxon, and an HOMD taxonomy database in Excel format can be downloaded from the links provided in the HOMD web pages.

Taxonomy

Compilation of the HOMD Taxa

The HOMD describes information linked to oral microbe species. For bacteria, or archaea, that have not been validly named, there is no definition of “species.” Molecular methods to identify novel species generally have used 16S rRNA sequencing of isolates or 16S rRNA-based analysis of clone libraries. These strains or clones can then be clustered into phylotypes or taxa based on their 16S rRNA sequences. Phylotype can be defined for any similarity cutoff. In HOMD, a cutoff of 98.5 % 16S rRNA sequence similarity was used to cluster the 16S rRNA sequences at the species level to define novel oral bacterial phylotypes. Each validly named species and novel phylotype cluster was given a unique integer number called Human Oral Taxon (HOT) ID.

The original collection of oral microbial taxonomy information came from a combination of literature, primarily reports from Forsyth Institute investigators (Dzink et al. 1985, 1988; Socransky and Haffajee 1994; Tanner et al. 1979, 1998) and from Lillian Holderman Moore and Ed Moore (Moore et al. 1982, 1983; Moore and Moore 1994) formerly at the Anaerobe Laboratory at the Virginia Polytechnic Institute. 16S rRNA sequences for these named species came either from sequences obtained in our laboratory or from GenBank. Over the past 20 years, our laboratory constructed and sequenced over 600 16S RNA gene libraries and obtained over 35,000 clone sequences. The cloning, sequencing, aligning, treeing, and clustering methods used to create HOMD are described elsewhere (Dewhirst et al. 2010). In brief,

sequences were manually aligned in a secondary structure-based database using the program RNA (Paster and Dewhirst 1988). Distance matrices and neighbor-joining trees were generated to determine the clustering of sequences. Sequences with similarity equal to or greater than 98.5 % were grouped together into a single taxon. Sequences were extensively checked for chimeras and several sequences and some provisional taxa were removed. As a result, several hundred apparently novel full 16S rRNA sequences were identified this way.

To share the information of both the named and novel human oral microbial taxa with the research community, we decided to build a database and designed web query interfaces and tools. When the HOMD was publicly launched in 2010, there were a total of 619 Human Oral Taxa in the initial release of the HOMD database. The 753 reference 16S rRNA gene sequences upon which this analysis was done have been released publicly for download on the HOMD website as version 10. At the time of writing this chapter, the total number of taxa described in the HOMD taxonomy database has grown to 688, represented by a total of 833 reference 16S rRNA sequences (HOMD RefSeq Version 13.1).

Navigating the HOMD Taxa

The HOMD taxonomy information can be viewed and retrieved in several different ways. The information can be viewed online directly in a web browser or downloaded as text files. For the online web browser viewing, the taxonomy pages can be searched with keywords or by visual navigation with the Taxon Table (Fig. 2) and the Taxonomic Hierarchy (Fig. 3). The Taxon Table can also be downloaded in Excel and tab-delimited plain text file from the Tools & Download page or through the HOMD FTP site. The keyword search can be done through the Meta-Database Search box on the home page or on the Taxon Table page. Both search boxes look for input keyword(s) in all text fields of the HOMD taxonomy database table.

On the Taxon Table page, all the human oral microbial taxa are listed in a table ordered

All Human Oral Microbial Taxa [Previous page](#)

Navigate Taxa by Alphabet: [A](#)|[B](#)|[C](#)|[D](#)|[E](#)|[F](#)|[G](#)|[H](#)|[I](#)|[J](#)|[K](#)|[L](#)|[M](#)|[N](#)|[O](#)|[P](#)|[Q](#)|[R](#)|[S](#)|[T](#)|[U](#)|[V](#)|[W](#)|[X](#)|[Y](#)|[Z](#)|[All Entries](#)

Search: All fields Matching ANY words Partial field

Total: **688 taxa** Table display: 20 50 100 All items per page

Items 1 - 688 Named Species Unnamed cultivated Uncultured phylotypes All

Oral Taxon ID (HOT)*	Genus *	Species *	Status	Flag	Taxon Link	Genome Link
842	<i>"Bacteroides"</i>	<i>ureolyticus</i>	Named	1	Taxon Description	View Genome
389	<i>Ablotrophia</i>	<i>defectiva</i>	Named		Taxon Description	View Genome
343	<i>Achromobacter</i>	<i>xylosoxidans</i>	Named		Taxon Description	View Genome
554	<i>Acinetobacter</i>	<i>baumannii</i>	Named		Taxon Description	View Genome
408	<i>Acinetobacter</i>	<i>sp. oral taxon 408</i>	Phylotype		Taxon Description	View Genome
183	<i>Actinobaculum</i>	<i>sp. oral taxon 183</i>	Unnamed		Taxon Description	View Genome
850	<i>Actinomyces</i>	<i>cardiffensis</i>	Named		Taxon Description	View Genome
888	<i>Actinomyces</i>	<i>dentalis</i>	Named		Taxon Description	View Genome
617	<i>Actinomyces</i>	<i>georgiae</i>	Named		Taxon Description	View Genome
618	<i>Actinomyces</i>	<i>gerencseriae</i>	Named		Taxon Description	View Meta Info
866	<i>Actinomyces</i>	<i>graevenitzii</i>	Named		Taxon Description	View Genome
645	<i>Actinomyces</i>	<i>israelii</i>	Named		Taxon Description	View Meta Info
849	<i>Actinomyces</i>	<i>johnsonii</i>	Named		Taxon Description	View Genome
852	<i>Actinomyces</i>	<i>massiliensis</i>	Named		Taxon Description	View Genome
671	<i>Actinomyces</i>	<i>meyeri</i>	Named		Taxon Description	View Genome
176	<i>Actinomyces</i>	<i>naeslundii</i>	Named		Taxon Description	View Genome
688	<i>Actinomyces</i>	<i>naeslundii II</i>	Named		Taxon Description	View Genome
701	<i>Actinomyces</i>	<i>odontolyticus</i>	Named		Taxon Description	View Genome
708	<i>Actinomyces</i>	<i>oricola</i>	Named		Taxon Description	View Genome

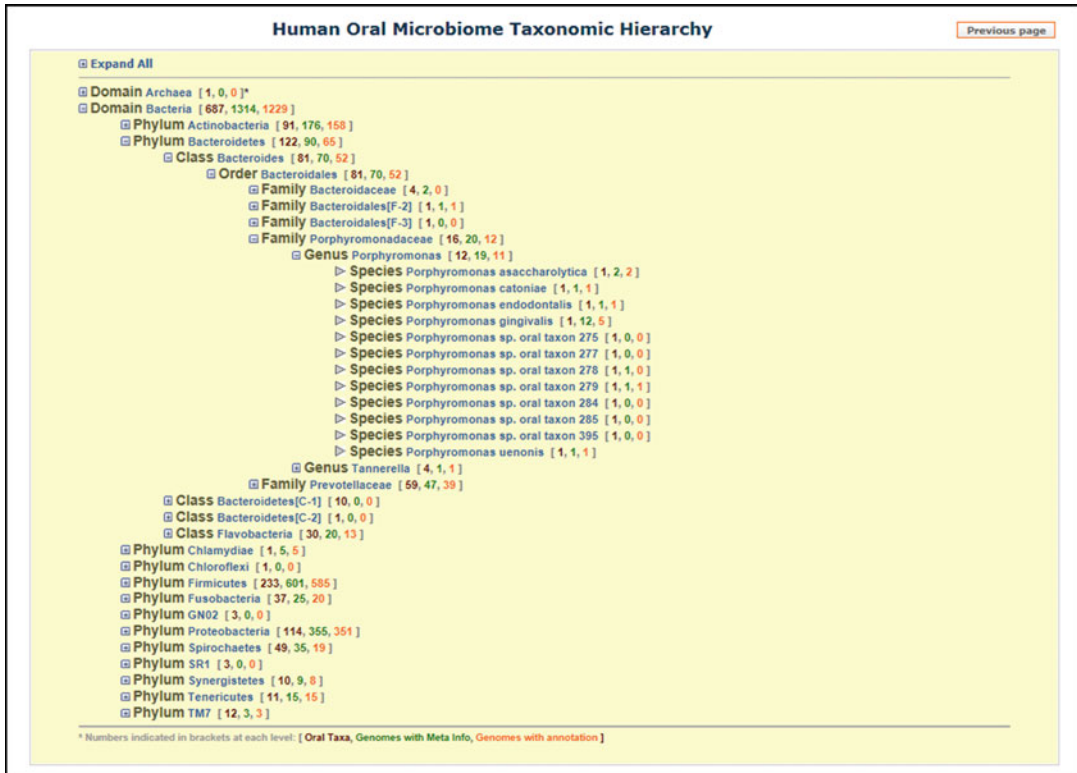
Human Oral Microbiome Database (HOMD), Fig. 2 Screenshot of the Taxon Table

alphabetically by organism names. The order can be changed by clicking the column name HOT IDs, Genus, or Species names, to toggle the display sort order. Three commonly used filters are also provided to show only those taxa with “named species,” “unnamed cultivated species,” or “uncultured phylotypes.” Each taxon listed in the table contains links to the individual Taxon Description page (described later) and to the genomic information, if available.

The taxa can also be viewed in the taxonomic hierarchical order, i.e., from domain, phylum, class, order, family, genus, to species levels, on the Taxonomic Hierarchy page (Fig. 3). The hierarchical tree is fully collapsed by default and can be dynamically expanded at any given level (or all levels). The link, at the species level, brings users to the detailed Taxon Description page. The designation of each level is followed

by two numbers enclosed in the square brackets indicating the number of taxa and genome sequences. For example, “Phylum Proteobacteria [107, 144]” indicates that in the phylum Proteobacteria, 107 taxa were identified in the oral cavity and 144 strains have genomic sequences available at HOMD. If a strain has been sequenced by multiple groups, or multiple strains sequenced for a species, we provide each sequence when available.

Another way to check the summary of the HOMD taxa is to view the number of taxa at various taxonomy levels. The **Taxonomic Level** page provides a list of taxa and the number of taxa at the next lower level for each of the 7 taxonomic levels: Currently, the numbers are Domain (2), Phylum (14), Class (24), Order (40), Family (83), Genus (183), and Species (688).



Human Oral Microbiome Database (HOMD), Fig. 3 Screenshot of the Taxonomic Hierarchy expanded at the order level *Bacteroidales*

Taxon Description

The HOMD Taxon Description page (Fig. 4) provides comprehensive information for a specific human oral microbial taxon. Information provided can be summarized in four categories: Taxonomic Hierarchy, biological characteristics, references, and community comments. Throughout the page, clickable dynamic cross-links are provided for additional information. The taxon page can be edited and curated by designated curators upon their logging-in. The page also allows input and comments provided by the users in the research community. Information described on this page are the following:

Human Oral Taxon (HOT) ID – The Human Oral Taxon ID is a unique numeric ID representing a particular taxon. The taxon can be unambiguously referred to from other sources of scientific literature. The taxon can be accessed on the web with an easy universal

resource locator (URL) format, <http://www.homd.org/taxon=NNN>, where NNN is the HOT ID. The Human Microbiome Project Data Analysis and Coordination Center (DACC; accessible at <http://www.hmpdacc.org>) is using HOT IDs to designate taxonomic identity isolates of the oral cavity with URLs cross-referenced to HOMD. These URLs are embedded in the data provided by DACC so that user can track down to the more comprehensive information for individual genome. The HOT IDs were also embedded in the GenBank sequence records for the 35,000 clone sequences that were used to build the initial collections of the HOMD taxa. The text embedded in the GenBank records has the syntax `/db_xref="HOMD:tax_NNN,"` in which NNN is the numeric HOT ID. If the GenBank sequence is viewed in the web browser through the NCBI website, the

Human Oral Microbiome Taxon Description

[Previous page](#)
[Link to this page](#)

Streptococcus mutans

Human Oral Taxon ID (HOT):	686		
Status:	Named species	Synonym:	
Type Strain:	ATCC 25175, NCTC10449		
	More info at StrainInfo		
Classification:	Domain: Bacteria Phylum: Firmicutes Class: Bacilli Order: Lactobacillales Family: Streptococcaceae Genus: Streptococcus Species: mutans	NCBI Taxonomy ID:	1309 [NCBI Taxonomy Link]
16S rRNA Sequence:	AJ243965 [Entrez Link]	PubMed Search:	8155 [PubMed Link]
16S rRNA Alignment:	View Alignment Download Alignment Note	Nucleotide Search:	37007 [Entrez Nucleotide Link]
Phylogeny:	View 16S rRNA tree View all Tree files	Protein Search:	178151 [Entrez Protein Link]
Prevalence by Molecular Cloning:	Clones seen = 1500 / 34879 = 4.30% Rank Abundance = 3	Genome Sequence	19 View Genomes

Hierarchy Structure: [Show](#)

General Information:
 Belongs to a phenotypic group called the mutans streptococci which include *S. sobrinus*, *S. ferus*, *S. cricetus*, *S. rattus* and serotype Th⁺ [4]. Phylogenetically distinct from other species of *Streptococcus*.

Cultivability:
 Colonies are whitish about 0.5 to 1 mm that stick to the agar. When media is supplemented with sucrose, puddles of liquid (e.g., soluble extracellular polysaccharide) surround the colonies.

Phenotypic Characteristics:
 Facultatively anaerobic, Gram positive cocci (0.5 to 0.75 µm in diameter), which occurs in pairs, or short chains [4]
 Produces extracellular polysaccharides from sucrose by glucosyltransferase (e.g., glucans) and fructosyltransferases (e.g., fructans). These polysaccharides promotes binding to cell surfaces.
 Glucose is fermented to L-lactic acid with no gas. Final pH in glucose broth cultures is 4.0 to 4.3. Growth is not inhibited by lower pH.
 Strains of *S. mutans* can be distinguished serologically. Peptidoglycan contains glutamic acid, alanine, lysine, glucosamine and muramic acid [4]

Prevalence and Source:
 Commonly detected on human teeth in supragingival plaque, and usually associated with caries. Has been isolated from human feces.

Disease Associations:
 Strong association with human dental caries, considered the primary cause of caries, although caries can occur in the absence of *S. mutans*. Also associated with endodontic lesions, odontogenic infections, infectious endocarditis and cardiovascular disease [2]
 Is cariogenic in experimental animals (rats, hamsters, gerbils, mice and monkey).
 Vaccines which target *S. mutans* are being developed to prevent caries formation [1].

References:
PubMed database:
 [1] Taubman MA, Nash DA. The scientific and public-health imperative for a vaccine against dental caries. *Nat Rev Immunol* 2006 Jul;6(7):555-63 [\[PubMed\]](#)
 [2] Nakano K, Inaba H, Nomura R, Nemoto H, Takeeda M, Yoshioka H, Matsue H, Takahashi T, Taniguchi K, Amano A, Ooshima T. Detection of cariogenic *Streptococcus mutans* in extirpated heart valve and atheromatous plaque specimens. *J Clin Microbiol* 2008 Sep;44(9):3313-7 [\[PubMed\]](#)
 [3] van Ruyven FO, Lingström P, van Houte J, Kent R. Relationship among mutans streptococci, "low-pH" bacteria, and iodophilic polysaccharide-producing bacteria in dental plaque and early enamel caries in humans. *J Dent Res* 2000 Feb;79(2):775-84 [\[PubMed\]](#)
Non-PubMed database:
 [4] Hardie JM. Genus *Streptococcus*. *Bergey's Manual of Systematic Bacteriology* 1986; Vol. 2, pp. 1043-1063.

Curator:	Creation Info:	Latest Modification: wenhan, 2008-01-17 12:42:02	
----------	----------------	--	--

Human Oral Microbiome Database (HOMD), Fig. 4 Screenshot of the Taxon Description page

portion of the text “tax_NNN” is also clickable and links to the corresponding taxon page on the HOMD website. For example, the GenBank record for the partial 16S rRNA sequence of the *Alloprevotella rava* clone GB024 (Accession No. GU409552, <http://www.ncbi.nlm.nih.gov/nuccore/GU409552>) contains the text `/db_xref=“HOMD:tax_302,”` because the HOT ID for *A. rava* is 302.

Clicking “tax_302” in this GenBank record in the web browser will bring the user to the corresponding taxon page on HOMD (<http://www.homd.org/taxon=302>). NCBI embeds external database reference IDs in the GenBank records for cross-database referencing. More information can be found at this link http://www.ncbi.nlm.nih.gov/genbank/collab/db_xref.

Status – This field displays the culturing status for the taxon. A taxon can be either a validly named cultivated species, an unnamed cultivated species, or an unnamed uncultured phylotype. This status is shown in this field and will be updated upon the change of actual status of the taxon.

Type strain/reference strain – If the taxon’s status is validly named cultivated species, the Type Strain is listed here; if the taxon is an unnamed isolate, the strain information will be listed as Reference Strain. If no cultivated strain is available yet, the Reference Strain field will be listed as “None, not yet cultivated.”

Classification – The Taxon Description page lists the nomenclatures of each taxonomic level from Domain to Species. This classification is defined by HOMD and may be different from the NCBI Taxonomy. The NCBI Taxonomy can be accessed using a dynamic link. The HOMD taxonomy is based on analysis of where each taxon falls in phylogenetic trees generated using several treeing methods and including over 100 non-oral reference taxa identified by searching the “greengenes” 16S rRNA gene database (<http://greengenes.lbl.gov>). For example, in HOMD, an organism such as *Eubacterium saburreum* is placed in the family *Lachnospiraceae* (because that is where it falls phylogenetically), rather than in the family *Eubacteriaceae* (because its incorrect genus name “Eubacterium” has not yet been revised). Synonyms of the taxon that are currently in use or were used before in the literature or publications are also provided.

16S rRNA gene sequence – GenBank accession number and link to NCBI corresponding Entrez record to one or more 16S rRNA gene sequences associated with the taxon.

16S rRNA gene sequence alignment – This field provides the link to the downloadable clone sequences preliminarily aligned to the reference sequence to which the clones belong. The current set contains the approximately 35,000 clone sequences (Dewhirst et al. 2010) aligned for each taxon. The clone alignments are provided concatenated FASTA format with the

reference sequence(s) on top which were used as the template for alignment. To view the alignment in color format and for further adjustment, third-party alignment viewing software may be used, such as SeqView (<http://pbil.univ-lyon1.fr/software/seaview.html>) and BioEdit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). Because some pairs of clone sequences may be nonoverlapping (i.e., 500-base sequences at opposite end of the molecule), this file must be used with caution for tree construction.

Phylogeny – A phylogenetic tree showing the position of this taxon among related HOMD taxa is provided here. The tree images are in PDF format and can be viewed or downloaded with the link provided in this field. A link to a list of all the downloadable phylogenetic tree images encompassing all the HOMD taxa is also provided.

Prevalence by molecular cloning – The number of clones found for this taxon in an analysis of approximately 35,000 clones (Dewhirst et al. 2010). Based on the number of clones found, the rank abundance of the taxon (out of 619) is given.

Synonyms – Lists previous names for the organism if validly named. Isolate or clone designations are given as synonyms when they have appeared in the literature as “names” for the taxon, such as “BU063.” (Zuger et al. 2007).

NCBI taxonomy – For validly named species, there is a link to the NCBI Taxonomy. NCBI has no taxonomy for unnamed taxa; hence, the reason HOMD was created.

PubMed search – The number of hits when the name (genus plus species) of this taxon is used in the PubMed search. HOMD automatically and periodically updates this hit number every 2 weeks. To get a most up-to-date search, simply click the “PubMed Link” to pull up the search result live from NCBI PubMed site. In general, there are no results for unnamed taxa, hence the need for HOMD. When articles referencing these taxa (often through clone numbers) are found by HOMD curators or community members, they are manually added to the Taxon Description.

Nucleotide search – Similar search as above using NCBI Entrez “nucleotide” as reference database. The latest result (hit count) is displayed with link to NCBI for most updated search.

Protein search – Similar search as above using NCBI Entrez “protein” as reference database. The latest result (hit count) is displayed with link to NCBI for most updated search.

Genomic sequence – Number of genomes that have been sequenced is indicated here with a link to a detailed list of these genomes.

Hierarchy structure – An expandable/collapsible view of a dynamically displayed taxonomy tree indicating the position of the taxon on the page.

Cultivability – Conditions and media for growing strains of this taxon, if available.

Phenotypic characteristics – Generic phenotypic description of the taxon if the taxon has cultivated member(s).

Prevalence and source – Describes the frequency and source of clones and isolates from different oral sites and states of health or disease when known.

References – Literature and publications referencing this taxon. These references are manually curated with up to ten key references which may also include older references not indexed in PubMed.

Community comments – Registered and logged-in users can provide their feedbacks related to this taxon. The comment requires the approval of the HOMD curators before it is shown to the public.

Identification of 16S rRNA Gene Sequence by BLAST Search

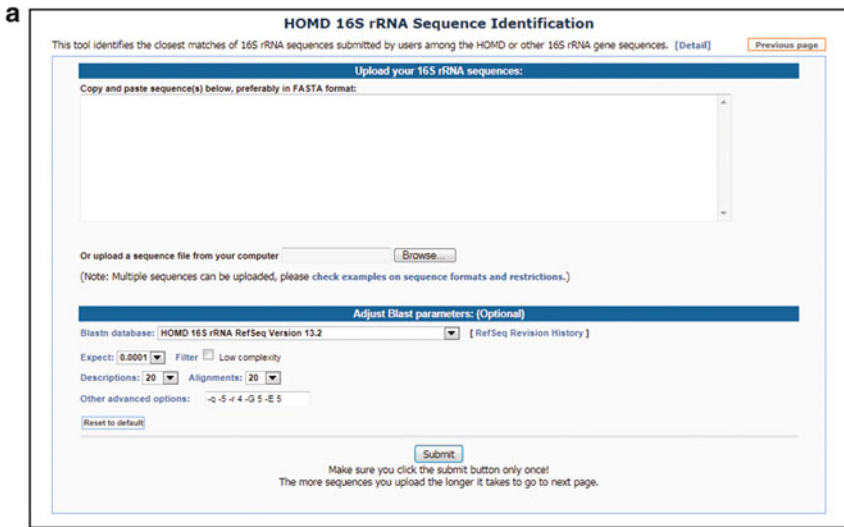
One of the most used HOMD software tools is the customized BLAST search specifically designed to identify user-provided 16S rRNA sequences against the comprehensive collection of the 16S rRNA reference gene sequences. Currently there are a total of 688 taxa defined based on version 13.1 of the 16S rRNA reference sequences. Since a phylotype can include members with up to 1.5 % sequence divergence (23 bases for a full 1,500-base sequence), multiple reference sequences have been selected where we have

sequences diverging by more than 10 bases within a taxon.

HOMD provides two primary sets of 16S rRNA gene reference sequence (RefSeq) for download and for BLAST search. The first set is the HOMD 16S rRNA RefSeq. This set contains sequences representing all currently named and unnamed oral taxa. In the latest reference sequence set (version 13.1 at the time of writing), there are 834 reference sequences representing the 688 taxa. The second is the HOMD 16S rRNA Extended RefSeq. This set contains additional 16S rRNA reference gene sequences that are distinctively different from existing taxa but have not yet been assigned with a taxon ID.

The HOMD reference sequences are corrected consensus sequences. Many have been corrected and extended based on alignment with other sequences for that taxon and Ns and indels removed. Therefore, for many sequences, there will be differences between the reference sequence and the GenBank sequence listed in the header information. We have not yet updated our own GenBank sequences and cannot update those from other depositors. We believe these are currently the best reference sequences available and, for the purposes of BLAST analysis, have the advantage of being of a uniform length.

On the HOMD 16S rRNA Sequence Identification page (Fig. 5), users can copy and paste the query sequences in the text field or upload from user’s computer. The query sequences should be in the concatenated FASTA format. The maximal number of query sequences allowed to upload in a single search is 5,000. Since viewing of the BLAST results in the web browser for over 5,000 query sequences becomes very slow, for search over 5,000 sequences, please contact the HOMD team. The HOMD 16S rRNA BLAST online tool was only designed for a modest number of sequences, up to a couple of thousands, which can be submitted in several batches. It is not capable of handling larger numbers of sequence reads, such as hundreds of thousands of reads from the next-generation sequencing pipeline. For larger numbers of sequences, the search can be done on a collaboration basis. HOMD provides secure FTP (sFTP) upload for large batches of user sequences,



b

HOMD 16S rRNA Sequence Identification

Search results: 3 query sequence(s) searched against HOMD 16S rRNA RefSeq Version 13.2

Query ¹	Query length(nt)	HIT HOMD files ²	HOMD clone name	Identities (%) ³	Mismatch ⁴	Identities (%) ⁵	Mismatch ⁶	Score (bits)	Query start ⁷	Subj start ⁷	Query End ⁷	Subj End ⁷
gi_5578899_emb_AJ243965_1_...	1512	688_3965	<i>Streptococcus mutans</i> HOT_688 Strain_NCTC 10449 AJ243965 Named	100	0/1512	100	0/1512	2358	1	1	1512	1512
		622_3931	<i>Streptococcus gordonii</i> HOT_622 Strain_ATCC 10558 AF003931 Named	94.7	88/1516	95.2	72/1508	2068	1	1	1512	1512
		758_3928	<i>Streptococcus sanguinis</i> HOT_758 Strain_ATCC 10556 AF003928 Named	93.6	94/1513	94.5	83/1502	2018	1	1	1512	1503
gi_5139007_obj_AB029087_1_...	1557	768_3966	<i>Streptococcus sobrinus</i> HOT_768 Strain_NCTC 12279 AJ243966 Named	93.7	95/1515	94.1	88/1509	2018	1	1	1512	1512
		642m3087	<i>Scardovia inopinata</i> HOT_642 Strain_DSM 10107 AB029087 Named	98.4	25/1521	99.9	1/1497	2241	21	1	1540	1498
		195_8626	<i>Scardovia wiggiae</i> HOT_195 Strain_C1A_55 AY278626 Named	93.1	105/1526	95.5	67/1488	1947	21	1	1540	1495
gi_9988917_gb_AY095053_1_...	1434	195CX010	<i>Scardovia wiggiae</i> HOT_195 Clone_CX010 AF287758 Named	93	107/1524	95.2	71/1488	1940	21	1	1540	1493
		568_9331	<i>Parascardovia denticolens</i> HOT_568 Strain_DSM 10105 D89331 Named	92.8	110/1523	95.5	66/1479	1934	21	1	1540	1483
		291AH005	<i>Prevotella denticola</i> HOT_291 Clone_AH005 AY095053 Named	99.7	4/1427	99.9	1/1424	2210	9	1	1434	1426
gi_9988917_gb_AY095053_1_...	1434	291A005	<i>Prevotella denticola</i> HOT_291 Strain_ATCC 35308 AY223524 Named	99.2	12/1427	99.4	8/1423	2183	9	1	1434	1426
		291A0036	<i>Prevotella denticola</i> HOT_291 Clone_A0036 AY095054 Named	98.2	25/1427	98.5	21/1423	2138	9	1	1434	1426
685_2483			<i>Prevotella multiformis</i> HOT_685 Strain_PPAP21 AB182483 Named	96.3	53/1429	95.6	48/1424	2039	9	1	1434	1428

Search ID: 86m716b9e32cca86avqrvpm3 Run Time: 6 seconds Date/Time: May 15, 2013, 2:38 pm

1 Check individual or all results to be downloaded. 2 Total mis-match nt / total match nt.
 2 Click to toggle the sort order of the results by query ID 3 Total mis-match nt / total match nt (excluding gaps and non-AGCTU).
 3 Access individual query sequence or original blast result. 4 Start or end positions of query or subject in the alignment.
 4 The table will show the top four hit results for each query sequence. * Percent identities >= 98.5% are highlighted in red

Human Oral Microbiome Database (HOMD), Fig. 5 HOMD 16S rRNA Sequence Identification. (a) Query sequence input interface; (b) Result page

and the search will be sent manually to the HOMD BLAST server cluster on user’s behalf and results made downloadable through the sFTP site. The upload page also provides options for adjusting the BLAST search parameters although the default setting should be sensitive enough to pick up matches with even short oligo sequences.

Once the query sequences are submitted, the sequences are uploaded to the HOMD computer servers and queued for the BLAST search. Once all the searches are done, the results are presented back to submitter in a tabularized format. Results containing up to 20 top matches for each query sequence can be downloaded in text or Excel file formats. Original full BLAST results including the alignments can also be accessed from the

result page. The match identity is presented as straight BLAST results and as an adjusted percent identity (API) calculated as

$$API = 100 \times M / (M + MM)$$

where M is the matched (identical) and MM the mismatch sequence length between the query and the reference sequence, respectively. This calculation excludes any gaps introduced during the alignment process of the BLAST search. We have found that this correction gives much better values for single primer sequence reads where the sequence adjacent the primer often includes indels. The top hits are ordered by their API rank, and sequences with alignment shorter than

95 % of query sequence are excluded from ranking. The top four matched reference sequences are listed by this method, and the table shown on the web page contains links to the original BLAST results as well as to the Taxon Description pages for reference sequences. The results for the 20 top matches can be downloaded as plain text or in Microsoft Excel format.

Genomics

Genomics Tools Overview

Complimentary to the taxonomy information, the HOMD also provides comprehensive information and tools for studying genomes of the human oral microbes. HOMD genomics database serves as the curated repository for the molecular sequences of human oral microbiome, including complete and partial genomics sequences, as well as 16S rRNA mentioned in the previous section. Genomic sequences available at HOMD can be fully assembled genomes, high-coverage genomes, or genome surveys. HOMD also keeps tracks of the status of ongoing genome sequencing projects for human oral microorganisms. A Sequence Meta Information page is created to hold relevant genomics and sequence meta information if a sequencing project for a human oral microbe is announced and available in the NCBI Genome Project Database. The genome project status is updated biweekly based on information collected from the NCBI Genome Project Database with an automatic query script. Once genomic sequences are publicly released, they are dynamically annotated by HOMD (Dynamic Annotation). Annotation done by other data centers, if available, is termed “static annotation” and is viewable in a separate panel in the Genome Viewer (described below). Relevant tools are provided for viewing and searching the annotation. These tools were first developed as part of the Bioinformatics Resource for Oral Pathogens (BROP: <http://www.brop.org>; Chen et al. 2005). The programs and the data-mining schemes used in HOMD are designed for both finished and unfinished (collections of multiple contigs) genome sequences. The tools are integrated with the HOMD website and are

conveniently accessible by users. Icons or links to available tools pertaining to a specific genome are automatically presented on relevant page to users. Important genomic data and bioinformatics tools provided by HOMD are described below. Additional information on tools is also available in the previous publication (Chen et al. 2005).

Genome Table

HOMD organizes genomes in three viewing options: Taxa with Annotated Genomes, Taxa with Genomes in Progress, and View All Genomes. The first option lists the oral taxa with annotated (static or dynamic) genomic information and provides links to all the genomes available for each taxon. The View Genome button links to the Genome Table showing all the available genomes of a specific taxon. The Genome Table shows the Oral Taxon ID (HOT), the Genus and Species names, Strain Culture Collection, HOMD Sequence ID (SEQ ID), number of contigs and singlets, combined sequence length, and links to available tools and information. The second option (Taxa with Genomes in Progress) lists those oral taxa with genomic sequencing project still in progress but no sequence is yet available. The third option shows all the genomes in the alphabetical order and provides searching and sorting function for easy navigation. Each genome listed has a link to the Sequence Meta Information page described next.

Sequence Meta Information

The Sequence Meta Information page provides detailed biological, molecular biological, genetic, genomic, and taxonomic as well as annotation information for a particular strain that has been, is being, or will be sequenced (Fig. 6). Information on these pages is semiautomatically updated. Updated information from both Genomes OnLine and NCBI Genome Project Database is retrieved biweekly and compared with the existing database automatically. New or modified Genomic Project information are then added to the Sequence Meta Information pages with confirmation by curators. The Sequence Meta Information page contains the following human-curated information related to

Sequence Meta Information: *Abiotrophia defectiva* ATCC 49176 [NCBI] [Previous page](#) [Link to this page](#) [Taxon page](#) [Annotation](#)

Info available in HOMD database: ⓘ

1 Oral Taxon ID	389
2 HOMD Sequence ID	SEQF1595
3 HOMD Name (Genus, Species)	<i>Abiotrophia defectiva</i>
4 Genome Sequence Name (Name associated with genomic sequence)	<i>Abiotrophia defectiva</i>
5 Comments on Name	NA
6 Culture Collection Entry Number	ATCC 49176
7 Isolate Origin	NA
8 Sequencing Status	High Coverage
9 NCBI Genome Project ID	33011
10 NCBI Taxonomy ID	592010
11 Genomes Online Goldstamp ID	G03551
12 NCBI Genome Survey Sequence Accession ID	ACIN00000000
13 JCVI (TIGR) CMR ID	NA
14 Sequencing Center	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine
15 Comments on GC Percentage	NA
16 ATCC Medium Number	NA
17 Non-ATCC Medium	NA

18 16S rRNA Gene Sequence

```
>SEQF1595 Abiotrophia defectiva strain ATCC 49176; oral taxon 389, 16S ribosomal RNA gene, partial sequence
GAGTTTAACTTCGGCTCAGGACACACTCGGCGCTGCTTAAACATGGAAGTCGAAAGCCGACACTAGTCTGTCACATTTGTCAGGTAGTTCGACAGCGTCACTAA
CACCTGGTAACTTACCTCATAATGCGGGATACACGTCGGAACGACTCTAAATACCCGATAGGACATGGGATCACATGATTTCTGGGAAATGTCGCGCAATCCCTAA
GAGATGGACCGCGTCAATTAGCTAGTGGTAAAGGCGCTACCAAGCGGATGATGACATGACCGACCTGAGAGGGTGTCCGCGCACATTTGGACTGAGACACCGCCAAA
CTCTTAGGGGAGGACGATAGGAAATCTTCCGCAATGACGCAAGTCTACGAGCAGACCGCGCTGATGAGAGAGTCTTCCGATGTTAAAGTCTCTTTTAGAGAGAA
CACCCATAGAGTAACTGCTATCCGCTGACGGTATCTAAACAGAAAGCCAGCGCTAACCTACGCTCCAGCAGCCCGGTTAAATACGATGGTGGCGGCTCTTCCGATGATTC
GGCTAAAGGGATGTAGCGGCTCTTTTAACTGATGTGTAATGAAAGCCAGCGCTCAACCTGGAGGGTCAATGGAACTGGAGCTTAACTGACAGAGAGAGAGCGAAATTC
ATTTTATGCGGTTAAAGCTGATATATGAGACACCGTGGGAAAGCGGCTCTTCTGCTTTAACTGACGCTGGGCTCGAAAGCTGGGAGCAAAAGATTAATTA
CCTTGTATGACCGGCTTAAAGATGATGCTGATGCTTGGAGCGGCTCCAGCTTCACTGCTCGATGCTTAAACGAAATAGGACATCCGCTGGGATAGCGCCGAAAGCTGA
AACTCAAAGGAAATTCAGCGGACCCGCAAGCGTGGACATGTTGTTAAATGGAAGCAACCGGAGACCTTACAGGCTCTTACATCCCGACGCGCTCTAGAGATAGA
GTTTTTCTGGAGCTGCTGACAGGTGGATGTTGCTGACGCTGPTCTGATGATTTGGTTAAATCCCGAAGCGGAAAGCTTAATATGATTTCCGAGATTT
GAGATGGGACTCTAATAGACTGGCGTGCAGAAACCGGAAAGTGGGATGACCTCAAATCATATGCCCTTATAGCTGGGCTTACACATGCTGATAGATGATGATGAC
ACGACGACGACCTCGAGGTTAGCGAATCTCTAAAGACCACTTCACTGCTGATTTGATGCTGCAACTGCACTACATGAGCGGAAATGCGTATGATCCGGATCAGCAC
CGCCCGGATATGCTCCGCGGCTTTTATACACACCGGCTTCAACACAGAGATTTTATACACCGGAGCGCGCTGCTTAACTTTTAAAGGAGGACCGCTCGAAGTGGAT
ACATGATGGGTTGAAGCTTAAAGATGACCGCTATCGGAGGTTCCG
```

19 Comments for 16s rRNA Gene Sequence

NA

Human Oral Microbiome Database (HOMD), Fig. 6 Screenshot of the Sequence Meta Information page

the target organism: Oral Taxon ID, HOMD Sequence ID (SEQ ID), Organism Name (genus, species), Culture Collection Entry Number, Isolate Origin, Sequencing Status, NCBI Genome Project ID, NCBI Taxonomy ID, Genomes Online Goldstamp ID, NCBI Genome Survey Sequence Accession ID, JCVI (previously TIGR) CMR ID, Sequencing Center, number of contigs and singlets, combined length (Kbp), GC percent, DNA molecular summary, ORF annotation summary, and 16S rRNA gene sequence. In addition, original external information such as NCBI Genome Project Database, NCBI Taxonomy Database, Genomes OnLine Database, and rRNA in NCBI Nucleotide Database, if available, is parsed into separate tables below the Sequence Meta Information for convenient referencing.

Full and High-Coverage Genomes

Full genomes are the oral microbial genomes that have been fully assembled, while the high-coverage genomes are not fully assembled but represent coverage of most of the genomes.

Both types of genomes are annotated and deposited in a public database such as GenBank. HOMD aims to provide frequently updated genomic annotation for oral bacterial genomes (see below). In addition, HOMD provides graphical genomic viewing for static annotations done by other public data centers such as NCBI or JCVI.

Genome Surveys

One of the original major goals of the NIH-funded project “A Foundation for the Oral Microbiome and Metagenome,” DE016937, was to partially sequence up to 100 representative human oral microbial species. A total of 12 low-coverage partial genomic sequences were sequenced and deposited in NCBI before this project fused with the Human Microbiome Project. The genome information for these 12 surveys is still maintained on HOMD even though they currently also have complete or high-coverage genomes (The Forsyth Metagenomic Support Consortium and IZARD 2010). Since the launch of the Human Microbiome Project, the HOMD team has been providing genomic DNA from

human oral microbes to the four HMP sequencing centers for high coverage rather than survey sequencing (The Forsyth Metagenomic Support Consortium and Izard 2010).

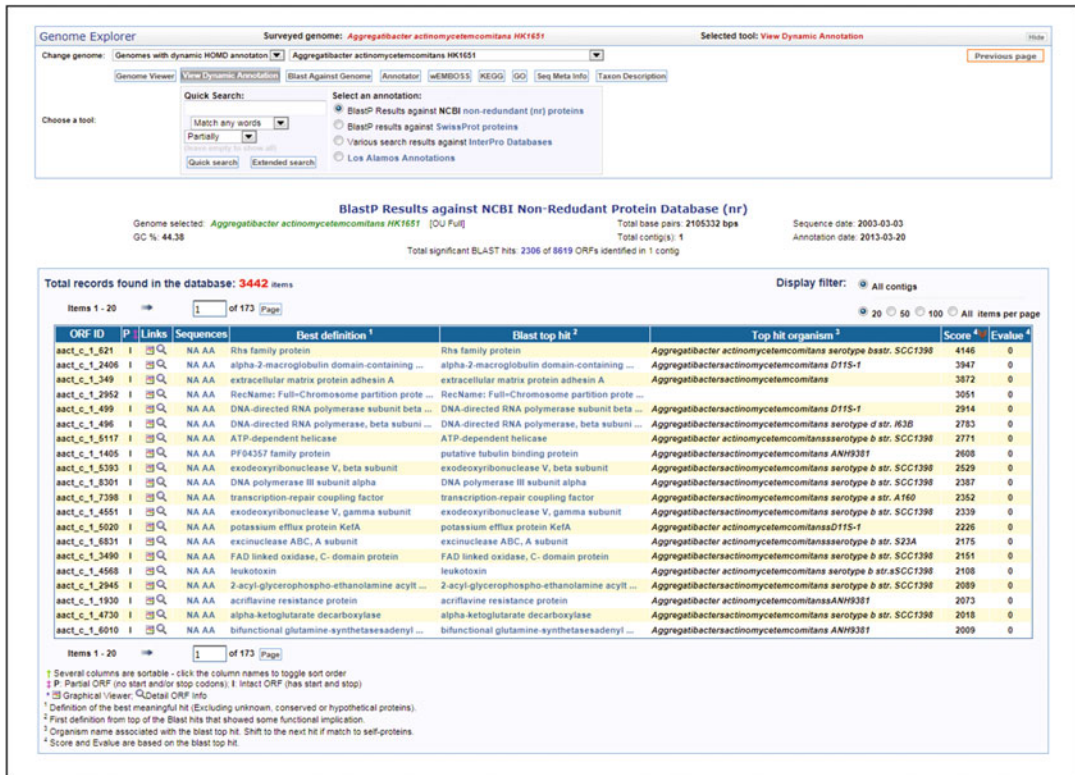
Dynamic Annotation of Genomic Sequences

One of the major features of the HOMD Genomic Database is the automatic and frequent updating of genomic annotation pipeline for genomes of oral isolates. Although the amount of sequence data is still growing rapidly, the computational power needed for bioinformatic analysis of this data is catching up and the cost and energy consumption per CPU decreasing due to the availability of multi-core CPU formats. The lower cost of computational power has made it feasible for us to set up a small computation farm dedicating to the annotation of human oral microbial genomes. HOMD recruited a cluster of multi-core multi-node computer servers to frequently update the annotation. Current HOMD genome annotation algorithms include (i) BLASTP (<http://www.ncbi.nih.gov/BLAST>; Altschul et al. 1997) search against weekly updated NCBI nonredundant protein data (<ftp://ftp.ncbi.nih.gov/blast/db/>), (ii) BLASTP search against Swiss-Prot protein data (<http://us.expasy.org/sprot/>; Boeckmann et al. 2003), and (iii) InterProScan search against various sequence databases (Zdobnov and Apweiler 2001; <http://www.ebi.ac.uk/interpro/>). To provide data on functional potential of genomes, BLASTP search results against Swiss-Prot are further processed for the construction of KEGG metabolic pathways and Gene Ontology Trees. We take advantage of the fact that the well-annotated Swiss-Prot protein sequence descriptions contain interlinks to the ENZYME (Bairoch 2000) and Gene Ontology (Camon et al. 2003). The dynamic genome annotation is running full time daily on the dedicated computer cluster except during the weekend, when the latest NCBI nonredundant protein database, Swiss-Prot, and InterPro databases are being downloaded to and updated on our server. Currently a total of 324 genomes representing 306 taxa are being repeatedly annotated by this pipeline. On average, each genome takes ~ 3 h to be annotated; thus, the current re-annotation

frequency is approximately a month for all the 300+ genomes. Additional genomes are being added to the annotation pipeline as more sequences are made available by other public sequencing projects such as the Human Microbiome Project (<http://www.hmpdacc.org>). A live update status of the genome annotation is provided on the HOMD home page indicating the latest genome annotated or updated. HOMD aims to maintain frequent and dynamic computer annotation for genomic sequence of at least one isolate from each oral taxon whenever sequences are made publicly available, as well as static annotation of all annotated releases.

Genome Explorer

Genome Explorer is the centralized web interface that interconnects all the genomics resources in HOMD (Fig. 7). The front end of Genome Explorer is a user-friendly interface that allows investigators to navigate among all the genomics information provided at HOMD. HOMD Genomics Tools can be accessed either by selecting the tool or the genome first. If the user chooses the desired tool first, the user is then directed to the Genome Explorer interface for selecting genomes. Once a target genome is chosen, the interface dynamically presents all the tools, including linked external databases, available for the selected genome. Currently available tools include Genome Viewer, Dynamic Annotation, BLAST, Annotator, EMBOSS, KEGG pathways (Kanehisa 2002), Gene Ontology Tree (Ashburner et al. 2000), Genomewide ORF Alignment, and Sequence Download. The back end of Genome Explorer is a searchable annotation database that integrates all the results generated from the data-mining pipeline described below. The search result is presented in a paginated and sortable table that also provides web links to (i) a summary page for individual ORF, (ii) Genome Viewer to show the exact location of the target ORF in the genome, and (iii) the original BLAST or InterProScan results. The summary page provides all the information and tools available for a specific ORF, including all the data-mining results mentioned above, as well as convenient links to other web tools for



Human Oral Microbiome Database (HOMD), Fig. 7 HOMD Genome Explorer displaying results of Dynamic Annotation for the genome *Aggregatibacter actinomycetemcomitans* HK1651

performing fresh search and analysis. In short, Genome Explorer is a one-stop interface for all the genomic information available for each target genome or gene.

Genome Viewer

Genome Viewer is a unique graphical genomic sequence viewer developed originally for the BROP project (Chen et al. 2005) (Fig. 8). The Genome Viewer was designed to alleviate the inconvenience encountered when comparing two different sets of annotations for the same genome. Genome Viewer provides a graphical, six-frame translational view of the same region of the genome with individual panels showing different sets of annotations. It has easy navigating features including zooming, centering, and searching by gene ID. For example, the genome *Porphyrromonas gingivalis* W83 has been annotated by JCVI (TIGR), Los Alamos National Laboratory, and NCBI separately. These

different annotations can be viewed and compared side by side in the Genome Viewer (<http://www.homd.org/index.php?name=GenomeExp&org=pgin&gprog=gview>).

HOMD Genomic BLAST

With the increasing number of genomes being sequenced, the output of a high-throughput BLAST search can be very complex and time-consuming to interpret, with many redundant results. We recently developed a graphic tool based on newly improved BLAST+ (Camacho et al. 2009) that allows the user to customize BLAST searches by dynamically selecting a group of any combination of the genomic sequences available in HOMD. The HOMD Genomic BLAST provides a visual taxonomy-based navigation interface (Fig. 9) for easy and dynamic selection of a set of genomes for sequence homology search. The selection can be a combination of individual genomes and/or



Human Oral Microbiome Database (HOMD), Fig. 8 HOMD Genome Viewer displaying multiple sources of annotations for *Aggregatibacter actinomycetemcomitans* HK1651

a group of genomes related at any taxonomic level (species, genus, etc.). The BLAST parameters are dynamically presented after the genome selection, and the results are available on the web and for download in multiple formats.

The HOMD Genomic BLAST query interface starts with the selection of the genomes to be searched against. All the HOMD genomes available for search are displayed and selectable in a collapsible tree based on the taxonomy

hierarchy. As shown in Fig. 9, upon starting the HOMD Genomic BLAST, the taxonomy hierarchical tree is fully expanded by default and can be dynamically collapsed at any given level. The links, at the species level or genomes level, lead to the detailed Taxon Description or Sequence Meta Information page, respectively. Numbers indicated in the square brackets at each level are the numbers of oral taxa, genomes with meta information, genomes with HOMD annotation,

HOMD Genomic BLAST against selected human oral microbial genomes Previous page

Select: all genomes one (1st) genome per taxon Include: HOMD annotated genes NCBI annotated genes

Total genomes selected: 310 (BLAST options will be shown on next page)

Numbers indicated in brackets at each level: [Oral Taxa, Genomes with Meta Info, Genomes with HOMD annotation, Genomes with NCBI Annotation]

- Domain Archaea [1,0,0,0]
 - Phylum Euryarchaeota [1,0,0,0]
 - Class Methanobacteria [1,0,0,0]
 - Order Methanobacteriales [1,0,0,0]
 - Family Methanobacteriaceae [1,0,0,0]
 - Genus Methanobrevibacter [1,0,0,0]
 - Species Methanobrevibacter oralis [1,0,0,0] [Taxon Page]
- Domain Bacteria [687,1314,329,1212]
 - Phylum Actinobacteria [91,176,50,157]
 - Class Actinobacteria [91,176,50,157]
 - Order Actinomycetales [64,109,32,94]
 - Family Actinomycetaceae [37,34,16,23]
 - Genus Actinobaculum [1,1,0,0]
 - Species Actinobaculum sp. oral taxon 183 [1,1,0,0] [Taxon Page]
 - Genus Actinomyces [34,28,16,18]
 - Species Actinomyces cardiffensis [1,1,1,1] [Taxon Page]
 - Actinomyces cardiffensis F0333 [Meta Page]
 - Actinomyces cardiffensis F0333 [Meta Page]
 - Species Actinomyces dentalis [1,0,0,0] [Taxon Page]
 - Species Actinomyces georgiae [1,1,1,1] [Taxon Page]
 - Actinomyces georgiae F0490 [Meta Page]
 - Actinomyces georgiae F0490 [Meta Page]
 - Species Actinomyces gerenceriae [1,1,0,0] [Taxon Page]
 - Species Actinomyces graevenitzi [1,2,1,1] [Taxon Page]
 - Actinomyces graevenitzi C83 [Meta Page]
 - Actinomyces graevenitzi C83 [Meta Page]
 - Species Actinomyces israelii [1,1,0,0] [Taxon Page]
 - Actinomyces johnsonii F0330 [Meta Page]
 - Actinomyces johnsonii F0330 [Meta Page]
 - Species Actinomyces massiliensis [1,1,2] [Taxon Page]
 - Actinomyces massiliensis F0489 [Meta Page]
 - Actinomyces massiliensis F0489 [Meta Page]
 - Actinomyces massiliensis 4401292 [Meta Page]
 - Species Actinomyces meyeri [1,0,0,0] [Taxon Page]
 - Species Actinomyces naeslundii [1,3,1,1] [Taxon Page]
 - Actinomyces naeslundii MG1 [Meta Page]
 - Actinomyces naeslundii str. Howell 279 [Meta Page]

Human Oral Microbiome Database (HOMD), Fig. 9 Screenshot of the HOMD Genomic BLAST tool – the genome selection page showing 107 *Bacteroides* genomic sequences selected for BLAST Search

and genomes with NCBI annotation, respectively. The genome selection is flexible and can be a single genome, any randomly selected individual genomes, a group of genomes at any taxonomy level (from Domain to Species), all the genomes dynamically annotated at HOMD, all the genomes with static annotations by NCBI, or a representative genome from all the species. The total number of genomes selected is shown on top of the page.

After the genomes are selected, users are directed to the next page for providing the query sequence and options for BLAST search (Fig. 10). A summary of the selected genome(s) is presented on top of this page with an option for going back and modifying the selection. Below the summary is the query sequence form.

The query sequence, in FASTA format, can be copied and pasted into the sequence field or uploaded directly from user's computer. Multiple sequences are allowed with the limit of ten sequences. BLAST parameters are dynamically changed based on the type of query and subject sequences. The query sequences can be either nucleotide or protein sequences. The subject can be whole genomic DNA sequences or nucleotide or amino acid sequences of the annotated proteins of the selected genomes. Once the sequence type (nucleotide or protein) is selected by user for both query and subject sequences, suitable BLAST programs are dynamically displayed for selection. For example, if both query and subject sequences are proteins, only BLASTP is available for search; likewise, if both queries and

HOMD Genomic BLAST

Against selected human oral microbial genomes

Subject Sequences

Total genomes selected: 310 -- click to view summary [▶](#) [Review/Modify](#) selection

Search against: Protein sequences (Amino Acid) of selected genomes
 Protein coding sequences (DNA)
 Genomic contigs (DNA)

Query Sequences

Query sequence type: DNA/RNA Protein

Enter query sequence below in **FASTA** format (limit 10 sequences): [help](#) [Clear sequence](#)

```

>query protein
MSETKYRLGIDIGSTTVKVALIDNDLKVLFSDYQRHYANIQETLASLLHDAIKVCGNAEV
YAMITGSGGLTLSKHLDIPFVQEVIAVATALKTFAPQTDVAIELGGEDAKIIFTGGIEQ
RMNGICAGGTGSFIDQMASLLKTDAAAGLNEYAKSYQSIYPIAARCGVFAKSDIQPLINDG
ATKPDLAASIFQAVVNQTTISGLACGKPIRGNVAFLGGPLHFLTELQAAFRTLKLGPENI
VAPEGSHLFAAMGSAMSSKYEKATTLANLHNTLHNKVSMDFEVARLERLFNSEEDYEAFK
AEHKKADEVKGNLADYHGKCFGLIDAGSTTTKVAVVAEDGTLTLYSPYSSNNGSPLKTSIK
AFNEIHELMPKDKCIARSCSTGYGEALLKAAFLLEGEVETVAHYTAASFFNPKVDCILD
IGGQDMKCIKIDGTVDGIQLNEACSSGCGSFIFAKSLNVEVADFAKVALLAENPIDL
CSRCTVFMNSKVKOAKFGASVADISSGI AYSVIKNAI I KVIKI TDPKDI GFNIWVQGGT

```

Or upload sequences from your computer: [Browse...](#)

Select a BLAST Program:

- BLASTP (protein-protein BLAST) [help](#)
- PSI-BLAST (Position-Specific Iterated BLAST) [help](#)
- PHI-BLAST (Pattern Hit Initiated BLAST) [help](#)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST) [help](#)

[Start BLAST](#)

General Parameters

Expect value: [help](#)

Word size: [help](#)

No. of Descriptions: [info](#)

No. of Alignments: [info](#)

Scoring Parameters

Matrix: [help](#)

Gap costs: Existence: 11 Extension: 1 [help](#)

Compositional adjustments: [help](#)

Filter and Masking Options

Filter: Low complexity regions [help](#)

Masks: Mask for lookup table only [help](#)
 Mask lower case letters [help](#)

Advanced Options [▶](#)

Human Oral Microbiome Database (HOMD), Fig. 10 The HOMD Genomic BLAST tool – query sequence input and BLAST parameter adjustment page

subjects are nucleotides, the search can be done with BLASTN, BLASTX, or TBLASTX. Furthermore, alternative algorithms are available for nucleotide to nucleotide searches, including MegaBLAST (Morgulis et al. 2008) and Discontiguous MegaBLAST (Morgulis et al. 2008). Similarly, for protein to protein searches, available algorithms are BLASTP, PSI-BLAST (Altschul et al. 1997), PHI-BLAST (Zhang et al. 1998), and DELTA-BLAST (Boratyn et al. 2012). For each BLAST program, only the parameters and options corresponding to the selected program type and algorithm appear on this page. Detailed information about BLAST parameters is available under the link “Help.” For the advanced users, the command-line style BLAST+ parameters can be added in Advanced Option section (Camacho et al. 2009).

Upon submission of the BLAST search, the requested job is sent to the back-end service for processing. The back-end service consists of a computer cluster to handle multiple requests from the query interface. The selected genomes/nucleotides/proteins are dynamically compiled to a virtual sequence database searchable by the BLAST programs, using the “blastdb_aliastool” tool provided by BLAST+ (Camacho et al. 2009). The searched jobs are distributed to the computer nodes of the cluster, which is managed by the TORQUE resource manager (<http://www.adaptivecomputing.com/products/open-source/torque>). During the search process, user is presented with an intermediate page to monitor the job status. This status page reports a summary of the job as well as time/duration elapsed since submission. The status page periodically refreshes itself, effectively polling the server while the job runs. BLAST result is automatically presented when the job completes.

BLAST results are presented dynamically in the output interface (Fig. 11). Users can check the details of BLAST job information and choose to download the results in different formats, such as HTML, archive, text, tabular, CSV, and XML. Additional jobs can also be submitted for the same queries and subjects with modified

parameters. The search strategy including the query, subject, and BLAST parameters can be saved or downloaded for future reference. The actual BLAST results are presented in a manner similar to the typical HTML format. They include a Graphical Overview section (Fig. 3) to display the alignment of the “high-scoring pairs” (HSPs) between the query and the subject sequences. HSPs are plotted against the query sequence and highlighted by different colors based on alignment scores. Every HSP on the plot is hyperlinked with the corresponding pairwise alignment in the Alignment section. Subject sequences that matched the query are listed in the Descriptions section, sorted by the expected (*e*) values. The Alignment section presents the alignments of the HSPs as a series of pairwise alignments. Each alignment contains a hyperlink to the corresponding HOMD- or NCBI-annotated gene, if such information is available.

To provide the research community with satisfactory experience with and the convenient features of the HOMD Genomic BLAST, we currently allow up to ten query sequences to be searched in a single job request. Since the time needed for the computation is linear-proportional to the numbers of both query and subject sequences, we expect the maximal waiting time to be no longer than 10 min, provided no previous job is waiting in the job queue. In fact, when a total of ten protein sequences with the size of 500 amino acids in length were submitted to an empty queue to search against all the protein sequences of all HOMD genomes, the job was completed in about 400 s, without any prior jobs waiting in the cluster queue. Special requests may be considered for jobs containing more query sequences than the current limit, on the collaboration basis.

The number of the genomes hosted by HOMD database has been growing from approximately 600 genomes at launch (June 2011) to nearly 1,200 genomes towards the beginning of 2013. We expect the number continue to grow, in concordance with the growth of the NCBI microbial genomes, as well as the progress of the Human

HOMD Genomic BLAST Result

[Genomic BLAST Home]

Display detail BLAST job information and additional options
 Bookmark Result*

Start Over	<input type="button" value="Search Again"/> <input checked="" type="radio"/> against same set of genomes <input type="radio"/> select different genomes													
JOB ID	20130515150922WEHDZQ													
Submitted Date/time	2013-05-15 15:10:32													
Time Elapsed	1 min 9 sec													
Genomes Selection	<table border="1" style="font-size: small; border-collapse: collapse;"> <thead> <tr> <th>Annotation</th> <th>Genomes Selected</th> <th>Searchable</th> </tr> </thead> <tbody> <tr> <td>HOMD</td> <td>310</td> <td>310</td> </tr> <tr> <td>NCBI</td> <td>0</td> <td>0</td> </tr> <tr> <td>Combined</td> <td>310</td> <td>310</td> </tr> </tbody> </table> <input type="button" value="View Detail Table"/>	Annotation	Genomes Selected	Searchable	HOMD	310	310	NCBI	0	0	Combined	310	310	
Annotation	Genomes Selected	Searchable												
HOMD	310	310												
NCBI	0	0												
Combined	310	310												
BLAST Parameters	<pre style="font-size: x-small; margin: 0;">-evalue 0.00001 -num_descriptions 100 -num_alignments 50 -task blastp -matrix BLOSUM62 -word_size 3 -gapopen 11 -gapextend 1 -comp_based_stats 2 -seg no -soft_masking false</pre>													
Available Options	Download result in different formats: <input type="button" value="HTML"/> <input type="button" value="Archive"/> <input type="button" value="Text"/> <input type="button" value="Tabular"/> <input type="button" value="CSV"/> <input type="button" value="XML"/>													
	Download search strategy: <input type="button" value="Strategy"/>													

* The result will be kept on the HOMD server for 1 week, afterwards the bookmark will no longer work.

BLASTP 2.2.26+

Reference:
 Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Reference for composition-based statistics:
 Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Res.* 29:2994-3005.

Database: Combined Database: 310 genomes
 3,962,770 sequences; 568,039,506 total letters

Query query	0	500	1000
adef_c_1_1722			
lot107_c_4_589			

Human Oral Microbiome Database (HOMD), Fig. 11 The HOMD Genomic BLAST tool result summary page showing different download option for the BLAST search results

Microbiome Project. To keep pace with this foreseeable growth and the computing power necessary for Genomic BLAST and other tools, we will continue the efforts to enhance the capabilities of HOMD's computer backbone.

Conclusions

The goal of creating the HOMD website and tools has been to create a community resource for those interested in obtaining information on human oral

bacteria and their genomes. We have attempted to create a useful provisional taxonomic scheme so that investigators can refer to phylogenetically defined taxa rather than unanchored clones or OTUs. We provide full-length reference sequences and BLAST tools tied to our taxonomic scheme. Finally, we provide access to all genomes completed for human oral bacteria.

References

- Aas JA, et al. Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol.* 2005;43:5721–32.
- Alcaraz LD, et al. Identifying a healthy oral microbiome through metagenomics. *Clin Microbiol Infect.* 2012;18 Suppl 4:54–7.
- Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402.
- Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25:25–9.
- Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res.* 2000;28:304–5.
- Belda-Ferre P, et al. The oral metagenome in health and disease. *ISME J.* 2012;6:46–56.
- Bik EM, et al. Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* 2010;4:962–74.
- Boeckmann B, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003;31:365–70.
- Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct.* 2012;7:12. doi: 10.1186/1745-6150-7-12.PMID:22510480.
- Camacho C, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421.
- Camon E, et al. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* 2003;13:662–72.
- Chen T, et al. The bioinformatics resource for oral pathogens. *Nucleic Acids Res.* 2005;33:W734–40.
- Chen T, et al. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford).* 2010;2010:baq013.
- Dewhirst FE, et al. The human oral microbiome. *J Bacteriol.* 2010;192:5002–17.
- Dzink JL, et al. Gram negative species associated with active destructive periodontal lesions. *J Clin Periodontol.* 1985;12:648–59.
- Dzink JL, et al. The predominant cultivable microbiota of active and inactive lesions of destructive periodontal diseases. *J Clin Periodontol.* 1988;15:316–23.
- Human Microbiome Project Consortium. A framework for human microbiome research. *Nature.* 2012a;486: 215–21.
- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012b;486:207–14.
- Kanehisa M. The KEGG database. *Novartis Found Symp.* 2002;247:91–101. discussion 101–103, 119–128, 244–152.
- Martin J, et al. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS One.* 2012;7: e36427.
- Moore WE, Moore LV. The bacteria of periodontal diseases. *Periodontol.* 1994;2000(5):66–77.
- Moore WE, et al. Bacteriology of severe periodontitis in young adult humans. *Infect Immun.* 1982;38:1137–48.
- Moore WE, et al. Bacteriology of moderate (chronic) periodontitis in mature adult humans. *Infect Immun.* 1983;42:510–5.
- Morgulis A, et al. Database indexing for production MegaBLAST searches. *Bioinformatics.* 2008;24: 1757–64.
- Paster BJ, Dewhirst FE. Phylogeny of campylobacters, wolliellas, *Bacteroides gracilis*, and *Bacteroides ureolyticus* by 16S ribosomal ribonucleic acid sequencing. *Int J Syst Bacteriol.* 1988;38:56–62.
- Socransky SS, Haffajee AD. Evidence of bacterial etiology: a historical perspective. *Periodontology.* 1994;5:7–25.
- Tanner AC, et al. A study of the bacteria associated with advancing periodontitis in man. *J Clin Periodontol.* 1979;6:278–307.
- Tanner A, et al. Microbiota of health, gingivitis, and initial periodontitis. *J Clin Periodontol.* 1998; 25:85–98.
- The Forsyth Metagenomic Support Consortium, Izard J. Building the genomic base-layer of the oral “omic” world. In: Sasano T, Suzuki O, editors. *Interface oral health science 2009: proceedings of the 3rd international symposium for interface oral health science.* New York: Springer; 2010.
- Xie G, et al. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol Oral Microbiol.* 2010; 25:391–405.
- Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17:847–8.
- Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 1998;26(17):3986–90.
- Zuger J, et al. Uncultivated *Tannerella* BU045 and BU063 are slim segmented filamentous rods of high prevalence but low abundance in inflammatory disease-associated dental plaques. *Microbiology.* 2007;153:3809–16.

Insights into Environmental Microbial Denitrification from Integrated Metagenomic, Cultivation, and Genomic Analyses

Stefan J. Green¹, Lavanya Rishishwar²,
Om Prakash³, I. King Jordan⁴ and Joel Kostka⁵

¹University of Illinois at Chicago, Chicago, IL, USA

²Bioinformatics, Georgia Institute of Technology, Atlanta, GA, USA

³National Centre for Cell Science, Pune, Maharashtra, India

⁴School of Biology, Georgia Institute of Technology, Atlanta, GA, USA

⁵School of Biology and Earth & Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

Genome sequencing; Metagenomic technology; Nitrogen cycling

Definition

The high sequence diversity of microbial functional genes can hinder cultivation-independent molecular analyses. Likewise, cultivation-based approaches also provide a distorted picture of *in situ* microbial communities. When cultivation and cultivation-independent molecular approaches are

acquired in tandem, deeper insights into community structure of organisms catalyzing specific metabolic functions can be obtained. Coupled cultivation, amplicon, genome, and metagenome sequence data, targeting denitrifying bacteria from a highly contaminated subsurface environment, were analyzed to reveal novel denitrifier diversity and the extent of bias associated with commonly used PCR primer sets targeting denitrification genes. Furthermore, genome sequencing revealed that some denitrifiers are incapable of denitrification from nitrate and demonstrated the need for integrated molecular and cultivation approaches to characterization of microbial communities.

Introduction

The advent of next-generation sequencing platforms and the subsequent increased availability of genomic and metagenomic sequence data have revolutionized environmental microbiology. However, though our eyes have been opened to the vast genotypic and metabolic potential of microbial communities in nature, exploration of the role of specific microbial groups in ecosystem function still requires the application of cultivation-based approaches. In fact, the verification of microbial phenotypes through cultivation is arguably more critical than ever as metagenomic information now allows for the generation of boundless hypotheses based on the metabolic potential represented by complex

microbial communities. Although the advances in cultivation-independent molecular analyses of microbial communities have been well advertised (e.g., high-throughput amplicon sequencing (e.g., Caporaso et al. 2011), metagenomics (e.g., Tringe et al. 2005), and metatranscriptomics (e.g., Poretsky et al. 2009)), parallel advances in cultivation have also been made, including the use of lower organic carbon media, extended incubation, single-cell encapsulation approaches, and overall improved mimicking of natural conditions within a culture vessel (e.g., Bollmann et al. 2007; Kaeberlein et al. 2002; Zengler et al. 2002). Here, data from metagenomic sequencing and isolation, physiological testing, and whole-genome sequencing of denitrifying bacteria from the highly contaminated subsurface of the Oak Ridge Integrated Field Research Challenge (ORIFRC) site are considered and the implications of this analysis on understanding the environmental distribution and ecological niche of denitrifying bacteria.

The ORIFRC Site

The ORIFRC site is highly contaminated with spent uranium and a wide variety of other contaminants (e.g., other radionuclides, heavy metals, and volatile organic contaminants) as a result of long-term uranium enrichment for nuclear weapons, coupled with improper disposal in unlined ponds (S-3 ponds) (Brooks 2001; Kostka and Green 2011; NABIR 2003; Watson et al. 2004). Although the ponds have been subsequently drained, much of the contaminant has migrated into the subsurface, where it serves to feed a plume migrating down-gradient across the site (Watson et al. 2005). Uranium is the priority contaminant of concern, though the nitrate in the near-source zone (adjacent to the former S-3 ponds) reaches extraordinarily high concentrations (in the range of 10–1,000 mM) due to the use of nitric acid in the processing of uranium. The high level of nitrate complicates remediation strategies at the site by inhibiting microbial reduction of soluble hexavalent uranium to an insoluble mineral form of tetravalent uranium (e.g., Finneran et al. 2002; Kostka and Green 2011; Shelobolina et al. 2003). The moderately

high acidity in the source zone (pH 3–4) also suppresses microbial activity and diversity (Fields et al. 2005; Hemme et al. 2010). Despite the restrictive conditions, there is evidence for significant nitrous oxide production in the near-source zone (Spalding and Watson 2008). As the low pH is ameliorated down-gradient of the source zone, nitrate, nitrous oxide, and soluble uranium are attenuated without active remediation, due to both microbial and geochemical processes (Kowalsky et al. 2011).

The contaminant levels in the near-source zone are alarming, and source zone remediation strategies have been examined, with limited success (Wu et al. 2007). The extraordinary levels of nitrate must be removed before microbial reduction of U(VI) to U(IV) can proceed (Akob et al. 2008; Luo et al. 2005; Wu et al. 2006, 2010), and down-gradient remediation has been more effective as nitrate is essentially absent (e.g., Gihring et al. 2011). The presence of nitrous oxide in the source zone wells suggested the presence of *in situ* denitrification, and thus grew an interest in microorganisms capable of nitrate reduction at *in situ* pH, with the hope that stimulation of these native organisms could aid in the long-term removal of uranium from the site groundwater. Initial studies revealed significant diversity in nitrite reductase genes in groundwater at the site, including both genes encoding for copper-containing (*nirK*) and cytochrome (*nirS*) forms (Palumbo et al. 2004; Yan et al. 2003). Based on metagenomic analysis of acidic groundwater from the site, Hemme et al. (2010) hypothesized that denitrification comprised the predominant form of metabolism in the near-source zone microbial community due to the low oxygen and lack of fermentation genes observed there. The overabundance of nitrate/nitrite antiporters in the metagenome was interpreted as a further indication of the strong effect of the elevated nitrate on the source zone microbial community.

Prior to the metagenome sequencing of the acidic groundwater at the ORIFRC site, cultivation-independent molecular surveys had been performed to track denitrifying organisms. As the denitrification phenotype is a polyphyletic

trait, and can be acquired readily via lateral gene transfer, ribosomal RNA gene sequencing is not suitable for identifying and tracking denitrifying organisms. Functional genes assays – targeting nitrate, nitrite, nitric oxide, and nitrous oxide reductases – have been performed for this purpose. Yan et al. (2003) and Palumbo et al. (2004) performed site-wide surveys of nitrite reductase genes at the ORIFRC site. No clear pattern relating the composition and relative abundance of nitrite reductase genes with groundwater geochemical conditions was observed, however. For example, a principal component analysis of clusters of *nirK* (gene encoding for copper-containing nitrite reductase) sequences grouped all wells across the pH gradient together, with the exception of one high nitrate groundwater sample. In all wells, the most abundant *nirK* sequences were most similar to the *nirK* gene sequence derived from *Hyphomicrobium zavarzini*, and all sequences were most similar to gene sequences derived from Proteobacteria. Thus, although a substantial diversity of nitrite reductase genes was observed, with many novel gene sequences recovered, more recent data from genome and metagenome sequencing indicates that the predominant denitrifiers were not detected in single-gene surveys (Green et al. 2010, 2012; Hemme et al. 2010).

Combined Cultivation and Direct Molecular Studies of Denitrifying Bacteria

The study of denitrifying microorganisms at the ORIFRC field site was approached in a multipronged fashion, including (a) site-wide microbial community characterization using DNA extraction from sediment and groundwater, coupled with high-throughput bacterial ribosomal RNA (rRNA) gene amplicon sequencing, (b) quantitative PCR (qPCR) analyses of bacterial small subunit (SSU) rRNA and nitrite reductase (*nirK*) gene abundance in groundwater and sediment samples, (c) cultivation and physiological testing of denitrifying bacteria from sediment and groundwater, and (d) *de novo* whole-genome sequencing of denitrifying isolates. Subsequently, genomic DNA (gDNA) samples from the site were reanalyzed with novel primers

targeting unique *nirK* genes, and whole-genome sequences were also recovered from non-denitrifying reference strains related to organisms isolated from the field site.

Bacteria from six distinct genera of denitrifiers were isolated, including strains of *Hyphomicrobium* (Alphaproteobacteria), *Afipia* (Alphaproteobacterium), *Pseudomonas* (Gammaproteobacteria), *Rhodanobacter* (Gammaproteobacteria), *Bacillus* (Firmicutes), and *Intrasporangium* (Actinobacteria) (Green et al. 2010). Under laboratory conditions, all strains were capable of growth with nitrate as the sole electron acceptor, though the Gram-positive strains produced only nitrous oxide as a terminal product, while *Rhodanobacter* spp. produced a mixture of nitrous oxide and nitrogen gas. Physiological and genetic characterization of the isolates from the genus *Rhodanobacter* was prioritized, as these organisms had been detected in great abundance in acidic groundwater as well as sediments from the near-source zone (Green et al. 2010, 2012). Bacteria from this genus were revealed to have extraordinarily high relative abundance in the near-source zone, over multiple sampling seasons, and were sometimes the only active organisms detected in RNA-based analyses of groundwater samples (Green et al. 2012). Highly similar strains were independently isolated from ORIFRC site sediment using a diffusion chamber approach (Bollmann et al. 2010), and in a metagenomic survey of acidic groundwater from the site, one of the dominant organisms detected (so-called FW106 γ I) is clearly a member of the genus *Rhodanobacter* (Hemme et al. 2010). This organism contained a full denitrification pathway.

Despite the apparent numerical abundance of members of the genus *Rhodanobacter* in the acidic source zone, these organisms were not detected in prior molecular surveys of denitrification pathway genes at the ORIFRC site (Palumbo et al. 2004; Yan et al. 2003). Nor could PCR amplification of *nirS* (cytochrome cd 1-containing nitrite reductase), *nirK*, or *nosZ* (nitrous oxide reductase) genes be achieved using standard primer sets (Green et al. 2010). Similar challenges were presented by the other

isolated strains, excepting *Afipia*. For the *Hyphomicrobium* strain, a novel primer set targeting *nirK* was designed based on a reference gene available in GenBank, but no similar reference sequences were available for the other strains. Subsequently, metagenome sequence data from acidic groundwater acquired at the site (Hemme et al. 2010) was surveyed, and two novel *nirK* sequences were identified. Using these *de novo* assembled sequences, primer sets were developed that allowed the amplification of a *nirK* gene from the *Rhodanobacter* isolates and from putative *Rhodanobacter* organisms from environmental genomic DNA (Green et al. 2010, 2012). Quantitative PCR analysis was utilized to quantitate SSU rRNA and *nirK* gene abundance in groundwater from across the watershed, and this analysis revealed that *nirK* genes were present in abundance across the ORIFRC site, including *nirK* genes derived from *Rhodanobacter* (Green et al. 2012). Coupled with relative abundance measurements derived from qPCR of rRNA genes and from rRNA gene amplicon sequencing, this analysis revealed that *Rhodanobacter* were the most abundant organisms in the near-source zone, that *nirK* genes most similar to those from *Rhodanobacter* strains were most abundant in the near-source zone, and that *Rhodanobacter* organisms were active, not just present in the near-source zone. Coupled with *in vitro* analysis of the physiological capabilities of *Rhodanobacter* strains in pure culture, these data led to the hypothesis that bacteria from the genus *Rhodanobacter* are the dominant near-source zone denitrifiers at the ORIFRC site. This hypothesis is supported by studies conducted in other ecosystems which demonstrate that *Rhodanobacter* spp. dominate under low pH, denitrifying conditions (e.g., van den Heuvel et al. 2010).

Direct PCR amplification of nitrite reductase genes from *Rhodanobacter* and other denitrifiers isolated from the site was not successful using standard primers, and subsequently, *de novo* shotgun genome sequencing and draft assembly of these bacterial denitrifiers was performed. The initial draft sequences of *Rhodanobacter* and *Intrasporangium* recovered complete *nirK* genes

and helped determine the cause of PCR amplification failure. First, the putative nitrite reductase genes from these organisms were highly divergent from many sequences present in gene databases, and the sequences contained a large number of mismatches with the most commonly used primer sets for targeting bacterial *nirK* genes (e.g., 10 and 11 mismatches, respectively, between primer R3Cu and first and second *nirK* gene of *R. denitrificans* 2APBS1 (Green et al. 2010; Hallin and Lindgren 1999)). In addition, most *Rhodanobacter* spp. have two highly divergent *nirK* genes located in different positions in the genome (Green et al. 2010; Kostka et al. 2012). Two strains of *Rhodanobacter* independently isolated (Bollmann et al. 2010) similarly contain two *nirK* genes apiece, and both are nearly (>99% similar) or completely identical to *nirK* genes from *R. denitrificans* 2APBS1^T. Both forms of *nirK* are expressed under denitrifying conditions in *R. denitrificans* 2APBS1^T, but the purpose of two copies of the gene is not yet clear (Green et al. 2012). One copy of the gene, colloquially called “*nirK*-B,” is most similar to *nirK* genes from certain Proteobacteria, including Betaproteobacteria from the genera *Burkholderia* and *Ralstonia*. The second copy, called “*nirK*-V,” is most similar to the *nirK* gene from *Opitutus terrae* PB90-1, within the phylum Verrucomicrobia.

To examine this phenomenon on a broader phylogenetic scale, Green et al. (2010) recovered complete *nirK* and *nosZ* genes from a number of microorganisms which had been sequenced by the Joint Genome Institute. These genes were aligned and primer binding sites were identified. This analysis revealed that the difficulty in amplifying *nirK* genes from ORIFRC site isolates is symptomatic of a broader difficulty in detecting denitrifying bacteria through single primer set amplification due to large numbers of mismatches between primer and gene sequences. The commonly used primer sets (including quantitative PCR primer sets) target a relatively narrow range of organisms, primarily within the Proteobacteria (Green et al. 2010). Thus, molecular approaches that depend upon single primers, even heavily degenerate primers, cannot be used

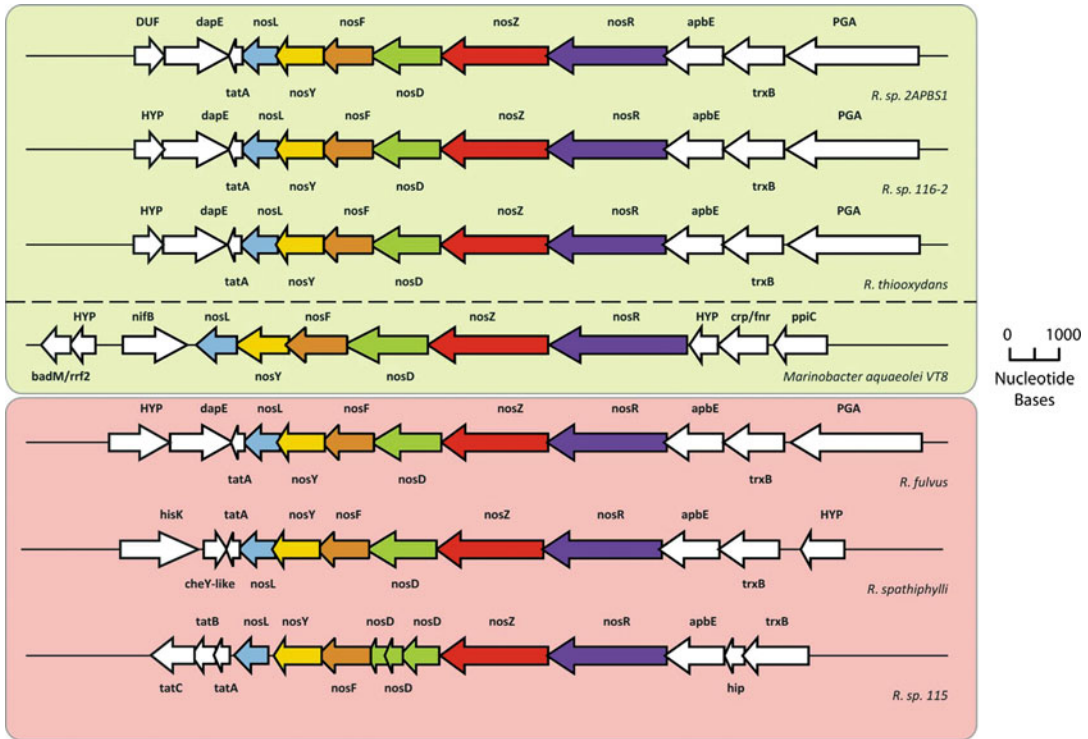
suitably to detect or quantify denitrifiers in environmental samples, and the true diversity and abundance of denitrifiers is most likely greatly underestimated from current surveys. Alternate approaches, which utilize the full availability of reference sequence data derived from *de novo* genome sequencing and from shotgun metagenome sequencing of environmental samples, must be developed to more fully assess the distribution of these important organisms.

Although the nitrite reductase gene is a particularly dramatic example, it is not unique in this regard, and other functional genes of significance to biogeochemical processes have shown similar levels of sequence diversity. The sequence diversity of *nirK* may be in part due to the multiple physiological roles for nitrite reduction (detoxification, respiration), different conditions under which the enzymes may be active (e.g., prior to anoxic conditions, after total anoxia), and multiple locations for nitrite reductases (periplasm, inner membrane) and for the different forms of the gene (copper nitrite reductase, *nirK*, and cytochrome-cd1 *nirS*). This broad sequence divergence but with retained function is present in other functional genes, including other genes in the denitrification pathway (e.g., *nosZ*; Green et al. 2010; Jones et al. 2013; Sanford et al. 2012).

Although many *Rhodanobacter* spp. isolated from the ORIFRC site subsurface were capable of complete denitrification, some members of the genus were incapable of growth on nitrate. Similarly, in a survey of the literature regarding *Rhodanobacter*, most strains were identified as aerobic bacteria, incapable of nitrate reduction. Strains isolated independently from the ORIFRC site were observed to be acid tolerant (arrest of growth was observed at pH 3.5–4), tolerant of high levels of nitrate (up to 250 mM), and moderately tolerant of various heavy metals, including uranium (Bollmann et al. 2010). The initial description of *R. thiooxydans*, the closest relative of *R. denitrificans*, indicated that the organism was capable of nitrate, but not nitrite, reduction (Lee et al. 2007). Subsequent work, however, demonstrated that these organisms are capable of complete denitrification from nitrate

(Prakash et al. 2012; van den Heuvel et al. 2010). More recently, a novel species, *R. caeni*, was described as capable of nitrate reduction to nitrite, but no evidence for complete denitrification was demonstrated (Woo et al. 2012). Likewise, *R. sp.* strain A2-61, shown to form intracellular uranium-phosphate complexes, was unable to reduce nitrate (Sousa et al. 2013).

To understand the genetic basis of the differences in physiology with respect to denitrification, the genomes of five additional strains of bacteria from the genus *Rhodanobacter* were sequenced (Kostka et al. 2012). In total, three strains of denitrifying *Rhodanobacter* were sequenced (*R. denitrificans* 2APBS1^T, *R. denitrificans* 116-2, *R. thiooxydans*) alongside three strains of apparent non-denitrifying (from nitrate) *Rhodanobacter* (*R. fulvus* Jip2 (Im et al. 2004), *R. spathiphylli* B39 (De Clercq et al. 2006), and *R. sp.* 115, isolated from the ORIFRC site (Kostka et al. 2012)). Preliminary analysis of the genomes of the six *Rhodanobacter* strains revealed that all members of the genus contained nearly complete denitrification pathways, including two copies of the nitrite reductase gene *nirK* (excepting *R. spathiphylli*, with only a single copy). All denitrifying isolates contained many genes in the dissimilatory denitrification pathway, but non-denitrifying isolates were missing several key genes involved in nitrate respiration, such as nitrate reductase genes (i.e., *narG*, *narH*, *narJ*, and *narI*). The genomic context of these genes was further examined, and it was observed that the nitrous oxide genes (e.g., *nosZ*) showed the greatest synteny among all six genomes (Fig. 1). Since relatively few organisms conduct nitrous oxide reduction alone, it may be supposed that the high level of synteny in this gene and the lower synteny in other parts of the denitrification pathway favor the hypothesis that the ancestral common ancestor of the bacteria within the genus *Rhodanobacter* likewise contained a full denitrification pathway, with subsequent rearrangement of the genes in the pathway. Further clarity will be obtained with additional whole-genome sequences of related organisms from the Xanthomonadaceae.



Insights into Environmental Microbial Denitrification from Integrated Metagenomic, Cultivation, and Genomic Analyses, Fig. 1 Gene order in the genomic region of the nitrous oxide reductase gene (*nosZ*) in denitrifying and apparent non-denitrifying strains of bacteria from the genus *Rhodanobacter*.

Strong gene synteny is observed between denitrifying (highlighted in green) and apparent non-denitrifying lineages (highlighted in pink). Gene order in *Marinobacter aquaeolei* VT8 (Gammaproteobacteria, Alteromonadaceae), capable of anaerobic growth on nitrate, was included as an out-group organism with a complete genome sequence. Gene symbols: *apbE*, ApbE family lipoprotein; *cheY-like*, two-component system sensor histidine kinase-response regulator hybrid protein; *dapE*, succinyldiaminopimelate desuccinylase; *DUF*, protein of

unknown function DUF2165; *hip*, high potential iron-sulfur protein; *hisK*, sensor histidine kinase; *HYP*, hypothetical protein; *nosD*, periplasmic copper-binding protein; *nosF*, ABC transporter related protein; *nosL*, NosL protein; *nosR*, nitrous oxide expression regulator, NosR; *nosY*, ABC-type transport system involved in multi-copper enzyme maturation, permease component; *nosZ*, nitrous oxide reductase; *PGA*, peptidase S45 penicillin amidase; *tatA*, twin-arginine translocation protein, TatA/E; *tatB*, twin-arginine-targeting protein translocase TatB; *tatC*, twin-arginine-targeting protein translocase subunit TatC; *trxB*, thioredoxin reductase oxidoreductase; *badM/Rrf2*, BadM/Rrf2 family transcriptional regulator; *nifB*, molybdenum cofactor biosynthesis protein A; *ppiC*, PpiC-type peptidyl-prolyl *cis-trans* isomerase

Conclusions Regarding *Rhodanobacter*

Bacteria from the genus *Rhodanobacter* appear to fill a relatively specific ecological niche, but under appropriate conditions, these organisms can dominate to an extreme extent. Conditions which appear to enable bacteria from the genus *Rhodanobacter* to dominate include low pH, high nitrate, low/variable oxygen concentrations, and heavy metal contamination. Although data in the literature are not particularly abundant for

Rhodanobacter, what is present suggests that heavy metal tolerance is a common feature of these organisms. Bollmann et al. (2010) isolated two strains of *Rhodanobacter* that are tolerant of 200 micromolar uranium (as well as other heavy metals), and most recently Sousa et al. (2013) described *R. sp.* strain A2-61, tolerant of up to 500 micromolar uranium, under aerobic conditions. *R. denitrificans* strains are capable of tolerating 1 mM uranium (data not shown).

Interestingly, *R. sp.* strain A2-61 was capable of forming intracellular uranium-phosphate complexes, presumably a detoxification strategy. In a survey of the genome of *R. denitrificans* 2APBS1^T, multiple genes involved in metal resistance have been detected, and these genes are strongly associated with horizontal gene transfer as indicated by low lineage probability scores (LPI), anomalous nucleotide compositions, and association with putative mobile genetic elements such as transposons and integrons (data not shown).

The presence of a near-complete denitrification pathway in “non-denitrifying” strains of bacteria from the genus *Rhodanobacter* suggests that denitrification capability is an inherent trait of all members of the genus but that denitrification by these organisms often requires nitrite rather than nitrate. Since nitrite is often available where there is nitrate, and a number of organisms are capable of nitrate-to-nitrite reduction, but cannot reduce nitrite further, the lack of a nitrate reductase may not be overly limiting for facultative anaerobes such as members of the *Rhodanobacter*. For example, in a study of denitrification capabilities in bacteria from the genus *Bacillus*, most-probable-number assays of a soil sample revealed nearly an order of magnitude greater abundance of organisms capable of nitrate-to-nitrite reduction relative to complete denitrifiers (Verbaendert et al. 2011). A further confounding observation is the presence of two putative *nirK* genes in almost all *Rhodanobacter*, including the non-nitrate reducers. It may be that the multiple nitrite reductases are involved in tolerance of high nitrate/nitrite conditions, stressful conditions that are further exacerbated by low pH (Spain and Krumholz 2012). The nitrite reductases may also represent two different strategies relating to denitrification by *Rhodanobacter* under fluctuating aerobic/anaerobic conditions, such as those found in the ORIFRC site subsurface. As described by Bergaust et al. (2011), bacteria can employ complex strategies to maximize energy generation, but provide insurance in case of sudden changes in environmental condition. Thus, while in the presence of oxygen, denitrifying bacteria (which are nearly always facultative

anaerobes) will favor the use of oxygen as terminal electron acceptor, and repress nitrogen oxyanion reduction to avoid loss of ATP-generation capability through a truncated respiratory pathway, and “entrapment” under anoxic conditions without capability to continue respiration (Bergaust et al. 2011). It has been hypothesized that an earlier onset of denitrification (in terms of oxygen concentration) is an indication of the likelihood for nitrous oxide production by the strain (Bergaust et al. 2011; Zumft and Kroneck 2007). This is consistent with the initial characterization of *R. denitrificans*, in which both nitrous oxide and dinitrogen accumulated during pure culture growth conditions in vitro, while other isolates from the site completed denitrification to dinitrogen (*Afipia*, *Hyphomicrobium*) or nitrous oxide only (Gram positives; *Bacillus* and *Intrasporangium*) (Green et al. 2010). Further work is needed to determine the regulatory strategy taken by *Rhodanobacter* in the subsurface under aerobic/microaerophilic/anaerobic conditions.

Are *Rhodanobacter* extremophiles? Based on the current data, it is not clear that they are. Although members of the genus can grow at pH values below pH 4, the optimum growth pH for *R. denitrificans* 2APBS1 is pH 6 (Bollmann et al. 2010; Prakash et al. 2012). However, even at circumneutral pH with excess organic carbon, growth by *R. denitrificans* is slow (generation time ~24 h). This may represent another strategy by *Rhodanobacter* strains leading to dominance in contaminated/extreme environments, but low relative abundance in more ameliorated conditions. It appears most likely that *Rhodanobacter* retain a variety of physiological capabilities – anaerobic growth, metal tolerance and detoxification, denitrification phenotype, and broad carbon substrate utilization capability (including acetate) – that under specific environmental conditions provides them with the opportunity for dominance.

Conclusions Regarding Denitrification

The ORIFRC, with nitrate-replete groundwaters, represents an ideal natural laboratory for investigation of the microbial populations that mediate

denitrification. Through a close coupling of cultivation-based and molecular approaches, characterization of denitrifying bacteria from the ORIFRC site has significant implications not just for broader characterization of denitrifying organisms but also for the application of PCR-based approaches to characterize microbial functional groups. With specific reference to denitrification, it was observed that the most commonly used primers targeting functional genes within the dissimilatory denitrification pathway were highly biased to a select group of genes largely derived from bacteria within the Proteobacteria and the genes from organisms outside this group could not conceivably be targeted with PCR due to the excessively large number of mismatches between primer and gene sequence. Thus, results generated from single-gene primer (even degenerate) sets must be interpreted carefully. A similar finding has been obtained for nitrous oxide genes as well (Sanford et al. 2012). Since *de novo* genome and shotgun metagenome sequences generate gene sequences that are clearly identifiable as nitrite (or nitrous oxide) reductases but also impossible to target with common primers, new strategies must be developed to detect a broader collection of denitrifiers in the environment. As the organisms capable of denitrification are broadly distributed and are polyphyletic, functional gene analyses will continue to be essential to identify and quantify denitrifying microorganisms and to characterize denitrifying microbial communities.

One of the essential extrapolations of these findings is that the true abundance of denitrification capability in bacterial lineages is underestimated due to two processes revealed in this study. First, the high sequence divergence present in functional genes in the denitrification pathway limits the detection of denitrification genes from isolates through PCR and sequencing. Second, the partial pathway observed in *Rhodanobacter* strains suggests that when searching for denitrification capabilities, other electron acceptors besides nitrate should be tested. In a sense, cultivation approaches and physiological testing of *Rhodanobacter* strains

have been partially misleading regarding the potential ecological niche for these organisms, and only when coupled with whole-genome sequencing has the putative *in situ* functional capability of these organisms been revealed. In an analysis of *Bacillus* isolate and culture-collection strains, Verbaendert et al. (2011) revealed that nitrate was not always a suitable electron acceptor for verification of denitrification capability and that 20 % of denitrifying strains could use nitrite but not nitrate-to-initiate denitrification. They opine that the true abundance of denitrifiers is underestimated because typically only nitrate is used as an electron acceptor when testing for denitrification capability, and this is consistent with observations of isolates of the genus *Rhodanobacter*. Remarkably, they also observed that growth conditions can also affect electron acceptor utilization, and this can further lead to missing identification of physiological capability. No doubt analogous situations for other genes, organisms, and functions are with us, waiting to be identified. Thus, it seems clear that for more robust physiological characterization of bacterial strains, genome-guided physiological testing must be implemented. Such an approach will have profound implications for the assessment of the ecological role of bacteria taxa.

Prior to the acquisition of multiple genomes from the genus *Rhodanobacter*, the denitrification phenotype in *Rhodanobacter* strains was hypothesized to result from a relatively recent lateral gene transfer rather than from vertical transmission, as appears to be the case (Green et al. 2010). Hemme et al. (2010) also opined that the inferred lateral gene transfer events most likely occurred after the introduction of contamination at the site. With multiple genomes in hand, phylogenetic analysis of the nitrite reductase genes from the whole-genome sequences of multiple *Rhodanobacter* strains revealed a phylogeny consistent with that of the rRNA genes from the same organisms. If there were lateral gene transfer events, these predated the last common ancestor of the genus *Rhodanobacter*, with the most parsimonious

interpretation being that nitrate reduction capability was later lost from certain members of the genus. The evolutionary history of the full denitrification pathway, however, appears to be fragmented – for example, the *nirK* genes do appear to be derived from a lateral gene transfer, but this transfer is not recent and certainly is independent of the ORIFRC site. The *Rhodanobacter nosZ* genes are more consistent with other Gammaproteobacterial denitrifiers. It is possible, though entirely speculative, that *Rhodanobacter* previously had type (or class) I soluble periplasmic nitrite reductases, like those present in *Pseudomonas denitrificans*, and these have been subsequently replaced by type II cytoplasmic membrane nitrite reductases. The ecologic benefit derived from this is not clear yet, but may relate to activity under aerobic and anaerobic conditions, as has been observed for nitrate reductases (Bedzyk et al. 1999).

Summary

A combination of approaches to the study of denitrifying bacteria in a contaminated subsurface environment, including cultivation and physiological testing of denitrifying bacteria, *de novo* whole-genome sequencing, and shotgun metagenome sequencing, revealed key limitations to the application of more straightforward molecular approaches. Commonly used PCR primers targeting functional genes in the denitrification pathway are shown to be incapable of detecting a broad diversity of environmental denitrifiers. Likewise, some denitrifiers are incapable of nitrate reduction from nitrate and may be misidentified in routine physiological testing of bacterial isolates. Bacteria from the genus *Rhodanobacter*, which can be abundant in highly contaminated environments with low pH, appear to be native denitrifiers, while metal resistance genes appear to have been acquired via lateral gene transfer. Overall, *Rhodanobacter* dominate in certain environments with low pH, heavy metal contamination, and conditions favoring denitrification phenotype.

Cross-References

- ▶ [Culture Collections in the Study of Microbial Diversity, Importance](#)
- ▶ [Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing](#)
- ▶ [GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities](#)
- ▶ [Lateral Gene Transfer and Microbial Diversity](#)

References

- Akob DM, Mills HJ, Gihring TM, Kerkhof L, Stucki JW, Anastacio AS, Chin KJ, Kusel K, Palumbo AV, Watson DB, Kostka JE. Functional diversity and electron donor dependence of microbial populations capable of U(VI) reduction in radionuclide-contaminated subsurface sediments. *Appl Environ Microbiol.* 2008;74:3159–70.
- Bedzyk L, Wang T, Ye RW. The periplasmic nitrate reductase in *Pseudomonas* sp. strain G-179 catalyzes the first step of denitrification. *J Bacteriol.* 1999;181:2802–6.
- Bergaust L, Bakken LR, Frostegard A. Denitrification regulatory phenotype, a new term for the characterization of denitrifying bacteria. *Biochem Soc Trans.* 2011;39:207–12.
- Bollmann A, Lewis K, Epstein SS. Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. *Appl Environ Microbiol.* 2007;73:6386–90.
- Bollmann A, Palumbo AV, Lewis K, Epstein SS. Isolation and physiology of bacteria from contaminated subsurface sediments. *Appl Environ Microbiol.* 2010;76:7413–9.
- Brooks SC. Waste characteristics of the former S-3 ponds and outline of uranium chemistry relevant to NABIR Field Research Center studies. Oak Ridge: NABIR Field Research Center; 2001.
- Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4516–22.
- De Clercq D, Van Trappen S, Cleenwerck I, Ceustermans A, Swings J, Coosemans J, Ryckeboer J. *Rhodanobacter spathiphylli* sp. nov., a gammaproteobacterium isolated from the roots of *Spathiphyllum* plants grown in a compost-amended potting mix. *Int J Syst Evol Microbiol.* 2006;56:1755–9.

- Fields MW, Yan TF, Rhee SK, Carroll SL, Jardine PM, Watson DB, Criddle CS, Zhou JZ. Impacts on microbial communities and cultivable isolates from groundwater contaminated with high levels of nitric acid-uranium waste. *FEMS Microbiol Ecol.* 2005;53:417–28.
- Finneran KT, Housewright ME, Lovley DR. Multiple influences of nitrate on uranium solubility during bioremediation of uranium-contaminated subsurface sediments. *Environ Microbiol.* 2002;4:510–6.
- Gihring TM, Gengxin Z, Brooks SC, Campbell JH, Watson DB, Brandt CC, Yang Z, Criddle CS, Lowe K, Overholt WA, Wu W-M, Mehlhorn T, Kostka JE, Green SJ, Schadt CW. A limited microbial consortium is responsible for longer-term bioreduction of uranium in a contaminated aquifer. *Appl Environ Microbiol.* 2011;77:5955–65.
- Green SJ, Prakash O, Gihring TM, Akob DM, Jasrotia P, Jardine PM, Watson DB, Brown SD, Palumbo AV, Kostka JE. Denitrifying bacteria from the terrestrial subsurface exposed to mixed waste contamination. *Appl Environ Microbiol.* 2010;76:3244–54.
- Green SJ, Prakash O, Overholt WA, Cardenas E, Hubbard D, Akob DM, Tiedje JM, Watson DB, Jardine PM, Brooks SC, Kostka JE. Denitrifying bacteria from the genus *Rhodanobacter* dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. *Appl Environ Microbiol.* 2012;78:1039–47.
- Hallin S, Lindgren PE. PCR detection of genes encoding nitrite reductase in denitrifying bacteria. *Appl Environ Microbiol.* 1999;65:1652–7.
- Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, Barry K, Tringe SG, Watson DB, He Z, Hazen TC, Tiedje JM, Rubin EM, Zhou J. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J.* 2010;4:660–72.
- Im WT, Lee ST, Yokota A. *Rhodanobacter fulvus* sp. nov., a beta-galactosidase-producing gamma-proteobacterium. *J Gen Appl Microbiol.* 2004;50:143–7.
- Jones CM, Graf DR, Bru D, Philippot L, Hallin S. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *ISME J.* 2013;7:417–26.
- Kaeberlein T, Lewis K, Epstein SS. Isolating “uncultivable” microorganisms in pure culture in a simulated natural environment. *Science.* 2002;296:1127–9.
- Kostka JE, Green SJ. Microorganisms and processes linked to uranium reduction and immobilization. In: Stolz JF, Oremland RS, editors. *Microbial metal and metalloid metabolism: advances and applications.* Washington, DC: ASM Press; 2011.
- Kostka JE, Green SJ, Rishishwar L, Prakash O, Katz LS, Marino-Ramirez L, Jordan IK, Munk C, Ivanova N, Mikhailova N, Watson DB, Brown SD, Palumbo AV, Brooks SC. Genome sequences for six rhodanobacter strains, isolated from soils and the terrestrial subsurface, with variable denitrification capabilities. *J Bacteriol.* 2012;194:4461–2.
- Kowalsky MB, Gasperikova E, Finsterle S, Watson D, Baker G, Hubbard SS. Coupled modeling of hydrogeochemical and electrical resistivity data for exploring the impact of recharge on subsurface contamination. *Water Resour Res.* 2011;47.
- Lee CS, Kim KK, Aslam Z, Lee ST. *Rhodanobacter thiooxydans* sp. nov., isolated from a biofilm on sulfur particles used in an autotrophic denitrification process. *Int J Syst Evol Microbiol.* 2007;57:1775–9.
- Luo J, Cirpka OA, Wu WM, Fienen MN, Jardine PM, Mehlhorn TL, Watson DB, Criddle CS, Kitanidis PK. Mass-transfer limitations for nitrate removal in a uranium-contaminated aquifer. *Environ Sci Technol.* 2005;39:8453–9.
- NABIR. Bioremediation of metals and radionuclides. . . What it is and how it works. Berkeley: Lawrence Berkeley National Laboratory; 2003.
- Palumbo AV, Schryver JC, Fields MW, Bagwell CE, Zhou JZ, Yan T, Liu X, Brandt CC. Coupling of functional gene diversity and geochemical data from environmental samples. *Appl Environ Microbiol.* 2004;70:6525–34.
- Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol.* 2009;11:1358–75.
- Prakash O, Green SJ, Jasrotia P, Overholt WA, Canion A, Watson DB, Brooks SC, Kostka JE. *Rhodanobacter denitrificans* sp. nov., isolated from nitrate-rich zones of a contaminated aquifer. *Int J Syst Evol Microbiol.* 2012;62:2457–62.
- Sanford RA, Wagner DD, Wu QZ, Chee-Sanford JC, Thomas SH, Cruz-Garcia C, Rodriguez G, Massol-Deya A, Krishnani KK, Ritalahti KM, Nissen S, Konstantinidis KT, Löffler FE. Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proc Natl Acad Sci U S A.* 2012;109:19709–14.
- Shelobolina ES, O’Neill K, Finneran KT, Hayes LA, Lovley D. Potential for in situ bioremediation of a low-pH, high-nitrate uranium-contaminated groundwater. *Soil Sediment Contam.* 2003;12:865–84.
- Sousa T, Chung AP, Pereira A, Piedade AP, Morais PV. Aerobic uranium immobilization by *Rhodanobacter* A2–61 through formation of intracellular uranium-phosphate complexes. *Metallomics.* 2013;5(4):390–397.
- Spain AM, Krumholz L. Cooperation of three denitrifying bacteria in nitrate removal of acidic nitrate- and uranium-contaminated groundwater. *Geomicrobiol J.* 2012;29:830–42.
- Spalding BP, Watson DB. Passive sampling and analyses of common dissolved fixed gases in groundwater. *Environ Sci Technol.* 2008;42:3766–72.

- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM. Comparative metagenomics of microbial communities. *Science*. 2005;308:554–7.
- van den Heuvel RN, van der Biezen E, Jetten MS, Hefting MM, Kartal B. Denitrification at pH 4 by a soil-derived *Rhodanobacter*-dominated community. *Environ Microbiol*. 2010;12:3264–71.
- Verbaendert I, Boon N, De Vos P, Heylen K. Denitrification is a common feature among members of the genus *Bacillus*. *Syst Appl Microbiol*. 2011;34:385–91.
- Watson DB, Kostka JE, Fields MW, Jardine PM. The Oak Ridge field research center conceptual model. NABIR Field Research Center Report, Oak Ridge; 2004.
- Watson DB, Doll WE, Gamey TJ, Sheehan JR, Jardine PM. Plume and lithologic profiling with surface resistivity and seismic tomography. *Ground Water*. 2005;43:169–77.
- Woo SG, Srinivasan S, Kim MK, Lee M. *Rhodanobacter caeni* sp. nov., isolated from sludge from a sewage disposal plant. *Int J Syst Evol Microbiol*. 2012;62:2815–21.
- Wu WM, Carley J, Fienen M, Mehlhorn T, Lowe K, Nyman J, Luo J, Gentile ME, Rajan R, Wagner D, Hickey RF, Gu BH, Watson D, Cirpka OA, Kitanidis PK, Jardine PM, Criddle CS. Pilot-scale in situ bioremediation of uranium in a highly contaminated aquifer. 1. Conditioning of a treatment zone. *Environ Sci Technol*. 2006;40:3978–85.
- Wu WM, Carley J, Luo J, Ginder-Vogel MA, Cardenas E, Leigh MB, Hwang CC, Kelly SD, Ruan CM, Wu LY, Van Nostrand J, Gentry T, Lowe K, Mehlhorn T, Carroll S, Luo WS, Fields MW, Gu BH, Watson D, Kemner KM, Marsh T, Tiedje J, Zhou JZ, Fendorf S, Kitanidis PK, Jardine PM, Criddle CS. In situ bioreduction of uranium (VI) to submicromolar levels and reoxidation by dissolved oxygen. *Environ Sci Technol*. 2007;41:5716–23.
- Wu W-M, Carley J, Green SJ, Luo J, Kelly SD, Nostrand J, Lowe K, Mehlhorn T, Carroll S, Boonchayanant B, Löffler FE, Watson DB, Kemner KM, Zhou J, Kitanidis PK, Kostka JE, Jardine PM, Criddle CS. Effects of nitrate on the stability of uranium in a bioreduced region of the subsurface. *Environ Sci Technol*. 2010;44:5104–11.
- Yan TF, Fields MW, Wu LY, Zu YG, Tiedje JM, Zhou JZ. Molecular diversity and characterization of nitrite reductase gene fragments (*nirK* and *nirS*) from nitrate- and uranium-contaminated groundwater. *Environ Microbiol*. 2003;5:13–24.
- Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM, Keller M. Cultivating the uncultured. *Proc Natl Acad Sci U S A*. 2002;99:15681–6.
- Zumft WG, Kroneck PM. Respiratory transformation of nitrous oxide (N₂O) to dinitrogen by Bacteria and Archaea. *Adv Microb Physiol*. 2007;52:107–227.

Integrated Database Resource for Marine Ecological Genomics

Renzo Kottmann

Max Plank Institute for Marine Microbiology,
Bremen, Germany

Synonyms

Database; Environmental data; Environmental genomics; GIS; Integration; Marine; Metagenomics

Definition

Megx.net, the integrated database resource for marine ecological genomics, is the first database to integrate bacterial and archaeal genes, genomes, and metagenomes from the marine environment with curated contextual metadata, as well as environmental data from heterogeneous resources.

Introduction

Over the last years, microbial ecology and environmental microbiology have undergone a paradigm shift, moving from a single experiment science to a high-throughput endeavor. Although the genomic revolution is rooted in medicine and biotechnology, it is currently the environmental sector, specifically the marine, which delivers the greatest quantity of data (Gilbert and Dupont 2011). Marine ecosystems, covering >70 % of the Earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Microorganisms are known to be the “gatekeepers” of these processes, and insights into their lifestyle and fitness can enhance our ability to monitor, model, and predict future changes.

Recent developments in sequencing technology have made routine sequencing of whole

microbial communities from natural environments possible. Prominent examples in the marine field are the Global Ocean Sampling (GOS) campaign (Rusch et al. 2007), ICOMM, TaraOceans, Malaspina, and the Ocean Sampling Day 2014 of the Micro B3 project.

These large-scale sequencing projects bring new challenges to data management and software tools for assembly, gene prediction, and annotation, which are fundamental steps in genomic analysis. Several dedicated database resources have emerged to tackle the current need for large-scale metagenomic data management and analysis, among which are CAMERA (Sun et al. 2010), IMG/M (Markowitz et al. 2008), and MG-RAST (Meyer et al. 2008). Nevertheless, it is increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the geographic and environmental context of the sequence data is considered. The metadata describing a sample's geographic location and environment, the details of its processing, from the time of sampling to sequencing and subsequent analyses are important for modeling species' responses to environmental change or the spread and niche adaptation of bacteria and viruses. Megx.net's unique integration of contextual and sequence data allows microbial ecologists and marine scientists to better compare biological data to understand the complex interplay between organisms, genes, and their environment.

Database Structure and Content

The Microbial Ecological Genomics Database (MegDB), the backbone of megx.net, is a centralized database based on the PostgreSQL database management system. The georeferenced data concerning geographic coordinates and time are managed with the PostGIS extension to PostgreSQL.

Sequences in MegDB are retrieved from the International Nucleotide Sequence Database Collaboration (INSDC). Currently, MegDB contains 1,832 prokaryote genomes (940 incomplete or draft) and 80 marine shotgun metagenomes

from the GOS microbial dataset. Finally, megx.net also incorporates all sequenced marine phage genomes in MegDB, which is the first step towards integrating viral genomic and biogeochemical data (Duhaime et al. 2011).

In an effort towards integrating microbial diversity with specific sampling sites, megx.net includes georeferenced small and large subunit rRNA gene sequences from the SILVA rRNA gene databases project (Quast et al. 2013). As of SILVA release 102, only 9 % (16S/18S) and 2 % (23S/28S) of over one million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced.

All genomic sequences in megx.net are supplemented with contextual data from GOLD (Pagani et al. 2012), NCBI Genome Projects, and Moore Foundation's Marine Microbial Genome Sequencing Project.

The main environmental data is retrieved from three sources:

1. World Ocean Atlas: a set of objectively analyzed (one decimal degree spatial resolution) climatological fields of in situ measurements
2. World Ocean Database: a collection of scientific, quality-controlled ocean profiles
3. SeaWiFS chlorophyll a data

These data are described at 33 standard depths for annual, seasonal, and monthly intervals. Together, the location and time data (x, y, z, and t) serve as a universal anchor and link environmental data to the sequence and contextual data.

Standards Compliance and Interoperability

Standards are an important means of enhancing data exchange and interoperability between different database resources. MegDB is designed to store all contextual data recommended by the Genomics Standards Consortium and is thus compliant with the Minimum Information about any (x) Sequence (MIxS) standard (Yilmaz et al. 2012). However, most sequence data is missing contextual metadata. Therefore, numerous bacterial and archaeal genomes were manually curated to assign geographic coordinates to

reveal their environmental origin. Even with careful curation, a geographic origin could not be assigned to the majority of genomes. In order to give at least an indication of the environmental origin of sequence data, they were manually curated with terms of the Habitat-Lite subset of the Environmental Ontology (Hirschman et al. 2008).

Functionalities

Genes Mapserver

The Genes Mapserver gives a sample-centric view of the georeferenced MegDB content. The map is interactive, offering user-friendly navigation and an overlay of the MegDB environmental data layers to display sampling sites on a world map in their environmental context. Sample site details and interpolated data can be retrieved by clicking the sampling points on the map.

The GIS Tools of the Genes Mapserver allow extraction of interpolated values for several physicochemical and biological parameters, such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified monthly, seasonally, or annually intervals.

Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages, and rRNA gene sequence data using the BLAST algorithm (Altschul et al. 1990). The Geographic-BLAST tool permits the alignment of query sequences against five databases instead of the standard BLAST query database:

- Prokaryotic genomes
- Global Ocean Sampling Metagenomes, which are publicly available metagenomes from the Global Ocean Sampling expedition
- 16S/18S rRNA
- 23S/28S rRNA
- Marine phage genomes

The results are reported according to the sample locations (if available) of the database hits and plotted on the Genes Mapserver world map, where they are labeled by the number of hits per site. Standard BLAST results are shown in

a table, which also provides direct access to the associated contextual data of the hits (Fig. 1).

GIS Tools

The GIS tools allow post-factum retrieval of interpolated environmental parameters, such as temperature, nitrate, or phosphate for any location in the ocean waters based on profile and remote sensing data.

Two GIS tools are currently available:

- World Ocean Atlas Extractor, comprised of analyzed climatological fields of physicochemical parameters and biological layers obtained at monthly, seasonal, and annual samplings
- World Ocean Database Extractor, comprised of time series measurements of physicochemical parameters and biological layers

Both GIS tools make use of Inverse Distance Weighted (IDW) interpolation to estimate the environmental data at a given geographic location, time, and depth in the ocean.

MetaBar

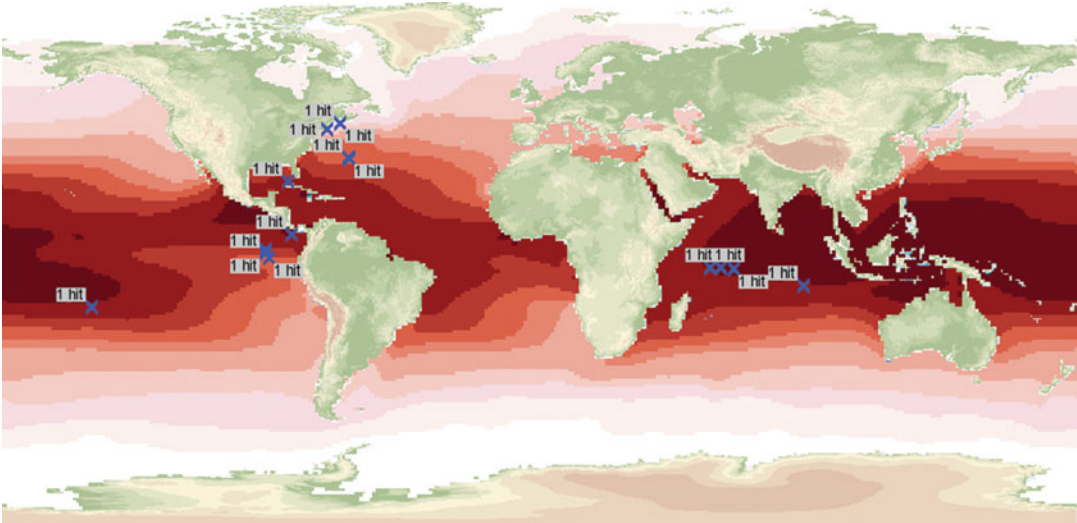
MetaBar aims to support investigators to efficiently capture, store, and submit contextual metadata gathered in the field. It is a spreadsheet-based sample data collection tool designed to support the complete workflow from the sampling event up to the metadata-enriched sequence submission to an INSDC database (Hankeln et al. 2010).

CDinFusion

Megx.net hosts a public installation of CDinFusion, a Web-based tool to combine MIxS compliant contextual and sequence data in (Multi)FASTA formatted files prior to submission (Hankeln 2011). It creates submission ready files for the NCBI submission system. However, CDinFusion is not (yet) appropriate for preparing data for the Sequence Read Archive (SRA) submission system.

Web Services

Megx.net offers programmatic access via Web services for experienced users and software developers. All geographical maps can be



Integrated Database Resource for Marine Ecological Genomics, Fig. 1 Geographic distribution of BLAST results of a proteorhodopsin from *Dokdonia sp. PRO95*.

Blue crosses and label indicating the number of significant BLAST hits in the GOS metagenome samples. The map is generated using the web service of the Genes Maps server

retrieved via simple Web requests, as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is <http://www.megx.net/wms/gms>, where one can also find a tutorial on how to use this service. Megx.net also provides access to MIXS reports in Genomic Contextual Data Markup Language (GCDML) XML files for all marine phage genomes through similar HTTP queries, e.g., http://www.megx.net/gcdml/Prochlorococcus_phage_P-SSP7.xml (Kottmann et al. 2008).

Current and Future Developments

Currently, megx.net is further developed within the FP7 EU project Micro B3 as an open source project to become an integral part of the Micro B3 Information System. This information system builds on a handful of long-established data resources that span marine science. These data resources include SeaDataNet and its network of National Oceanographic Data Centers (oceanographic data), EurOBIS (macrobiological data), and EBI's European Nucleotide Archive (EBI-ENA; molecular sequence data). While these resources exist to broaden and simplify

access to data in their domains, integration of their data across domains requires megx.net to develop a set of new tools and Web services to facilitate seamless interoperability between the different data domains.

Summary

Megx.net's unique integration of environmental and sequence data allows microbial ecologists and marine scientists to better contextualize and compare biological data, using, e.g., the Genes Maps server and GIS tools. The integrated datasets facilitate a holistic approach to understanding the complex interplay between organisms, genes, and their environment. As such, megx.net is continuously improved to serve as a fundamental resource in the emerging field of ecosystems biology.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [SILVA Databases](#)

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- Duhaime MB, Kottmann R, Field D, Glöckner FO. Enriching public descriptions of marine phages using the MIGS standard: a case study assessing the contextual data frontier. *Stand Genomic Sci.* 2011;4(2):1.
- Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci.* 2011;3(1):347–71. *Annual Reviews.*
- Hankeln W, Buttigieg PL, Fink D, Kottmann R, Yilmaz P, Glöckner FO. MetaBar – a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinforma.* 2010;11:358.
- Hankeln W, Wendel NJ, Gerken J, Waldmann J, Buttigieg PL, Kostadinov I, et al. CDinFusion – submission-ready, on-line integration of sequence and contextual data. *PLoS ONE.* 2011;6(9):e24797. Highlander SK, editor.
- Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, et al. Habitat-lite: a GSC case study based on free text terms for environmental metadata. *OMICS J Integr Biol.* 2008;12(2):129–36.
- Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T, et al. A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS J Integr Biol.* 2008;12(2): 115–21.
- Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008. PMID:17932063
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9(1):386.
- Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 2012;40(Database issue):D571–9.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41(Database issue): D590–6.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 2007;5(3):e77.
- Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* 2010; 39(Database):D546–51.
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol.* 2012;29(5):415–20. doi: 10.1038/nbt.1823.

Integrans as Repositories of Genetic Novelty

Bridget Mabbutt¹, Chandrika Deshpande¹,
Visaahini Sureshan¹ and Stephen J. Harrop²

¹Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia

²School of Physics, University of New South Wales, Sydney, NSW, Australia

Synonym

Novel proteins engaged for LGT within the gene cassette/integron system

Definition

An important vehicle for lateral (or horizontal) gene transfer in bacteria is the integron: it enables the capture and expression of genes as small mobile elements, or gene cassettes. These mobile gene cassettes encompass a vast pool of genetic novelty, ostensibly for purposes of adaptation. In most cases, their functional annotation is obscured by their characteristically high sequence novelty. Our isolation and solving of protein structures encoded by the cassette metagenome reveals a relatively high proportion of completely novel folds. These newly defined crystal structures are found to encompass diverse topologies and fold families and delineate new protein domains.

Introduction

Bacteria dominate the planet; they are omnipresent, inhabiting a wide range of environments, including those appearing too extreme or inhospitable for life (Rothschild and Mancinelli 2001). Lateral gene transfer (LGT) is known to contribute to the enormous genetic diversity of this microbial world. Rendering the bacterial genome in a constant state of flux, LGT can be said to produce a gene pool that is collectively owned, leading to the concept of a mobile prokaryotic metagenome (Koonin and Wolf 2008).

One important mediator of LGT involves the integron system (Boucher et al. 2007; Cambray et al. 2010; Hall 2012), which allows bacteria to capture and express genes occurring in the environment as small mobile elements, named gene cassettes. Although originally identified as the vehicle for the spread of antibiotic resistance, it is now clear that the integron/gene cassette system is not just limited to the clinical context, but plays a wider role in shaping niche advantage (Labbate et al. 2012). While most integrons contain a small number of gene cassettes (generally up to ~10), in some instances multiple insertion events assemble large cassette arrays, particularly notable within chromosomes of *Vibrio* species (Boucher et al. 2006; Joss et al. 2009).

It is immediately obvious that the cassette metagenome comprises a repertoire of distinctly novel genes, with sequence homologs (if any) sparsely represented or not annotated in current databases. This is true for both isolated gene cassettes and gene cassette arrays derived from cultivated bacterial strains (Rowe-Magnus et al. 2003; Boucher et al. 2006), as well as for wider metagenomic surveys (Elsaied et al. 2007; Koenig et al. 2008).

With the cassette metagenome extending beyond the coverage of conventional sequencing, protein structure provides a first functional inference for many gene cassettes through determination of three-dimensional fold homology relationships (Sureshan et al. 2013). This approach has resulted in the structural definition of many new proteins, although a large subset includes entirely novel folds. It is now

established that the gene cassette metagenome encodes fully folded and functional proteins and includes new enzymes and protein-binding factors (Robinson et al. 2005; Robinson et al. 2008). This newly expanding group of protein folds and structures reveals the extraordinary genetic novelty encoded by the cassette metagenome.

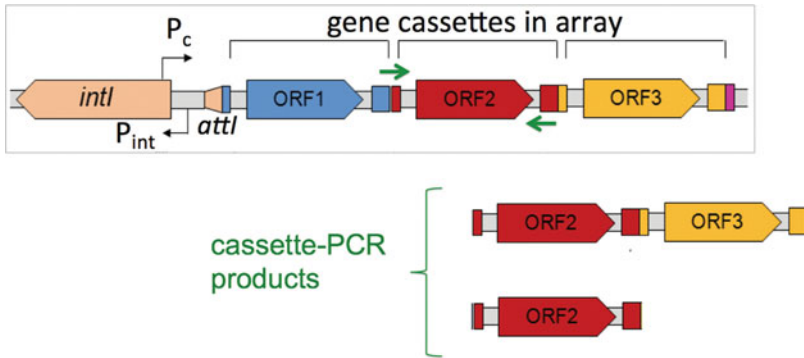
This entry focuses on cassette-encoded proteins directly recovered by the technique of cassette PCR (outlined in Fig. 1) (Stokes et al. 2001; Boucher et al. 2007). The method has been exploited for uncultured bacteria present within environmental samples, as well as for strain isolates of *Vibrio cholerae* and the related *V. metecus* (formerly *V. paracholerae*).

Novel (Currently Unique) Gene Cassette Structures

Examination of protein structures encoded by the cassette metagenome reveals a relatively high proportion to display a completely novel fold (Sureshan et al. 2013). These newly defined three-dimensional structures encompass diverse topologies and fold families and impact beyond specific gene cassettes to delineate new protein domains and their sequence homologs. Although it is not possible to yet identify specific substrates or biochemical properties for these first members of new families, their molecular features and organizations (see Fig. 2) contribute currency to the ongoing discussion assessing the degree to which function and/or protein network capacity favors mobilization of genes (Cohen et al. 2011; Labbate et al. 2012).

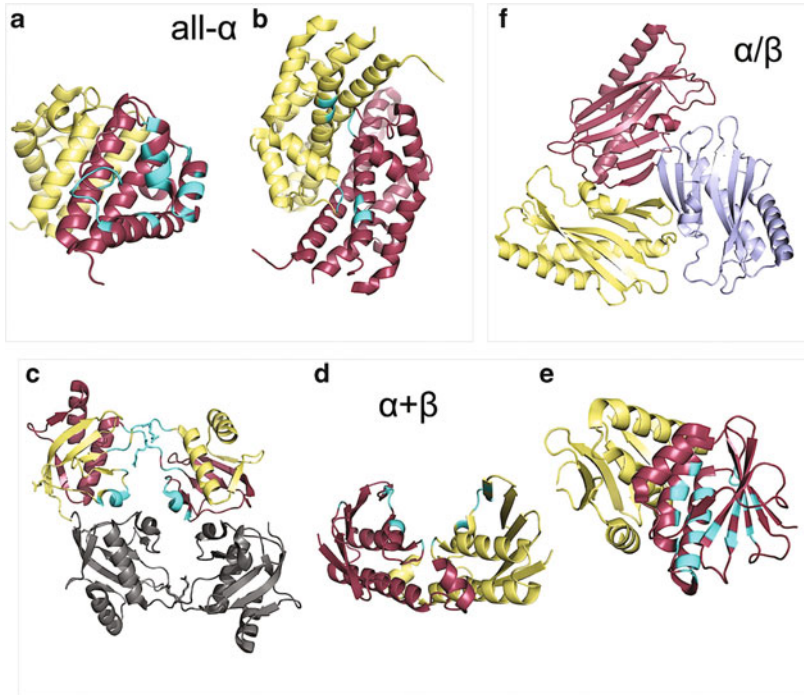
All- α Fold Members

The crystal structure determined for a gene cassette isolated from a sewage outfall (**Hfx_cass2**, PDB 3FXH) depicts a dimeric protein incorporating a compact fold of six helical segments. The homodimer is stabilized by a hydrophobic interface engaging two helices from each chain (Fig. 2a). Exposed on the external face of each subunit is a triangular-shaped hydrophobic crevice flanked by two acidic residues and



Integrans as Repositories of Genetic Novelty, Fig. 1 Recovery of gene cassettes from integrin arrays. The structure of an integrin, showing core features including the *intI* gene (*beige*) with its *P_{int}* promoter, the *attI* attachment site and the *P_c* promoter. Three integrated gene cassettes (*blue, red, and yellow*) are shown. By

using primers (*green arrows*) targeting the 59-be elements, cassette PCR has the capacity to recover gene cassettes and arrays independently of any specific encoded sequence. This allows recovery of entirely novel gene cassettes (Adapted from Boucher et al. 2007 and Stokes et al. 2001)



Integrans as Repositories of Genetic Novelty, Fig. 2 Ribbon depiction of novel cassette-encoded protein structures: (a) *Hfx_cass2*, (b) *Vpc_cass2*, (c) *Hfx_cass5*, (d) *Vch_cass3*, (e) *Vch_cass14*, (f)

Hfx_cass1. Each subunit within the oligomeric organization is indicated in a different color. Putative binding sites, for interaction with either small molecule ligands or, potentially, other protein partners, are highlighted in *cyan*

a flexible loop. Pronounced acidic surface features extend perpendicular to each cavity due to Glu and Asp side chains of an outer helix. This unique binding groove, presented twice on

opposing faces of the dimer and possibly gated by residues of the flexible loop, appears highly appropriate for hydrophobic and/or basic substrates or protein partners.

A distinct all-helical protein had also been identified in a gene cassette recovered from a *V. metecus* strain, **Vpc_cass2** (PDB 3JRT). The fold incorporates a four-helix bundle with helical extensions wrapping about at midpoint (Fig. 2b); orthogonal packing of two chain pairs creates a globular-shaped dimer. Sequence homologs (*Shewanella baltica* and *Moritella* genomes, at ~50 % identity) highlight preservation of exposed residues (Lys63, Glu66, His109', Val110') clustered across the dimeric interface, indicating a possible substrate-binding site. This fold is weakly related to the substrate-binding domain of the kanamycin nucleotidyltransferase (KNTase-C) clan of proteins, yet the shape of the dimeric interface in **Vpc_cass2** is distinct to that found in its closest KNTase-C relatives (e.g., HI0074 from *Haemophilus influenzae*). Lehmann and workers have documented substrate-binding/nucleotide-binding module pairs prevalent in bacterial genomes, particularly from harsh conditions and pathogens (Lehmann et al. 2003). Thus, the mobile gene cassette **Vpc_cass2** may comprise one half of a bipartite system with the capacity to organize with a nucleotidyltransferase domain into a functional enzyme.

$\alpha + \beta$ Fold Members

A gene cassette also isolated from a sewage outfall (Halifax, Canada), **Hfx_cass5**, occurs as two domain-swapped $\alpha + \beta$ dimers organized into a tetramer (PDB 3IF4; Fig. 2c). Across the center of the tetramer, 3_{10} helices of two opposing subunits stack via polar and charged groups. The flattened nature of the tetramer and the asymmetrical interactions of its component dimers result in two large faces with markedly different surface features. A small group of sequence homologs (55–71 % identity) include gene cassettes from contaminated environments: a geographically distinct sewage outfall in Canada and an Australian industrial site (Stokes et al. 2001). Residues mediating the tetrameric organization are preserved across all members of this emerging sequence family, indicating this to be the functional form. Also conserved is the inter-module linker segment, which presents

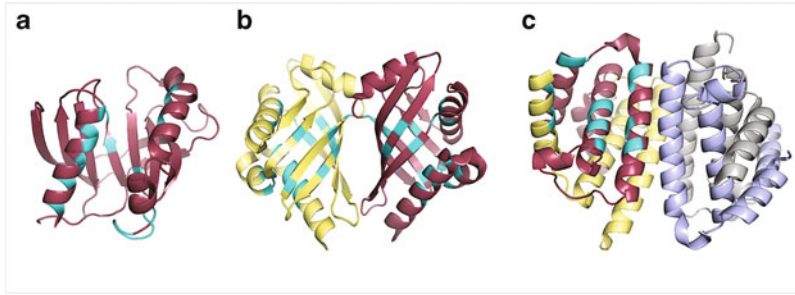
basic groups that line the pronounced surface clefts on both faces of the tetramer.

Derived from a strain of *V. cholera*, the structure of **Vch_cass3** (PDB 3FY6) reveals an unusual two-layered $\alpha + \beta$ organization. Within the dimer, central helices stack end-to-end, so separating and exposing two distinct sheet components (Fig. 2d). A long pronounced surface cleft is enclosed between the outer edge strands of these two sheets, flanked by acidic side chains. To date, two sequence homologs (~40 % identity) have been detected: within *Desulfatibacillus alkenivorans* from polluted water and a metagenomic sample of Antarctic bloom-forming cyanobacterium. These sequence relatives do not, however, retain the distinctive Asp/Glu residues surrounding the proposed binding cleft within the **Vch_cass3** structure.

Another *V. cholera*-derived gene cassette, **Vch_cass14**, also incorporates an $\alpha + \beta$ dimer, in this case within a two-layer sandwich fold (PDB 3IMO, Fig. 2e). Sequence relatives of this gene cassette have been found in the genomes of several soil- and water-dwelling bacteria. A particularly long and deep ligand cavity is internalized within this protein, appropriate for a linear hydrophobic substrate (e.g., fatty acid or alcohol). The features of this binding cavity are retained across all sequence relatives; 20 of the **Vch_cass14** internal residues are conserved in its two closest homologs. A high degree of conservation is also seen among residues responsible for mediating dimerization of the module, pointing to a dimeric functional protein. A notable feature of the dimer, possibly of functional importance, is the projection of positively charged surface clusters from the two exposed β -sheets.

α/β Fold Members

An unusual trimeric protein is encoded by **Hfx_cass1**, a gene cassette extracted from a salt marsh environment (Koenig et al. 2008). Although there are no sequence homologs in current databases, the unique three-layered α/β fold bears some topological relationship to the zinc transporter CziB of *Thermus thermophilus*. This new cassette-encoded protein presents three clefts at each inter-subunit interface across the



Integrans as Repositories of Genetic Novelty, Fig. 3 Ribbon depiction of cassette-encoded new variants of known folds: (a) **Cass2**, (b) **Bal32a**, (c) **iMazG**. Each subunit within the oligomeric organization is

indicated in a different color. Putative binding sites, for interaction with either small molecule ligands or, potentially, other protein partners, are highlighted in *cyan*

flattened trimer surface (Fig. 2f). The clefts are polar in nature, occupied in the crystal structure by water, and surrounded by pronounced acidic loops. Although the chemical organization of the binding site is unique to **Hfx_cass1**, some components are common to active site chemistry of enzymes known to engage with adenosine- and/or nicotinamide-based cofactors.

New Variants of Known Folds Encoded by Gene Cassettes

Cationic Drug-Binding Module

The structure (PDB 3GK6) of gene cassette **Cass2** derived from environmental *V. cholera* has identified an independent binding module related to domains of the AraC/XylS transcription activator system (Deshpande et al. 2011). Sequence analysis identifies the cassette-encoded protein to be representative of a group of independent binding modules undergoing lateral gene transfer within *Vibrio* and related species. Closest structural relatives of the **Cass2** β -barrel (Fig. 3a) occur as domains of multidrug-binding proteins (including BmrR), incorporating a hydrophobic binding pocket with a signature glutamate side chain. **Cass2** has been demonstrated to bind a range of cationic drug compounds. The structure of this module depicts a surface proximal to the drug-binding cavity with features homologous to those engaged for protein interaction within multidomain transcriptional regulators.

Thus, it can be proposed that the **Cass2** family has the capacity to form functional transcription regulator complexes and possibly represents evolutionary precursors to multidomain regulators of cationic compounds.

$\alpha + \beta$ Barrel Transporter

A gene cassette derived from industrially polluted soil has yielded a new member of the highly adaptable $\alpha + \beta$ barrel family of transport proteins and enzymes (Fig. 3b). The dimeric structure of **Bal32a** (PDB 1TUH (Robinson et al. 2005)) features cone-shaped binding pockets within each barrel, common to this superfamily for engaging small hydrophobic substrates or peptides. The **Bal32a** structure is, however, unique in that each of its central cavities is unusually deep and isolated from solvent by a flexible loop. A potential catalytic site of clustered polar groups within the barrel is equivalently positioned to corresponding active sites within structurally related enzymes. Although these enzymes likely share a common evolutionary ancestry, with preservation of active site features internal to the barrel, their very low overall sequence relationship to **Bal32a** (<20% identity) suggests a wide adaptation of the $\alpha + \beta$ barrel fold for varied demands. Within its originating cassette array, the **Bal32a** gene cassette was immediately adjacent to a second cassette, **Bal32b**, encoding a likely membrane-associated protein. This suggests the two components may well possibly function in concert as a combined binding and transport system.

MazG Enzyme Subfamily

As part of an ongoing investigation of an intact integron array of 116 gene cassettes located in *Vibrio rotiferianus* DAT722 (Boucher and Stokes 2006; Chowdhury et al. 2011), a new type of MazG nucleoside triphosphate pyrophosphohydrolase (NTP-PPase) has been described (Robinson et al. 2007). This cassette-encoded protein, **iMazG**, has close sequence relatives (some within gene cassettes) only within *Vibrio* sp. and other aquatic γ -proteobacteria. The structure of the **iMazG** tetramer (PDB 2Q5Z) (Fig. 3c) shows the typical α -helical hairpin fold of the general enzyme family in “closed” and “open” states, as well as its essential Mg^{2+} -coordination site. However, this new class of MazG enzymes contains significant variation, with unique loop and β -turn features connecting the four helices of the scaffold, creating a distinct substrate site adjacent to the divalent metal. Functional assays demonstrated that this single-domain type of MazG cleaves phosphates of dNTP substrates, with a preference for dCTP and dATP. Thus, **iMazG** has the capacity to act as a house-cleaning enzyme capable of removing noncanonical dNTPs.

Gene Cassettes Encode Novel Protein Folds with Distinct Binding Features

Regardless of the degree of novelty displayed, all gene cassette-derived structures appear to be consistent with adaptive functions (e.g., secondary metabolism, DNA modification) and possibly selective advantage (e.g., drug resistance). A tendency to form homo-oligomers has been a consistent observation across this structural survey of cassette proteins, with only one exception to date (the cationic drug-binding protein **Cass2** from *Vibrio* (Deshpande et al. 2011)). This clear preference for oligomerization may be a consequence of the relatively short sequence lengths of genes cassettes within arrays, stabilizing small protein modules which can perhaps also be readily and flexibly mixed for different functions. Such modules may readily combine with

appropriate catalytic, binding, or membrane domains as adaptive pressure selects more specific biochemical or regulatory networks (Bornberg-Bauer and Alba 2013). Certainly, the surface features described for each of the cassette protein structures have potential to act as heterogeneous protein interfaces within multidomain or multi-protein systems.

Summary

Our structural studies continue to enforce the notion that the highly novel gene cassette metagenome is not merely a repository of sequence divergent variants of known proteins, but in fact mobilizes a repertoire of genes belonging to poorly characterized protein families. Thus, to fully scope and understand the global proteome, it remains essential to continue to independently target structural investigation of the metagenomic element.

Cross-References

- ▶ [Lateral Gene Transfer and Microbial Diversity](#)
- ▶ [Metagenomic Potential for Understanding Horizontal Gene Transfer](#)

References

- Bornberg-Bauer E, Alba MM. Dynamics and adaptive benefits of modular protein evolution. *Curr Opin Struct Biol.* 2013;23(3):459–66.
- Boucher Y, Stokes HW. The roles of lateral gene transfer and vertical descent in vibrio evolution. In: Fabiano Lopes Thompson BA, Swings JG, editors. *The biology of vibrios*. Washington, DC: ASM Press; 2006. p. 84–94.
- Boucher Y, Nesbo CL, Joss MJ, Robinson A, Mabbutt BC, Gillings MR, et al. Recovery and evolutionary analysis of complete integron gene cassette arrays from *Vibrio*. *BMC Evol Biol.* 2006;6:3.
- Boucher Y, Labbate M, Koenig JE, Stokes HW. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.* 2007;15(7):301–9.

- Cambray G, Guerout A, Mazel D. Integrons. *Annu Rev Genet.* 2010;44:141–66.
- Cohen O, Gophna U, Pupko T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol.* 2011;28(4):1481–9.
- Deshpande CN, Harrop SJ, Boucher Y, Hassan KA, Di Leo R, Xu X, et al. Crystal structure of an integron gene cassette-associated protein from *Vibrio cholerae* identifies a cationic drug-binding module. *PLoS One.* 2011;6(3):e16934.
- Elsaied H, Stokes HW, Nakamura T, Kitamura K, Fuse H, Maruyama A. Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environ Microbiol.* 2007;9(9):2298–312.
- Hall RM. Integrons and gene cassettes: hotspots of diversity in bacterial genomes. *Ann N Y Acad Sci.* 2012;1267:71–8.
- Joss MJ, Koenig JE, Labbate M, Polz MF, Gillings MR, Stokes HW, et al. ACID: annotation of cassette and integron data. *BMC Bioinformatics.* 2009;10:118.
- Koenig JE, Boucher Y, Charlebois RL, Nesbo C, Zhaxybayeva O, Bapteste E, et al. Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol.* 2008;10(4):1024–38.
- Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 2008;36(21):6688–719.
- Labbate M, Boucher Y, Luu I, Chowdhury PR, Stokes HW. Integron associated mobile genes: Just a collection of plug in apps or essential components of cell network hardware? *Mob Genet Elements.* 2012;2(1):13–8.
- Lehmann C, Lim K, Chalamasetty VR, Krajewski W, Melamud E, Galkin A, et al. The HI0073/HI0074 protein pair from *Haemophilus influenzae* is a member of a new nucleotidyltransferase family: structure, sequence analyses, and solution studies. *Proteins.* 2003;50(2):249–60.
- Robinson A, Wu PS, Harrop SJ, Schaeffer PM, Dosztanyi Z, Gillings MR, et al. Integron-associated mobile gene cassettes code for folded proteins: the structure of Bal32a, a new member of the adaptable alpha + beta barrel family. *J Mol Biol.* 2005;346(5):1229–41.
- Robinson A, Guilfoyle AP, Harrop SJ, Boucher Y, Stokes HW, Curmi PM, et al. A putative house-cleaning enzyme encoded within an integron array: 1.8 Å crystal structure defines a new MazG subtype. *Mol Microbiol.* 2007;66(3):610–21.
- Robinson A, Guilfoyle AP, Sureshan V, Howell M, Harrop SJ, Boucher Y, et al. Structural genomics of the bacterial mobile metagenome: an overview. *Methods Mol Biol.* 2008;426:589–95.
- Rothschild LJ, Mancinelli RL. Life in extreme environments. *Nature.* 2001;409(6823):1092–101.
- Rowe-Magnus DA, Guerout AM, Biskri L, Bouige P, Mazel D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res.* 2003;13(3):428–42.
- Roy Chowdhury P, Boucher Y, Hassan KA, Paulsen IT, Stokes HW, Labbate M. Genome sequence of *Vibrio rotiferianus* strain DAT722. *J Bacteriol.* 2011;193(13):3381–2.
- Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KM, Mabbutt BC, et al. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol.* 2001;67(11):5240–6.
- Sureshan V, Deshpande CN, Boucher Y, Koenig JE, Stokes HW, Harrop SJ, et al. Integron gene cassettes: a repository of novel protein folds with distinct interaction sites. *PLoS One.* 2013;8(1):e52934.

IPRStats, Overview

Iddo Friedberg

Department of Microbiology, Miami University,
Oxford, OH, USA

Abbreviations

EBI	European Bioinformatics Institute
GO	Gene Ontology
IMG/M	Integrated Microbial Genome/ Metagenomics
pHMM	Profile hidden Markov model
PSSM	Position-specific scoring matrix
SQL	Structured Query Language
XML	Extensible Markup Language

Definition

IPRStats is a lightweight platform-independent open-source licensed software package for storing and visualizing metagenomic data annotated by InterProScan. IPRStats is unique in that it provides the user with the same annotation choices offered by the popular open reading frame annotation pipeline, InterProScan. IPRStats can be installed either as a Web server or as a stand-alone software.

Introduction

The functional annotation of open reading frames (ORFs) in metagenomic data is a highly challenging problem. The problem is difficult enough with regular genomic data. When functionally annotating metagenomic data, one is confronted with the additional problems arising from sequence fragmentation, imperfect assemblies, unmitigated sequencing errors, partially identified ORFs, and higher rates of error in ORF calling. One way to overcome these problems is to do away with ORF calling altogether. Instead, assembled metagenomic sequences are translated in six open reading frames. Those that produce proteins above a certain minimal length threshold (say, 100aa) are subjected to functional analysis. The rationale behind such a strategy is that there is a very low probability that a sequence which is (1) long enough and (2) found in a database of protein signatures is not a true ORF or a partial ORF. Each such sequence is then treated as a member in a population, with biological function attributes assigned to it. The storage, visualization, and analysis of metagenomic data can be handled using common database, statistical and visualization tools used in population analyses. Here such a package, IPRStats, which is based on the popular InterProScan tool, is described.

Annotation of Translated Sequences

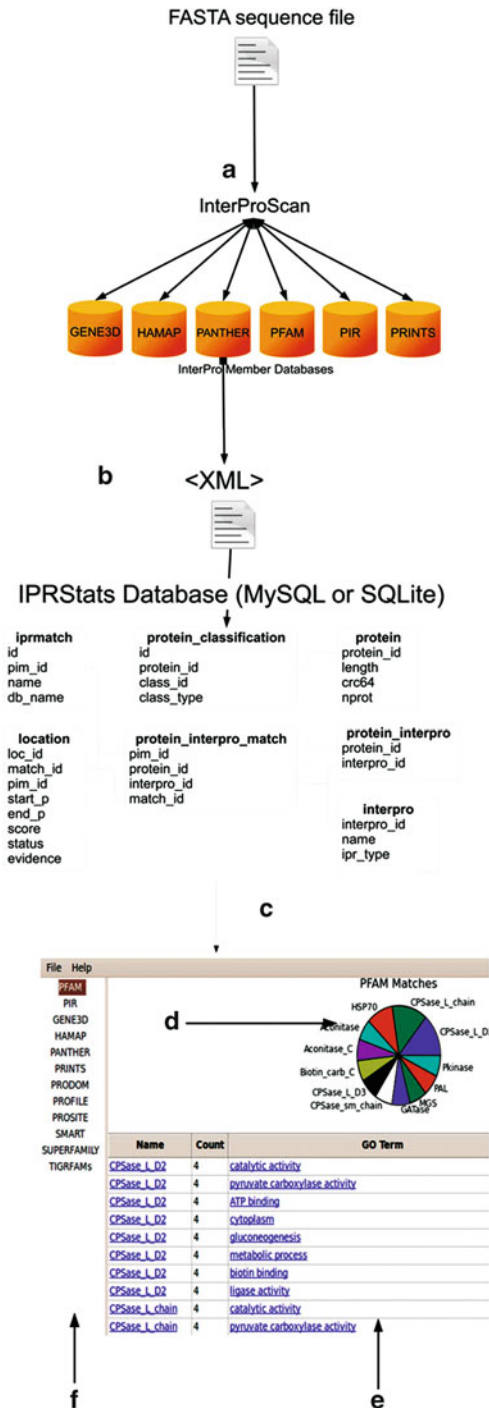
Homology-based transfer algorithms require, first and foremost, a comprehensive, accurately annotated, and up-to-date reference sequence database, but no single database can boast all three traits at 100 % (Schnoes et al. 2009). This is true for pairwise sequence alignment algorithms and simple sequence-motif algorithms as well as for the more complex profile hidden Markov models (pHMM) (Eddy 1998) and position-specific score matrix (PSSM) similarity-based algorithms. Therefore, several function annotation programs are typically used to functionally annotate ORFs. The rationale is that by using more than one algorithm to functionally annotate

a protein, the lack of sensitivity that may result from using only one program can be overcome. Additionally, a consensus method can help weed out false positives, by picking only those annotations on which there is a plurality agreement, or some other voting mechanism. InterProScan (Zdobnov 2001) is a function annotation program that compares query protein sequences against a repository of collected and annotated protein signatures. These InterPro (McDowall and Hunter 2011) *member databases* employ a variety of motif, pHMMs, and position-specific score matrices (PSSMs) to describe protein families. Those include PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR superfamily, SUPERFAMILY, Gene3D, PANTHER, and HAMAP. These also include the associated software used to query these databases: pfscan, FingerPRINTScan, HMMer3.0, HMMER 2.3, and BLAST. More information on current member databases and search software employed in InterPro, including updated references, can be found at <ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/README.html>

Visualization and Management of Metagenomic Function Annotations from InterProScan Using IPRStats

InterProScan can be installed on computer clusters and therefore can handle large amounts of sequence data. However, when analyzing large amounts of sequence data, as in metagenomic data, there are two needs which InterProScan does not provide: first, a visualization of the results to make them comprehensible and, second, a simple data storage and retrieval mechanism for further analysis.

To implement both goals, each translated sequence is treated as a member in a population, which is assigned one or more functional attributes by the member programs of InterProScan. IPRStats (Kelly 2010), or InterProScan STATistics, uses the output of InterProScan as its input and quickly produces charts and tables enabling a visualization of the functional potential of the



IPRStats, Overview, Fig. 1 Overview of IPRStats. (a) Protein sequence information as a single FASTA file submitted to InterProScan (one or more proteins). (b) InterProScan XML output imported into IPRStats SQL database. (c) Display of sequence signature statistics.

sequences analyzed. It also stores the results in a simple SQL schema (Fig. 1b), which can be used by other applications for downstream data analysis and presentation.

Figure 1 describes the information flow in IPRStats. The output of an InterProScan run is stored in XML format. The XML file is parsed into a 7-table SQLite or a MySQL database. The tables follow the data structure outlined by the InterProScan XML schema. After reading the tables, IPRStats displays the information alphabetically and graphically. The tabs in the sidebar of the main program screen toggle between the displays of results for each sequence signature program called by InterProScan. The results display includes a chart (Fig. 1d) and a table (Fig. 1e). The chart is either a pie chart or a bar chart, which shows the count of different sequence signatures from the relevant program in the analyzed sequence population. Chart drawing is implemented using either Google Chart Tools or matplotlib. Google Chart Tools is a web-based API that dynamically generates charts using a URL string, so when drawing using Google Chart Tools, an active Internet connection is required. Alternatively, matplotlib may be used: matplotlib is a Python-based clone of MatLab, which can be used for chart graphics as well, and does not require an Internet connection.

Availability

IPRStats is written in Python, with a graphic user interface (GUI) based on wxWidgets, a cross-platform toolkit for graphic user interfaces. Relying on platform-independent fully open-source infrastructure ensures that we maximize portability of IPRStats. Currently IPRStats has been tested on Windows XP/7, Max OS x 10.6, and Ubuntu GNU/Linux 9.10 and 10.04. IPRStats is

IPRStats, Overview, Fig. 1 (continued) (d) Graphic display. (e) Table display. (f) Toggle between results from different InterPro member databases (Reproduced from seven under BMC CC 2.0 license, copyright owned by authors)

downloadable from GitHub at <http://github.com/idoerg/IPRStats>. Packages for Windows, Mac OSX, and Linux are available at <http://github.com/idoerg/IPRStats/downloads>. Community participation and further development of this tool are strongly encouraged.

References

- Eddy S. Profile hidden Markov models. *Bioinformatics*. 1998;14(9):755–63.
- Kelly RJ, Vincent DE, Friedberg I. IPRStats: visualization of the functional potential of an InterProScan run. *BMC Bioinformatics*. 2010;11:S13.
- McDowall J, Hunter S. InterPro protein classification. *Methods Mol Biol*. 2011;694:37–47. doi: 10.1007/978-1-60761-977-2_3. PMID:21082426.
- Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*. 2009;5(12):e1000605. doi: 10.1371/journal.pcbi.1000605. Epub 2009 Dec 11. PMID:20011109.
- Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001;17(9):847–8.

I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments

Algorithms for Efficient In Silico Identification and Classification of Ribosomal RNA Gene Fragments in Metagenomic Datasets

Sharmila Mande, Tarini Shankar Ghosh and Mohammed Monzoorul Haque
Biosciences R & D, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Limited, Pune, MH, India

Synonyms

Classification of 16S rRNA gene fragments; In silico identification of 16S rRNA gene fragments

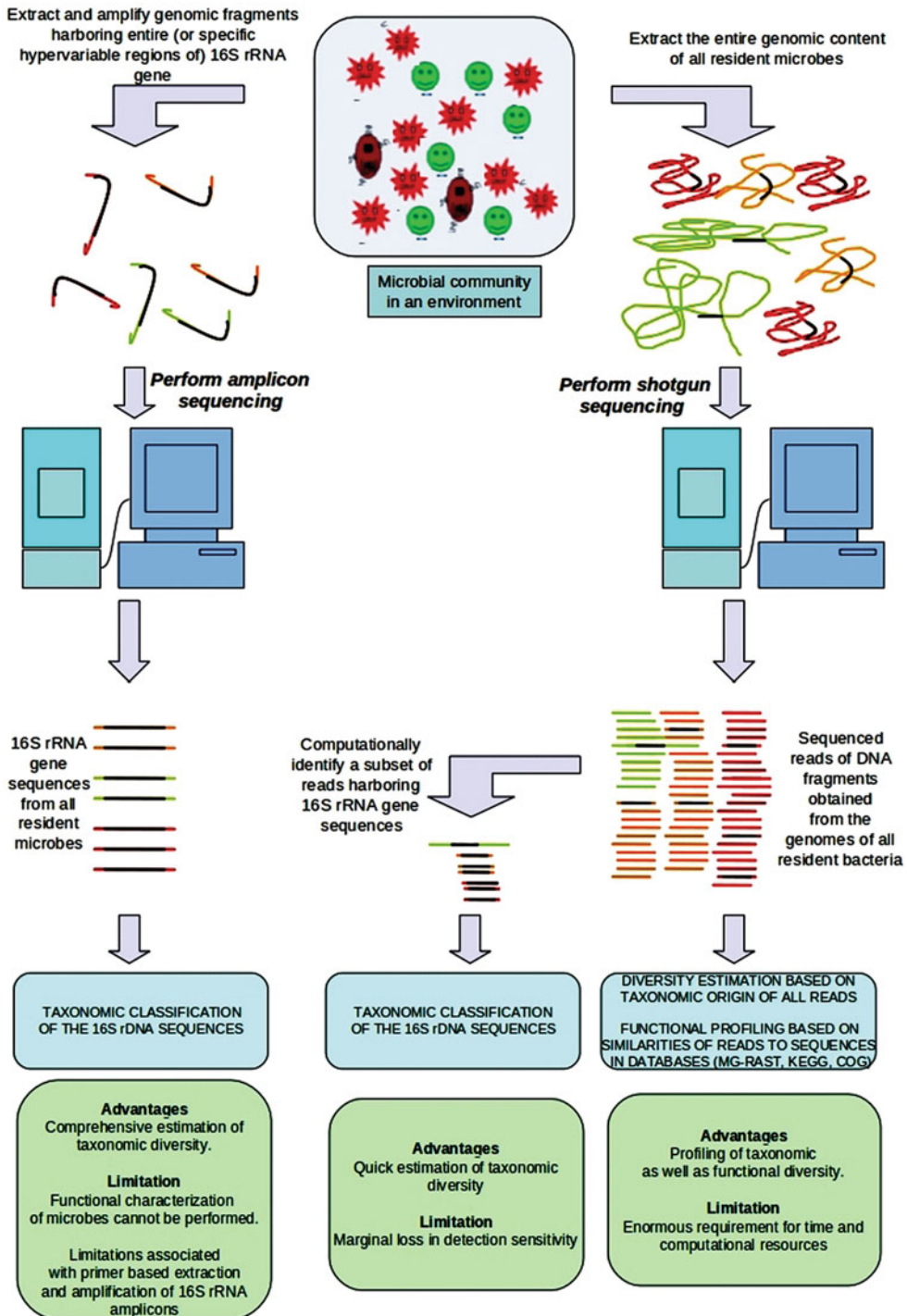
Definition

Estimation of microbial diversity in an environment by efficiently identifying and classifying 16S rRNA gene fragments in metagenomic datasets using computational methods.

Introduction

Recent advances in high-throughput sequencing technologies have enabled life-science researchers to rapidly sequence and characterize the entire genomic content of microbial communities residing in diverse ecological niches. A key advantage of characterizing microbial communities in this fashion is that it enables the concomitant characterization of several microbes (constituting the community), most of which cannot be studied using traditional culture-based genomic approaches. Moreover, this approach (referred to as “Metagenomics”) is useful in understanding the interaction patterns between the resident microbes as well as between the microbes and the environment.

Characterizing and comparing the taxonomic as well as functional diversity of microbial communities (obtained from varied ecological niches) are the broad objective of metagenomic projects. These objectives are attained using two well-established approaches (Fig. 1). In the first approach (commonly referred to as the amplicon-based approach), a quick snapshot of taxonomic diversity of a given environmental sample is obtained by specifically amplifying, cloning, and sequencing gene or gene fragments corresponding to one or more phylogenetic marker genes. The 16S rRNA gene is the most widely used phylogenetic marker gene employed in such amplicon-based approaches. Subsequently, bioinformatic approaches are used for taxonomically classifying these sequenced genes or gene fragments. The relative proportions of various taxonomic groups present in the metagenomic dataset (representing a given environmental sample) are then obtained from the identified taxa. In the second approach



I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Fig. 1 An overview of different approaches adopted by metagenomic projects for profiling the taxonomic and/or functional diversity of

a given environment. Advantages and limitations of each approach are also summarized. *Black* regions depicted in the genomic fragments correspond to entire or a fragment of 16S rRNA gene

(commonly referred to as the shotgun-sequencing approach), the genomic content of a given environmental sample is extracted and sequenced. Genomic fragments (referred to as “reads”) obtained from the sequencing platforms are then computationally analyzed in terms of taxonomy and function. Since the shotgun-sequencing approach generates millions of reads originating from random positions/locations within the genomes of various microbes constituting a given environmental sample, a subset of these reads (hereafter referred to as 16S rDNA fragments) are expected to originate from genomic regions that specifically encompass the 16S rRNA genes of the resident microbes. Identifying 16S rDNA fragments (from within millions of reads constituting a typical metagenomic dataset) and subsequently classifying them is therefore expected to aid in quickly deciphering the taxonomic diversity of a given metagenomic dataset. The following sections describe two algorithms, namely, i-rDNA and C16S, which are used for the identification and taxonomic classification of 16S rDNA fragments in metagenomic datasets, respectively.

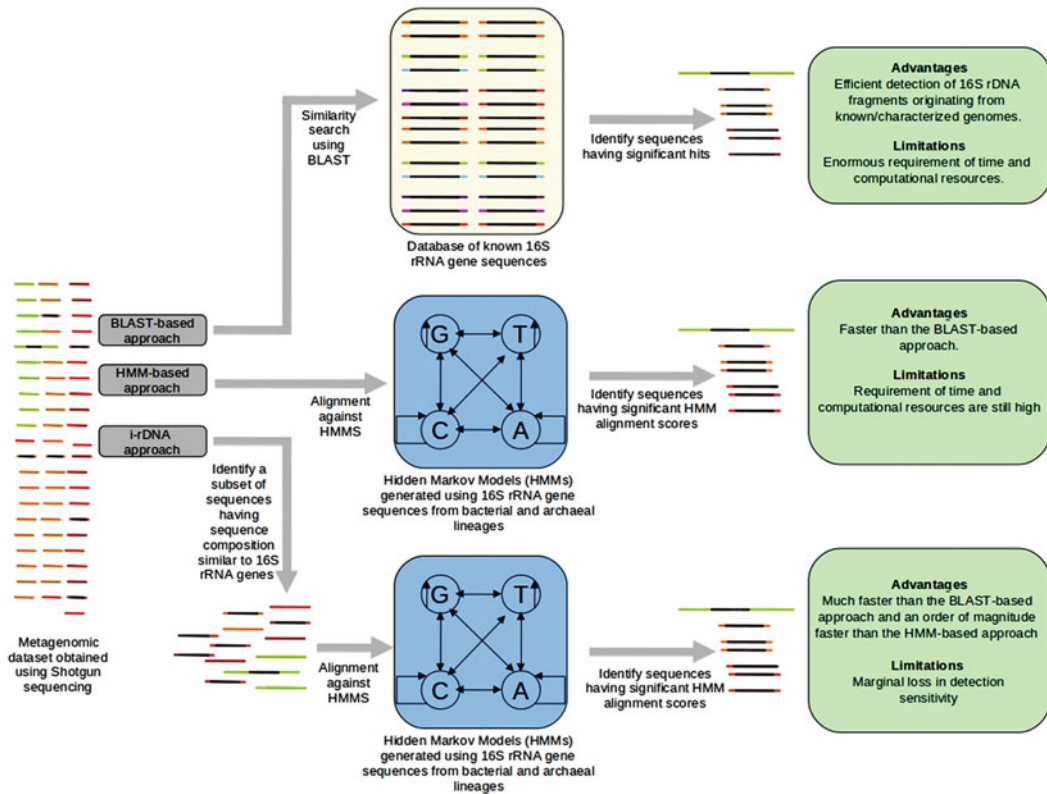
i-rDNA: Algorithm for Identification of 16S rRNA Gene Fragments in Metagenomic Datasets

One of the simplest ways of identifying 16S rDNA fragments in a metagenomic dataset is by performing similarity searches of all reads constituting the dataset against a database containing known 16S rRNA gene sequences. Such similarity searches are typically performed using popular algorithms such as BLAST (Altschul et al. 1990) and BLAT (Kent 2002). Similarity of a read with sequences in the database is evaluated based on how it aligns with these sequences. Reads having significant similarity (similarity being defined in terms of alignment parameters such as e-value, identity, and alignment length) with database sequences are identified as 16S rDNA fragments. Since this approach enables identification as well as taxonomic classification of 16S rDNA fragments, it is currently incorporated as a standard procedure in popular metagenomic analysis platforms such as

MG-RAST (Meyer et al. 2008) and CAMERA (Seshadri et al. 2007). Given the robustness of BLAST/BLAT algorithms, this approach has high sensitivity in identifying/classifying 16S rDNA fragments (even for reads with lengths <100 bp) originating from known and characterized genomes.

The BLAST-based approach, although identifies 16S rDNA sequences with high sensitivity, requires huge compute power for performing alignments of millions of metagenomic reads with thousands of reference 16S rRNA gene sequences. This makes it unsuitable for practical use in research labs lacking access to high-end computational infrastructure. Another alignment-based methodology attempts to address/overcome this limitation by employing hidden Markov models (HMMs) that represent the universally conserved sequence architecture of the 16S rRNA gene (Huang et al. 2009). These HMMs, built separately for bacterial and archaeal kingdoms, reflect the sequence conservation pattern observed within the 16S rRNA genes of microbes belonging to these two lineages. For identification of 16S rDNA fragments, reads in a metagenomic dataset are individually aligned to these two HMMs. Reads obtaining significant alignment scores are then tagged as 16S rDNA fragments. Given that the alignments of individual reads are done only against two HMMs, rather than against thousands of individual reference 16S rRNA gene sequences (as in the case of the BLAST-based approach), the execution time as well as the requirements of compute power are significantly reduced. Moreover, this approach is observed to achieve similar levels of detection sensitivity as that of BLAST-based approach.

Though the above-described HMM-based approach represents a rapid way of identifying 16S rDNA fragments (as compared to the BLAST-based approach), it still involves performing alignments of each individual read (in metagenomic datasets) against two HMMs. Consequently, adopting the HMM-based approach (on a standard work-station) for identification of 16S rDNA fragments within huge metagenomic datasets (e.g. the Human Microbiome Project containing more than



I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Fig. 2 Available approaches for identification of 16S rRNA gene fragments in metagenomic datasets obtained using shotgun

sequencing. Advantages and limitations of each approach are also summarized. *Black* regions depicted in the genomic fragments correspond to entire or a fragment of 16S rRNA gene

32 million sequences) is expected to take several hours to a few days. The recently published i-rDNA method (Mohammed et al. 2011) has addressed this issue by employing a sequence composition-based step prior to the similarity search step performed against the bacterial and archaeal HMMs (Fig. 2). This precursor step is based on the following premise/observations. Given that significant portions of 16S rRNA gene sequences are universally conserved across all prokaryotic lineages, genomic regions encompassing 16S rRNA genes are characterized by distinct sequence compositions (in terms of oligonucleotide usage patterns) as compared to the other regions of the genome. The i-rDNA method utilizes this observation to first identify a subset of reads which have an oligonucleotide composition similar to that of 16S rRNA gene

sequences and subsequently provide this small subset of reads as input to the HMM-based approach. This step of prefiltering data (based on compositional characteristics) essentially reduces the volume of data which are provided as input to the HMM alignment step. The finer algorithmic details of the i-rDNA method are explained in the subsequent paragraphs.

The i-rDNA method first captures the oligonucleotide usage patterns which are specific to 16S rRNA gene sequences. This procedure is performed as a one-time preprocessing step. For this purpose, genomic fragments (of lengths 1,000 bp each) from all completely sequenced prokaryotic genomes are first obtained. Each fragment is then represented as a 256-dimensional vector containing the frequencies of all possible tetranucleotides. Subsequently, vectors

corresponding to all fragments are clustered (using k-means clustering algorithm) based on their tetranucleotide frequency patterns. This generates a feature vector space with a number of clusters. Centroids of the clusters are then calculated based on the fragments contained in them. Each cluster in the feature vector space is thus represented by its centroid. Given the unique sequence composition of the 16S rRNA gene, genomic fragments encompassing this gene are localized to a subset of these clusters. In the preprocessing step of i-rDNA method, clusters containing significant proportions of 16S rRNA gene fragments (as compared to other clusters) are identified and tagged as “probable 16S” clusters (Fig. 3). This information is stored in the form of a mapping file that contains cluster centroids along with their respective tags (either probable 16S or non-16S).

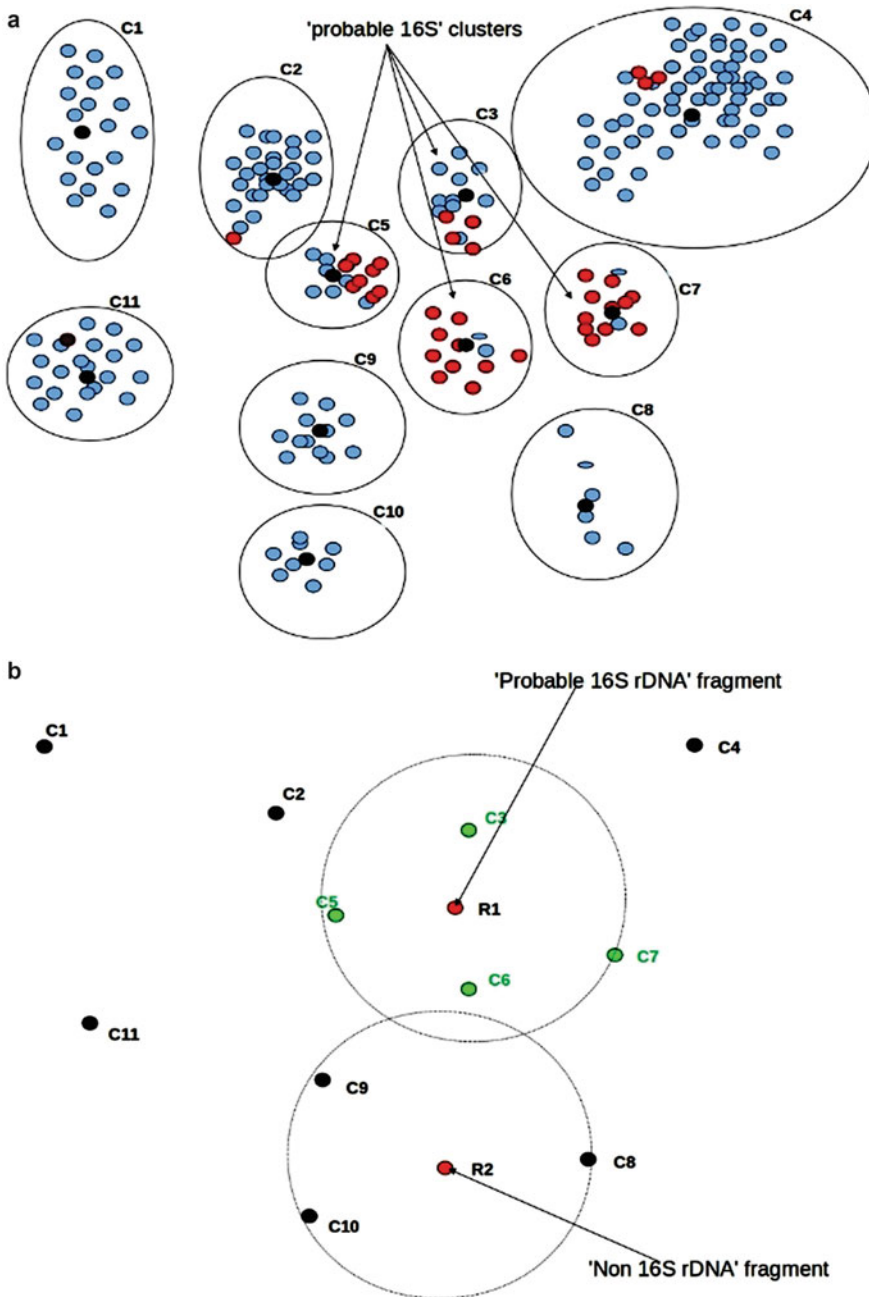
The i-rDNA method identifies 16S rDNA fragments (from amongst all reads constituting a given metagenomic dataset) in the following manner. For each read, the distances of its tetranucleotide frequency vector to all the cluster centroids in the mapping file (obtained as described in the previous paragraph) are first computed. This step helps in identification of a set of clusters having tetranucleotide composition most similar to that of the read. If a significant proportion of the identified clusters are observed to be pre-tagged (in the mapping file) as “probable 16S,” the read is classified as a “probable 16S rDNA” fragment (Fig. 3). Only those reads classified as “probable 16S rDNA” fragments are provided as input to the downstream HMM search. Adoption of the above strategy in the published study (Mohammed et al. 2011) indicated a six to ten times reduction in the number of sequences provided as input to the HMM search, thereby drastically reducing the overall time for identifying 16S rDNA fragments (in metagenomic datasets). Furthermore, this noticeable reduction in the overall analysis time was observed to be achieved without any significant loss in detection sensitivity (Mohammed et al. 2011).

Table 1 provides an additional comparison of detection sensitivity and execution time for three

approaches, namely, BLAST based, HMM based, and i-rDNA, for four simulated metagenomic datasets. These datasets were generated by providing 35 prokaryotic genomes as input to the MetaSim sequence simulator software (Richter et al. 2008). Sequences in each of these datasets simulated the lengths and error rates of four popular sequencing platforms, viz., Sanger (sequence length approximately 800 bp), 454-titanium (~400 bp), 454-standard (~250 bp), and Illumina (~110 bp). These comparative evaluations were performed on a standard Linux workstation having a 2.33 GHz dual core processor and 2GB RAM memory. Results in this table indicate the utility of the i-rDNA method in reducing the overall time taken for identification of 16S rDNA fragments in metagenomic datasets. The i-rDNA method is observed to be 50 and 8 times faster in identifying 16S rDNA fragments as compared to the BLAST-based and HMM-based meta-rna program, respectively. As can be observed, this reduction in time for identification is not accompanied by a noticeable decrease in detection sensitivity.

C16S: Algorithm for Taxonomic Classification of 16S rRNA Gene Fragments in Metagenomic Datasets

Extraction and classification of 16S rRNA gene fragments is one of the quickest ways to estimate taxonomic diversity of any microbial community. Due to the presence of several characteristic features, the 16S rRNA gene has been used as an ideal taxonomic marker. Primarily, this gene is ubiquitously present within the genomes of all prokaryotic organisms. Secondly, given its role in key cellular processes (e.g., protein synthesis), the probability of this gene being involved in lateral gene transfer events is also minimal (Jain et al. 1999; Daubin et al. 2003). This property enables its use as a phylogenetic marker to study the evolutionary patterns in diverse prokaryotic lineages with high confidence. Furthermore, 16S rRNA genes are characterized by highly conserved regions (U1-U8) that flank hypervariable regions (V1-V9) (Jonasson et al. 2002). Universal/customized primers designed against these conserved stretches (which are adjacent to the



I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Fig. 3

A conceptual overview of the framework used by the i-rDNA method. (a) A schematic representation of the preprocessing step of i-rDNA method. A feature vector space is generated by performing a k-means clustering (using tetranucleotide frequencies) of genomic fragments from all completely sequenced microbial genomes. In this feature vector space, clusters C3, C5, C6, and C7,

containing genomic fragments harboring portions of 16S rRNA gene in significant proportions, are tagged as "probable 16S" clusters. *Red dots*: fragments originating from genomic regions harboring portions of 16S rRNA gene. *Blue dots*: fragments not containing any portion of the 16S rRNA gene. *Black dots*: centroids corresponding to each of the clusters in the feature vector space. (b) Identification workflow of the i-rDNA method. Tetranucleotide frequency vectors corresponding to query reads

hypervariable regions) facilitate specific isolation, PCR-based amplification and subsequent sequencing of the entire length (or specific portions) of 16S rRNA genes. The hypervariable regions within the sequenced 16S rRNA gene fragments are specific to each organism and thus serve as “taxonomic barcodes.” These “barcodes” can be used to classify 16S rRNA gene fragments sampled from a given environment into different taxonomic groups.

Various strategies are currently employed for classification of 16S rDNA fragments (Fig. 4). Overall, these strategies involve comparing the sequences and/or the compositions of 16S rDNA fragments with sequences/models corresponding to known taxonomic groups. Details of these strategies are described below. The BLAST-based approach (described in the previous section) is also employed for classifying 16S rDNA fragments. For this purpose, 16S rDNA fragments having significant hits with reference 16S rRNA gene sequences (from known and characterized microbes) are assigned to the taxa corresponding to the best hit(s). In this process, the quality of the BLAST hit (obtained between the query 16S rDNA fragment and reference 16S rRNA gene sequence) is judged based on user-specified thresholds of alignment parameters such as bit score, e-value, identity percentage, etc. Apart from the huge compute power requirement (for performing the alignment step), the BLAST-based approach has the following limitation. In a given metagenomic dataset, a large proportion of query 16S rDNA fragments typically originate from hitherto unknown taxa. Such sequences may belong to an entirely new species or genus or family or order or class or even a new phylum. Attempting to map such novel query 16S rDNA fragments to known taxonomic groups is expected to result in incorrect taxonomic

classification. Although using a stringent set of BLAST thresholds (for evaluating alignment quality prior to assignment) is expected to reduce the misclassification rate (to some extent), a large number of 16S rDNA fragments may remain unassigned/unclassified. It may be noted that various read mapping algorithms, e.g., BWA (Li and Durbin 2010), Bowtie (Langmead et al. 2009), etc., have also been used for aligning query 16S rDNA fragments with sequences in reference databases. The premise and the overall methodology for inferring the taxonomic origin of query sequences however remain the same as in the BLAST-based approach.

Inferring the taxonomic origin of query 16S rDNA sequences can also be performed by mapping/aligning them to precomputed multiple sequence alignments (MSAs). These MSAs are generated by pre-aligning well-annotated 16S rRNA gene sequences belonging to organisms of known taxonomic lineages. A detailed description of the methods adopting such strategies is provided in another review (Sun et al. 2011). MSA-based approaches, though observed to provide robust taxonomic inferences, are critically dependent on the quality and the taxonomic coverage of the reference sequences which are used for generating the precomputed alignments. Furthermore, given the algorithmic complexity of the process of performing/generating multiple sequence alignment(s), enormous amount of time and compute power are typically required for MSA-based analyses.

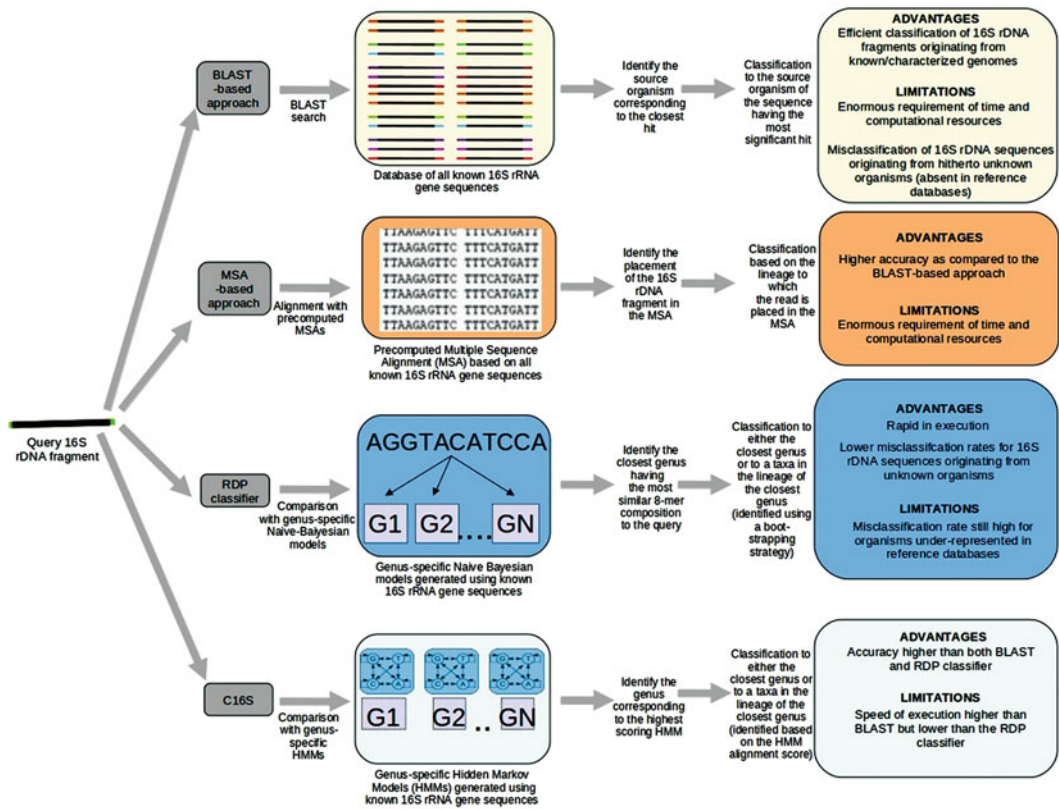
The widely popular RDP classifier (Wang et al. 2007) attempts to address the limitations associated with the above-described BLAST-based as well as MSA-based approaches. This method taxonomically classifies a query 16S rDNA fragment by comparing its compositional properties (e.g., oligonucleotide usage pattern)

I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Fig. 3 (continued) (R1 and R2) are first mapped to the feature vector space (generated in the preprocessing phase of i-rDNA as described above in (a)). Read R1 maps to an area (within the feature vector space) that is in close proximity to

cluster centroids C3, C5, C6, and C7 (all of which are pre-tagged as “probable 16S” clusters). Consequently, read R1 is identified as a “probable 16S rDNA” fragment. Read R2 is in close proximity to clusters C8, C9, and C10 (all of which are pre-tagged as “non-16S” clusters). Read R2 is therefore identified as a “non-16S rDNA” fragment

I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Table 1 Performance of i-rDNA, meta-rna (a HMM-based identification method) and BLAST in terms of detection sensitivity and execution time. The approximate length of reads constituting each of the four simulated test datasets is indicated in brackets

Test dataset	Number of reads	Detection sensitivity (%)			Execution time (in seconds)		
		i-rDNA	meta_ma	BLAST	i-rDNA	meta_rna	BLAST
Illumina (~110 bp)	1,000,000	93.1	94.6	98.1	102	1,110	6,317
454-Standard (~250 bp)	400,000	90.6	96.4	99.2	97	1,026	6,681
454-Titanium (~400 bp)	250,000	91.3	97.1	99.6	92	947	6,128
Sanger (~800 bp)	100,000	87.6	95.2	99.8	105	929	5,783



I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Fig. 4 Different approaches available for the taxonomic classification of 16S rRNA gene fragments

with models generated using compositional features of sequences of known taxonomic lineages. For this purpose, it first creates (as a preprocessing step) Naive Bayesian models that capture 8-mer oligonucleotide word frequencies in 16S rDNA sequences belonging to known genera. During the classification step, for a given

query sequence, the RDP classifier first identifies a model (and the corresponding genus) whose 8-mer word frequencies are “most” similar to that of the query sequence. The classifier then employs a bootstrapping procedure to compute a confidence score of assignment to each taxa belonging to the taxonomic lineage of the

identified genus. The query sequence is then assigned to a taxon (within this lineage) that is at the most specific taxonomic level and also generates a confidence score that exceeds the user-specified confidence score threshold. Besides being alignment-free, a major advantage of the RDP classifier is the bootstrapping procedure employed to compute the confidence scores. The overall strategy of this procedure ensures the accurate assignment of novel 16S rDNA sequences (i.e., originating from hitherto unknown organisms) to related taxa at appropriately higher taxonomic levels. However, it is important to note that the overall process of classification involves scoring and identifying the “best” taxonomic lineage corresponding to the query sequence. This scoring however does not take into account the actual level of compositional similarity between the model corresponding to the “best” taxonomic lineage and the query sequence. Consequently, in cases where 16S rDNA fragments originate from taxonomic lineages that have minor representation in existing 16S rDNA databases, the classification accuracy of the RDP classifier has been shown to decrease (Biers et al. 2009).

In contrast to all the three methods described above, the recently published C16S algorithm (Ghosh et al. 2012) employs genus-specific HMMs for the taxonomic classification of 16S rDNA fragments. The overall classification strategy is based on the following premise. 16S rDNA sequences contain alternating conserved and hypervariable regions. The latter regions are characterized by clade-specific sequence variation patterns. As a preprocessing step, the C16S algorithm captures these clade-specific patterns at the taxonomic level of genus. For this purpose, genus-specific HMMs are first generated and subsequently utilized for classifying query 16S rDNA fragments. During the classification phase, a query 16S rDNA fragment is first mapped to these precomputed genus-specific HMMs and the genus corresponding to the best scoring HMM is identified. The score obtained in this process is then utilized for dynamically

identifying an appropriate level of taxonomic assignment for each query sequence. The strategy of correlating the HMM score with the taxonomic level of assignment is based on the empirical observation that the HMM score decreases with increasing taxonomic divergence between the taxa corresponding to the query and the HMM (Ghosh et al. 2012).

The classification methodology adopted by C16S has the following advantages. First, employing representative genus-specific HMMs significantly reduces the time and compute power as compared to that typically required by BLAST-based or MSA-based classification approaches. Second, the use of precomputed threshold scores in C16S ensures assignment of query sequences (originating from unknown organisms) at appropriately higher taxonomic levels, thereby reducing its misclassification rate as compared to that by the RDP classifier. Finally, given that the identified taxonomic levels are specific to the extent possible, the overall specificity of assignments by C16S is not compromised.

The above observations (with respect to classification efficiency of C16S) are also reflected in the results of a comparative evaluation between the C16S algorithm and the RDP classifier (run with default parameters). This evaluation was performed using five simulated 16S rDNA datasets (each comprised of 30,000 sequences). While one of these datasets consisted of full-length 16S rRNA gene sequences from taxonomically diverse microbes, the others consisted of 16S rDNA fragments that mimicked the length and the sequencing error rates associated with four popular sequencing platforms, viz., Sanger, 454-titanium, 454-standard, and Illumina. Furthermore, for each dataset, evaluation was performed in four different simulated metagenomic scenarios, wherein the input 16S rDNA sequences mimicked those originating from entirely new genera, families, orders, and classes, respectively. These simulated scenarios were generated by progressively removing the models corresponding to the genus, family, order, and class of the source organisms

l-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments, Table 2 Distribution of taxonomic assignments obtained using C16S and RDP classifier for five simulated metagenomic datasets (each comprised of 30,000 sequences)

Assignment category	Database scenario							
	Minus genus		Minus family		Minus order		Minus class	
	C16S	RDP	C16S	RDP	C16S	RDP	C16S	RDP
Illumina dataset (average read length ~ 110 bp)								
Correct	86.7	81.2	91.2	86.2	88.7	86.2	85.3	84.9
Higher levels	16.7	10.2	21.2	16.6	39.2	34.7	65.7	63.9
Intermediate levels	56.4	56.7	52.2	48.5	25.6	26.3	0	0
Specific levels	13.6	14.3	17.8	21.1	23.9	25.2	19.6	21
454-Standard dataset (average read length ~ 250 bp)								
Correct	95.7	80.5	92.6	88.2	84.5	80.7	84.6	84.3
Higher levels	12.6	4.5	19.8	8.3	37.8	31.5	60.4	60.2
Intermediate levels	56.4	46.7	47.8	45.9	18.4	22.6	0	0
Specific levels	26.7	29.3	25	34	28.3	26.6	24.2	24.1
454-Titanium dataset (average read length ~ 400 bp)								
Correct	94.4	79.5	92.6	84.3	87.2	70.3	86.4	79.5
Higher levels	1.2	2.7	20.3	4.1	22.2	7.3	47.2	43.4
Intermediate levels	56.4	42.5	45.9	44.2	18.4	21.3	0	0
Specific levels	36.8	34.3	26.4	36.0	46.6	41.7	39.2	36.1
Sanger dataset (average read length ~ 800 bp)								
Correct	82.2	58.1	88.4	73.1	90.1	59.6	88.2	68.6
Higher levels	11.2	2.1	19.9	3.1	37.0	6.0	48.2	32.4
Intermediate levels	34.1	20.2	41.1	32.2	6.4	25.8	0	0
Specific levels	36.9	35.8	27.4	37.8	46.7	27.8	40.0	36.2
Dataset with full-length 16S rRNA gene sequences								
Correct	90.1	57.6	88.8	64.9	78.8	48.3	70.7	51.8
Higher levels	3.3	2.1	6.4	3.4	11.2	4.4	19.8	13.8
Intermediate levels	44	16	47.3	26.4	19.6	15.8	0	0
Specific levels	42.8	39.5	35.1	35.1	48.0	28.1	50.9	38.0

(corresponding to the query 16S rDNA fragments) from the databases utilized by RDP classifier as well as the C16S algorithm. Results of this evaluation (Table 2) indicate improved levels of classification accuracy of C16S algorithm as compared to the RDP classifier. Interestingly, the improvement in performance, with respect to both classification accuracy and specificity, is especially pronounced in simulated scenarios, wherein query sequences originate from hitherto unknown genomes lacking counterpart models at the levels of order and class (in the databases of both algorithms). Furthermore, for full-length and Sanger datasets, the classification accuracy

and specificity of C16S is observed to be noticeably better than the RDP classifier.

Correct assignments are assignments made to taxa lying in the path between the root and the source genus of the query sequence.

Correct assignments are further subcategorized into “specific levels,” “intermediate levels,” and “higher levels” as described below

(a) Specific levels: If HMMs corresponding to genus or family or order or class are absent from the reference database, assignment of a query sequence is classified as “correct” at “specific level,” only if the assignment is

made to a correct taxon at the immediate higher taxonomic level. For instance, in a “new family” simulated database scenario (wherein HMMs corresponding to the source family of the query 16S rDNA fragment are absent from the reference database), an assignment of the query sequence to the corresponding order is categorized as correct at specific level.

- (b) Intermediate levels: Correct assignments to taxa lying between the phylum level and the specific level (as described above) are classified as “correct” at “intermediate levels.”
- (c) Higher level: Assignments to root or cellular organisms or to superkingdom levels are categorized as correct assignments at “higher levels.”

Summary

One of the major objectives of most metagenomic projects is to profile and subsequently compare the spatial and temporal variations of microbial communities residing in diverse ecological niches. Analyzing such variations helps in the identification of microbial groups that confer specific characteristics to a given environment in terms of phenotype/function. Development of efficient *in silico* methods for identifying and classifying 16S rRNA genes (or gene fragments) from metagenomic datasets (obtained using amplicon-based or shotgun sequencing approach) is therefore an important computational problem. This article describes two recently reported methods, viz., i-rDNA and C16S, that cater to the tasks of identification and classification of 16S rDNA fragments in metagenomic datasets. The i-rDNA method represents an approach which is efficient in terms of execution speed as well as detection sensitivity. Given its ability to directly identify 16S rDNA fragments from metagenomic datasets (obtained using the shotgun sequencing approach), it holds the potential to completely bypass the experimental procedures (and the related costs of the same) associated with extraction, cloning, and sequencing of the 16S rRNA

gene or gene fragments. On the other hand, the relatively higher classification accuracy of the C16S method (as compared to other contemporary classification methods) is expected to provide an accurate picture of taxonomic diversity of microbial communities inhabiting any given environment.

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes](#)
- ▶ [Microbial Diversity, Bar-Coding Approaches](#)
- ▶ [Nucleotide Composition Analysis: Use in Metagenome Analysis](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data](#)

References

- Altschul SF, Gish W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
- Biers EJ, Sun S, et al. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol.* 2009;75(7):2221–9.
- Daubin V, Moran NA, et al. Phylogenetics and the cohesion of bacterial genomes. *Science.* 2003;301(5634):829–32.
- Ghosh TS, Gajjala P, et al. C16S - a Hidden Markov Model based algorithm for taxonomic classification of 16S rRNA gene sequences. *Genomics.* 2012;99(4):195–201.
- Huang Y, Gilna P, et al. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics.* 2009;25(10):1338–40.
- Jain R, Rivera MC, et al. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A.* 1999;96(7):3801–6.
- Jonasson J, Olofsson M, et al. Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments. *APMIS.* 2002;110(3):263–72.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12(4):656–64.

- Langmead B, Trapnell C, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589–95.
- Meyer F, Paarmann D, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics.* 2008;19(9):386.
- Mohammed MH, Ghosh TS, et al. i-rDNA: alignment-free algorithm for rapid in silico detection of ribosomal gene fragments from metagenomic sequence data sets. *BMC Genomics.* 2011;12 Suppl 3:S12.
- Richter DC, Ott F, et al. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One.* 2008;3(10):e3373.
- Seshadri R, Kravitz SA, et al. CAMERA: a community resource for metagenomics. *PLoS Biol.* 2007;5(3):e75.
- Sun Y, Cai Y, et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform.* 2011;13(1):107–21.
- Wang Q, Garrity GM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007;73(16):5261–7.

K

KEGG and GenomeNet, New Developments, Metagenomic Analysis

Masaaki Kotera, Yuki Moriya, Toshiaki Tokimatsu, Minoru Kanehisa and Susumu Goto
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

Synonyms

GenomeNet; Kyoto Encyclopedia of Genes and Genomes

Definition

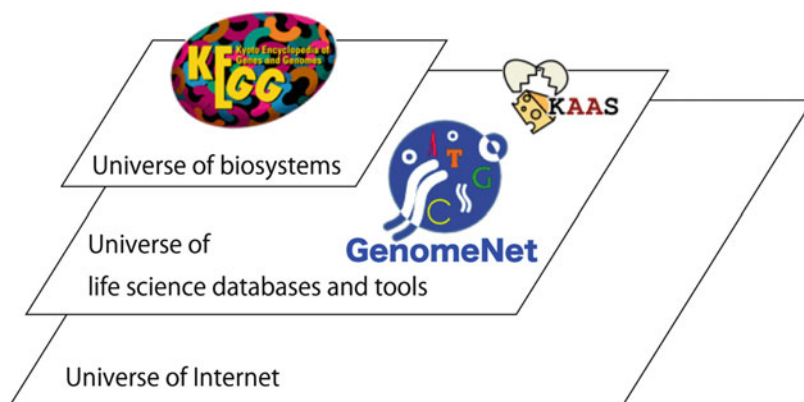
KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource representing biological systems, such as the cell, the organism, and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. GenomeNet is database and computational services for genome research and related research areas in biomedical sciences, operated by the Kyoto University Bioinformatics Center in Japan. Both services work in collaboration putting a special focus on the visualization and interpretation of large amount of data, such as metagenome sequence data, derived from high-throughput measurement techniques.

Introduction

The number of complete genomes has been increasing dramatically. From the completion of the influenza genome in 1995, it took about 13 years (1995–2008) to complete a total of 500 species. The number of complete genomes is expected to have quadrupled (~2,000) during the following 4 years (2009–2012). The total number of putative genes in these ~2,000 genomes is ~8 million. In contrast, recent prevailing technology such as Next Generation Sequencing produces even larger amount of data. One emerging field enabled by this advance in technology is referred to as metagenomics, i.e., genomic-scale sequencing of samples containing a mix of different species. The total amount of publicly available metagenomic data has already become larger than that of genomes: 139 metagenome samples are currently stored in the KEGG database (<http://www.kegg.jp/>; Kanehisa et al. 2012) and the total number of putative genes in these samples is ~14 million. The need arises for novel tools and interfaces to handle this flood of data, which is expected to exponentially increase in the foreseeable future.

KEGG have been storing complete genome, draft genome, and metagenomic data and given them additional functional annotations. The development of KEGG is the continuous effort to construct an integrative knowledgebase for widespread use in many fields, such as molecular biology. GenomeNet (<http://www.genome.jp/>; Kanehisa et al. 2002) is a database and

**KEGG and GenomeNet,
New Developments,
Metagenomic Analysis,**
Fig. 1 KEGG and
GenomeNet among
Internet resources



computational service for genome research and related research areas in biomedical sciences, operated by the Kyoto University Bioinformatics Center. It integrates KEGG with other databases that focus on genes, proteins, enzyme reactions, metabolic compounds, drugs, natural products, and other biological resources scattered all over the world. DBGET/LinkDB (<http://www.genome.jp/dbget/>; Fujibuchi et al. 1998) is an integrated database retrieval system for handling such molecular biology databases and is used as a backbone system in GenomeNet and KEGG. They also develop Web tools for functional analysis based on genome, metabolome, and metabolic reaction information and provide an integrated analysis environment for researchers and general public (Fig. 1). KEGG and GenomeNet depend on each other to provide the high-quality knowledge and sophisticated user interfaces that promote the interpretation of massive amounts of biological data.

Genomic and Metagenomic Contents in KEGG

The KEGG Organism pages list complete genomes, expressed sequence tag (EST) datasets, metagenomes, and pangenomes (set of sequences derived from a group of closely related strains, typically in bacterial phyla) in the following URLs:

Complete and draft genomes (http://www.genome.jp/kegg/catalog/org_list.html)

EST datasets (http://www.genome.jp/kegg/catalog/org_list2.html)

Metagenomes (http://www.genome.jp/kegg/catalog/org_list3.html)

Pangenomes (http://www.genome.jp/kegg/catalog/org_list1.html)

Genome sequences registered in the RefSeq database are incorporated in the KEGG GENES database, and additional annotation is given so that the genes have links to ortholog groups, pathways, etc. Annotation is processed manually with the help of the in-house KOALA (KEGG Orthology And Links Annotation) software, based on the bidirectional best-hit strategy of SSEARCH. Once the annotation is completed, the organism-specific pathway is automatically generated on the basis of the KEGG Orthology and reference pathway (explained below).

By June 2011, KEGG had incorporated two environmental metagenome samples retrieved from the ocean and 137 microbiome samples from human intestines (Fig. 2). KEGG gives organism codes for complete and draft genomes consisting of three or four characters (e.g., hsa for *H. sapiens*, human). The KEGG Organism codes specify organisms and are also used as the headers of the pathway map IDs (e.g., hsa00010 for glycolysis/gluconeogenesis pathway in *H. sapiens*). KEGG recently introduced an identifier system named “T numbers” that specify the sets of sequencing data (EST, metagenomes, and pangenomes). At the time of writing, KEGG has incorporated metagenome data from three sources (NCBI, Metagenome.jp, and MetaHIT).



KEGG Metagenomes

[Genomes | ESTs | Meta | Pan]

Environmental samples

Category	Project	Source
Ocean	T30001 Planktonic microbial communities from North Pacific Subtropical Gyre	NCBI
	T30002 Planktonic microbial communities from Monterey Bay, CA	NCBI
	T30003 Human gut metagenome collected from healthy human sample F1-S (male adult)	Metagenome.jp
	T30004 Human gut metagenome collected from healthy human sample F1-T (female adult)	Metagenome.jp
	T30005 Human gut metagenome collected from healthy human sample F1-U (infant female)	Metagenome.jp
	T30006 Human gut metagenome collected from healthy human sample F2-V (male adult)	Metagenome.jp
	T30007 Human gut metagenome collected from healthy human sample F2-W (female adult)	Metagenome.jp
	T30008 Human gut metagenome collected from healthy human sample F2-X (male child)	Metagenome.jp
	T30009 Human gut metagenome collected from healthy human sample F2-Y (female child)	Metagenome.jp
	T30010 Human gut metagenome collected from healthy human sample In-A (male adult)	Metagenome.jp
	T30011 Human gut metagenome collected from healthy human sample In-B (male infant)	Metagenome.jp
	T30012 Human gut metagenome collected from healthy human sample In-D (male adult)	Metagenome.jp
	T30013 Human gut metagenome collected from healthy human sample In-E (male infant)	Metagenome.jp
	T30014 Human gut metagenome collected from healthy human sample In-M (infant female)	Metagenome.jp
	T30015 Human gut metagenome collected from healthy human sample In-R (female adult)	Metagenome.jp
	T30016 MH0001 MetaHIT sample from healthy Danish female	MetaHIT
	T30017 MH0002 MetaHIT sample from healthy Danish female	MetaHIT
	T30018 MH0003 MetaHIT sample from healthy Danish male	MetaHIT
	T30019 MH0004 MetaHIT sample from healthy Danish male	MetaHIT
	T30020 MH0005 MetaHIT sample from healthy Danish male	MetaHIT
	T30021 MH0006 MetaHIT sample from healthy Danish female	MetaHIT

KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 2 Screenshot of KEGG Metagenomes page

The examples of T numbers include T30001 for planktonic microbial communities from North Pacific Subtropical Gyre (retrieved from NCBI), T30003 for human gut metagenome collected from a healthy Japanese adult male F1-S (retrieved from Metagenome.jp), and T30016 for human gut microbial gene sample from healthy Danish female (retrieved from MetaHIT).

For users interested in an organism (identified by the KEGG Organism code) or a sample (identified by the T number), embedded links make it is easy to jump to the corresponding summary pages. Clicking the “T30003”, for instance, in the KEGG Metagenomes page takes the user to the summary page specific for the sample T30003. KEGG provides this type of

pages for all genomes, metagenomes, pangenomes, and EST datasets. Also, users can search for genes of interest and jump to pathway maps, functional hierarchy, modules, etc.

KEGG PATHWAY Maps and BRITE Functional Hierarchy

KEGG PATHWAY maps (<http://www.kegg.jp/kegg/pathway.html>) and BRITE functional hierarchy (<http://www.kegg.jp/kegg/brite.html>) generally do not focus on a specific organism. BRITE contains a number of hierarchical classifications of vocabularies used in journal articles and other public data in academic communities. The “reference” pathway maps are the combined

pathways present in a number of organisms and are consensus among many published articles. Only the reference pathway map is manually drawn with in-house software called KegSketch, whereas all other organism-specific maps are computationally generated. The user can conduct a search limited to an organism of interest as well as a comprehensive search throughout all of the genome-sequenced organisms. In the pathway maps, rectangles and circles represent gene products (mostly proteins) and other molecules (mostly metabolites), respectively. The maps are colored in black and white in reference pathways, i.e., when no organism has been specified. When the user can specify an organism of interest, the organism-specific pathways include some colored rectangles indicating that the specified organism possesses the corresponding genes or proteins in the genome (Fig. 3, left). White rectangles indicate that no genes have been annotated to the corresponding function. This does not necessarily mean the organism does not possess the corresponding genes, but it is possible that the genes have not been identified yet.

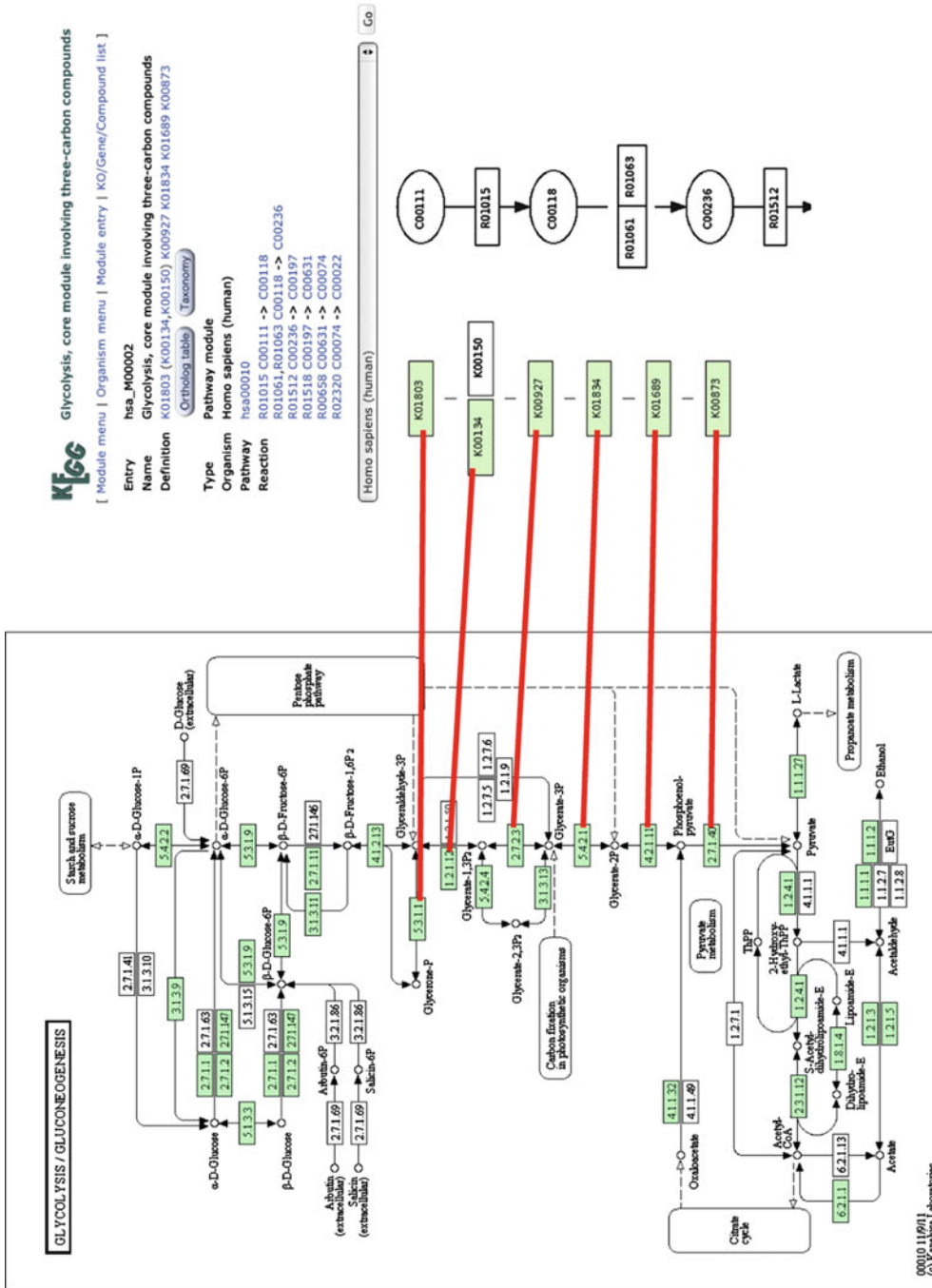
KEGG Module

KEGG has three different levels of resolutions for visualizing pathways: global maps (Fig. 6), (conventional) pathway maps (Fig. 3, left), and pathway modules (Fig. 3, right). Mapping genes to global maps helps users to grasp the overview of the sample. Mapping genes to pathway maps is useful to check the functional capability of the genome or metagenome. There are some cases where the smaller functional units, as defined in KEGG Modules, are more helpful to conduct the detailed analysis. KEGG Modules include consecutive reaction steps, operon or other regulatory units, and phylogenetic units by genome comparison. KEGG have recently been focusing effort on the development and annotation of KEGG Modules, leading to the increase of the number of entries. KEGG Module (<http://www.kegg.jp/kegg/module.html>) collects functional units classified into the following four categories:

- (1) pathway modules – representing smaller pathway units than KEGG PATHWAY maps, such as M00002 (glycolysis, core module involving three-carbon compounds; see Fig. 3, right);
- (2) structural complexes – often forming molecular machineries, such as M00072 (oligosaccharyltransferase);
- (3) functional sets, for other types of essential sets, such as M00360 (aminoacyl-tRNA synthetases, prokaryotes); and
- (4) signature modules, as markers of phenotypes, such as M00363 (EHEC pathogenicity signature, Shiga toxin).

KEGG Orthology (KO)

Coloring the rectangles in the organism-specific pathways, i.e., estimating the presence/absence in the respective genes in pathway maps, is determined based on the KEGG Orthology (KO). KO collects the groups of orthologous genes having a common function and the same evolutionary origin. A group of orthologous genes (a KO entry) is given an identification number (K number) and in principle corresponds to more than one gene derived from more than one organism. Genes assigned to the same K number correspond to the same rectangle in a PATHWAY map (Fig. 3, left), MODULE (Fig. 3, right), and BRITE hierarchy. The top page of KO (<http://www.kegg.jp/kegg/ko.html>) provides the form to obtain an ortholog table (Fig. 4), which shows currently annotated genes in individual genomes for a given set of K numbers, together with coloring of adjacent genes on the chromosome. Each KEGG Module also contains a link to the corresponding ortholog table. The ortholog table is a useful tool to check completeness and consistency of genome annotations. KO entries for complete genomes are manually defined and annotated by the KEGG expert curators based on the phylogenetic profiles and functional annotations of the genes. On the other hand, KO for draft genomes, metagenomes, pangenomes, and EST datasets are automatically annotated by KAAS (KEGG Automatic Annotation Server), one of the GenomeNet tools.



KEGG and GenomeNet, New Developments, Metagenomic Analysis, (right). Rectangles colored in green indicate that human genome possesses the corresponding genes. KEGG Orthology entries are used to define KEGG Modules, which is part of pathway maps, as indicated by the red lines

Fig. 3 Mapping human genome onto glycolysis pathway and module. Human genome mapped onto glycolysis pathway map00010 (left) and module M00002

Ortholog table

Eukaryotes Page: 1

Organism	K01803 (TPI) [227]	K00134 (GAPDH) [317]	K00927 (PGK) [232]	K01834 (gpmA) [180]	K01689 (ENO) [244]	K00873 (PK) [366]
hsa	7167	2597	5232 5230	5223 5224 441531	2027 2026 2023	5315
ptr	451799	451783 467251	462757 473678	737767 494122 450650	454457 457913	748700
pon	100172948	100172694	100434395 100174172	100462476 100174215 100435315	100461201 100173448	100174114
mcc	714090 706960	574353 712582	706939 706325	694418 706211 694946 695799 696959 720615	714208 709378	710798 697742
mmu	21991	14433 100042025 100048117	18655 18663	18648 56012	13806 13807 13808 433182	18746
rno	24849 246267 500959 498731	498019 24383 498123 290604 688783 295423 685186 498099 500983 306115 498881 293876 303448	24644 499525 316265	24642 24959	24333 25438 24334	25630 100364062
cfa	477711	403755 481027 477441 487478 481849	480964 474933	477786 475495	479469 100856683 479597	403874

KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 4 Screenshot of the ortholog table for module M00002. Ortholog tables contain

links to the genes corresponding to the orthologs (K numbers) in genome-sequenced species. *Columns* and *rows* represent orthologs and species, respectively

KAAS Automatic Annotation

The KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007) is one of the genome analysis tools available in GenomeNet (<http://www.genome.jp/tools/kaas/>) and has been developed for annotating draft genomes, metagenomes, pangenomes, and EST datasets in the framework of KEGG. KAAS accepts any groups of gene sequences and helps users annotate these genes if the genes are derived from organisms that are not yet a member of the KEGG Organisms, or users obtain the gene IDs otherwise (Fig. 5). After submitting the sequence data, it may take a long time to complete the calculation. Therefore, users are requested to input their e-mail addresses, and the URL to access the calculation result will be informed

later. The result contains the corresponding KO list, links to automatically colored PATHWAY pages and the BRITE pages. It is recommended that the users download the result, since these results will be removed from GenomeNet server after a few days.

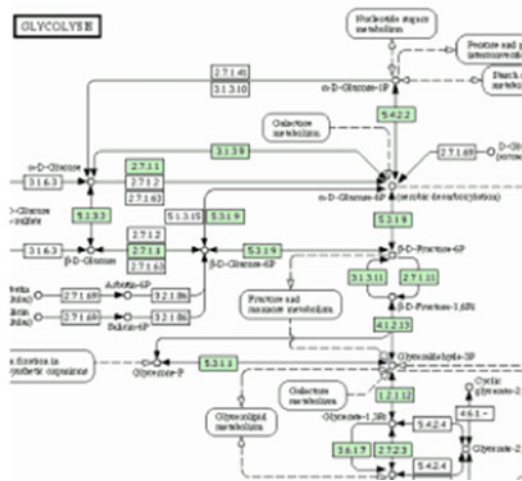
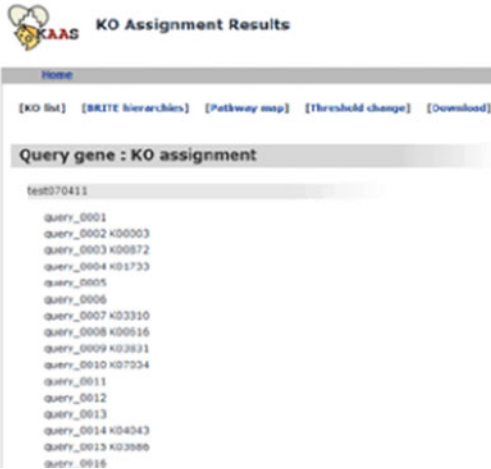
Mapping Metagenome Data on KEGG PATHWAY

It is possible to color KEGG PATHWAY/BRITE in a user-defined manner by using KEGG Mapper (<http://www.kegg.jp/kegg/mapper.html>). This will become more valuable for the interpretation of metagenome and pangenome studies. KEGG Mapper has an option to specify multiple organisms at a time. This option is particularly helpful



complete or draft genomes
 partial genomes
 EST datasets
 metagenomic samples

automatic KO assignment + KEGG pathway mapping

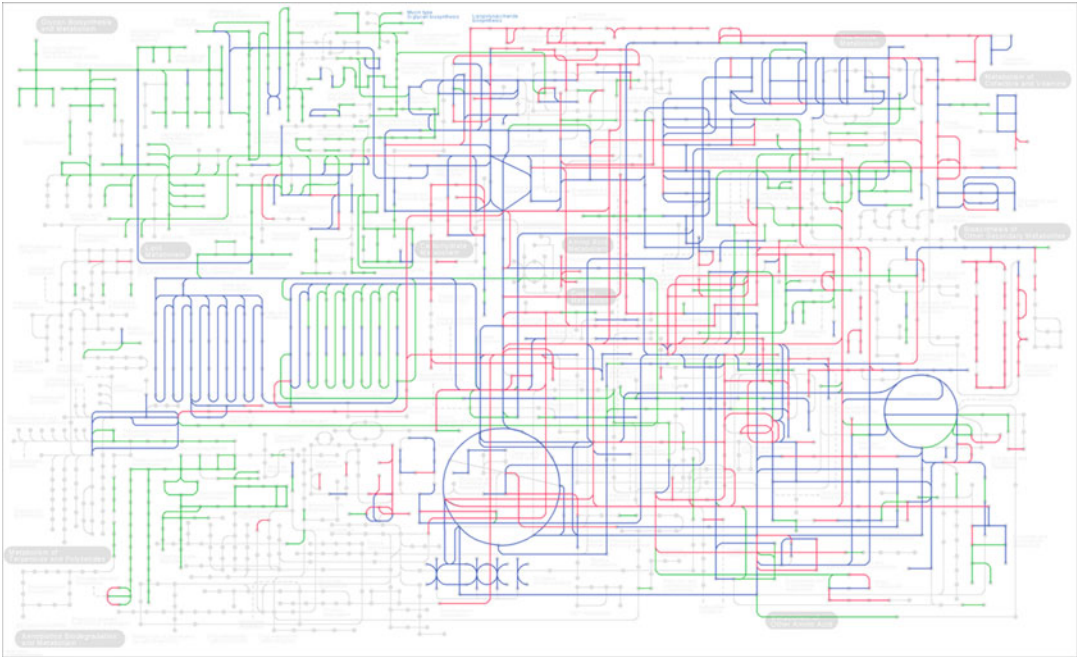


KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 5 KEGG Automatic Annotation Server (KAAS)

not only for comparing genomes but also for visualizing host-microbiome relationship such as in human gut microbiome, host-symbiont relationship, and host-pathogen relationship. If a user inputs “hsa + pfa”, meaning human (*Homo sapiens*) plus a pathogen (*Plasmodium falciparum 3D7*), the resulting pathways will be double colored. The two colors would represent the gene products from the two organisms. This option accepts any combinations up to a total of ten genomes. For instance, the query “hsa + mmu + dme”, which means human (*Homo sapiens*) + mouse (*Mus musculus*) + fruit fly (*Drosophila melanogaster*), provides the three-colored map.

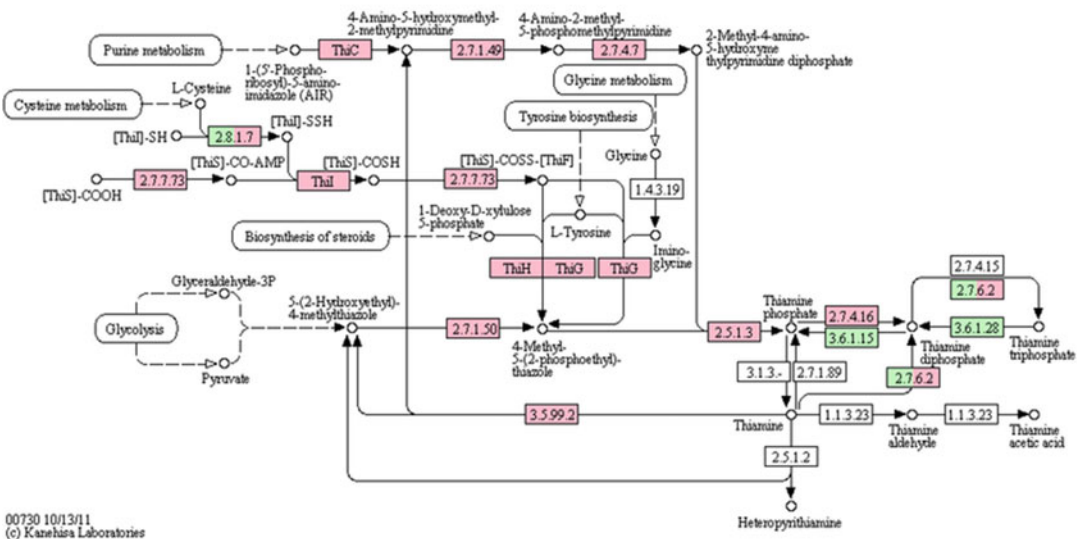
Metagenomes can also be viewed with KEGG Mapper. Figure 6 shows the human genome and a human intestine metagenome mapped onto

a global map, where green lines indicate genes that the human genome (only) possesses, red lines indicate gut metagenome (only) genes, and blue lines indicate genes possessed by both. Figure 7 shows an example of the reconstructed thiamine metabolism pathway by mapping human genome (*hsa*, colored in green) and human intestine metagenome (T30003, colored in pink). Thiamine diphosphate shown in this pathway works as an essential nutritional factor for human, but this cannot be synthesized without the help of the symbiotic bacteria in human intestine. By clicking one of the pink-colored rectangles (e.g., ThiC), a user can see the list of corresponding genes in the metagenome (Fig. 8). The possible common sets of functions between human genome and human gut



KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 6 Mapping of human genome and human intestine metagenome on a global map

THIAMINE METABOLISM



00730 10/13/11
(c) Kanehisa Laboratories

KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 7 Mapping of human genome and human intestine metagenome on thiamine metabolism

KEGG T30003 (Metagenome): id06895 Help

Entry	id06895 CDS T30003
Gene name	F1-S_1629.1_B_pred_1_+_57_1553_0_C
Definition	
Orthology	K03147 thiamine biosynthesis protein ThiC
Taxonomy	Bacteroidetes [GN:bvu] Taxonomy
Pathway	00730 Thiamine metabolism
AA seq	498 aa AA seq DB search MEIDLKKGKLPMMRESWIIGRGDVEKLPISITSEYGMRRDDKSLDHLRFEHIALPYRAKAG KAITQMAYAKAGIVTPEMEYVAIRENMNCRELGIDTFITPEFVRDEIAAGRAVL PANINH PESEPMIIGRNFLVKINTNIGNSATSSIDEVEKAVWWSCKWGGDTLMDLSTGDNIHETR EWIVRNCPPVGTVPYQALEKVNKVENLNWEIYKDTLIEQCEQGVDFYFIHAGIRRQN VHLADKRLCGVSRGGSIMSKWCLVHDKESFLYEHFDDICDILAQYDVAVSLGDGLRPGC IADANDEAQFAELDTMGELVLRWANKNVQAFIEGPGHVPLHKIKENMERQISHCHNAPFY TLGPLVTDIAPGYDHITSAIGAAQIGWLGTAAML CYVTPKEHLGLPNKEDVRIGVITYKIA AHAADLAKGHPGAQJRDNALSKARYEFRWRDQFHLSLDPDRAL EYFNEGRHTDGEYCTMC GPNFCAMKLSRDLKNVGN

KEGG T30003 (Metagenome): id10171 Help

Entry	id10171 CDS T30003
Gene name	F1-S_2876.1_B_pred_2_+_444_1808_0_C
Definition	
Orthology	K03147 thiamine biosynthesis protein ThiC
Taxonomy	Firmicutes [GN:ere] Taxonomy
Pathway	00730 Thiamine metabolism
AA seq	454 aa AA seq DB search MTSTLLFRDKLSFGGTQRMQTYTTQMDAARKGITPEMEIVAKKEYRTTEEIRQWVAEGK VAIPANKHHKCLNPEGVGSMLRTKINVNLGVSRDCKDYNIEMQKVMSAVNMGAEAIMDLS SHGNTQPFQRLKTHECPVMIGTVPVYDSVIHYQRDLAELTAQDFIDVRLHAEDGVDFVT LHCGITRKTIEQIRKHKRKMNIVSRGSSLVFAWMSMTGNENPFYEHFDEICEICAEHDVT ISLGDACRPGCLADATDVCQEELVRLGELTKRAWAHNVQVMVEGPGHVPLNQVAANMEV QKSCIMGAPFYVGLPLVTDIAPGYDHITAAIGGAVAAASGA AFLCYVTPAEHLALPNVDD VKQGIVASKIAAHAADIAGKIPHARDIDDKMGDARRVLDWDAQFACALDPETA KAIRDAR LPEDDHS DTCSMCGKFC AVRSMNKALAGEYIDIL

KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 8 Examples of the metagenome sequences annotated in the place of ThiC

metagenome can also be compared in terms of the possessed KEGG Module entries. From the top page of a metagenome samples (e.g., T30003), the user can jump to the module page, where the thiamine biosynthesis module (M00127) is present (Fig. 9). In contrast, the human genome also has the corresponding page, but there is no such module, meaning that no gene in human

genome has been annotated to have such a function.

Conclusion

This review introduced the KEGG and GenomeNet resources, putting emphasis on the

KEGG Metagenome: T30003

The screenshot shows the KEGG Metagenome: T30003 interface. The 'Module' tab is selected, displaying a list of KEGG pathway modules assigned to the metagenome sample. The list is categorized into Energy metabolism, Nitrogen metabolism, Methane metabolism, Sulfur metabolism, and Cofactor and vitamin biosynthesis.

KEGG pathway modules

Pathway module

Energy metabolism

- Carbon fixation
 - M00165 Reductive pentose phosphate cycle (Calvin cycle) [PATH: T30003_00710]
 - M00166 Reductive pentose phosphate cycle, RuBP + CO₂ => glyceraldehyde-3P [PATH: T30003_00710]
 - M00167 Reductive pentose phosphate cycle, glyceraldehyde-3P => RuBP [PATH: T30003_00710]
 - M00168 CAM (Crassulacean acid metabolism), dark [PATH: T30003_00710]
 - M00169 CAM (Crassulacean acid metabolism), light [PATH: T30003_00710]
 - M00170 C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type [PATH: T30003_00710]
 - M00171 C4-dicarboxylic acid cycle, NAD⁺-malic enzyme type [PATH: T30003_00710]
- Nitrogen metabolism
 - M00175 Nitrogen fixation, nitrogen => ammonia [PATH: T30003_00910]
 - M00531 Assimilatory nitrate reduction, nitrate => ammonia [PATH: T30003_00910]
 - M00530 Dissimilatory nitrate reduction, nitrate => ammonia [PATH: T30003_00910]
 - M00529 Denitrification, nitrate => nitrogen [PATH: T30003_00910]
 - M00528 Nitrification, ammonia => nitrite [PATH: T30003_00910]
- Methane metabolism
 - M00174 Methane oxidation, methylotroph, methane => CO₂ [PATH: T30003_00680]
- Sulfur metabolism
 - M00176 Sulfur reduction, sulfate => H₂S [PATH: T30003_00920]

...

Cofactor and vitamin biosynthesis

- M00127 Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P [PATH: T30003_00730]
- M00125 Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD [PATH: T30003_00740]
- M00124 Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P [PATH: T30003_00750]
- M00115 NAD biosynthesis, aspartate => NAD [PATH: T30003_00760]
- M00119 Pantothenate biosynthesis, valine/L-aspartate => pantothenate [PATH: T30003_00770]
- M00120 Coenzyme A biosynthesis, pantothenate => CoA [PATH: T30003_00770]
- M00123 Biotin biosynthesis, pimeloyl-CoA => biotin [PATH: T30003_00780]
- M00126 Tetrahydrofolate biosynthesis, GTP => THF [PATH: T30003_00790 T30003_00670]
- M00121 Heme biosynthesis, glutamate => protoheme/siroheme [PATH: T30003_00860]
- M00129 Ascorbate biosynthesis, animals, glucose-1P => ascorbate [PATH: T30003_00040 T30003_00053]
- M00114 Ascorbate biosynthesis, plants, glucose-6P => ascorbate [PATH: T30003_00010 T30003_00051 T30003_00053]

KEGG and GenomeNet, New Developments, Metagenomic Analysis, Fig. 9 KEGG Module entries assigned for a metagenome sample

usage for metagenomics studies. Their focus on metagenomes has just begun; however, they plan on developing novel user-oriented tools designed for discovery and analysis of metagenomic data. For further reading, some other publications are recommended (Wheelock et al. 2009a, b; Tokimatsu et al. 2011; Kotera et al. 2012) explaining other contents that are not mentioned in this review. The authors appreciate any suggestions, questions, and comments on KEGG and GenomeNet. Please send a message through the feedback form (<http://www.genome.jp/feedback/>).

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Customizable Web Server for Fast Metagenomic Sequence Analysis](#)
- ▶ [Genome Portal, Joint Genome Institute](#)
- ▶ [GHOSTM](#)
- ▶ [Human Gut Microbial Genes by Metagenomic Sequencing](#)

- ▶ [Human Oral Microbiome Database \(HOMD\)](#)
- ▶ [MEMOsys: Platform for Genome-scale Metabolic Models](#)
- ▶ [MetaBioME](#)
- ▶ [MEtaGenome Analyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)
- ▶ [PhyloPythia\(S\)](#)
- ▶ [Variable Selection to Improve Classification of Metagenomes](#)
- ▶ [Viral MetaGenome Annotation Pipeline](#)

References

- Fujibuchi W, Sato K, Ogata H, Goto S, Kanehisa M. KEGG and DBGET/LinkDB: integration of biological relationships in divergent molecular biology data. In: Knowledge sharing across biological and medical knowledge based systems, Technical report WS-98-04. AAAI Press; 1998. p. 35–40. <http://www.aaai.org/Papers/Workshops/1998/WS-98-04/WS98-04-006.pdf>
- Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucl Acids Res.* 2002;30(1):42–6.
- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40(Database issue):D109–14. Epub 2011 Nov 10.
- Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Chapter 2 In: Wang J, Choon Tan A, Tian T, editors. *Next generation microarray bioinformatics*. Springer; 2012. ISBN 978-1-61779-399-8. doi:10.1007/978-1-61779-400-1_2 [PMID: 22130871]. http://link.springer.com/protocol/10.1007%2F978-1-61779-400-1_2
- Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 2007;35:W182–5.
- Tokimatsu T, Kotera M, Goto S, Kanehisa M. KEGG and GenomeNet resources for predicting protein function from omics data including KEGG PLANT resource. Chapter 14. In: Kihara D, editor. *Protein function prediction for omics era*. Springer; 2011. p. 271–288. http://link.springer.com/chapter/10.1007%2F978-94-007-0881-5_14
- Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, van Erk M, Kleemann R, Haeggstrom JZ, Goto S. Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol Biosyst.* 2009a;5:588–602.
- Wheelock CE, Goto S, Yetukuri L, D’Alexandri FL, Klukas C, Schreiber F, Oresic M. Bioinformatics strategies for the analysis of lipids. *Methods Mol Biol.* 2009b;580:339–68.

Krona: Interactive Metagenomic Visualization in a Web Browser

Brian D. Ondov, Nicholas H. Bergman and Adam M. Phillippy
National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

Abbreviations

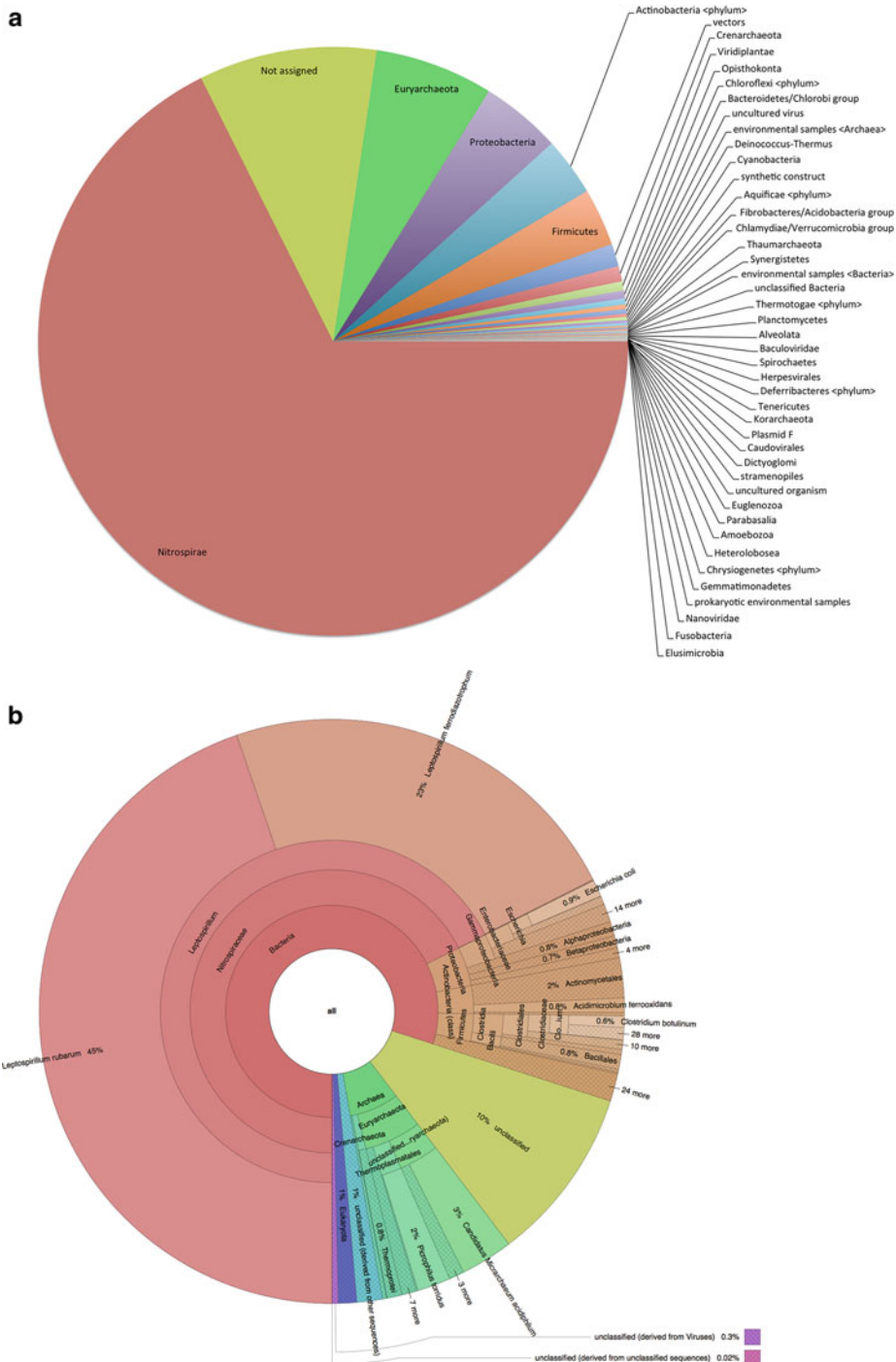
BLAST	Basic Local Alignment Search Tool
HTML	HyperText Markup Language
NCBI	National Center for Biotechnology Information
RDP	Ribosomal Database Project
XML	eXtensible Markup Language

Definition

Krona is an interactive visualization tool for exploring the composition of metagenomes within a Web browser.

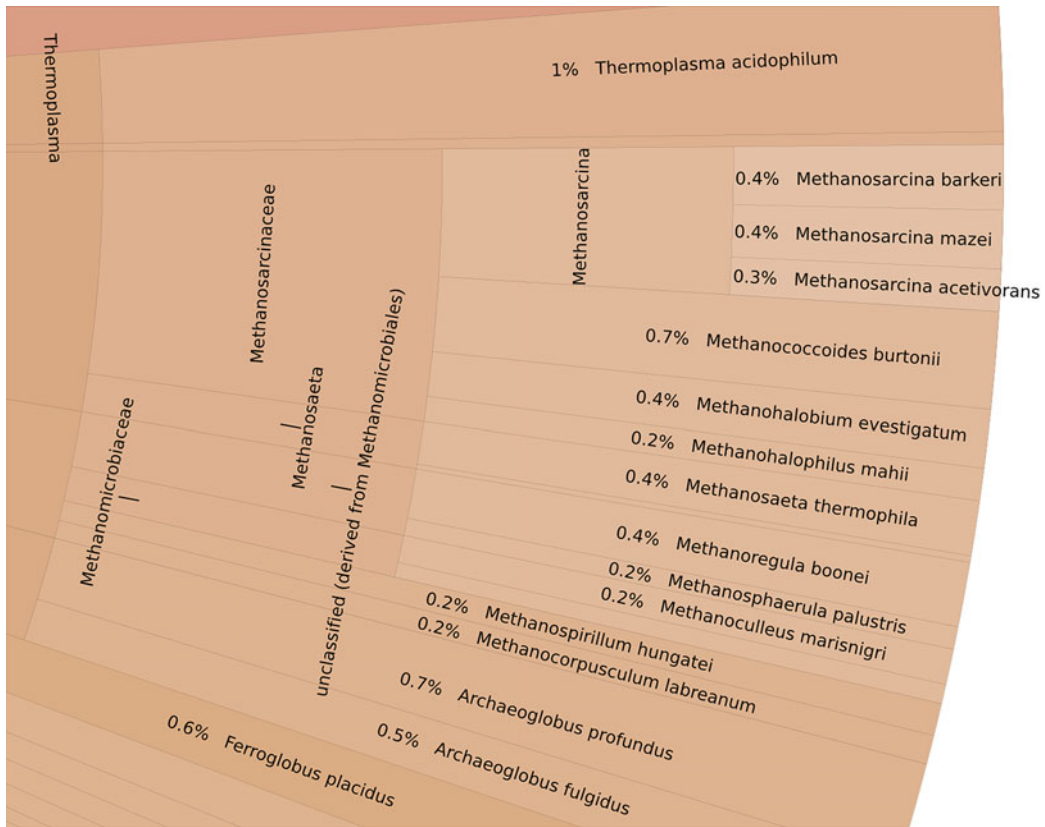
Introduction

Much of the research performed in metagenomics is exploratory, making visualization a prominent aspect of the field. Graphically representing a metagenome, however, is not a trivial task. A single sample can easily contain too many species to represent in one figure, and classifications are not always specific. This often forces visualizations to summarize the sample at higher ranks, such as genus or family, trading details for a meaningful overview. Though user interaction can typically reveal more specific classifications,



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 1 Types of overviews. The traditional pie chart (a) shows abundances of organisms in a metagenome, summarized at the phylum level. Many phyla are still too small to compare, while genus- and species-level classifications for the larger phyla cannot

be seen even though they would be large enough. The multilayer pie chart (b) depicts ranks more dynamically, dividing high-level classifications into more specific ones toward the outside of the circle. This allows more details to be shown for large phyla while small phyla are grouped and labeled



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 2 Zoomed multilayer pie chart. Standard zooming can show more detail for a region of a multilayer pie chart, but can move the center off screen

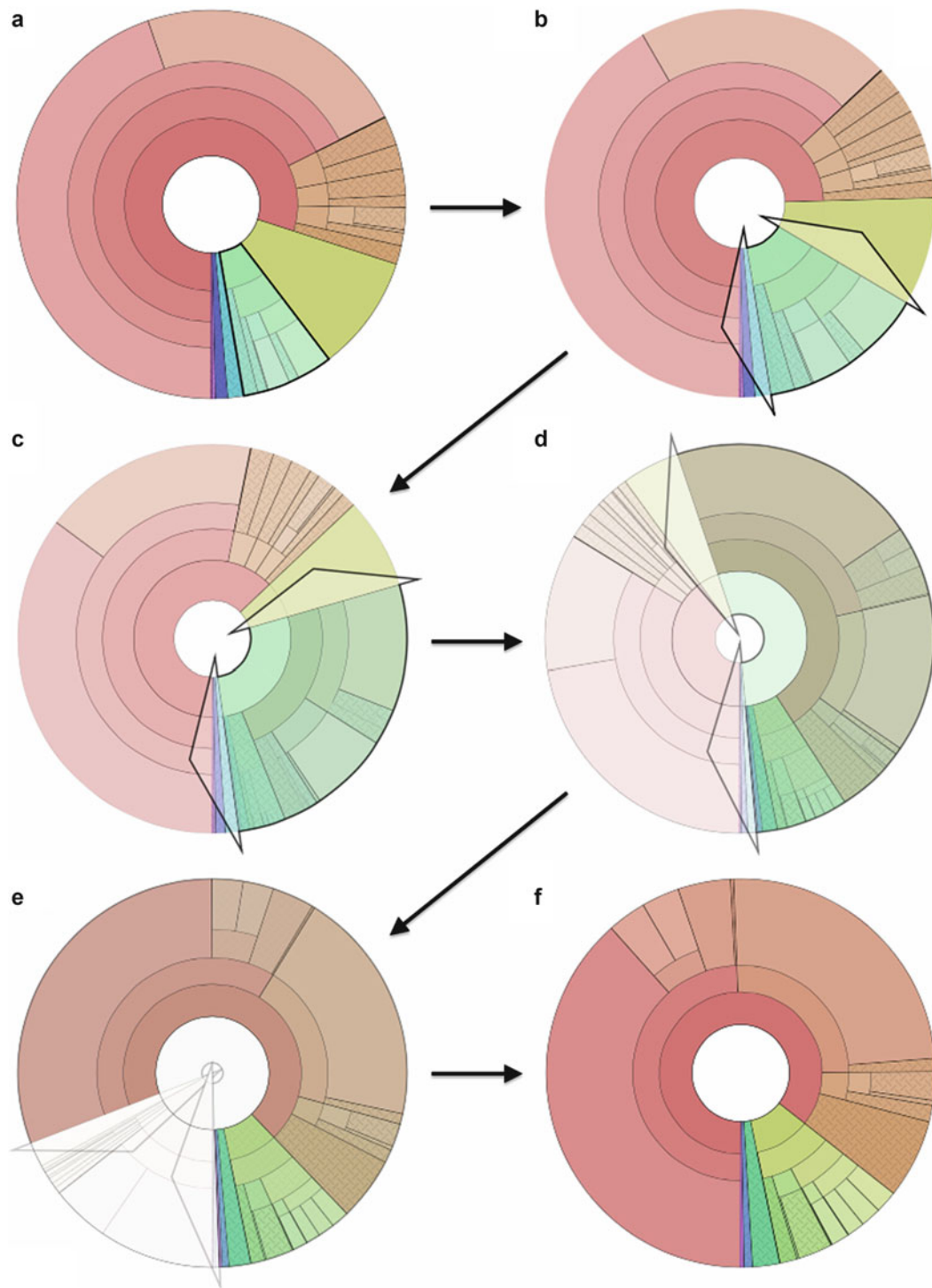
and cause wedges to become nearly rectangular. As a result, it is less intuitive to discern relative abundances and hierarchical organization

there is still a trade-off between comparing the most abundant organisms and viewing their most specific classifications (Huson et al. 2007; Meyer et al. 2008). **Krona** uses multilevel pie charts to visualize both the most abundant organisms and their most specific classifications (Fig. 1). Rather than hiding lower ranks in its overview, **Krona** hides low-abundance organisms, which can be expanded interactively. Additionally, **Krona**'s browser-based implementation allows it to be much more portable than other interactive metagenomic visualization tools.

Overviews and Details

Interactive visualizations can make complex results more accessible by providing both

high-level overviews and detailed views of specific portions as needed (Shneiderman 2002). Though an overview can (and usually must) omit some complexity, this view helps users determine which areas to view in further detail and provides context as they browse between sections. Multilevel pie charts are a good option for metagenomic overviews because they can convey hierarchy implicitly, nesting lower-level wedges within higher ones (Draper et al. 2009). This allows the abundances of multiple levels to be shown in the same view and using the same scale. As in other metagenomic visualizations, some nodes will have to be hidden for the overview to be informative. The benefit of multilevel pie charts is that the nodes are hidden based on abundance rather than specificity of their classifications. This gives priority to nodes that make



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 3 Polar zooming. Zooming in polar space allows the zoomed region to retain the intuitive properties of the original multilayer pie chart.

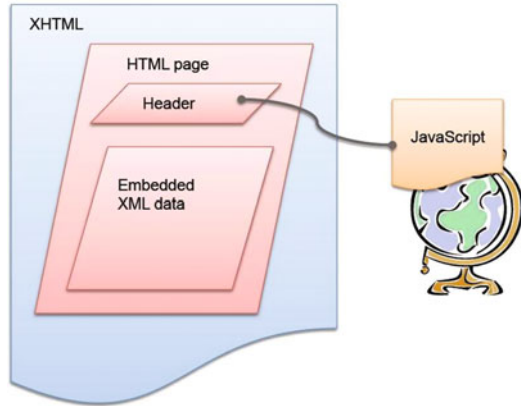
A wedge in the overview (a, green) is stretched around the center (b–e) until it fills the entire circle (f). The detailed view also serves as a new overview from which the process can be repeated with smaller wedges

up the greatest portion of the sample, which are typically of the most interest. A potential drawback, however, is that simply zooming in on the smaller nodes would cause them to lose their resemblance to a pie chart (Fig. 2). Krona avoids this problem using polar zooming, in which a wedge is stretched around the center until it forms a new multilayer pie chart (Fig. 3). The zoomed in view also serves as a new overview for further zooming, allowing even complex hierarchies to be explored with only a small amount of navigation.

Interactivity Without Installation

Since researchers often use visualizations to convey data to others, portability is an essential feature of visualization software. In the past, interactive features were typically at odds with portability because they required software to be installed. However, thanks to technologies such as JavaScript and HTML5, the modern Web browser has become a ubiquitous, standardized platform for interactivity. Many software packages are now entirely Web based, hosting both tools and data on centralized servers. While this “cloud computing” model offers many advantages, it also creates a dependency on those servers and an obligation for the software developers to maintain and scale them. Furthermore, it requires researchers to store their data remotely, which may not always be desirable. Krona offers a compromise in which each chart is a locally stored Web page, in the form of a single HTML file with embedded XML data. When this file is opened in a Web browser, viewing code is fetched from the Internet, allowing the data to be viewed interactively without installing software or using remote storage (Fig. 4). Krona charts can easily be shared with anyone that has an Internet connection and a modern Web browser. They can also be embedded in existing Web pages without modifying the server. For cases in which an Internet connection is not available, Krona charts can still be viewed locally, but require installation.

Krona Chart



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 4 Krona architecture. Embedding XML data within an HTML document and linking to remote JavaScript allows a hybrid of Web-based interactivity and locally stored data

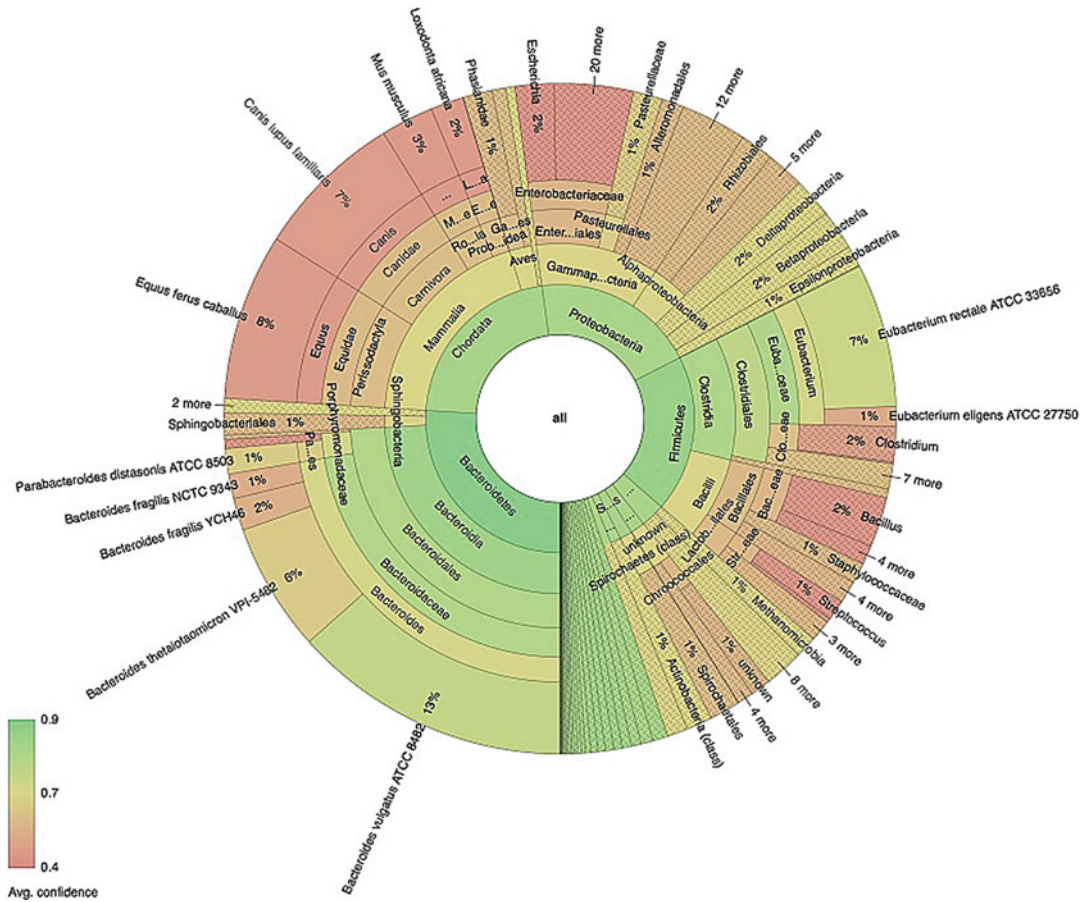
Showing More Information

Confidence

Metagenomic classification algorithms are constantly improving, but their results still come with a significant degree of uncertainty. Only a small fraction of the tree of life is represented in reference databases, and this causes widespread bias in classifications (Wooley et al. 2010). It is thus important to consider classification confidence, whenever it is available, when analyzing classificatory results. Krona can vary wedge coloring to visualize classification confidence in tandem with abundances (Fig. 5).

Comparison

Metagenomic studies often compare differences in metagenomes sampled from multiple locations or times. Though direct comparison of samples is infeasible for multilayer pie charts, it is possible to convey differences through animation and color. To show animated differences, the chart can be morphed from one sample to the next, causing wedges that change significantly in size to draw attention from the user (Fig. 6). To show the differences with color, each wedge can be colored based on how much it varies between



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 5 Classification confidence. Classification confidence is mapped to a gradient from

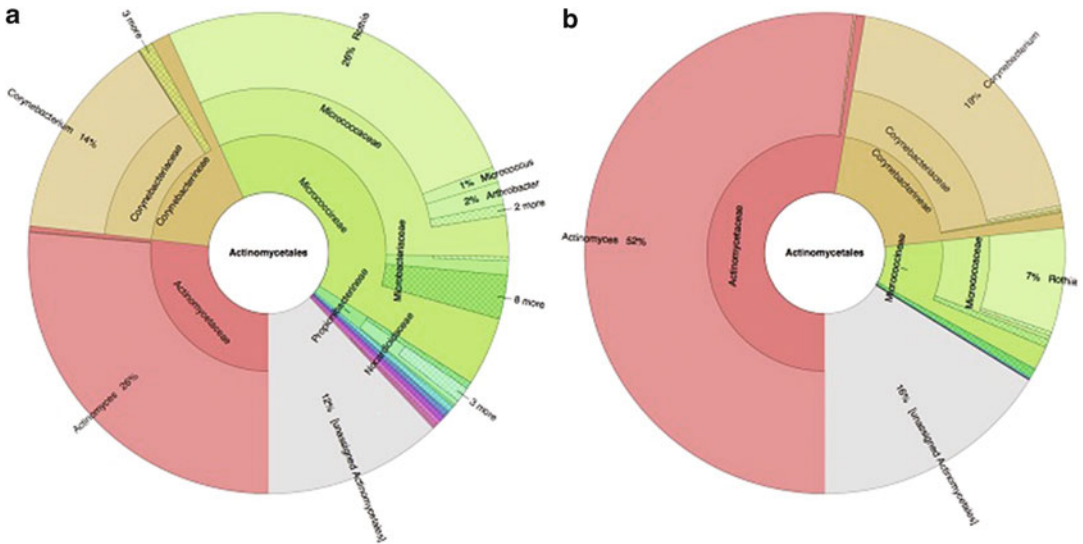
red (signifying low confidence) to *green* (signifying high confidence), allowing it to be depicted in tandem with abundance and hierarchy

samples. These two methods can also be combined to provide a clearer picture of sample variation.

Applications in Metagenomics

Metagenomic analyses typically produce data from one of the two categories: taxonomic and functional. Taxonomic classifications, which place sequences on the tree of life, are inherently hierarchical because of the various ranks in the tree (species, genus, etc.). Functional classifications, which describe the roles of predicted proteins, are often made hierarchical by grouping specific functions into more general ones. Since

both types of data focus on quantities within hierarchies, both are suited to visualization with Krona charts. To create Krona HTML files from these data, many common formats can be imported with **KronaTools**, a software package for Unix-based systems. Classifications can be directly imported from the RDP Classifier, Phymm/PhymmBL, FCP, MG-RAST, or the Web-based bioinformatics platform Galaxy. For raw BLAST results downloaded from NCBI or the METAREP metagenomic repository, KronaTools performs MEGAN-like (lowest common ancestor) classification using NCBI taxonomy information. When importing



Krona: Interactive Metagenomic Visualization in a Web Browser, Fig. 6 Comparing datasets. Differences between samples are shown with an animated transition from one sample (a) to the next (b). The persistence

of wedge coloring between samples helps the user keep track of individual wedges and draws attention to ones that change by significant amounts

K

classifications from RDP and PhymmBL, a color gradient can be used to represent the average reported confidence of assignments to each node. For MG-RAST, METAREP, and raw BLAST results, the nodes can be colored by e-value, score, or percent identity. Since classifications can sometimes be performed on assembled contigs rather than reads, KronaTools can be given contig magnitudes to more accurately convey abundance in the chart. To extend KronaTools to formats that are not yet supported, it can also import generic tabular files containing NCBI Taxonomy Identifiers or Enzyme Commission numbers. Other types of classifications can be imported from basic text files or an Excel template detailing lineage and magnitude. Finally, a custom XML file can be imported to gain complete control over the chart, including custom attributes and colors for each node. Since node attributes can contain HTML and hyperlinks, XML import allows Krona to be deployed as a custom data browsing and extraction platform in addition to a visualization tool.

Summary

Krona enables the interactive visualization of complex metagenomic data without installed software or cloud computing resources. It uses multilayer pie charts to provide overviews that emphasize the most abundant members of a sample, while its polar zooming intuitively provides details for the least abundant. Supplementary data, such as classification confidence and sample variation, can be conveyed through color and animation. Krona's hybrid Web/local architecture allows each interactive chart to be a single file, viewable on any computer with an Internet connection and a modern Web browser. Charts can be created from common metagenomic and generic file formats using KronaTools, a software package for Unix-like systems. Both Krona and KronaTools are freely available under a BSD open-source license and available from <http://krona.sourceforge.net>.

Acknowledgments This publication was developed and funded under agreement no. HSHQDC-07-C-00020 awarded by the US Department of Homeland Security for the management and operation of the National

Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Homeland Security. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication.

Cross-References

- ▶ [METAREP, Overview](#)
- ▶ [MEtaGenome ANalyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [Novel Alkalistable and Thermostable Xylanase-Encoding Gene \(Mxyl\) Retrieved from Compost-Soil Metagenome](#)

References

- Draper G, Livnat Y, Riesenfeld R. A survey of radial methods for information visualization. *Vis Comput Graph IEEE Trans.* 2009;15(5): 759–76.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17(3): 377–86.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics.* 2008;9:386.
- Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. *Visual languages*, 2002.
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol.* 2010;6(2): e1000667.

L

Lateral Gene Transfer and Microbial Diversity

Tania Nasreen, Rebecca J. Case and Yan Boucher
Department of Biological Sciences, University of
Alberta, Edmonton, AB, Canada

Synonyms

Horizontal gene transfer (HGT); Lateral gene transfer (LGT)

Definition

LGT is genetic changes within an individual or a population that occur through the acquisition of DNA from individuals that are not an organism's direct cellular parent or progenitor. One of its effects on microbial populations is to alter diversity through the acquisition of genetic material or by homogenization of a population. If molecular sequences are available for a community, statistical estimators can be used to calculate its total diversity and structure so that it can be compared to other ecosystems.

Introduction

The “uncultured majority” (Whitman et al. 1998) of prokaryotes capture the imagination, as it suggests a seemingly limitless potential for

biological diversity. This estimate of less than 1 % of prokaryotes being represented by cultures (Torsvik et al. 1990) suggests that exploration of the untapped diversity of microbial species, genes, pangenomes, metabolism, behaviors, and complex interactions will be a fruitful endeavor. However, how we discover and understand microbial diversity has been heavily influenced by LGT. We can compare any bacterial or archaeal 16S *rRNA* gene directly recovered from the environment to the comprehensive public sequence databases. This allows the identification of this gene's host based on its similarity and phylogenetic placement relative to sequences from described (and therefore cultured) prokaryotes. These sequences can also be compared to all the other 16S *rRNA* gene sequences directly retrieved from the environment; however, without a described culture, little can be inferred about their hosts physiology and thereby their role in an ecosystem. This is further complicated by the prevalence of LGT making inferences of few if any phenotypic characteristics of a species, genera, family, or phylum impossible. Therefore, a sequence rarely tells us anything about the ecology of an organism and its real value is that it can tell us something about the biological diversity of an ecosystem.

Impact of LGT on Measurements of Microbial Diversity

Diversity has been used as a metric by ecologists for decades and can be correlated with other

information to describe an ecosystem (Gravel et al. 2011). The diversity of a system is not simply the number of organisms or unique DNA sequences identified. Probability-based estimators can be used to extrapolate the total diversity from subsampling the diversity of operational taxonomic units (OTUs) defined as a similarity threshold of the 16S *rRNA* gene sequence. This can be done for populations with parametric (e.g., a rarefaction) or nonparametric (e.g., Chao1) distributions. This is analogous to capture-recapture methods of determining the population size of animals. For example, to determine the population size of swamp wallabies, several wallaby's ears are tagged within a population and subsequent sampling of the population can be used to estimate its size by calculating the probability of recapturing tagged wallabies among non-tagged wallabies. Molecular microbial ecology is much more powerful than such macroecology studies as it rarely focuses on a single species, but rather the total bacterial and/or archaeal community and the numerous populations that encompass thousands of species. Diversity estimators are then used to calculate the total diversity and structure of the community using indices such as Simpson's Diversity Index (proportional distribution of all species), species evenness (distribution of individuals among species), and Shannon Index (entropy of community measured from the richness and evenness of community). These indices allow us to compare natural and experimental communities to identify factors that influence diversity such as the volume of water in tree holes (Bell et al. 2005) or a chronosequence within a lichen (Mushegian et al. 2011).

Microbial systems rarely have a perceived intrinsic value in that people do not marvel at a termite's hindgut as they do old growth forests. Their value is in what they do, their function. Diversity is a powerful measurement in microbial ecology as it has a major influence on the productivity and stability (or resilience) of an ecosystem (Gravel et al. 2011). The indices described above are useful in characterizing these systems as it can be used to compare their productivity. However, in microbial systems, the productivity is not

often studied (with the exception of phytoplankton in aquatic systems). Often a specific process is of interest, such as degradation of xenobiotics or denitrification. This presents one of the biggest dilemmas for microbial ecologists as they cannot study a phylogenetic group (for which there are many 16S *rRNA*-based primers and probes that could be used in targeted studies) and infer the function of the group (Case et al. 2007). Macroecologists can infer that plants are primary producers at the base of the food web and provide shelter for other species as habitat-forming species, which is not possible for microecologists. This is the result of LGT, as this phenomenon facilitates the movement of genes among phylogenetically distant organisms. This means that phylogeny based on universal marker genes such as 16S *rRNA* is not a predictive tool of function in microbiology.

Molecular methods have been adapted to circumvent this conundrum so that functional genes (such as *hupL* for hydrogen oxidation) can be directly targeted through PCR (Balskus et al. 2011). Such functional genes can then be used in community fingerprinting, clone libraries, or CARD-FISH, which has been adapted to identify mRNA to look at expression of specific genes inside cells. Such gene-omic (sequencing of a single marker gene directly from an environmental sample) approaches are popular for targeted studies and can be adapted to high-throughput sequencing techniques. Datasets that include deep sequencing of a gene involved with a specific function can be used to identify redundancy in a system. Such redundancy is important for the stability of an ecosystem through environmental change, as genetic redundancy represents the diversity of organisms able to perform a function within a system. The alternative to gene-omic approaches is metagenomics, whose popularity has been greatly influenced by the disconnect created by LGT between phylogeny and function. Metagenomics retrieves large nontargeted sequence datasets from an environment such that metabolic networks and interactions can be inferred from the community's metagenome. This method can be



coupled to metatranscriptomics (RNA) and/or metaproteomics (proteins) to move beyond the genetic potential of a metagenome to the transcribed and translated. These methods, however, have their greatest power when targeted or used in low-diversity systems (Hugenholtz and Tyson 2008).

Mechanisms Responsible for the Generation of Genetic Diversity in Microbes

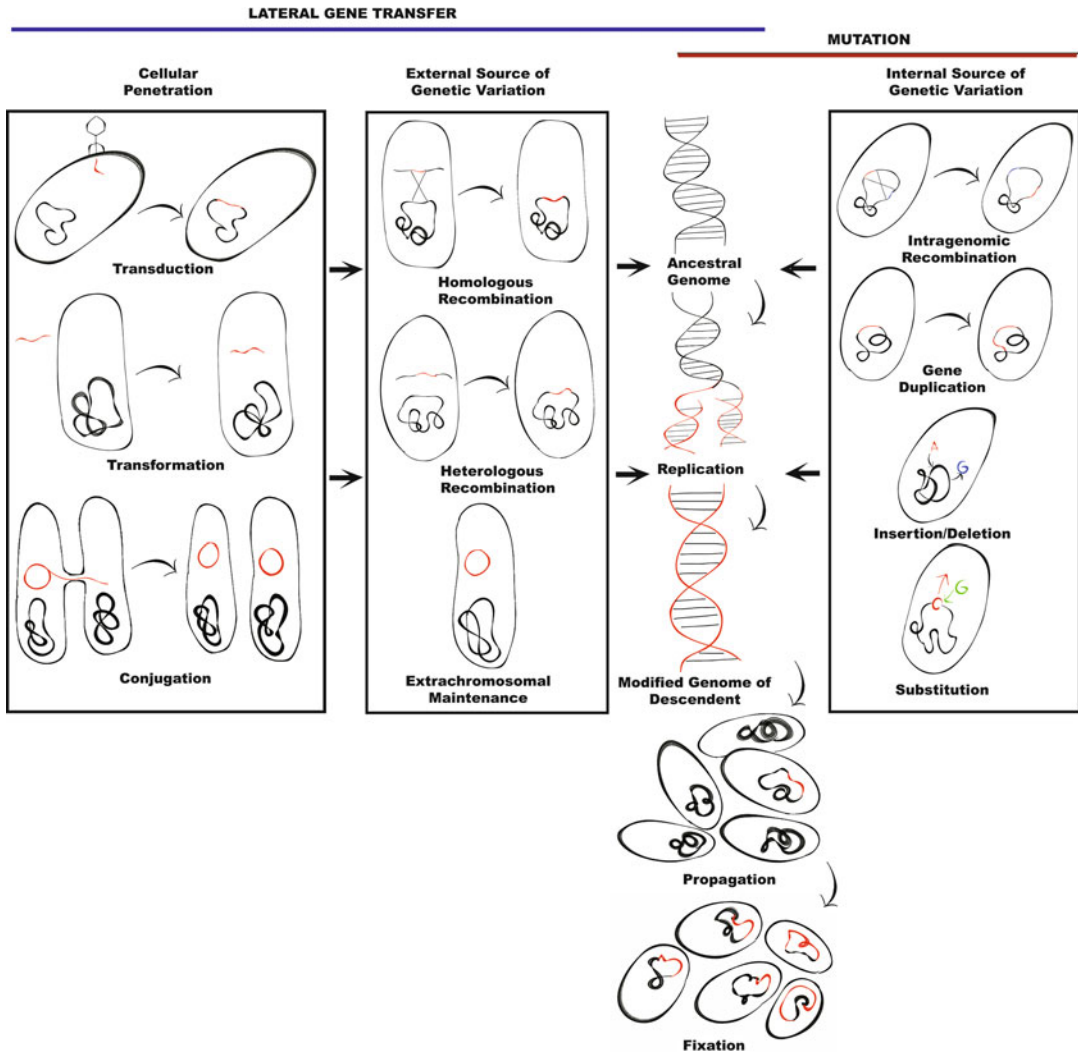
What is measured through gene-omic approaches such as the 16 *rRNA* gene or *nifH* clone libraries is *nucleotide sequence* diversity. The latter, however, is not the only type of genetic diversity. Metagenomic or genomic approaches allow the measurement of *gene content* diversity, which is the measurement of differences in the genes found in various genomes or metagenomes. Both of these are strongly affected by LGT, which influences not only the rate at which they change but also how they change.

Sequence Diversity. The only force responsible for de novo creation of genetic diversity is mutation. It can be defined as changes in the DNA sequence of a genome that is inherited from a progenitor. The nature of such changes can vary: base pair substitutions, insertion/deletion of one or more nucleotide(s), as well as larger or more complex changes (such as chromosomal rearrangement or gene duplication) (Fig. 1). The physical causes of mutations are also diverse: unforced DNA replication errors, errors during proofreading or post-replication mismatch repair, and DNA damage leading to replication errors or inaccurate repair. Although mutation is responsible for *creating* diversity, it is not the only phenomena *introducing* variation in particular groups or lineages of microbes. Genetic changes within an individual or a population can occur through the acquisition of DNA from individuals that are not an organism's direct cellular progenitor. This process is LGT. In bacteria and archaea, it has two main steps. First, foreign DNA penetrates the cellular envelope in one of three ways:

transformation (the uptake of DNA directly from the environment or from a membrane vesicle), conjugation (cell-to-cell contact mediated by the apparatus encoded on a conjugative element or by a cytoplasmic fusion), or transduction (introduction of DNA by a phage) (Fig. 1). Second, integration into the new host genome is required, which can be achieved by homologous recombination (i.e., this requires a homologous region of DNA between the donor and recipient), heterologous recombination (i.e., that does not require a homologous region of DNA between the donor and recipient DNA), or extrachromosomal maintenance and replication.

We can now obtain minimal LGT estimates through quantification of homologous recombination. This type of LGT directly affects sequence diversity and is usually simply termed "recombination" in most molecular population studies. This is because mathematical models currently used in population genetics can only take into account changes in genetic material that is present in all members of the population, therefore excluding acquisition of novel genetic material through heterologous recombination and as extrachromosomal elements. Population recombination rates therefore only include events in which foreign DNA, through replacement of a homologous locus by recombination, is integrated in the host genome. Studies that have compared population mutation and recombination rates in various prokaryotic lineages have found a relatively even split between those in which mutation introduces most of the changes and those where homologous recombination is responsible for most nucleotide variations (sequence diversity).

Gene Content Diversity. LGT also (if not predominantly) introduces change through the acquisition of novel genetic material through heterologous recombination. This, in combination with gene loss and gene duplication, leads to changes in the gene content of an organism. For example, strains of the marine heterotrophic bacterial genera, *Vibrio*, which are identical at one or more protein-coding housekeeping gene, can be differentiated by genome size (up to 800 kb



Lateral Gene Transfer and Microbial Diversity, Fig. 1 Description of the processes generating genetic diversity in bacteria and archaea

variation) (Thompson et al. 2005). Also strains of the nitrogen-fixing soil bacteria *Frankia* that are more than 97 % identical in their rRNA gene sequences – the conventional cutoff value for a bacterial species – can differ by as many as 3,500 genes, which represents nearly half of their 7.5 Mb genomes (Normand et al. 2007).

Although gene content and sequence diversity are often correlated, it is not always the case. According to empirical data, the correlation is

hypothesized to hold for bacteria that partially overlap in their ecological niche (Konstantinidis et al. 2006). Sequence diversity dominates for bacteria with identical or almost entirely overlapping niches (little change in gene content), and gene content diversity is more pronounced when bacteria occupy separate niches. Ecological adaptation is therefore directly linked with gene content diversity but less so with sequence diversity.



Impact of LGT on the Phenotypic Diversity of Microbes

Microorganisms exhibit great diversity in their cellular structures, metabolic properties, interactions, and ecological niches. It is well established that mutation (sequence diversity) has contributed to this phenotypic diversifications of microorganisms. However, growing numbers of genomic studies suggest that LGT influences the acquisition of novel functions through its effect on gene content, not sequence, diversity. For example, recent studies of the genomic context and phylogenetic relatedness of proteorhodopsin genes suggested that they had been transferred by LGT from marine Archaea to Proteobacteria. This *single gene* is hypothesized to provide its host with a competitive advantage by allowing it to harness light energy for cellular function. As these organisms reside in the photic zone of the ocean, proteorhodopsin allows them to take full advantage of available UV energy (Frigaard et al. 2006).

In some species, most of the genetic variation and adaptation occurs through LGT. Although *Prochlorococcus* species have a conserved core of genes, they show a significant variation in the genes present on *genomic islands*. These represent the evolutionary hot spots inside their genomes. It is hypothesized that these genomic islands are acquired by LGT and undergo extensive rearrangement, suggesting a common mechanisms of niche differentiation in microbial species. The pathogenicity islands of pathogenic bacteria also share the same characteristics (Coleman et al. 2006). Some genomic island associated LGTs are thought to be mediated by *phages*, since they can carry host genome fragments. For example, the cholera toxin gene in *Vibrio cholerae* that is actually encoded within a bacteriophage (CTX ϕ) genome that necessarily needs the toxin co-regulated pilus (TcpA), an intestinal colonization factor, as its receptor. TcpA is encoded within the pathogenicity island named VP1. However, this VP1 region mainly constitutes the genome of another bacteriophage

(Faruque and Mekalanos 2012). Thus, two individual LGT events involving these two phages have the potential to make almost any *Vibrio cholerae* strain into a potent human pathogen.

Various metabolic properties, virulence, and antibiotic resistance traits can also be carried on *plasmids* or *transposons* or a combination of the two. This makes these genes more likely to be transferred through LGT. For example, Tn10 is a transposon consisting of a pair of IS10, a tetracycline determinant and a regulatory gene. Similarly, transposon Tn5 consists of two IS50 elements and a three-gene operon that attributes resistance to kanamycin, bleomycin, and streptomycin. Both of these transposons can be incorporated into the chromosomes of phylogenetically diverse groups of bacteria. Plasmids are the other major mediator of antibiotic resistance gene acquisition by LGT. Not only are plasmids themselves transfer agents, but they can also change rapidly through LGT. For example, based on gene organization and sequence similarity, plasmid pKF3-140 found in *Klebsiella pneumoniae* has been speculated to have originated from *Escherichia coli* (plasmids p1ESCUM and pUTI89) and further modified by acquiring resistance genes from different enteric bacteria by LGT.

Another genetic element facilitating LGT and phenotypic diversity is the integron. This genetic element carries genes for site-specific recombination known as mobile gene cassettes in the host genome. It has been found that about 17 % of the sequenced bacterial genomes have integrons. For example, many species of *Pseudomonas* contain integrons with a variable number of gene cassettes (10–32) that are considered to have been obtained by LGT at the late stage of species segregation (Vaisvila et al. 2001).

These are only a few representative examples of the contribution of LGT to the phenotypic and genotypic diversity of microbial populations. Importantly, this diversity is not only driven by natural selection. Microbes have evolved the ability to sense the environments and generate

diversity as a response to a stressor. For example, the transfer of genomic islands encoding specific metabolic properties is sometimes controlled by quorum sensing mechanisms. The genomic island ICEMISym^{R7A} of *Mesorhizobium loti* strain R7A encodes proteins required for symbiotic nitrogen fixation and that regulate the transfer of plasmid by quorum sensing to nonsymbiotic mesorhizobia (Ramsay et al. 2009). Another example of stressor-generated genotypic diversity is CRISPRs. These elements are considered to be an acquired immune system against virus and plasmids by which the host identifies foreign DNA in a sequence specific manner (Horvath and Barrangou 2010). Experimental evidence of CRISPR-mediated immunity to bacteriophages has been shown in *Streptococcus thermophilus*. After exposure to a phage to which *S. thermophilus* was susceptible, only a small fraction of cells survived, but the genome of the survivors had acquired novel sequences in their CRISPR loci identical to the DNA of the infecting phage. This a genomic change directly triggered by an environmental factor. Similarly, the SOS response, a global regulatory network that is activated in response to DNA damage, has recently been discovered to induce recombination activity integrons. This causes an increased acquisition of gene cassettes, potentially encoding novel phenotypes. This creates a link between environmental factors inducing the SOS responses such as oxidative stress, pH change, and exposure to antibiotics and genetic diversity (Guerin et al. 2009).

Cross-References

- ▶ [Metagenomic Potential for Understanding Horizontal Gene Transfer](#)
- ▶ [Mining Metagenomic Datasets for Antibiotic Resistance Genes](#)
- ▶ [Novel approaches to Pathogen Discovery in Metagenomes](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies](#)

References

- Balskus EP, Case RJ, Walsh CT. The biosynthesis of cyanobacterial sunscreen scytonemin in intertidal microbial mat communities. *FEMS Microbiol Ecol.* 2011;77(2):322–32.
- Bell T, Ager D, Song JI, et al. Larger islands house more bacterial taxa. *Science.* 2005;308(5730):1884.
- Case RJ, Boucher Y, Dahllöf I, Holmstrom C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol.* 2007;73(1):278–88.
- Coleman ML, Sullivan MB, Martiny AC, et al. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science.* 2006;311(5768):1768–70.
- Faruque SM, Mekalanos JJ. Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. *Virulence.* 2012;3(7):556–65.
- Frigaard NU, Martinez A, Mincer TJ, DeLong EF. Proteorhodopsin lateral gene transfer between marine planktonic bacteria and archaea. *Nature.* 2006;439(7078):847–50.
- Gravel D, Bell T, Barbera C, et al. Experimental niche evolution alters the strength of the diversity-productivity relationship. *Nature.* 2011;469(7328):89–92.
- Guerin E, Cambray G, Sanchez-Alberola N, et al. The SOS response controls integron recombination. *Science.* 2009;324(5930):1034.
- Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science.* 2010;327(5962):167–70.
- Hugenholtz P, Tyson GW. Microbiology: metagenomics. *Nature.* 2008;455(7212):481–3.
- Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. *Philos Trans Roy Soc London B Biol Sci.* 2006;361(1475):1929–40.
- Mushegian AA, Peterson CN, Baker CC, Pringle A. Bacterial diversity across individual lichens. *Appl Environ Microbiol.* 2011;77(12):4249–52.
- Normand P, Lapiere P, Tisa LS, et al. Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res.* 2007;17(1):7–15.
- Ramsay JP, Sullivan JT, Jambari N, et al. A LuxRI-family regulatory system controls excision and transfer of the *Mesorhizobium loti* strain R7A symbiosis island by activating expression of two conserved hypothetical genes. *Mol Microbiol.* 2009;73(6):1141–55.
- Thompson JR, Pacocha S, Pharino C, et al. Genotypic diversity within a natural coastal bacterioplankton population. *Science.* 2005;307(5713):1311–3.
- Torsvik V, Goksoyr J, Daae FL. High diversity in DNA of soil bacteria. *Appl Environ Microbiol.* 1990;56(3):782–7.
- Vaisvila R, Morgan RD, Posfai J, Raleigh EA. Discovery and distribution of super-integrons among pseudomonads. *Mol Microbiol.* 2001;42(3):587–601.
- Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A.* 1998;95(12):6578–83.



Lessons Learned from Simulated Metagenomic Datasets

Germán Bonilla-Rosso
Laboratorio de Evolución Molecular
y Experimental, Instituto de Ecología UNAM,
Universidad Nacional Autónoma de México,
Mexico City, Mexico

Definition

A simulation is the dynamic modeling of a real process over time. A simulated metagenomic dataset is the product of a single simulation iteration of the sequencing process of a microbial community under a specific set of sequencing-platform model parameters.

Summary

The use of simulations to produce model metagenomic datasets allows to test the performance of technological methodologies and the testing of theoretical hypothesis that cannot be achieved by empirical experimentation. Methodologically, it has been used to evaluate the performance of assembly programs and the effect of differences of read length and error rate on the quality of the resulting datasets. Theoretically, it has revealed biases and heterogeneity in the estimation of several diversity metrics from metagenomic samples. However, the full potential of the implementation of simulated datasets to metagenomics is still to be revealed.

Introduction

The complexity of microbial communities, and the nature of the metagenomic datasets resulting from their sequencing, belongs to systems with high nested complexity. To analyze them, there is a growing need to test the robustness of new methodological and analytical tools (Angly et al. 2012). The evaluation of these tests could in theory be done by the construction of *in vitro*

communities (Morgan et al. 2010), but this approach is expensive, time-consuming, and limited to communities of reduced complexity, so the alternative presented is to apply mathematical models and simulations to test the robustness of the tools for their analysis (Caswell 1988).

A simulation is the imitation of a natural time-ordered sequence of states a system takes in a given time period with another that is the product of a representative model (Peck 2008). In other words, they are the dynamic imitation of natural processes that follow the changing states of a system under a particular theoretical model. Simulations are used principally because the equations in the models cannot be followed in time, but the individual states in the processes defined by the model are. The models create virtual worlds, with rules defined by the model parameters, that can be modified and followed in ways that would be too costly or unethical in real systems, and the simulations that can be run in these modeled worlds can be seen as individual experimental systems (Winsberg 2003). Most commonly, simulations are theoretical models used to explain natural phenomena and test the outcome of theoretical hypotheses (Caswell 1988), often used in computational biology to numerically estimate the behavior of a system that is too complex to be resolved by analytical solutions by generating a sample of scenarios that represent stochastically distinct moments of the same state of a modeled system under particular conditions (Peck 2008).

Since no ecological community (microbial or otherwise) has been sampled to exhaustion, and no completely and accurately annotated metagenome is available (Mende et al. 2012), the construction of simulated datasets rely on the available genomic data from the complete genomes of individual datasets. These simulated datasets have been used to date for two main purposes: the test of the sequencing performance of different platforms and their processing pipelines and the analysis of the accuracy of a diverse set of alpha and beta diversity estimations.

Simulation of Metagenomic Datasets

To simulate the production of a metagenomic dataset, a program needs three basic components: a pool of reference sequences, usually annotated complete genomes, from where sequences will be drawn; a profile that details the composition (taxonomic assignment of species) and structure (relative abundance of species) of the designed source community; and an error model that specifies how variability will be introduced to the simulation, and usually accounts for sequencing-platform-associated errors and rates for mutation-introduction. In recent years, several different simulation software programs have been made available that differ in the type of sequencing platform supported and the adjustable parameters to model errors. The nature of the first and most commonly used simulated dataset, and the two most commonly used and representative software programs available for metagenomic dataset simulations, is reviewed in the following section.

The FAMeS Dataset

The first metagenomic simulated dataset was produced by the group led by Konstantinos Mavrommatis at Department of Energy Joint Genome Institute (JGI), with the objective of benchmarking the alternative metagenomic processing pipelines commonly used in the JGI sequencing facility (Mavrommatis et al. 2007). They randomly selected reads from the complete genome projects of 113 isolates sequenced at JGI as their pool of sequences. Since these reads were derived from the real clone libraries in the shotgun sequencing process, they incorporate real errors derived from the Sanger sequencing method. Three different artificial source communities were designed with contrasting structure and composition reflecting the following: a low complexity community with a single dominant near-clonal population like that found in bioreactors (simLC), a moderately complex community with few dominant populations and several low abundance ones like those observed in the acid mine drainage biofilms (simMC), and a high complexity community with no dominant populations such as those observed in soils and

microbial mats (simHC). These datasets were deposited and made available online as part of the Fidelity of Analysis of Metagenomic Samples program (FAMeS:<http://fames.jgi-psf.org/index.html>) as an attempt to standardize the benchmarking of metagenomic assembly and annotation tools.

These datasets are very atypical in that they were simulated from real sequencing reads, so that the sampling step from genomes was not simulated. As such, they only model the error distribution of the Sanger sequencing platform as implemented by the JGI's particular shotgun sequencing process and prevent their extrapolation to other sequencing platforms and error models. Moreover, the lack of replication, the fixed species richness (the 113 isolate genomes from the pool), and the reduced and arbitrary complexity range of the source community profiles prevent their use for testing more ecological hypothesis regarding contrasting either species richness or gradients in structure complexity. These datasets, however, introduced the concept of benchmarking metagenomic analysis pipelines with simulated datasets, and more recent studies have used their community profiles for the construction of new simulations with replications and their extrapolation to different sequencing platforms (Mitra et al. 2010; Charuvaka and Rangwala 2011; Pignatelli and Moya 2011).

MetaSim

One of the first computer programs developed specifically to simulate metagenomic datasets was developed and is maintained by the group led by Daniel Huson at the University of Tübingen (Richter et al. 2008). It has been widely used both because of its efficient algorithm and the benefit of having a GUI. MetaSim uses complete genomes as a reference pool of sequences and by default can take advantage of the complete genomes available at the NCBI RefSeq database. This also allows the use of NCBI's taxonomy to construct the source community profiles, either by providing a relative abundance matrix or by using an interactive graphic taxonomy tree to select the genomes to be included. Finally, MetaSim provides a large array of adjustable error-model configurations



such as read length, sequencing depth, error rate, and error distribution. MetaSim includes a variety of default error models for the three main sequencing platforms (Sanger, 454, Illumina) that can be easily modified.

MetaSim allows the user to easily develop complex designs of source community profiles by specifying the richness, structure, and composition of a community via a species-abundance matrix. These profiles can be saved and several simulations can be run, allowing the comparison of simulated datasets from the same sample under different sequencing platforms and error models. As such, its main limitation is then its dependency on the available reference genomes and their associated taxonomy at NCBI, which to July 2012 contains more than 2,000 genomes.

Finally, MetaSim includes a tool to simulate sampling from a set of “evolved” genome offsprings derived from the reference genomes using an evolutionary tree. That is, it simulates real metagenomic datasets that usually contain populations of organisms with different degrees of relatedness to the available reference genomes, in a way that it simulates genetic variability in real, natural populations.

Grinder

While most metagenomic dataset simulators were developed under a vision of metagenomic process benchmarking, Grinder, developed and maintained by Florent Angly at the Australian Centre for Ecogenomics (Angly et al. 2012), is the first simulator with a more ecologically oriented perspective. Its main novel feature is that it can also simulate amplicon datasets, addressing the need to benchmark the tools for the analysis of 16S rRNA amplicon datasets that are widely used in microbial ecology. As a pool of reference sequences, Grinder can use any sequence database with FASTA format, like the NCBI RefSeq genomes database for metagenome datasets, and GreenGenes or SILVA for the amplicon datasets. Grinder supports error models for the three main sequencing platforms (Sanger, 454, Illumina) and allows the implementation of user-defined error models. It allows for the adjustment of error-model configurations such as genome size bias,

read length, sequencing depth, substitution and error distribution, and homopolymer and read end error rates for metagenomic datasets, and chimera production and gene copy number for amplicon datasets (Angly et al. 2012).

Grinder accepts two different methods to provide community profiles. The first is the canonical species-abundance matrix where the user simultaneously defines community composition and structure. The second one is by defining the community richness and a rank abundance model for the relative abundance distribution of species. Composition will be, however, selected randomly from the species list. Moreover, multiple datasets can be produced simultaneously from the same profile, both for replication purposes when source communities are identical and to simulate the sampling of related communities with a defined percentage of shared species (beta diversity) (Angly et al. 2012).

Lessons Learned from Simulated Metagenomic Datasets

Benchmarking of Technical Aspects

As explained above, the first simulated metagenome comprising the FAMEs dataset (Mavromatis et al. 2007) was developed to evaluate the fidelity of the sequencing processing pipeline regarding the assembly and gene prediction of metagenomes derived from shotgun sequencing. They revealed that the application of common single-isolate genome assemblers resulted in a low incorporation of reads into contigs and a high degree of chimeric contigs, which in turn can lead to up to 20 % of inaccurately called genes in metagenomes and errors in functional and taxonomic annotations (Mavromatis et al. 2007). Although the pipelines and sequencing platform addressed by Mavromatis et al. (2007) are outdated, several recent studies have confirmed their findings on the low performance of metagenome assemblers with communities that are more complex than a few dominant clonal populations, either with new sequencing platforms (Pignatelli and Moya 2011; Mende et al. 2012) or alternative assembly

methods (Pignatelli and Moya 2011; Charuvaka and Rangwala 2011; Mende et al. 2012).

The effect of average read length on gene annotation has been addressed by Wommack et al. (2008). They simulated the subsampling of existing Sanger-sequenced metagenomic datasets producing shorter (<400 bp) reads characteristic of the next-generation sequencing technologies 454 and Illumina. Their simulations revealed that short reads can miss up to 72 % of the annotated functions revealed by longer (~750 bp) Sanger reads and can detect only highly conserved sequences with phylogenetically close relatives in reference databases (Wommack et al. 2008). The simulations also indicate that even an increase in sampling depth with short reads (as promised by the Illumina platform) does not improve the annotation achieved by long reads. In addition, a related study using simulated datasets to assess the effect of sequencing error on gene prediction (Hoff 2009) revealed that all metagenomic gene prediction tools show a reduced accuracy at gene calling with increasing sequencing error rates and that their individual performance seems to be affected by the taxonomic composition of the samples, except when using Sanger reads with error rates below 0.15 % (Hoff 2009). Pignatelli and Moya (2011) adapted the FAMeS community profiles to the 454 and Illumina sequencing platforms and at a deeper sequencing coverage and demonstrated that all *de novo* assemblers produce a significant amount of chimeric contigs (up to 10 %) that have a profound impact on the functional and phylogenetic annotation of metagenomic sequences. Since domain and motif databases like Pfam and TIGRFam rely on short conserved sequences, they may give better annotations at a more functionally general annotation (Pignatelli and Moya 2011).

All these studies reveal that the assembly of metagenomic datasets is highly influenced by the community composition complexity, depth of sequencing coverage, and average length of the sequenced reads, discouraging the assembly of metagenomic datasets. Nevertheless, the recent development of software specifically designed

for the assembly of metagenomic datasets like Genovo (Laserson et al. 2011), IDBA-UD (Peng et al. 2012), and MetaVelvet (Namiki et al. 2012) shows a promising improvement in metagenomic assembly, although only for low complexity with communities with phylogenetically distant members. An approach that should be used in all assembly benchmarking studies is the comparison of the assembly obtained with the mixed simulated metagenomic dataset against the assembly obtained with an independent assembly of each species since most simulated datasets are produced from the annotated complete genomes from isolates, as done by Charuvaka and Rangwala (2011) and Namiki et al. (2012).

Evaluation of Ecological Aspects

Computer simulations have been long used in community ecology for modeling communities (Garfinkel 1962) and testing the performance of diversity indexes (e.g., Heltshe and Forrester 1983). But the use of computer-simulated datasets to study the diversity of microbial communities had to wait until molecular methods were available to study microbial communities (Liu et al. 1997; Bent and Forney 2008). Simulated communities are the only option to test the performance of diversity metrics on metagenomic datasets, since currently no natural community has been sampled to exhaustion and hence no real diversity measure is accurately known that we can compare our estimations against. The design of an artificial community *in vitro* and its subsequent sequencing (Morgan et al. 2010) is at best methodologically and economically unfeasible to test the performance of several replicated datasets (Angly et al. 2012). Three published studies exist that use simulated datasets to evaluate the performance of community diversity metrics, two of which deal with 16S rRNA amplicon-derived datasets (Kuczynski et al. 2010; Parks and Beiko 2012) and one with metagenomic datasets (Bonilla-Rosso et al. 2012).

Bent and Forney (2008) were the first to implement large-scale sequencing simulations to evaluate alpha diversity (species diversity in



individual samples) metrics from 16S rRNA amplicon clone libraries and T-RFLPs. They demonstrated that most alpha diversity metrics are sensitive to the number of rare and uncommon species, which are precisely the ones likely to be undersampled by 16S rRNA amplicon-based techniques (the so-called tragedy of the uncommons). Moreover, they show that different methods applied on the same community can produce radically different estimations for these metrics (Bent and Forney 2008). Using a replicated simulated dataset of nine communities in a cross-gradient of species richness and dominance, Bonilla-Rosso et al. (2012) demonstrated that the use of conserved protein genes in metagenomic datasets outperforms 16S rRNA genes at reflecting the original community. Moreover, they show that the most common alpha diversity metrics derived from metagenomic samples are biased because of insufficient sampling and variations in the taxonomic composition representation. These last two studies point toward the use of scale-dependent metrics such as Rényi's profiles or Hill's numbers as a better representation of alpha diversity's multidimensional nature.

Two studies have addressed the performance of beta diversity metrics (similarity in species composition between samples) with simulated datasets. The use of simulated datasets to test ecological hypotheses was first implemented with deep sequencing of 16S rRNA amplicons (Kuczynski et al. 2010). Addressing the effect of the environment on community structure, they simulated datasets to model communities that were either shaped along an environmental gradient or where the environment partitioned them into discrete clusters. They found that the patterns from environmental gradients were more easily detected than those from ecological clustering, specially when differences between clusters were subtle. Moreover, qualitative methods (richness based) performed better on clustered datasets, while quantitative methods (abundance based) performed better on gradients, so both types of methods should be applied if the underlying pattern is unknown. Finally, they

demonstrate that patterns are more readily identified with several low-coverage samples than with few deep-coverage datasets (Kuczynski et al. 2010). These results were further extrapolated for similarity metrics that incorporate phylogenetic information, and it was found that most distance metrics are highly intercorrelated, and highly robust to rooting, choice of threshold for defining OTUs and the presence of basal lineages (Parks and Beiko 2012).

Perspectives

Often obscured by the large amount of data produced, metagenomics is still a very young discipline where a consensus set of rigorously tested analytical tools is still lacking. Moreover, the rapid advance of sequencing technologies causes a constant development and diversification of their accompanying bioinformatic tools and approaches that require an objective quantification of their performance. This is worsened by the lack of theoretical understanding of the assembly, dynamics, and functioning of natural microbial communities. The use of simulated datasets after sequencing modeling is the best alternative to approach the benchmarking of technical and analytical methodologies as well as the testing of theories and hypotheses. However, a much more efficient benchmarking framework is still needed.

A set of source communities from where new datasets are to be simulated need to be consensually designed by the academic community as the minimal standard benchmarking start point, so that the comparison of the performance of bioinformatic tools across studies and sequencing platforms is achieved. This was the original intention of the FAMeS dataset (Mavromatis et al. 2007), but currently almost each new tool developed is tested against a tailored simulated dataset, in part because the three FAMeS communities cover a narrow range of community composition options. Ideally, this standard source community dataset should be designed in a way that spans a wide spectrum across three dimensions of

assembled communities consisting of number of species (richness), relative abundance (dominance), and taxonomic composition (phylogenetic relatedness). As an example, the effect of the presence of closely related strains on both the assembly and diversity estimation of a metagenomic sample is largely unknown.

Variability is a factor that should be more often considered in simulated datasets. There is a need to incorporate variation in platform-specific error models, and the incorporation of empirical thresholds for best and worst case scenarios in simulation software would greatly improve this. Moreover, due to their dynamic nature, two independent simulations from the same source community will produce a different set of datasets, and this sampling variability should be incorporated into the benchmarking and hypothesis testing process that allow the incorporation of variability in the models and the statistical testing of significant differences.

Finally, it should be noted that the potential of simulated datasets to metagenomics is far from explored, since they have mostly been used to test the performance of technical methodologies, and as mentioned by Caswell (1988), they can be readily applied for exploring the consequences of proposed ecological theories, finding simple explanatory models that can reproduce the observed patterns in natural communities, and aiding in the design of accurate future experiments. Furthermore, the implementation of replications to variability modeling will also permit the identification of theoretical thresholds for the detection of differences between communities and as such will help define the scopes and limits of metagenomics.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads](#)
- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Extraction Methods, Variability Encountered in](#)
- ▶ [Microbial Ecology in the Age of Metagenomics: An Introduction](#)
- ▶ [Mock Community Analysis](#)
- ▶ [Next-Generation Sequencing for Metagenomic Data: Assembling and Binning](#)

References

- Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 2012;40(12):e94.
- Bent SJ, Forney LJ. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J.* 2008;2(7):689–95.
- Bonilla-Rosso G, Eguiarte LE, Romero D, Travisano M, Souza V. Understanding microbial community diversity metrics derived from metagenomes: performance evaluation using simulated data sets. *FEMS Microbiol Ecol.* 2012;82:37–49. doi:10.1111/j.1574-6941.2012.01405.x.
- Caswell H. Theory and models in ecology: a different perspective. *Ecol Mod.* 1988;43(1–2):33–44.
- Charuvaka A, Rangwala H. Evaluation of short read metagenomic assembly. *BMC Genomics.* 2011;12 Suppl 2:S8.
- Garfinkel D. Digital computer simulation of ecological systems. *Nature.* 1962;194(4831):502–7.
- Heltshel JF, Forrester NE. Estimating species richness using the jackknife procedure. *Biometrics.* 1983; 39(1):1–11.
- Hoff KJ. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics.* 2009;10(1):520.
- Kuczynski J, Liu Z, Lozupone C, et al. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods.* 2010;7(10):813–9.
- Laserson J, Jovic V, Koller D. Genovo: *de novo* assembly for metagenomes. *J Comput Biol.* 2011; 18(3):429–43.
- Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol.* 1997;63(11):4516–22.
- Mavromatis K, Ivanova N, Barry K, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4(6):495–500.
- Mende DR, Waller AS, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One.* 2012;7(2): e31386.
- Mitra S, Schubach M, Huson DH. Short clones or long clones? A simulation study on the use of



- paired reads in metagenomics. *BMC Bioinformatics*. 2010;11(Suppl 1):S12
- Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One*. 2010;5(4):e10209.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucl Acids Res*. 2012;40:e155. doi:10.1093/nar/gks678.
- Parks DH, Beiko RG. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *ISME J*. 2012;7:173–83. doi:10.1038/ismej.2012.88.
- Peck SL. The hermeneutics of ecological simulation. *Biol Philos*. 2008;23(3):383–402.
- Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2012;28(11):1420–8.
- Pignatelli M, Moya A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One*. 2011;6(5):e19984.
- Richter DC, Ott F, Auch AF, Schmid R, Huson DH. Metasim – a sequencing simulator for genomics and metagenomics. *PLoS One*. 2008;3(10):e3373.
- Winsberg E. Simulated experiments: methodology for a virtual world. *Philos Sci*. 2003;70(1):105–25.
- Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. *Appl Environ Microbiol*. 2008;74(5):1453–63.

M

MEMOSys: Platform for Genome-Scale Metabolic Models

Stephan Pabinger^{1,2} and Zlatko Trajanoski¹

¹Division of Bioinformatics, Biocenter, Innsbruck Medical University, Innsbruck, Austria

²AIT – Austrian Institute of Technology, Health & Environment Department, Molecular Diagnostics, Vienna, Austria

Synonyms

Bioinformatics platform for genome-scale metabolic models

Definition

MEMOSys is a web-based platform for constructing, managing, and storing genome-scale metabolic models. It provides sophisticated query and data exchange mechanisms, offers an integrated version control system, and allows researchers to easily compare models. MEMOSys is freely available at <http://www.icbi.at/memosys> under the GNU Affero General Public License.

Introduction

Driven by recent innovations in sequencing technology, genome-scale metabolic models have

been compiled for a number of different organisms (Henry et al. 2010). Each model is in general a network consisting of metabolites that are connected by reactions. Genome-scale models include all reactions occurring in a living organism and are primarily reconstructed using the annotated genome and literature information. Metabolic models can be used to provide an alternative approach for integrating large amounts of data about biological systems to gain novel insights into their interconnected functionality (Kay and Wren 2009). Moreover, they have already been used for a variety of different purposes including strain engineering (Benedict et al. 2012), gene deletion studies (Choi et al. 2010), biofuel production (de Jong et al. 2011), and interpretation of gene and protein expression data (Gowen and Fong 2010).

The generation of new models is a well-documented iterative process comprising a multitude of different steps (Thiele and Palsson 2010), where often 10 % of construction time is needed to model 90 % of reactions and 90 % to collect the remaining 10 % (Rocha et al. 2008). Until the final version of a model is assembled, usually several intermediate revisions are generated. During this reconstruction process, simulated results are constantly compared to experimental data, and if they do not agree, the model is critically reevaluated (Baart and Martens 2012). It is therefore of great importance to be able to review all changes, extract previous versions, compare different versions of one model, and have access to easy to use software for creating and manipulating models.

The METabolic MOdel research and development System (MEMOSys) (Pabinger et al. 2011, 2014) has been developed to support the construction, modification, and management of genome-scale metabolic models. It is a web-based bioinformatics platform that uses an automatic version control system to store the complete developmental history of all model components. This allows researchers to access the entire model at any time during the iterative model building process. Furthermore, MEMOSys offers sophisticated query mechanisms and supports the exchange of models using standardized formats.

Model Management

Database Structure

MEMOSys has been designed to store all properties of a metabolic model in a database. The model itself is represented by a name, its unique model identifier, as well as containing reactions, genes, and metabolites. In addition, it is assigned to an organism and may contain references to an image that graphically represents the metabolic network. MEMOSys supports the upload of arbitrary additional data files, which can be directly linked to stored models. Such files may include experimental data sets that were used to validate the model during the reconstruction process. In addition, analysis results from external tools can be directly attached to the investigated model.

Each model has an arbitrary number of reactions, which are described by a multitude of properties, including name, Enzyme Commission (EC) number, reactants, products, and reversibility. Reactions can be linked to citations in order to provide primary literature evidence and are assigned to a subsystem, which is used to group reactions into metabolic pathways. MEMOSys supports the definition of lower and upper bound constraints, which are automatically included when the model is exported into a file and can then be directly used in constraint-based analyses.

Reactants and products of reactions contain the metabolite itself and the stoichiometric coefficient for that metabolite, and they are assigned to a compartment. Compartments are linked to the

corresponding Systems Biology Ontology (SBO) term and arranged in a hierarchy to support fine-grained compartmentalization when exporting models. SBO is a hierarchically arranged set of controlled, relational vocabularies of terms that are commonly used in mathematical modeling. MEMOSys uses an integrated balance check mechanism that validates the elemental composition of consuming and producing reactants. The check is automatically executed when reactions are modified, or during the import of a new model.

Each organism of a model can be annotated with the corresponding BioCyc (Karp et al. 2005) identifier. BioCyc is a biological database collection, which includes highly curated genome and pathway information for individual organisms. In order to facilitate the assignment process, MEMOSys dynamically fetches all available organisms from BioCyc and provides suggestions to select the correct identifier.

Genes and their relationship to other genes and reactions can be described using hierarchical structures and Boolean operators (e.g., [gene1 or gene2] and gene3). They are linked to the corresponding BioCyc pages if the organism identifier and the unique gene symbol are provided. In addition, for genes having a reference to the Universal Protein Resource (UniProt) (Magrane and Consortium 2011) database, MEMOSys offers a mechanism to download the amino acid sequence of the transcribed protein and provides an integrated system to fetch additional information from the UniProt entry. UniProt is a popular, freely accessible comprehensive resource containing protein sequence data as well as functional and annotation information.

Genome-scale metabolic models rely on annotations to unambiguously identify model components. History has shown that biologists have been using different notations and naming schemes for the same gene or protein. MEMOSys allows researchers to annotate reactions, metabolites, genes, and compartments with references to external databases using the minimum information requested in the annotation of biochemical models (MIRIAM) (Le Novère et al. 2005) notation. Every MIRIAM identifier is a single

unique string, which unambiguously references an object in an external resource and facilitates scientific collaborations and model comparability. MEMOSys automatically transforms MIRIAM annotations into web addresses and displays direct links to the external data sources.

Furthermore, the application includes a mechanism to easily define additional external databases, which can then be used by all model components to create further references and annotations.

Due to the iterative model building process, components may be modified several times by different members of the reconstruction team. To facilitate the discussion between researchers, MEMOSys features an integrated web board that allows attaching discussions to every model component. Associated threads are shown at each component page, and latest comments of all discussions are displayed on the home screen. In addition, global threads can be created to discuss general properties of models.

Querying System

MEMOSys uses enhanced lists to present and query data stored in the database. Every list can be customized to display a selection of available attributes. They are fully sortable and incorporate attributes from different tables into one view, which allows comprehensive data representations. MEMOSys supports fine-grained searches where different restrictions can be combined to query for a specific question. In addition, the application offers an easy to use quick search mechanism that allows users to easily search for reactions, metabolites, genes, and organisms.

As all model components are highly connected with each other, MEMOSys displays links to referenced components throughout the system and allows free navigation within and across all stored models.

Versioning

The construction of a metabolic model is an iterative task, which has been broken down into 96 steps (Thiele and Palsson 2010) generating several intermediate versions until the final model is established. Therefore, MEMOSys

integrates an automatic version control system, which creates a new revision for every modification of a model component. This system allows researchers to access the complete model history and query, compare, and export previous versions of a model. Each modification can be annotated with a comment, and the complete change history is displayed as a list at the respective component pages. The home screen of the application lets the user specify which version of a model should be used and lists the latest modifications for metabolites and reactions.

Data Access and Supervision

MEMOSys is a multiuser application using four different user classes to control data access: (a) unregistered visitors are allowed to view accepted, publicly available versions of models; (b) registered users are able to display in addition to publicly available models, accepted versions of assigned models; (c) editors have access to all versions of their own models and are able to create, update, and delete model components. In addition they are allowed to upload files to the application and import models; (d) administrators are editors with additional rights to access all models, change the public availability of models, and accept modifications.

Each modification of a model component is at first marked as pending and needs to be confirmed by an administrator. Upon approval of a modification, a new internal revision number is assigned to the model. In addition to the automatically set revision number, administrators can assign major version numbers to each model. MEMOSys differentiates between publicly available models, which are visible to all visitors and contain all accepted modifications, and restricted models that are only visible to registered users and editors of the assigned models.

Comparison

As the construction of a draft genome-scale metabolic model is getting more and more a routine

application, future developments will strongly rely on already existing reconstructions of related organisms. In addition, researchers are often interested in the subtle differences between organisms when exploring specific biological functionalities. Hence, MEMOSys offers a flexible and intuitive mechanism to assess the similarity between models allowing users to compare any version of different models. Furthermore, it is possible to compare two versions of the same model to identify development changes.

The first section of the comparison result presents Venn diagrams that graphically display the calculated differences for reactions, metabolites, and genes. Next, restrictions on the used models can be set to display only differences in selected compartments and subsystems. In addition to the graphical representation, the application shows detailed lists of unique and equal model components and uses tabs to switch between reactions, metabolites, and genes lists. Every model component is connected to the corresponding page where detailed information is presented.

Data Exchange

MEMOSys features the export of current metabolite and reaction query result lists into Excel or PDF files, where only the active result set is included. Since several methods and toolboxes which analyze genome-scale metabolic models have been published over the last years (Baart and Martens 2012), MEMOSys provides a sophisticated data exchange mechanism that allows the export of models into valid SBML files. The Systems Biology Markup Language (SBML) (Hucka et al. 2003) provides a common intermediate format that can be used to define models in regulatory networks, metabolic pathways, signaling pathways, and gene regulation networks.

The exported files are compliant with the consensus yeast format (Herrgård et al. 2008) or with the COBRA toolbox format (Schellenberger et al. 2011). Researchers are able to export all available versions of a model and restrict the set of exported reactions by either including only

reactions that are in certain subsystems or using the result of a reaction query as input for the export mechanism.

MEMOSys features three different ways to assign reactions and metabolites to compartments (compartmentalization), which allow researchers to directly use exported models in analysis tools that do not support a fine-grained assignment of reactions to compartments.

The system supports the import of models that are encoded in valid format as defined by the consensus yeast reconstruction group. In addition, existing models in SBML format can be used to improve the annotation of stored model components (see Fig. 1).

Installation

The application itself and the source code of MEMOSys are freely available under the GNU Affero General Public License. As MEMOSys is a web application, it is recommended installing it on a server system and set appropriate access permissions for potential users. A detailed user guide and installation instructions are available at the distribution website. MEMOSys is available for download at <http://www.icbi.at/MEMOSys>.

Summary

During the last years, numerous genome-scale metabolic models have been developed for a multitude of different organisms. They are a promising approach to systematically analyze complex cellular systems and have been successfully applied for improving gene annotation, increasing the product yield, and predicting the effect of gene deletions.

The web-based METabolic MOdel research and development System (MEMOSys) is a versatile bioinformatics platform for the management, storage, modification, and development of genome-scale metabolic models. It facilitates the construction of new models by providing a built-in version control system, which allows researchers to access the complete reconstruction history.

Improve Annotation

Metabolites Genes

[Set all to new](#) [Set all empty properties](#) [Save Metabolites](#)

Abbreviation	Name	CHEBI Id	InChI	Formula
<input type="radio"/> 2-Oxoglutarate	<input type="radio"/> 2-Oxoglutarate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:30916	<input type="radio"/> InChI=1/C5H6O5/c6-3(5(9)10)1-2-4(7)8/h1-2H2,(H,7,8)(H,9,10)/p-1/c5H5O5/h7H/q-1	<input type="radio"/> C5H4O5 <input type="radio"/> C5H6O5
<input type="radio"/> 4-Aminobenzoate	<input type="radio"/> 4-Aminobenzoate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:17836	<input type="radio"/> InChI=1/C7H7NO2/c8-6-3-1-5(2-4-6)7(9)10/h1-4H,8H2,(H,9,10)	<input type="radio"/> C7H6NO2 <input type="radio"/> C7H7NO2
<input type="radio"/> Acetate	<input type="radio"/> Acetate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:30089	<input type="radio"/> InChI=1/C2H4O2/c1-2(3)4/h1H3,(H,3,4)/p-1	<input type="radio"/> C2H3O2 <input type="radio"/> C2H4O2
<input type="radio"/> Adenine	<input type="radio"/> Adenine	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:16708	<input type="radio"/> InChI=1/C5H5N5/c6-4-3-5(9-1-7-3)10-2-8-4/h1-2H,(H3,6,7,8,9,10)/f/h9H,6H2	<input type="radio"/> C5H5N5
<input type="radio"/> Aminoacetaldehyde	<input type="radio"/> Aminoacetaldehyde	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:17628	<input type="radio"/> InChI=1/C2H5NO/c3-1-2-4/h2H,1,3H2	<input type="radio"/> C2H6NO <input type="radio"/> C2H5NO
<input type="radio"/> Anthranilate	<input type="radio"/> Anthranilate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:16567	<input type="radio"/> InChI=1/C7H7NO2/c8-6-4-2-1-3-5(6)7(9)10/h1-4H,8H2,(H,9,10)/p-1/c7H6NO2/q-1	<input type="radio"/> C7H6NO2 <input type="radio"/> C7H7NO2
<input type="radio"/> beta-Alanine	<input type="radio"/> beta-Alanine	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:16958	<input type="radio"/> InChI=1/C3H7NO2/c4-2-1-3(5)6/h1-2,4H2,(H,5,6)/f/h5H	<input type="radio"/> C3H7NO2
<input type="radio"/> Choline	<input type="radio"/> Choline	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:15354	<input type="radio"/> InChI=1/C5H14NO/c1-6(2,3)4-5-7/h7H,4-5H2,1-3H3/q+1	<input type="radio"/> C5H14NO
<input type="radio"/> Citrate	<input type="radio"/> Citrate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:16947	<input type="radio"/> InChI=1/C6H8O7/c7-3(8)1-6(13,5(11)12)2-4(9)10/h13H,1-2H2,(H,7,8)(H,9,10)(H,11,12)/p-3/c6H5O7/q-3	<input type="radio"/> C6H5O7 <input type="radio"/> C6H8O7
<input type="radio"/> CO2	<input type="radio"/> CO2	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:16526	<input type="radio"/> InChI=1/CO2/c2-1-3	<input type="radio"/> CO2
<input type="radio"/> Cyanate	<input type="radio"/> Cyanate	<input type="radio"/> urn:miriam:obo.dchebi:CHEBI:29195	<input type="radio"/> InChI=1/CHNO/c2-1-3/h3H/p-1/cNO/h3H/q-1	<input type="radio"/> CNO <input type="radio"/> CHNO

Navigation: << 1 2 3 4 5 >>

MEMOSys: Platform for Genome-Scale Metabolic Models, Fig. 1 Displayed is the user interface for improving the annotation of metabolite and genes. In addition, empty model component fields can be filled with new annotations, or all stored annotations can be replaced with the currently loaded ones.

Research on existing models is facilitated by a powerful search system, a feature-rich comparison mechanism, and standardized references to external databases.

MEMOSys provides customizable data exchange mechanisms using the SBML format to enable further analysis in external tools and supports different user roles and access rights to allow collaborations across departments and universities. The system is freely available at <http://www.icbi.at/MEMOSys>.

Cross-References

- ▶ [KEGG and GenomeNet, New Developments, Metagenomic Analysis](#)
- ▶ [New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE](#)

References

- Baart GJE, Martens DE. Genome-scale metabolic models: reconstruction and analysis. *Methods Mol Biol.* 2012;799:107–26.
- Benedict MN, Gonnerman MC, Metcalf WW, et al. Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A. *J Bacteriol.* 2012;194:855–65.
- Choi HS, Lee SY, Kim TY, et al. In silico identification of gene amplification targets for improvement of lycopenene production. *Appl Environ Microbiol.* 2010;76:3097–105.
- de Jong B, Siewers V, Nielsen J. Systems biology of yeast: enabling technology for development of cell factories for production of advanced biofuels. *Curr Opin Biotechnol.* 2011;23:624–30.
- Gowen CM, Fong SS. Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of *Clostridium thermocellum*. *Biotechnol J.* 2010;5:759–67.
- Henry CS, DeJongh M, Best AA, et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol.* 2010;28:977–82.
- Herrgård MJ, Swainston N, Dobson P, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol.* 2008;26:1155–60.
- Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics.* 2003;19:524–31.
- Kay E, Wren BW. Recent advances in systems microbiology. *Curr Opin Microbiol.* 2009;12:577–81.
- Le Novère N, Finney A, Hucka M, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol.* 2005;23:1509–15.
- Magrane M, Consortium U. UniProt knowledgebase: a hub of integrated protein data. *Database (Oxford).* 2011;2011:bar009.
- Pabinger S, Rader R, Agren R, et al. MEMOSys: bioinformatics platform for genome-scale metabolic models. *BMC Syst Biol.* 2011;5:20.
- Pabinger S, Snajder R, Hardiman T, Willi M, Dander A, Trajanoski Z. MEMOSys 2.0: an update of the bioinformatics database for genome-scale models and genomic data *Database.* 2014;bau004 doi:10.1093/database/bau004. published online February 14, 2014.
- Rocha I, Förster J, Nielsen J. Design and application of genome-scale reconstructed metabolic models. *Methods Mol Biol.* 2008;416:409–31.
- Schellenberger J, Que R, Fleming RMT, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc.* 2011;6:1290–307.
- Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.* 2010;5:93–121.
- Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucl Acids Res.* 2005;33(19):6083–89.

MetaBin

Vineet K. Sharma¹ and Todd D. Taylor²

¹MetaInformatics Laboratory, Metagenomics and Systems Biology Group, Department of Biological Sciences, Indian Institute of Science Education and Research, Bhopal, India

²Laboratory for Integrated Bioinformatics, Core for Precise Measuring and Modeling, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Synonyms

Taxonomic assignment; Taxonomic binning; Taxonomic classification

Definition

MetaBin: Taxonomic binning of metagenomic sequences.

Introduction

The first, and primary, challenge in metagenomic data analysis is to ascertain the genomic origin of metagenomic sequences and to make appropriate taxonomic assignments (Tringe and Rubin 2005; McHardy et al. 2007; Sharma et al. 2012). Composition- or homology-based classification of metagenomic sequences are the two main approaches that are currently used (McHardy et al. 2007; Huson et al. 2007). Among the two, homology-based methods are more sensitive and accurate but require a large amount of time to generate the BLAST alignments, which are used as an input for these programs. The composition-based approach is exploited by classification tools such as PhyloPythia, TETRA, and TACO, for taxonomic classification of metagenomic sequences (Diaz et al. 2009; McHardy et al. 2007; Teeling et al. 2004). These methods require prior training using longer reads (>800 bp) to carry out the classification, and thus the classifications remain limited to higher taxonomic levels. Homology-based approach assesses the taxonomic identity of a read from the results of a homology-based search against a known reference sequence database which is usually the NCBI non-redundant (NR) database (Sayers et al. 2011). Examples of some homology-based tools are MEGAN and SOrt-ITEMS (Huson et al. 2007; Monzoorul et al. 2009). Both of these carry out taxonomic binning based on the BLAST bit-score and lowest common ancestor (LCA) approach. If a read shows a match with multiple genomes, it is assigned to the common taxonomic ancestor (higher level) of the hits. Since these are only based on bit scores, they may lead to incorrect or nonspecific taxonomic assignment. The homology-based methods primarily depend on the representation of genomic sequences in the

reference databases and are able to carry out classification of metagenomic sequences at lower taxonomic levels (genus or species) when a comprehensive reference database is used. Another homology-based method, WebCARMA, scans for the presence of conserved Pfam domains and protein families in the metagenomic reads (Gerlach et al. 2009).

The motivation to develop a better homology-based algorithm for taxonomic classification came from the fact that none of the available methods are comprehensive in that they have not considered some key features of metagenomic sequences which could result in increased and more accurate taxonomic assignments. Therefore, in this entry, a novel algorithm called “MetaBin” is presented which exploits the information from all possible ORFs (complete or partial) for each sequence read while carrying out the taxonomic assignment. This algorithm is faster and results in much higher accuracy and sensitivity for taxonomic classification. It can be used for the taxonomic assignments of various read lengths (≥ 45 bp, single or paired end) which are commonly generated using available traditional and next-generation sequencing technologies.

Methods

Reference Database Construction and Simulated Reads

The non-redundant (NR) sequence database (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/>) was retrieved from NCBI (Sayers et al. 2011). In addition, genomic sequences of 25 completed bacterial genomes belonging to different taxonomic lineages were retrieved (<ftp.ncbi.nih.gov/genomes/Bacteria>). Local versions of the NR database were created, to test the performance of MetaBin, by removing all sequences belonging to the associated genus and family. This helps in assessing the performance of MetaBin on reads for which no genome of the genus (novel genome) is present in the NRminusFamily or NRminusGenus database. The reads created

from these genomes are similar to reads of novel or yet unknown genomes because the NRminusGenus or NRminusFamily databases do not contain any genome of that genus. Simulated read datasets were created using the MetaSim program to represent Sanger (read length ~800 bp) and 454 (read lengths of ~400 and ~250 bp) sequences (Richter et al. 2008). A Perl script was developed for generating 1,000 simulated reads of length ~75 bp and ~45 bp, respectively, from each of the bacterial genomes, since the option to create short reads was absent in MetaSim.

The metagenomic sequences for a real metagenomic dataset were taken from human gut samples from a single Spanish male individual generated by Illumina sequencing (V1.CD-2, age 49, BMI 27.76, 20,707,369 high-quality reads, library 090107) (ftp://public.genomics.org.cn/BGI/gutmeta/High_quality_reads/) (Qin et al. 2010). The sample data sequences (Sargasso Sea Subsample 1) for Sargasso Sea were downloaded from <http://www-ab.informatik.uni-tuebingen.de/software/megan/old-datasets> (Huson et al. 2007). This set contains the first 10,000 reads from Sample 1 of the Sargasso Sea dataset (Venter et al. 2004).

BLAST (version 2.2.22, <ftp://ftp.ncbi.nih.gov/blast/>) was downloaded from NCBI. MEGAN (version 3.8) (<http://www-ab.informatik.uni-tuebingen.de/data/software/megan/download/welcome.html>) (Huson et al. 2007), SOrt-ITEMS (<http://metagenomics.atc.tcs.com/binning/SOrt-ITEMS>) (Monzoorul et al. 2009), and TACO (version 1.0, <http://www.cebitec.uni-bielefeld.de/brf/taco/taco.html>) (Diaz et al. 2009) were retrieved from their respective sites. WebCARMA (version 1.0) was run from their Web server (<http://webcarma.cebitec.uni-bielefeld.de/cgi-bin/webcarma.cgi>) (Gerlach et al. 2009).

Algorithm Development

MetaBin provides significant improvements over currently existing homology-based methods for better taxonomic assignments. It reduces (up to 1,000-fold) the amount of time needed to

generate the alignments by implementing Blat (Kent 2002) as the faster alignment method in place of BLASTX which is comparable to the time taken by composition-based methods. This feature makes it practical to use a more accurate and sensitive homology-based approach for both Web- and console-based high-throughput analysis of large datasets.

A unique approach has been adopted which considers the taxonomic information from all verified complete or partial ORFs present in a read and then assigns a taxonomic bin. This helps to make correct assignments of reads of diverse lengths to different taxonomic bins. Since our procedure comprehensively considers all imaginable cases, the results are more accurate and specific, and the assignments are not limited by read length. (Details are provided in the manuscript, Sharma et al. 2012.)

The taxonomic binning of the simulated read datasets was carried out using MetaBin and MEGAN, and the assignments were counted at three levels, namely, “Genus,” “Phylum,” and “Higher.” The “correct assignments” were those where the assigned phylum was same as the expected phylum or simply if it was assigned to its own phylum. Only the intragenic reads were considered to calculate sensitivity and the positive predictive value (PPV) because the NR reference database contains only protein sequences, and thus the reads coming from known protein coding regions (intragenic) are expected to find a match. The following standard formulae were used to calculate sensitivity and PPV:

$$\text{Sensitivity (\%)} = (TP / (TP + FN)) \times 100$$

$$\begin{aligned} \text{Positive predictive value (PPV) (\%)} \\ = (TP / (TP + FP)) \times 100 \end{aligned}$$

True positive (TP) = number of reads assigned with correct (expected) phylum

False positive (FP) = number of reads assigned to other (incorrect) phylum

False negative (FN) = number of unassigned intragenic reads plus number of reads assigned above to the phylum level (higher)

The average sensitivity and PPV were calculated for all simulated read datasets aligned with the complete NR database or the NR-G versions.

MetaBin Development

The MetaBin algorithm was developed in Perl (version 5.10.1), and the dendrogram images were generated using the Perl GD module. Options are provided to change the different run parameters such as bin size, minimum bit-score, and bit-score range, to select hits and to create a dendrogram image after comparing the proportions of each taxonomic group in the selected metagenomes, and to display the respective proportions as a pie chart. The algorithm can be used for the taxonomic assignments of both single- and paired-end sequence reads. A user-friendly website (<http://metabin.riken.jp/>) was developed on our server including detailed instructions for installation, usage, and updating of the taxonomy database. A free stand-alone executable program is also provided and can be downloaded for different operating systems including Windows, Linux, and Mac.

Results

The overall performance of MetaBin was found to be superior to the other available tools such as MEGAN, SOrt-ITEMS, TACO, and WebCARMA for all read datasets. It assigned a higher percentage of reads to their correct genus and phylum, as compared to the other methods. Particularly for the short (<100 bp) Illumina reads, it assigned up to 18 % more reads to their correct taxonomic levels. This is a useful and unique ability of MetaBin to make more accurate assignments at the lower and more specific taxonomic levels. For all simulated read datasets, the average sensitivity and PPV of MetaBin was similar to or higher than those of MEGAN, especially for short reads. For ~75 bp reads, MetaBin showed up to 6 % and 16.8 % higher average sensitivity as compared to MEGAN and SOrt-ITEMS, respectively. For ~45 bp reads, MetaBin showed up to 32 % and

46 % higher average sensitivity as compared to MEGAN and SOrt-ITEMS, respectively.

The performance of MetaBin was also evaluated on real metagenomic data using the recent human gut data obtained by Illumina sequencing (short reads) from a European male individual and analyzed using MetaBin with Blat as the alignment program. Only those bins containing at least 10,000 reads were considered under default parameter conditions. The analysis of such a large metagenomic dataset proves the ability of MetaBin to work on real metagenomic datasets. In this analysis, Bacteroidetes was found as the most abundant phylum (77.4 %) followed by Firmicutes (16.8 %), Proteobacteria (3.5 %), Actinobacteria (1.7 %), Cyanobacteria (0.27 %), and Euryarchaeota (0.24 %). These results corroborate previous observations (Kurokawa et al. 2007).

The performance was also evaluated using longer (~800 bp) reads obtained from the Sargasso Sea dataset. Using this common dataset, the results of MetaBin, MEGAN, and SOrt-ITEMS were compared. MetaBin and MEGAN both predicted a similar number of bins; however, MetaBin assigned comparatively more reads (nearly twice the number of reads at the species level) to each of these common bins which shows its higher sensitivity and higher accuracy. The performance of SOrt-ITEMS was comparatively poor compared to both MetaBin and MEGAN. A brief comparison of MetaBin was also carried out with one of the composition-based methods (TACO) and with another method based on homology to protein families (WebCARMA) using the above dataset. Both the composition- and protein family-based methods showed limitations in making comprehensive taxonomic assignments and performed poorly as compared to homology-based methods.

The Web Server

Different pages are provided on the Web server with several options for carrying out online taxonomic analysis. The main page is the

MetaBin,

Fig. 1 Screenshot of “application” page using a sample query

“Application” page, where the user can submit and carry out taxonomic analysis of either sequence reads or Blastx output (Fig. 1).

Two options, BLAT and BLAST, are provided to generate the alignments. The input sequences should be submitted in FASTA format, for which the ORFs are predicted, and the qualified ORFs are aligned against the NCBI NR database using Blat. This output is used to classify the sequences into their appropriate taxonomic bins. Another option, BLAST, uses Blastx for generating the alignments and takes comparatively a much longer time for generating the alignments as compared to Blat. The input parameters such as minimum bit-score (Blat or Blastx output), bit-score range to select hits, and bin size (minimum number of reads needed to form a taxonomic bin) can be changed or used as default. The “Results” page provides the output files in tab-delimited format and displays thumbnail images of the taxonomic tree (*.png file) and functional annotation of the reads using COGs functional classes. The results can be downloaded from the website (Fig. 2).

The “Visualization” page provides several options for displaying the results and carrying out comparative analysis (Fig. 3).

An option to upload the resultant *.json file generated after using the stand-alone version for additional Web-based analyses is also provided. There are options to visualize the taxonomic tree and prepare a “composition chart” for a single dataset. The composition chart gives an overview of the microbial distribution in the dataset and

shows their abundance values. Another option is available to compare the taxonomic profiles of up to five metagenomic datasets by “Compare Metagenome Profiles.”

The stand-alone console-based version is provided to analyze large metagenomic datasets locally on the user’s system after installation. A free stand-alone executable program is available for download for several operating systems including Linux, Mac, and Windows.

Discussion

Homology-based approaches are more common and considered to be more specific and useful for diverse read length as compared to composition-based approaches. However, their implementation on large metagenomic datasets is limited due to the longer analysis time. The MetaBin algorithm overcomes this limitation and provides a significant improvement over the currently existing homology-based methods for better and faster taxonomic assignments by using a more specific ORF-based approach. It carries out more accurate and specific taxonomic assignments at both genus and species levels. The replacement of BLAST by Blat in MetaBin makes it possible to employ a more accurate and sensitive homology-based approach for the high-throughput analysis of large datasets and also for the development of a Web-based community server. The performance of this approach was validated using



Results

Uploaded file : sample.reads

Total reads in file : 1000

Total assigned reads : 939

Selected options - Alignment method : BLAT, Bin size : 2, Minimum bit-score : 17, Bit-score range : 90

You can download the following result files :

Files	Download All
sample.reads.blat.bin.reads	Download
sample.reads.blat.format.bestID	Download
sample.reads.blat.format.class	Download
sample.reads.blat.format.class.png	Download
sample.reads.blat.format.enrich	Download
sample.reads.blat.json	Download
sample.reads.blat.png	Download
sample.reads.blat.sum	Download
sample.reads.blat.sum.reads	Download

M

MetaBin, Fig. 2 Screenshot of results page for the sample query



MetaBin, Fig. 3 Screenshot of visualization page

both simulated reads and real metagenomic datasets. In addition, it can be a tool of choice for large metagenomic datasets as demonstrated in this entry. It can be used for the taxonomic assignment of sequence reads of diverse lengths (≥ 45 bp) derived from any existing sequencing

technology, and perhaps it is the only method which can be applied for the taxonomic binning of reads of lengths as short as 45–75 bp with high accuracy and sensitivity. Thus, the MetaBin Web server and program can be considered a significant improvement over currently

existing programs for carrying out the taxonomic binning of metagenomic sequences with high accuracy, speed, and sensitivity.

References

- Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinforma.* 2009;10:56.
- Gerlach W, Junemann S, Tille F, Goesmann A, Stoye J. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinforma.* 2009;10:430.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17:377–86.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12:656–64.
- Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 2007;14:169–81.
- McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4:63–72.
- Monzoorul HM, Ghosh TS, Komanduri D, Mande SS. -Sort-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics.* 2009;25:1722–30.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:59–65.
- Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One.* 2008;3:e3373.
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2011;39:D38–51.
- Sharma VK, Kumar N, Prakash T, Taylor TD. Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PLoS ONE.* 2012;7:e34030.
- Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinforma.* 2004;5:163.
- Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet.* 2005;6:805–14.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004;304:66–74.

MetaBioME

Computational Tool for Mining Metagenomic Datasets to Discover Novel Biocatalysts by Using a Homology-Based Approach

Vineet K. Sharma¹ and Todd D. Taylor²

¹MetaInformatics Laboratory, Metagenomics and Systems Biology Group, Department of Biological Sciences, Indian Institute of Science Education and Research, Bhopal, India

²Laboratory for Integrated Bioinformatics, Core for Precise Measuring and Modeling, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Synonyms

Biocatalysts; Commercially useful enzymes; CUEs

Definition

MetaBioME: Comprehensive metagenomic biomining engine.

Introduction

The relationship between man and microbes is as old as the age of man himself and it is no wonder that man carries around ten times more of these little friends than of his own cells (Gill et al. 2006). However, it has only been a few 1,000 years since man first learned to harness the power of microbes, initially to accomplish crude and trivial fermentations like brewing and curdling. With the evolution of man, today, these applications have been extended to almost all areas such as agriculture, pharmaceuticals, industry, biotechnology etc., where microbes have become indispensable. These applications have now become more refined, and the most remarkable change, which has happened, is that microbial enzymes have replaced whole microbes in many such processes.

These microbial enzymes a.k.a. “biocatalysts” offer ecologically friendly or “green” solutions for the implementation of biochemical processes at a reduced cost and produce a large variety of chemical substances without involving the use of polluting reagents that are often characteristic of chemical synthesis (Ferrer et al. 2005).

However, only a few enzymes are currently known which can be used as biocatalysts due to the limited number of sequenced microbes, which is principally limited by the fact that most (>98 %) of the microbes cannot be cultured, a necessary step for their sequencing by traditional methods (Amann et al. 1995). This, yet unculturable, majority of microbes potentially conceals an enormous treasure of unknown biological functions locked in their unidentified genes, proteins, and biochemical pathways. Therefore, approaches aimed at mining environmental genetic diversity can significantly enhance the enzyme repertoire and will be helpful in the discovery of novel biocatalysts with potential biotechnological applications.

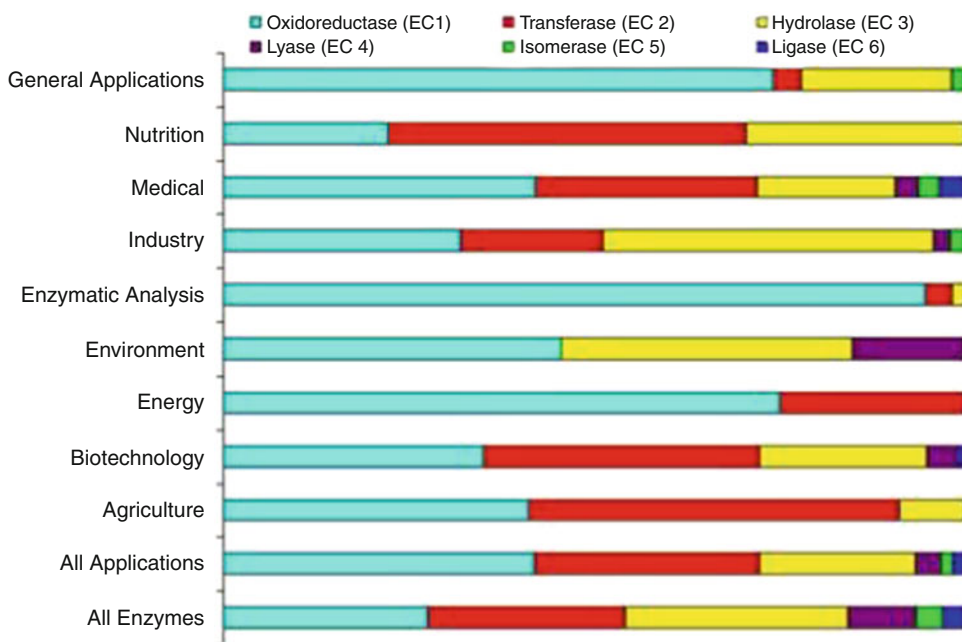
Another feasible, yet challenging method is to create novel biocatalysts by using *in silico* approaches and bioengineering is to reshuffle the 20 known amino acids and mutate the existing proteins. However, there exist nearly infinite possibilities for such an approach, and it is impractical and costly to test them all. In this scenario, nature appears the veteran since it began its bioengineering laboratory billions of years ago and has already created and tested an intriguing diversity of biochemical pathways and their constituent enzymes that perform numerous transformations of molecules in diverse biological systems with great precision and specificity. Therefore, it is conceivable that the ideal biocatalyst may already exist in nature and a wise strategy would be to augment our knowledge base by exploring the inherent diversity of nature.

To this end, metagenomics has emerged as a powerful culture-independent approach for exploring the complexity of microbial genomes in their natural environments (Tringe and Rubin 2005a). Many metagenomic projects have recently been conducted, such as metagenomic studies of soil, sea, acid mines, human gut, termite gut, whale

fall, etc. (Daniel 2005; Edwards and Rohwer 2005; Kurokawa et al. 2007; Tringe et al. 2005; Turnbaugh et al. 2006; Tyson et al. 2004; Venter et al. 2004a; Warnecke et al. 2007), and several large-scale worldwide metagenomic projects are currently under progress or in planning. From these metagenomic projects, some important biocatalysts have already been isolated such as lipases/esterases, proteases, nitrilases, β -lactamases, hydrolases, cellulases, α -amylases, xylanases, oxidoreductases, and dehydrogenases (Ferrer et al. 2005; Yun and Ryu 2005). Therefore, the upcoming information from further metagenomic projects holds enormous prospect for the discovery of novel genes, biocatalysts, and biochemical pathways, irrespective of the necessity for complete genomic sequences.

Novel biocatalysts can be detected in genomic or metagenomic libraries using three commonly used strategies: (i) homology-driven screening, (ii) substrate-induced gene expression screening, and (iii) activity-based analysis (Ferrer et al. 2005; Yun and Ryu 2005). While these methods have certain advantages like high specificity and reliability, they require extensive mining of large genomic or metagenomic libraries and result in a few positives per enzymatic screening (Ferrer et al. 2005). This is further limited by the low quality of DNA, low coverage, host bias, and need for better vector-host combinations for expression.

An alternative and promising approach which now exists involves direct shotgun sequencing of metagenomic libraries (Tringe and Rubin 2005b). This approach was earlier considered too expensive, since it required massive sequencing by conventional sequencers (Sanger). However, the recent availability of a new generation of sequencers, like Roche 454, Illumina HiSeq, Ion Torrent, etc., has made sequencing even more high-throughput, several orders less expensive, and most importantly cloning independent (Mardis 2008). Considering the sheer volume of metagenomic samples and implementation of such high-throughput sequencing methods, combined with high-throughput computational analysis, screening of potential biocatalysts is more promising and is likely to accelerate the process of biocatalyst discovery.



MetaBioME, Fig. 1 Distribution of enzymes (EC classes) into nine application categories

In the present entry, we describe a computational platform and resource to identify novel biocatalysts in metagenomic datasets using homology-based approaches. We have developed a comprehensive Metagenomic BioMining Engine (MetaBioME) platform (Sharma et al. 2010), which provides a unique resource for the identification of novel alternatives to the existing known biocatalysts and novel biocatalysts in metagenomic datasets, which can be used as leads for further experimental verification.

Results

The distribution of 510 biocatalysts in nine application categories indicates that the highest number (234, 46 %) of biocatalysts is present in the “Biotechnology” category and the lowest (3, 3 %) in the “Energy” category (Fig. 1).

Oxidoreductases (EC 1), which catalyze oxidation-reduction reactions, are most abundant in five out of nine applications, namely, Enzymatic Analysis (95 %), Energy (75 %), General Applications (74 %), Environment (45 %), and Medical

(42 %). Transferases (EC 2), which perform the transfer of functional groups from one molecule to another, are most abundant in three application categories, namely, Agriculture (50 %), Nutrition (48 %), and Biotechnology (37 %). Hydrolases (EC 3), which are involved in formation of two separate products from a single substrate by hydrolysis, are most abundant only in Industrial applications (45 %). It is clear from the above findings that oxidoreductases (EC 1) are most widely used as biocatalysts followed by transferases (EC 2) and hydrolases (EC 3). It is also noteworthy that although hydrolases (EC 3) constitute most of the enzymes among the six EC classes, they are not the most widely employed biocatalysts. The biocatalysts belonging to the remaining three EC classes (4, 5, and 6) were not as widely distributed or were completely absent from many of the nine application categories.

Gene Prediction in Metagenomic Datasets (Except HFV)

The average contig length in the metagenomic datasets varied between 0.8 and 1.8 kb with the exception of AMD (4.18 kb). The prediction of

ORFs by Glimmer and MetaGene showed considerable variation with MetaGene predicting up to twice the number of ORFs as compared to Glimmer. With the exception of AMD, having an average number of four ORFs per contig predicted by MetaGene and Glimmer, the average number of ORFs per contig for the remaining datasets was found to vary between 0.6 and 2.3. The median protein length in bacteria was reported in one study as 267 amino acids (801 base pairs) (Brocchieri and Karlin 2005). Since, in the above analysis, the average length of the contigs varies between 0.8 and 1.8 kb, and the average number of ORFs per contig varies between 0.6 and 2.3, it is likely that a significant portion of at least one ORF can be predicted in a contig of about 1 kb (Tringe et al. 2005). The ORFs predicted by Glimmer and MetaGene in all the metagenomic datasets were fed into the “Metabase” database, which is being used for the development of MetaBioME.

Identification of Potential Biocatalysts

Using MetaBioME’s homology-based approach, we identified 199 potential alternatives (49 % of total biocatalysts) to known biocatalysts in the metagenomic datasets using a stringent threshold of identity $\geq 50\%$ and coverage $\geq 90\%$. Among the nine application categories, novel alternatives to known biocatalysts could be predicted for 39–50 % of total biocatalysts in each category. We further relaxed the above cutoff (identity $\geq 30\%$ and coverage $\geq 90\%$) to identify an expanded list of potential alternate biocatalysts in the metagenomic datasets which could be used as leads for experimental verification. Using this relaxed cutoff, novel alternatives for a total of 305 (75 %) biocatalysts could be identified in the metagenomic datasets from all application categories. Among these potential biocatalysts, 20 were commonly found in all nine metagenomic datasets, while 64 biocatalysts were rare and could be found in any one of the nine metagenomic datasets.

Description of Web Resource: MetaBioME

We used the above strategy, data, and results to develop a comprehensive resource “MetaBioME,”

which can be queried using a publicly available Web interface available at <http://metasystems.riken.jp/metabiome> (Sharma et al. 2010). The key idea of MetaBioME is to develop a computational tool for mining metagenomic datasets by using homology-based approaches to discover novel biocatalysts and novel alternatives for existing biocatalysts, with advanced analysis options for facilitating the validation of results. Therefore, for comprehensive querying, we have developed the following query pages:

MetaSearch: It houses a pre-classified set of 510 biocatalysts in nine application categories that can be searched for in different metagenomic datasets.

MetaXplorer: It contains the complete set of EC enzymes and options to search for their homologous ORFs in metagenomic datasets.

MetaAlign: It allows users to submit a gene or protein sequence of interest and search for the existence of a homologous ORF in metagenomic datasets.

The details of these query pages are provided below.

MetaSearch: Search for Biocatalysts in Metagenomic Datasets

The “MetaSearch” query page is designed to identify novel biocatalysts, categorized into nine main application categories in metagenomic datasets (Fig. 2).

This pre-classification helps the user to select biocatalysts belonging to any application area and search for them in metagenomic datasets. A search can be made by selecting one or more of the application categories and a single metagenomic dataset. Since the metagenomic datasets contain volumes of information, the number of hits reported for each query is expectedly large; therefore, we have currently restricted the option to select and search in only one metagenomic dataset per query. The queries can also be made by selecting different attributes such as EC number, enzyme name, Swiss-Prot ID, biochemical pathway, and substrate or products. Multiple keywords can also be submitted using Boolean operators. An option is also provided to limit the number of results by selecting “Best hit” or “Best 10 hits.”

MetaBioME
Comprehensive Metagenomic BioMining Engine

Home | **Meta Search** | CUEs Explorer | Meta Explorer | Meta Align | Tutorial | Links

CUEs
Commercially Useful Enzymes

Home
Metagenome
Enzyme Info
Data Sources
RIKEN
Team
Feedback

Select CUEs of interest from the following categories

- Agriculture
Agriculture & related
- Energy
Biofuels, Fuel cells, etc.
- Food & Nutrition
Dairy, Wine, Nutrition, etc.
- Biotechnology**
Biotechnology, Molecular Biology, Synthesis, etc.
- Biosensor
Biosensor, Enzymatic estimation & assay, etc.
- Environment
Bio-degradation of toxic compounds, etc.
- Other Industries
Textile, Paper, Steel, Leather, Rubber, Fibers, Detergents, etc.
- Medical
Diagnostics, Drug Development, Pharmaceuticals, etc.
- Miscellaneous
Other Applications

Advanced search Select All Clear All

Other Search Options

EC Number Enzyme Name or Keywords

Biochemical Pathway Swiss-Prot ID

Substrates or Products

Select a (Meta)Genomic Dataset to Search for CUE Homologs

Choose Dataset : Metagenomic Source Metagenomic Project Completed Bacterial Genome

marine
sludge
mine drainage

Optimise Results :

% Coverage with Known CUEs E-value

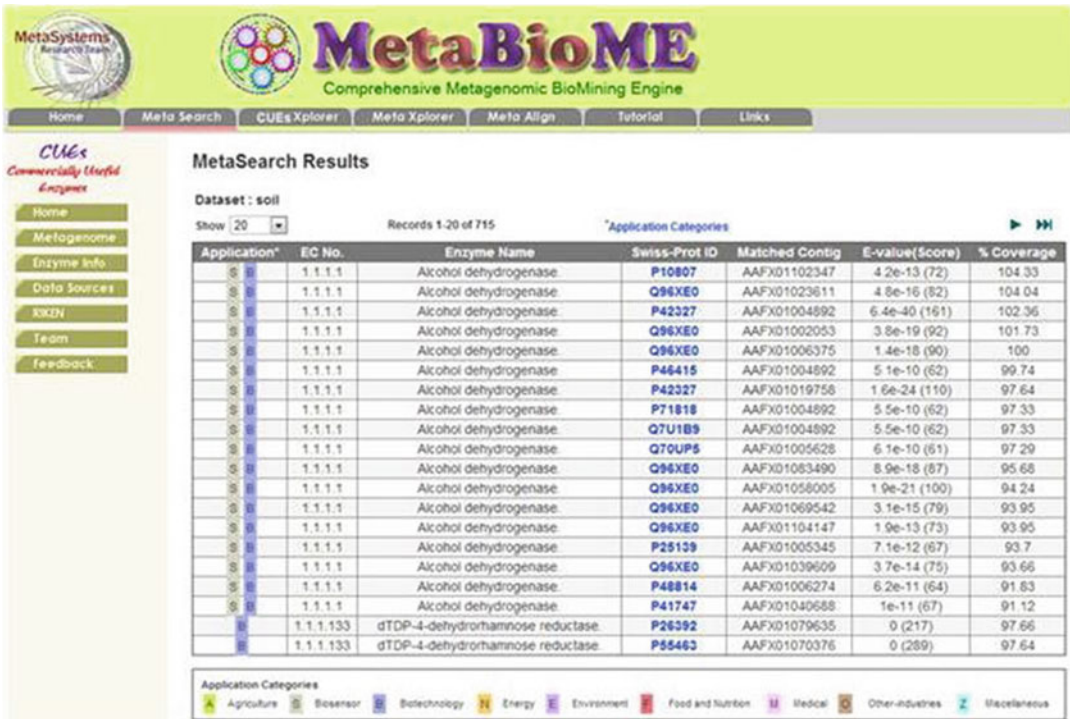
Reset Submit

MetaBioME, Fig. 2 Screenshot of “MetaSearch” page with a sample query

On submission of a query for a selected application category and metagenomic dataset, MetaBioME examines the alignments of all Swiss-Prot sequences known for all EC numbers present in that category with the ORFs predicted in contigs of the selected metagenomic dataset. The subsequent “MetaResults” page displays the qualified hits as a table sorted on the basis of percent coverage (completeness of the alignment) and provides a list of all matching Swiss-Prot IDs which

showed at least 50 % coverage with the matched metagenomic contigs (Fig. 3).

Comprehensive information for each match can be retrieved by clicking on the Swiss-Prot ID link on the Results page which opens up the “MetaBioME Profile” page. The profile page summarizes information on the enzyme properties, reaction performed, pathway information (as available in KEGG), links to related publicly available databases, queried dataset,



MetaBioME, Fig. 3 Screenshot of “MetaResults” page showing the results of the submitted sample query

and application category. This information is followed by a table of predicted ORFs, where the ORFs are segregated as commonly predicted by Glimmer and MetaGene and uniquely predicted by Glimmer or MetaGene, respectively. The ORF showing the best match with the Swiss-Prot sequence is highlighted in green. This table is followed by the contig view window displaying the predicted ORFs as directional arrows as per the orientation of the ORFs on the contig. The best match is displayed as a green-colored arrow. Each arrow can be clicked to retrieve the nucleotide and protein sequences of the predicted ORFs. This window is followed by a table providing summary information for the best matching ORF. The next table provides information on the closest available PDB structure and displays the 3-D protein structure.

In order to provide a useful indicator for the goodness of the results, we have provided a “MetaBioME Rating,” which rates the best matching ORF on a scale of 1–5 stars, with a single star for lowest match and five stars for

the best match. In the case of a good match, users are advised to carry out an “Advanced Search,” which helps to confirm the goodness of the results by using a suite of options. Users can check the alignment of the Swiss-Prot sequence of the selected biocatalyst with the best matching ORF. Since conserved motifs likely play a key role in the activity of an enzyme, all Swiss-Prot sequences belonging to the same EC number can be aligned together or with the best matching ORF to find the overall sequence homology among these sequences. This helps in the identification of conserved motifs and confirms if the best matching ORF also possesses any conserved motifs which may be present in the Swiss-Prot sequences. As another functional confirmation, users can also look for the presence of conserved domains in the best matching ORF by aligning the sequence against the NCBI Conserved Domains Database (CDD).

Additionally, the user can also check if the same Swiss-Prot sequence of the biocatalyst in question is present in any other metagenomic

MetaBioME
Comprehensive Metagenomic BioMining Engine

Home Meta Search **CUEsXplorer** Meta Explorer Meta Align Tutorial Links

CUEs
Commercially Useful Enzymes

Home
Metagenome
Enzyme Info
Data Sources
Riken
Team
Feedback

Explore curated CUEs by EC Classification Application Category [Download] / [View] complete list

Enzyme Commission #	Class	Function
<input checked="" type="radio"/> EC 1	Oxidoreductases	catalyze oxidation/reduction reactions
<input type="radio"/> EC 2	Transferases	transfer a functional group (e.g. a methyl or phosphate group)
<input type="radio"/> EC 3	Hydrolases	catalyze the hydrolysis of various bonds
<input type="radio"/> EC 4	Lyases	cleave various bonds by means other than hydrolysis and oxidation
<input type="radio"/> EC 5	Isomerases	catalyze isomerization changes within a single molecule
<input type="radio"/> EC 6	Ligases	join two molecules with covalent bonds

Select an enzyme for the EC class selected above

Select an enzyme

- 1.1.1.1 (Alcohol dehydrogenase.)
- 1.1.1.2 (Alcohol dehydrogenase (NADP(+)))
- 1.1.1.4 ((R,R)-butanediol dehydrogenase.)

MetaBioME, Fig. 4 Screenshot of “CUEsXplorer” page

dataset by carrying out a homology search against other metagenomic datasets. Another search option is provided to determine if the novel predicted ORF sequence is already present or has a close match with any protein from a known genome available in the Non-Redundant (NR) database. These additional options are helpful in confirming the uniqueness of the novel identified biocatalyst.

CUEsXplorer: Explore Commercially Useful Enzymes (CUEs) Database

This page provides options for exploring the CUEs database for any application category or EC classification. It provides details about enzyme function and the curation summary of any selected enzyme (Fig. 4).

MetaExplorer: Search for Enzymes in Metagenomic Datasets

This query page provides an option to select and search for any enzyme from the six EC classes in metagenomic datasets (Fig. 5).

On selecting any EC class, a list box containing all EC numbers belonging to that class opens up. Selecting an EC number from this list box reveals an expanded page with information on the enzyme name, EC number, Prosite ID, enzymatic reaction, KEGG pathway, and list of all Swiss-Prot IDs belonging to that EC

number. Any representative Swiss-Prot sequence can be selected and searched by TBLASTN in one or more metagenomic datasets selected from the drop-down menu. The results “MetaSearch Results” and profile “MetaBioME Profile” pages, for the submitted query, are similar to as explained in the earlier section. This query page provides users with an option to search all known enzymes, as available in EC, irrespective of their known role as biocatalyst, which is a subset of this set.

MetaAlign: Online Application to Search for Protein Sequences in Metagenomic Datasets

MetaAlign is an application powered by the BLAT (faster and less sensitive) and BLAST (slower and more sensitive) sequence alignment tools (Fig. 6).

It provides the user an option to carry out homology-based searches for single or multiple (multi-FASTA format) submitted nucleotide or protein sequences against the metagenomic sequences available in the ten metagenomic datasets. Larger files containing multiple sequences can also be uploaded, with an email being sent to the user on completion of analysis. The searches can be limited by selecting the threshold E-value and the number of resultant hits. The output format can also be specified as “tabular” or “full” (complete alignment).

Search for enzymes in metagenomic or bacterial genomic datasets

Enzyme Commission #	Class	Function
<input checked="" type="radio"/> EC 1	Oxidoreductases	catalyze oxidation/reduction reactions
<input type="radio"/> EC 2	Transferases	transfer a functional group (e.g. a methyl or phosphate group)
<input type="radio"/> EC 3	Hydrolases	catalyze the hydrolysis of various bonds
<input type="radio"/> EC 4	Lyases	cleave various bonds by means other than hydrolysis and oxidation
<input type="radio"/> EC 5	Isomerases	catalyze isomerization changes within a single molecule
<input type="radio"/> EC 6	Ligases	join two molecules with covalent bonds

Select an enzyme for the EC class selected above

Select an enzyme

- 1 1 1 1 (Alcohol dehydrogenase)
- 1 1 1 2 (Alcohol dehydrogenase (NADP(+)))
- 1 1 1 3 (Homoserine dehydrogenase)

MetaBioME, Fig. 5 Screenshot of “MetaXplorer” page

Discussion

There is so much richness and natural diversity inherent in the metagenomic data that the possibility of retrieving functional genes of interest is almost certain. This is further assured with the availability of more metagenomic datasets, deeper coverage, and completed genomic sequences. Therefore, a computational homology-based approach search engine such as MetaBioME has great potential to reveal novel alternatives for existing biocatalysts.

To look for an “ideal biocatalyst,” however, is not an easy task, since the requirements and conditions of the bioprocesses are not constant. Generally, an “ideal” catalyst is defined in terms of turnover number (k_{cat}) or, for a given process, in terms of the maximum specificity constant (k_{cat}/K_M) (Burton et al. 2002). However, from a bioprocess viewpoint, each bioprocess is constrained by a set of conditions governed by the specific properties of the substrates, products, and the bioconversion reaction. Thus, the currently used microbial biocatalysts, whose selection has been limited by the limited number of available genomes, may not be “ideal” and sometimes, the industrial processes have to be designed to fit only mediocre enzymes (Lorenz and Eck 2005).

Therefore, MetaBioME does not involve an exclusive approach in looking for ideal

biocatalysts, but employs an inclusive approach to identify all possible alternatives with reasonable criterion. For any given function (EC number), MetaBioME reports all possible ORFs (with stringent cutoff similarity) from the naturally existing diverse protein repertoire of yet unidentified microbial genomes which have evolved and survived in diverse environments. Thus, each resultant metagenomic ORF having significant similarity to a known biocatalyst is unique with distinct characteristics such as thermodynamic and pH stability, turnover frequency, specific activity, etc., offering a wide choice for their selection and employment as per the requirements for a given bioprocess. This approach is especially useful for pharmaceutical and supporting fine-chemical companies, both of which explore multiple diverse biocatalysts to construct their local databases for biotransformations (Lorenz and Eck 2005).

The alternative novel biocatalysts found using MetaBioME can serve as leads for further experiments involving cloning and expression to establish their enzymatic activity and commercial potential. Therefore, a combination of computational predictions of MetaBioME with activity-based mining and subsequent tailoring of these proteins using bioengineering techniques could provide a proficient prospect to replace chemical synthesis with biotechnological processes, which are ultimately more sustainable to mankind.

MetaSystems Research Team

MetaBioME
Comprehensive Metagenomic BioMining Engine

Home Meta Search CUEs Explorer Meta Explorer Meta Align Tutorial Links

CUEs
Commercially Useful Enzymes

Home
Meta Explorer
Enzyme Info
Data Sources
RIKEN
Team
Feedback

Please select an option

Search your enzyme sequences in (meta)genomic datasets

Search your (meta)genomic sequences for CUEs

Sequence Type

Search Using BLAST (Slow and sensitive) BLAT (Fast and less sensitive)

Paste your sequence(s) here : [Max. 300,000 characters or 50 sequences]

OR

Upload File No file chosen [max 10 MB]

E-value [Default is 10⁻⁶]

Metagenomic Project Completed Bacterial Genomes

Dataset

Output Format

MetaBioME, Fig. 6 Screenshot of “MetaAlign” page

Methods

Enzyme Database

We have used the Enzyme Commission number (EC number) as a numerical classification scheme for enzymes based on the chemical reactions they catalyze, with each EC class exclusively defining the function performed by the enzyme (Bairoch 2000). Information on the complete set of 4,877 enzymes annotated with an EC number was retrieved from the ENZYME nomenclature database, as available at ExPASy. Swiss-Prot sequences were retrieved from the

Swiss-Prot database (O’Donovan et al. 2002) for the different enzymes belonging to these EC numbers. The remaining EC numbers did not have any known Swiss-Prot sequence. An EC number in this analysis is used exclusively to refer to an enzyme and defines its function.

We curated a database of 510 microbial enzymes, with known or potential commercial applications as “biocatalysts,” by mining the information available at BRENDA (Barthelmes et al. 2007), NCBI (Wheeler et al. 2008), ExPASy (Gasteiger et al. 2003), and available literature. These biocatalysts were classified into nine broad

application categories, namely, Agriculture, Biotechnology, Energy, Environment, Enzymatic Analysis, General Applications, Industry, Medical, and Nutrition. These broad application categories were further subclassified into 21 more specific subcategories.

Other Resources

The Non-Redundant (NR) and **Conserved Domains Database (CDD)** were retrieved from NCBI (<ftp://ftp.ncbi.nih.gov/blast/db>), and Protein Data Bank (PDB) database was retrieved from the Worldwide Protein Data Bank (wwPDB) (<http://www.wwpdb.org/>). Protein structures were created using RasMol (version 2.6).

Mining the Metagenomic Databases

The publicly available metagenomic data from ten diverse environments is analyzed in the current version of the database. Of these, the Sargasso Sea [SSEA] dataset was retrieved from the J. Craig Venter Institute (<https://research.venterinstitutione.org/sargasso/>) (Venter et al. 2004b), and the remaining nine datasets, including sludge [SLUDGE] (Garcia et al. 2006), acid mine drainage [AMD] (Tyson et al. 2004), whale fall [WFALL] (Tringe et al. 2005), soil [SOIL] (Tringe et al. 2005), human gut (2 individuals) [HGUTI] (Gill et al. 2006), human gut (13 individuals) [HGUTII] (Kurokawa et al. 2007), mouse gut [MGUT] (Turnbaugh et al. 2006), termite gut [TGUT] (Warnecke et al. 2007), and human fecal virus [HFV] (Zhang et al. 2006), were retrieved from the DDBJ database (ftp://ftp.ddbj.nig.ac.jp/database/wgs/WGS_ORGANISM_LIST.html). The sequences available in these datasets are referred to as “contigs” by the authors; therefore, we have called them “contigs” in this analysis. However, we realize that several metagenomic sequences in these datasets are too short to be called contigs and are likely singletons.

Swiss-Prot protein sequences of known biocatalysts were aligned with their corresponding nucleotide sequences (contigs) in each metagenomic dataset using TBLASTN with a threshold of $E < 10^{-6}$, with only the best ten

matches being considered, and the output was generated in XML format.

Complete and partial ORFs (open reading frames) were predicted in the metagenomic sequences using the Glimmer (Delcher et al. 2007) and MetaGene (Noguchi et al. 2006) gene prediction programs with a minimum length of 50 amino acids (150 nucleotides). We adopted a self-training approach for implementing Glimmer by using the contig itself as the training sequence. Additional confidence for an ORF prediction is provided by integrating the results of MetaGene and Glimmer, using an in-house developed algorithm “SuperGene.” It called the ORFs as “Exact” (same start and end predicted by both methods), “End_match” (start is variable and only end is matching), and “Unique” (predicted by only one method). The “Exact” ORFs are certainly predicted with higher confidence with reliable start and end positions, because they were predicted by two independent methods. For the “End_match” cases, the longer ORF was kept in this analysis to ensure that no part of an ORF was left out, even if some extra part was included in the initial prediction. The exact start and end were further confirmed after alignment of the ORFs with their corresponding Swiss-Prot sequences. The ORFs lying at the terminals of the contigs were considered partial. The above data was imported into a MySQL database (Metabase).

Web Interface and Metabase Development

Apache (version 2.2.8), MySQL (version 5.0.45), PHP (version 5.2.4), and Perl (version 5.8.5) were used for development of the GUI. The back-end database was called as “Metabase.” The Web server was developed using Apache HTTP Server (version 2.2.8). Client-side scripting was done using XHTML, JavaScript, and AJAX, and server-side scripting was done using PHP and XML. The publicly available applications, BLAT (v34) (Kent 2002), BLAST (version 2.2.17) (Wheeler et al. 2008), and MAFFT (version 6.240) (Katoh et al. 2005), were used for additional analysis.

Cross-References

- ▶ [Binning Sequences Using Very Sparse Labels Within a Metagenome](#)
- ▶ [Challenge of Metagenome Assembly and Possible Standards](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [FragGeneScan: Predicting Genes in Short and Error-Prone Reads](#)
- ▶ [MetaBin](#)
- ▶ [MEtaGenome ANalyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [New Computational Methodologies to Understand Microbial Diversity](#)
- ▶ [Next-Generation Sequencing for Metagenomic Data: Assembling and Binning](#)
- ▶ [NGS QC Toolkit: A Platform for Quality Control of Next-Generation Sequencing Data](#)

References

- Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev.* 1995;59:143–69.
- Bairoch A. The enzyme database in 2000. *Nucleic Acids Res.* 2000;28:304–5.
- Barthelme J, Ebeling C, Chang A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.* 2007;35:D511–4.
- Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* 2005;33:3390–400.
- Burton SG, Cowan DA, Woodley JM. The search for the ideal biocatalyst. *Nat Biotechnol.* 2002;20:37–45.
- Daniel R. The metagenomics of soil. *Nat Rev Microbiol.* 2005;3:470–8.
- Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics.* 2007;23:673–9.
- Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol.* 2005;3:504–10.
- Ferrer M, Martinez-Abarca F, Golyshin PN. Mining genomes and ‘metagenomes’ for novel catalysts. *Curr Opin Biotechnol.* 2005;16:588–93.
- Garcia MH, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol.* 2006;24:1263–9.
- Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 2003;31:3784–8.
- Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. *Science.* 2006;312:1355–9.
- Katoh K, Kuma K, Miyata T, Toh H. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform.* 2005;16:22–33.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12:656–64.
- Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 2007;14:169–81.
- Lorenz P, Eck J. Metagenomics and industrial applications. *Nat Rev Microbiol.* 2005;3:510–6.
- Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 2008;24:133–41.
- Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 2006;34:5623–30.
- O’Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform.* 2002;3:275–84.
- Sharma VK, Kumar N, Prakash T, Taylor TD. MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets. *Nucleic Acids Res.* 2010;38(Database issue):D468–72.
- Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet.* 2005a;6:805–14.
- Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet.* 2005b;6:805–14.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, et al. Comparative metagenomics of microbial communities. *Science.* 2005;308:554–7.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 2006;444:1027–31.
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovvey VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004;428:37–43.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004a;304:66–74.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE,

- Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004b;304:66–74.
- Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. 2007;450:560–5.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2008;36:D13–21.
- Yun J, Ryu S. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact*. 2005;4:8.
- Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol*. 2006;4:e3.

MEtaGenome Analyzer (MEGAN): Metagenomic Expert Resource

Daniel H. Huson
Center for Bioinformatics, Algorithms in
Bioinformatics, University of Tübingen,
Tübingen, Germany

Synonyms

MEGAN = MEtaGenome ANalyzer

Definition

MEGAN is a tool for analyzing metagenomic sequencing data, allowing the user to interactively explore the taxonomic and functional content of a dataset. It also supports the comparison of multiple datasets. The program was originally published in (Huson et al. 2007), and the most recent version was published in (Huson et al. 2011). Written in Java, the program runs on all major operating systems. The program can be downloaded from <http://www-ab.informatik.uni-tuebingen.de/software/megan>.

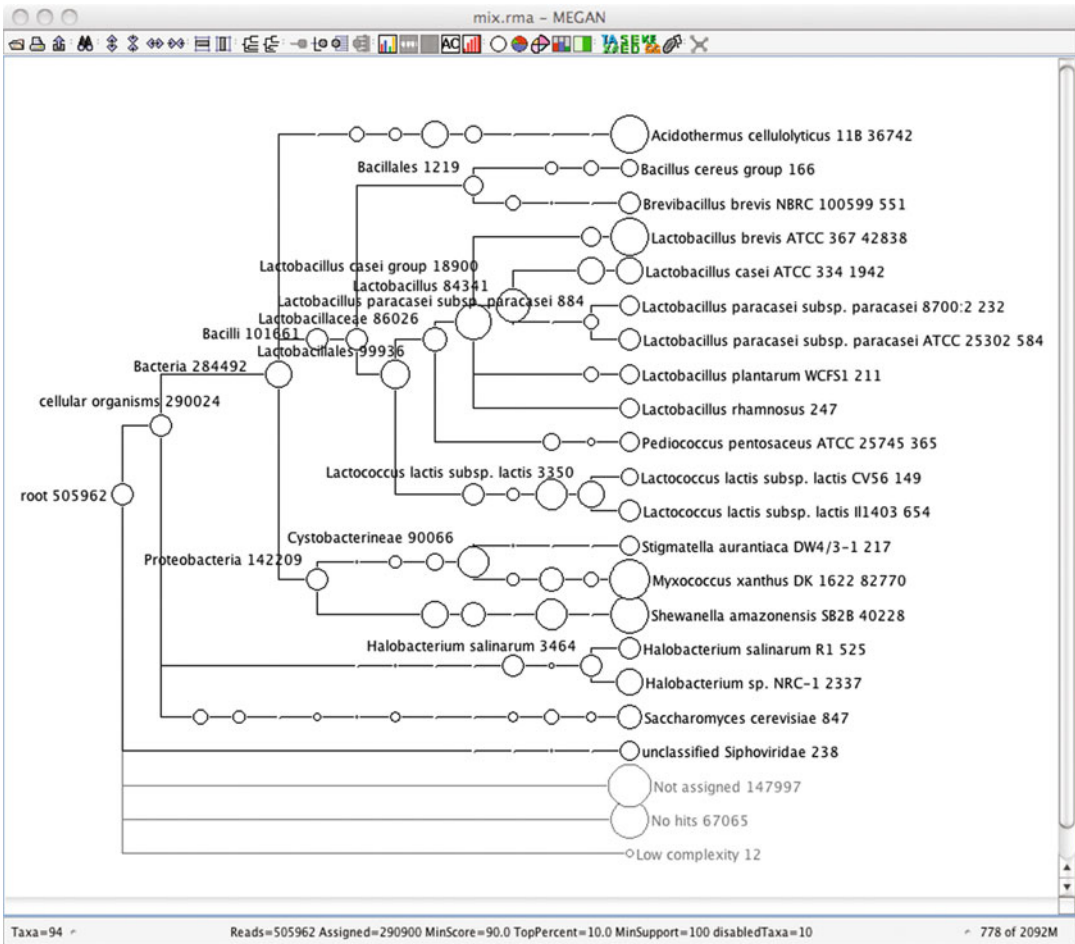
Introduction

Metagenomics is the study of uncultured organisms in their native environment using DNA sequencing (Handelsman et al. 1998). In a typical project, DNA (or, in the case of metatranscriptomics, cDNA reverse-transcribed from RNA) is extracted from an environmental sample and then shotgun sequenced. Once a metagenome dataset of DNA sequencing reads has been generated in this way, the first three main computation challenges are to (1) estimate the taxonomic content of the sample, (2) estimate its functional content, and (3) compare different samples.

To address these challenges, the first step is to align the set of sequencing reads against a database of known reference protein sequences such as NCBI-NR or RefSeq (Benson et al. 2005) using a pairwise alignment tool such as BLASTX (Altschul et al. 1990) or RapSearch2 (Zhao et al. 2012). A read is said to *hit* a given reference sequence, if a significant alignment is found in this process. The comparison of the sequencing reads against a reference database is usually the computationally most expensive step of analysis, and subsequent steps are based on the obtained alignments. Given the result of the alignment step, an analysis program such as MEGAN is then required to explore and analyze the data.

Taxonomic Analysis

To perform a taxonomic analysis of a metagenomic dataset, MEGAN attempts to place each read onto a node in the NCBI taxonomy, based on an analysis of its hits. The key idea is to use all ranks of the taxonomy so as to assign reads specific to a particular species near the leaves of the taxonomy and to map sequences that are conserved across a wider range of organisms to higher-level nodes. For example, a read that comes from a gene that only *Escherichia coli* has will be placed on the *E. coli* node, whereas a read that comes from a gene that is shared widely across different *Proteobacteria* will be assigned to the node labeled *Proteobacteria*.



MEtaGenome Analyzer (MEGAN): Metagenomic Expert Resource, Fig. 1 Taxonomy analysis of $\approx 500,000$ reads from an in vitro-simulated microbial community Morgan et al. (2010). Each *circle* represents a taxon in the NCBI taxonomy and is scaled

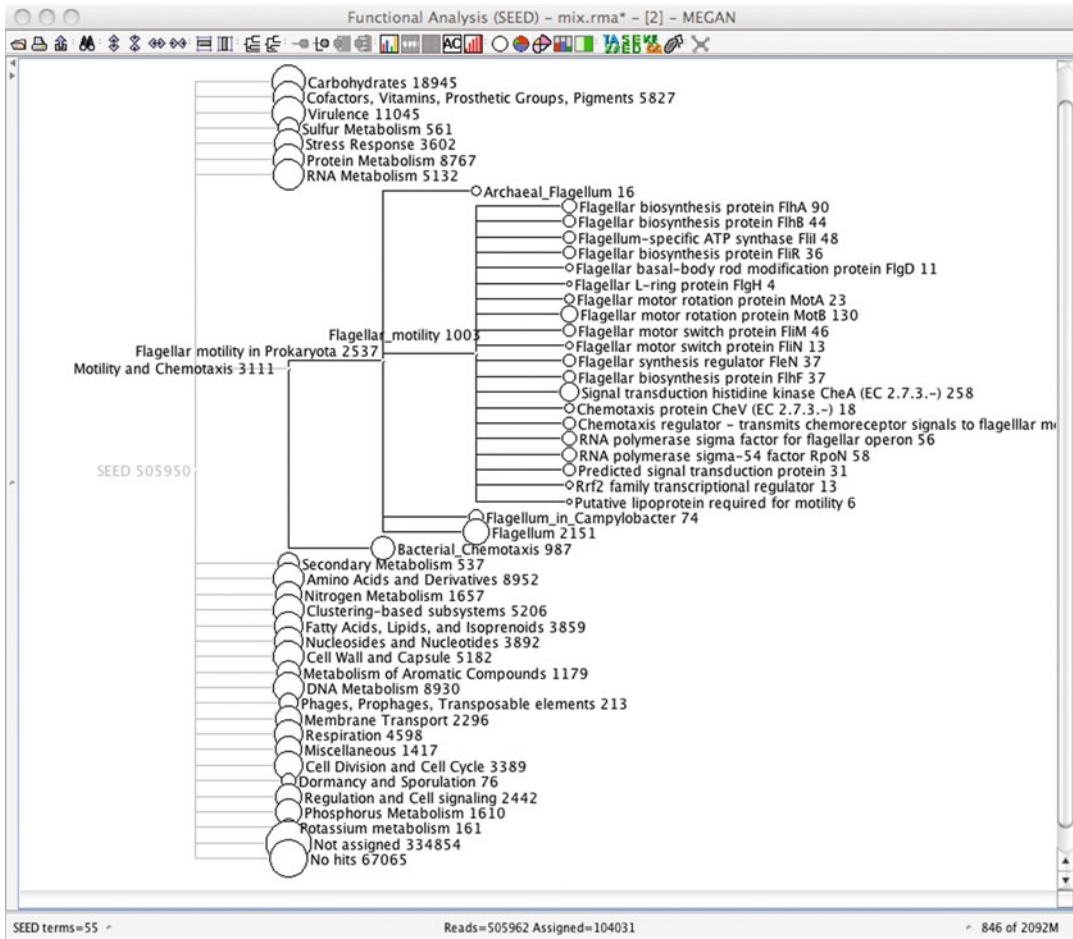
logarithmically to indicate how many reads have been assigned to it. In addition to the taxon name, each node is also labeled by the cumulative number of reads assigned to, or below, that node

The input to MEGAN is a file of DNA reads and a file containing all their hits in a reference database, usually in BLAST or SAM format. In addition, at start-up, MEGAN reads in the whole NCBI taxonomy. To perform a taxonomic analysis of a metagenome dataset, MEGAN processes each DNA read in turn, assigning each read to the node in the NCBI taxonomy that is the *lowest common ancestor* of the set of species associated with all reference sequences that were hit by the read. This approach is known as the LCA algorithm.

The LCA algorithm has a number of parameters, such as *minScore*, the minimum bit score that

an alignment must achieve to be considered; *minPercent*, a further filter to remove all those hits whose bit score differences by more than the given percentage from the top scoring hit for the given read; and *minSupport*, the minimum number of reads that a node in the NCBI taxonomy must attract before it is shown in the final output.

Reads that have no hits are assigned to a special node labeled *No Hits*, whereas reads that have hits but cannot be assigned to a taxon are mapped to a special *Unassigned* node. In addition, reads consisting of highly repetitive sequence are assigned to a *Low Complexity* node.



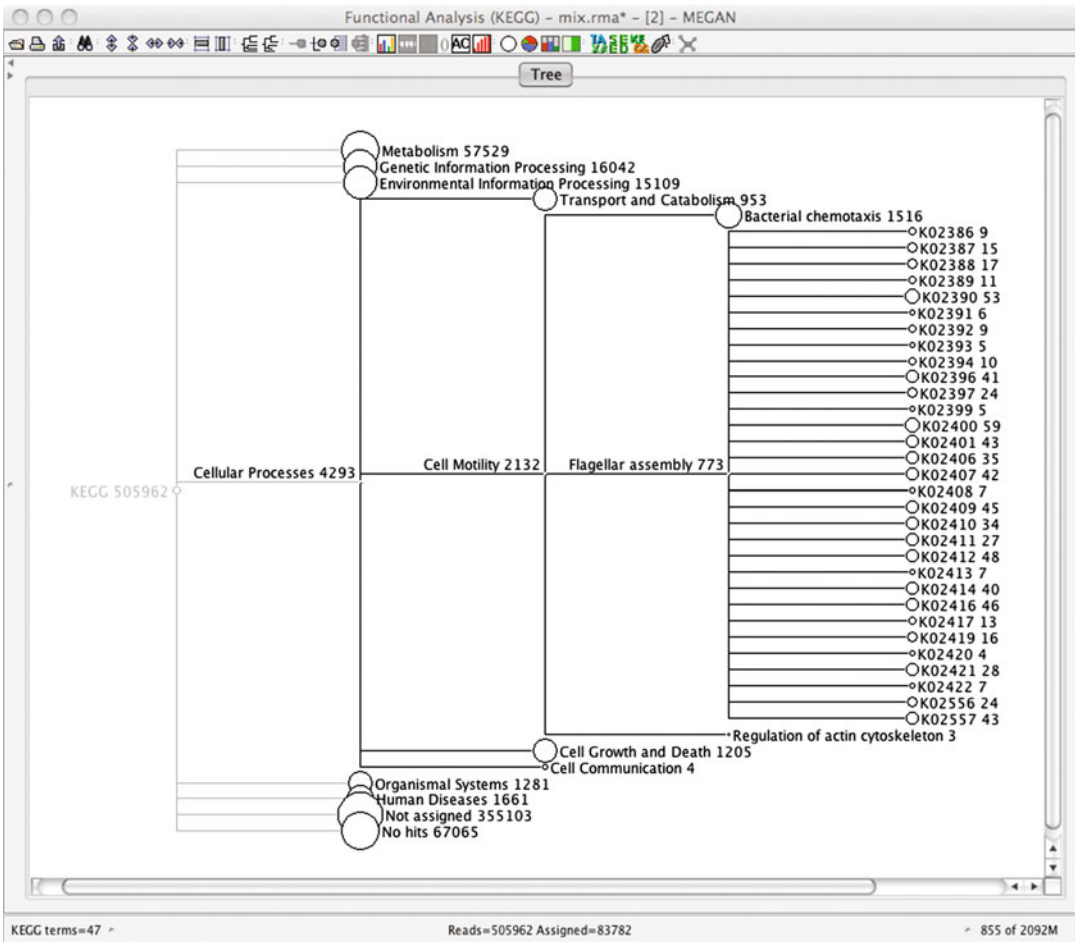
MEtaGenome Analyzer (MEGAN): Metagenomic Expert Resource, Fig. 2 SEED-based functional analysis of $\approx 500,000$ reads from an in vitro-simulated microbial community Morgan et al. (2010). The SEED classification tree has been partially expanded to show details on functional roles involved in flagellar motility.

The pertinent part of the NCBI taxonomy is displayed in the taxonomy viewer of MEGAN, and by default, each node is scaled logarithmically to represent the number of reads associated with it; see Fig. 1. Nodes can be interactively collapsed or expanded to show more or fewer details of the classification. The user can select nodes of interest and then either inspect the associated reads and alignments, or save them to a file, or chart them in a number of standard ways.

Each *circle* represents a SEED category and is scaled logarithmically to indicate the cumulative number of reads that have been assigned to it. In addition to the SEED name, each node is also labeled by the number of reads assigned to, or below, that node

Functional Analysis

MEGAN uses both the SEED (Overbeek et al. 2005) and the KEGG (Kanehisa and Goto 2000) classifications to analyze the functional content of a metagenome dataset. In essence, the SEED classification maps genes onto functional roles, and these appear in different subsystems. Similarly, KEGG maps genes onto KEGG orthology groups, or KO groups, which are associated with enzymes that appear in different



MEtaGenome Analyzer (MEGAN): Metagenomic Expert Resource, Fig. 3 KEGG-based functional analysis of $\approx 500,000$ reads from an in vitro-simulated microbial community Morgan et al. (2010). The KEGG classification tree has been partially expanded to show details on KO groups involved in flagellar assembly.

Each *circle* represents a KEGG category and is scaled logarithmically to indicate the cumulative number of reads that have been assigned to it. In addition to the KEGG name, each node is also labeled by the number of reads assigned to, or below, that node

pathways. In both cases, the classification can be represented as a tree with roughly 13,000 nodes.

To perform a SEED-based analysis, for each read in the input, MEGAN identifies the highest scoring hit to a reference sequence for which the corresponding functional role is known and then maps the read to that functional role. In a KEGG-based analysis, each read is mapped to a KO group in a similar fashion.

Both the SEED and KEGG classifications are displayed as trees in MEGAN, and the viewers provide the same interactive features as the

taxonomy viewer. In addition, the KEGG viewer allows one to see how reads map to different enzymes in a given pathway; see Figs. 2 and 3.

Sequence Alignment

As pointed out above, the main computational step is to compute pairwise alignments between the set of DNA reads and all sequences in an appropriate reference database. Based on this,



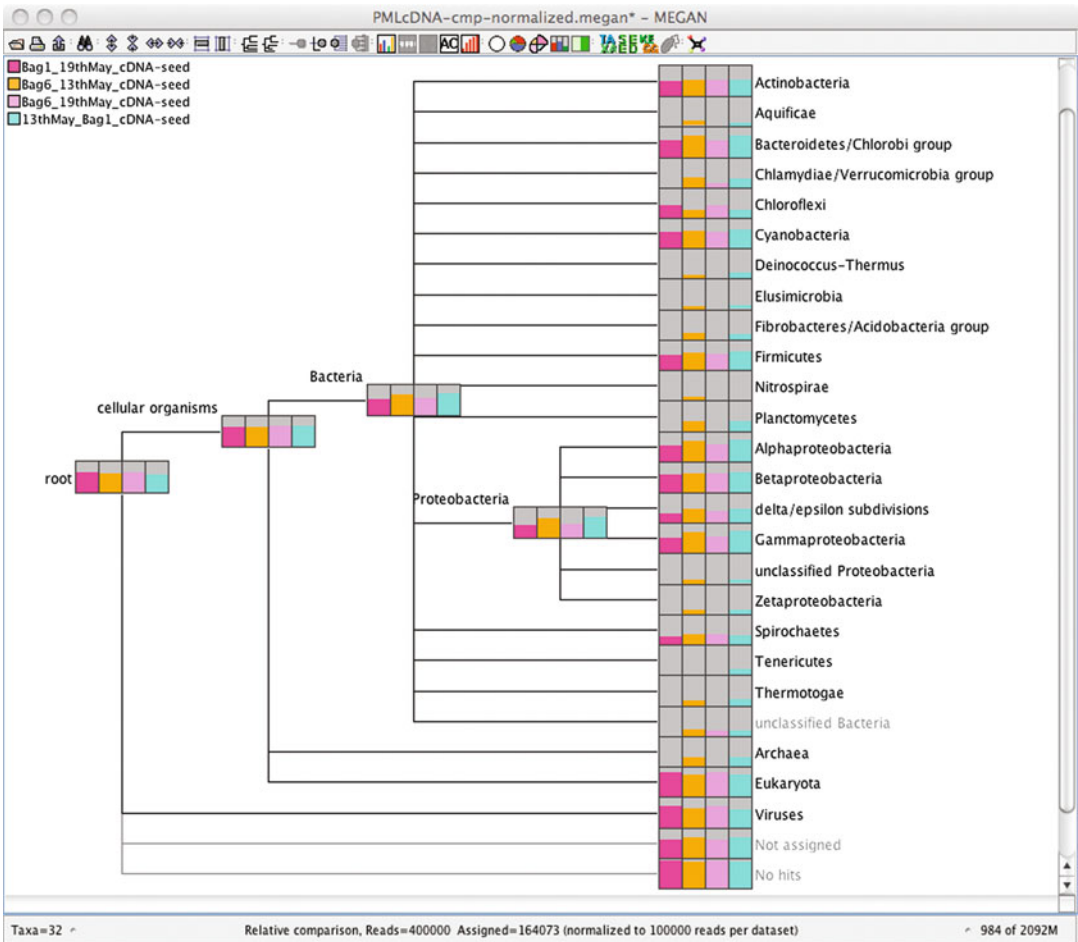
MEtaGenome ANalyzer (MEGAN): Metagenomic Expert Resource, Fig. 4 MEGAN’s alignment viewer constructs and displays a multiple sequence between all reads that map to the same reference sequence. The top

track shows the reference sequence and the main panel displays the aligned reads. Letters shown in gray belong to the reads but are not part of the alignment

it is possible to construct reference-guided multiple sequence alignments between all reads that hit the same reference sequence. This calculation is implemented in MEGAN in a new feature called the alignment viewer. Once the user has specified a node in the taxonomy, SEED or KEGG viewer for which the alignment viewer is to be launched, the program first collects all reference sequences that correspond to the given node and then, for each such reference sequence, the program determines all reads that hit it. The user can then select a reference sequence, and the corresponding sequence alignment is subsequently displayed; see Fig. 4.

Comparison of Datasets

To facilitate the comparison of datasets, MEGAN allows the user to open multiple datasets simultaneously, showing each dataset in a different window. The user can then select a number of open datasets to be combined into a single new comparison document. For such a document, the taxonomy, SEED, and KEGG viewers indicate how many reads were assigned to each node for each original input document by drawing the node as a pie chart or bar chart, for example, see Fig. 5. MEGAN also supports the calculation of standard ecological indices for a comparison



MEtaGenome ANALYZER (MEGAN): Metagenomic Expert Resource, Fig. 5 High-level comparison of taxonomic content of four different cDNA datasets from a seawater monitoring study (Gilbert et al. 2008). The four

different datasets are represented by different colors, and each node shows a bar chart that indicates the number of reads assigned to that node, on a logarithmic scale

document and then, based on this, the program can be used to compute a tree, network, or MDS plot (not shown here).

Handling Large Data

As sequencing technologies continue to improve, the size of analyzed datasets continues to increase. MEGAN was reportedly used to perform the taxonomic analysis of 124 human gut samples involving around 600 gigabases of sequence (Qin et al. 2010). In an ongoing study, MEGAN is currently being used to analyze a set

of 350 million reads with 1.3 billion BLASTX matches. While MEGAN is mainly designed for interactive use on a laptop or desktop computer, all features of the program can also be accessed in command-line mode, and thus analyses can also be performed on a server within the framework of a larger bioinformatic analysis pipeline.

Summary

MEGAN is an interactive tool for analyzing the taxonomic and functional content of metagenomic (and metatranscriptomic) datasets.

Input is a set of DNA reads and the result of comparing the reads against a reference database. Taxonomic analysis is performed by placing DNA reads onto nodes of the NCBI taxonomy, whereas functional analysis is based on mapping reads to SEED and KEGG categories. The program supports comparative analysis of multiple datasets. The program is written in Java and runs on all major operating systems. When run in command-line mode, the program can also be integrated into larger bioinformatic analysis pipelines.

Cross-References

► [Metagenomics, Metadata, and Meta-analysis](#)

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D. Genbank. *Nucleic Acids Res.* 2005;1(33):D34–8.
- Gilbert JA, Field D, Huang Y, Edwards R, Li W, Glinn P, Joint I. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One.* 2008;3:e3042.
- Handelsman J, Rondon M, Brady S, Clardy J, Goodman R. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:245–9.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17(3):377–86.
- Huson DH, Mitra S, Weber N, Ruscheweyh H-J, Schuster SC. Integrative analysis of environmental sequences using megan4. *Genome Res.* 2011;21:1552–60.
- Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
- Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE.* 2010;5.
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33(17):5691–702.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464(7285):59–65.
- Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics.* 2012;28(1):125–6.

Metagenome of Acidic Hot Spring Microbial Planktonic Community: Structural and Functional Insights

Diego Javier Jiménez¹ and María Mercedes Zambrano²

¹Department of Microbial Ecology, University of Groningen, Center for Ecological and Evolutionary Studies (CEES), Groningen, The Netherlands

²Molecular Genetics and Microbial Ecology, Corporación CorpoGen, Bogotá, DC, Colombia

Synonyms

The microbiome of Andean acidic hot springs

Definition

Metagenomic analyses were done to obtain a deeper view of the microbial community structure and to gain insight regarding the functional properties present in the planktonic fraction of these Neotropical high Andean acidic hot springs.

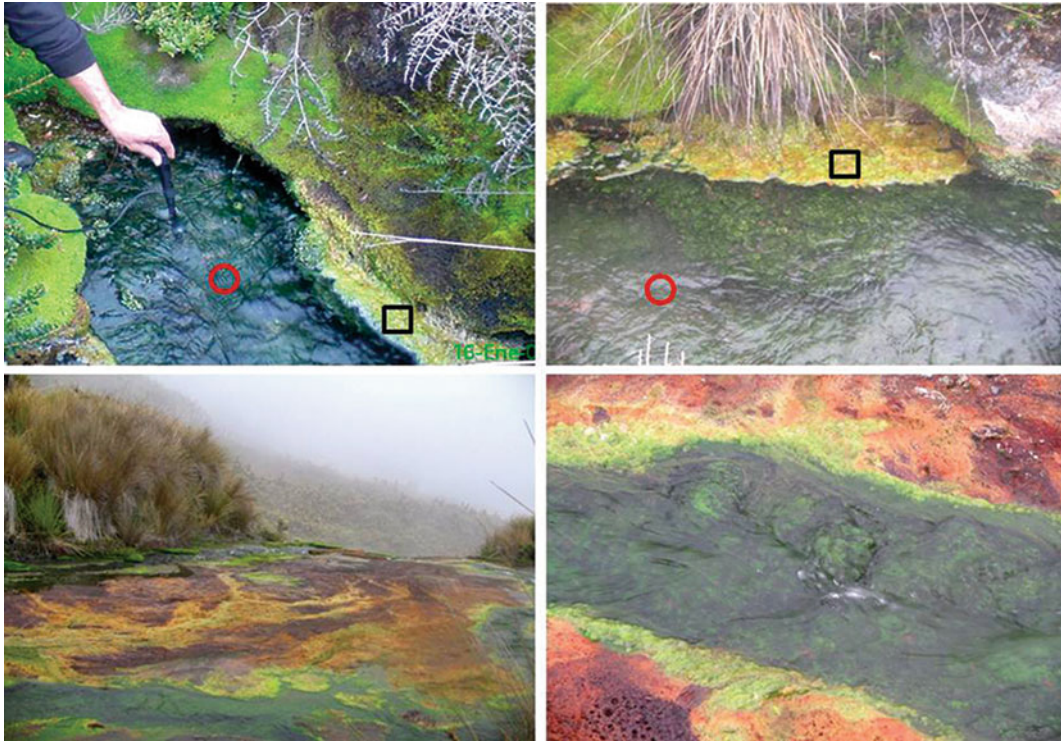
Introduction

High-mountain Andean ecosystems are rich in biodiversity and natural resources (Myers et al. 2000). The South American Andean region is part of what is known as the “Ring of Fire” and has several hot springs that represent unique and undisturbed extreme environments due to their high elevation and exposure to ultraviolet (UV) light. These springs are heated mainly by the underlying magma chamber from volcanic activity; they are oligotrophic and vary in their geochemistry, such as mineral content, temperature, and pH. Thus far, little is known regarding the microbiomes of these high-mountain ecosystems. A hot spring is characterized by discharge of hot water from a vent. There is, however, no universally accepted definition of “hot,” and the temperature for distinguishing a “warm spring” from a “hot spring” remains contentious (Rzonca and Schulze-Makuch 2003; Pentecost et al. 2003; Jones and Renaut 2011). Hot springs contain several microhabitats such as the planktonic fraction (which has low cell density), microbial mats, and sediments (with high cell density) each with different microbial assemblages. The microbial diversity in the planktonic fraction is dictated by environmental physicochemical characteristics as a pH, redox potential, temperature, and concentration of trace elements (Siering et al. 2006; Mathur et al. 2007). Metagenomic (total DNA), metataxonomic (16S rRNA and/or ITS sequences), meta-transcriptomic (mRNA), and PCR-target analyses have been extremely valuable for describing the microbial structure and functionality in different hot springs (López-López et al. 2013; Wemheuer et al. 2013; Liu et al. 2011). Cyanobacteria and Chloroflexi, for example, are abundant in low temperature-sediment samples from high-mountain hot springs located in the Tibetan plateau (Wang et al. 2013). A previous analysis of the planktonic microbial community in one Colombian acidic hot spring (*El Coquito*) located in the national park of *Los Nevados* showed that Bacteria rather than Archaea dominated the community, with predominance of Proteobacteria, Firmicutes, and Planctomycetes (Bohórquez et al. 2012a).

These acidic-hot ecosystems are also of interest as a source of potential biotechnological products, new species (Tirawongsaroj et al. 2008; Bouraoui et al. 2013), and features relevant to ecosystem maintenance and ecology such as horizontal gene transfer, UV damage, and biogeochemical cycles. The microbial planktonic community contained putative chemotrophic bacteria potentially involved in cycling of ferrous iron and sulfur-containing minerals. In extremely acidic and UV light-irradiated hot springs, primary production may also be mediated by phototrophic acidophiles (mainly eukaryotic micro-algae) (Aguilera et al. 2010). However, the presence of bacterial-rhodopsin photosystems has been reported to complement the chemotrophic lifestyle (Bohórquez et al. 2012b).

Metataxonomic Approach: Microbial Diversity Assessment by 16S rRNA Sequences

Microbial diversity in terrestrial hot springs has been extensively studied in locations as varied as Yellowstone National Park (YNP), Japan, New Zealand, Great Basin, Iceland, Thailand, the Philippines, Russia, and the Tibetan plateau. These surveys, done mostly by 16S rRNA gene analysis, have expanded our view of the microbial communities present in these extreme and difficult-to-study water ecosystems. An elegant multi-approach based on 16S rRNA analysis of an acidic hot spring in the Colombian Andes, called *El Coquito* (EC) (Fig. 1), was recently carried out using high-throughput sequencing, PhyloChip, and 16S rRNA clone libraries (Bohórquez et al. 2012a). The EC hot spring is located at 3,973 m above sea level and is characterized by an acidic pH (2.7), high solar radiation (~9–11 mW/cm² nm UV-B), and high sulfate content (1,003 mg SO₄⁻² L⁻¹). This spring is moderately hot, with a water temperature of approximately 29 °C, which is considerably higher than ambient temperature (~9 °C) (Rzonca and Schulze-Makuch 2003). Despite differences among the results obtained with the three strategies used to analyze the microbial diversity of



Metagenome of Acidic Hot Spring Microbial Planktonic Community: Structural and Functional Insights, Fig. 1 Photographs of the acidic hot spring

El Coquito (EC); red circle indicates the planktonic fraction and black square indicates the biofilm surface formation

M

this ecosystem, there was dominance of the orders Burkholderiales, Legionellales, Rhodospirillales, Rhodocyclales, Clostridiales, Planctomycetales, Nitrospirales, Rhizobiales, and Acidomicrobiales. The most abundant genera belonged to *Acidithiobacillus*, *Acidiphilium*, *Leptospirillum*, *Thiomonas*, *Acidocella*, and *Acidisphaera*. In general, the community was reminiscent of those found in hot and acidic environments with mesophilic organisms (Norris 2001; Stout et al. 2009). The high abundance of chemolithoautotrophic and heterotrophic acidophiles suggested that primary production in this community could be driven by solar energy at the surface and by inorganic chemicals that affect the biogeochemistry of iron and sulfur in the water. A more recent evaluation of 16S rRNA sequences present in EC hot spring, based on analysis of whole metagenome sequencing, which thus eliminates biases associated with PCR and cloning (Jiménez et al. 2012), showed consistent results

and prevalence of Gammaproteobacteria, Alphaproteobacteria, and Betaproteobacteria (25 %), followed by micro-algae chloroplast ribosomal DNA (15 %), Firmicutes (14 %), and Bacteroidetes (6 %). Both studies detected oxygenic eukaryotic phototrophs that could be present both in the planktonic fraction and in mat communities. Overall, the community was dominated by Bacteria rather than Archaea; it had a large proportion of novel and unclassified sequences and the presence of eukaryotic micro-algae. In addition, the presence of chemolithoautotrophic acidophiles in this high-mountain thermal spring suggested that primary production could be driven by chemical energy in the water, as well as by solar energy at the surface.

A comparative study of the planktonic microbial communities in five high-mountain hot springs was also carried out by 16S rRNA gene assessment. The springs, which varied in altitude, geographical location, and geochemical

characteristics, also showed differences in terms of diversity indexes. However, certain bacterial phyla showed predominance in all of them: Proteobacteria, Aquificae, Chloroflexi, Cyanobacteria, Firmicutes, Nitrospirae, and Thermotogae. Based on cluster analysis of the microbial populations, these spring communities grouped together in a manner consistent with sample physicochemical parameters, with pH and sulfate concentration being the parameters that most influenced the population structure. Some springs were also characterized by site-specific bacterial taxa that distinguished each community. Thus despite their geographic proximity and similar origins, the environmental factors at each location have resulted in marked differences in the microbial assemblages present.

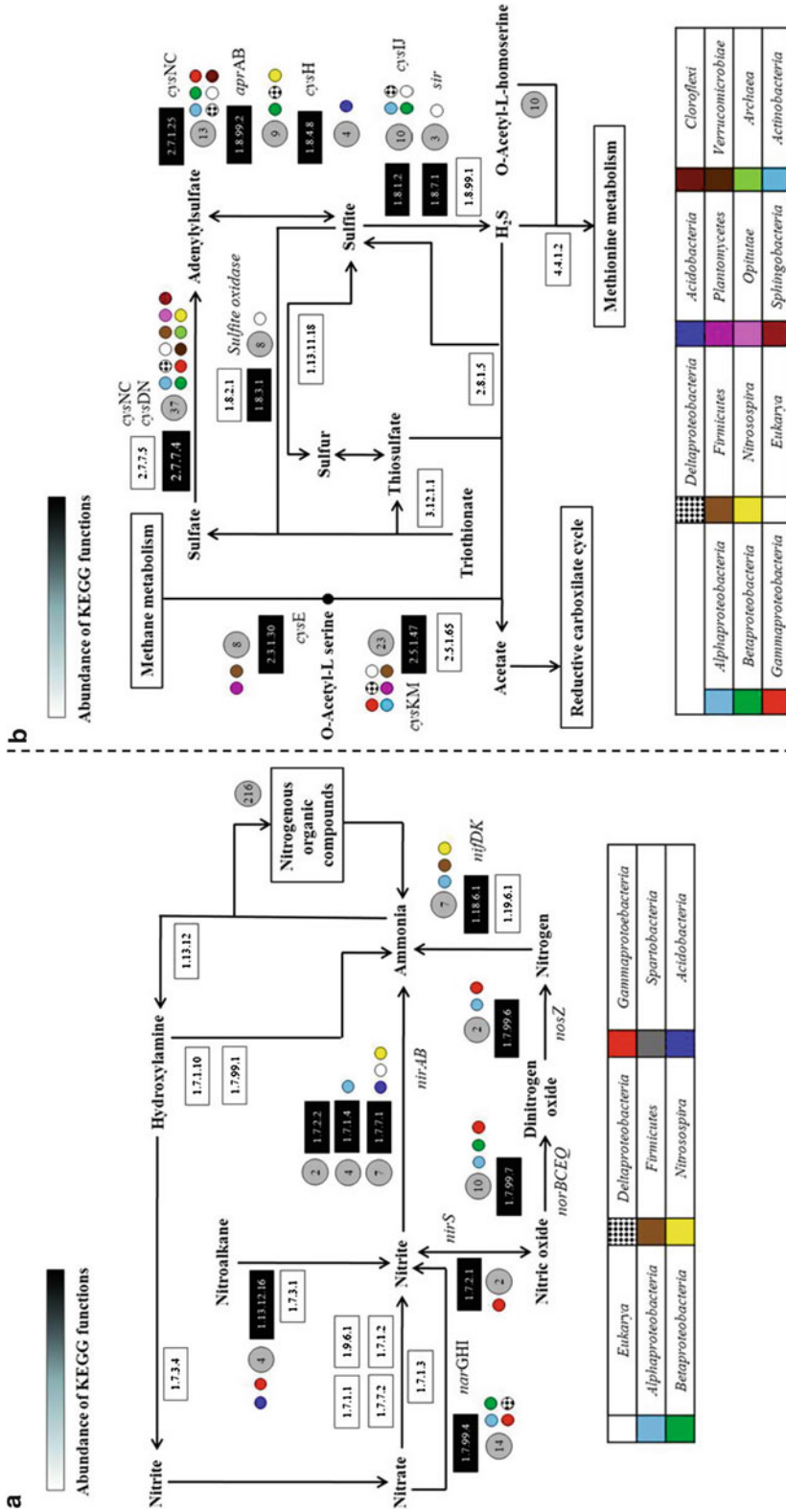
Metagenomic Approach: Taxonomic and Functional Assignment of Metagenome Sequences

Although 16S rRNA gene analysis is very useful for assessing microbial diversity, it does not provide ecologically relevant functional information. Thus a direct analysis of total metagenomic sequences becomes relevant. The current and most frequently used tools for taxonomic and functional classification of metagenomic reads are based on local alignments (BLAST) against different databases and associating best hits to taxa, specific genes, functional identifiers, or metabolic pathways (Montaña et al. 2012). An analysis was therefore carried out with 53 Mb of metagenomic information retrieved from a planktonic fraction of the EC hot spring (Jiménez et al. 2012). However, only 8,121 reads (2.9 %) of the total reads could be assigned to a taxonomic category, suggesting a great amount of newly described sequences or a large amount of noncoding DNA present in these genomes (especially in microeukaryotes). A high number of sequences were related to Acidithiobacillales (represented by sequences related to *Acidithiobacillus caldus*, *Acidithiobacillus ferrooxidans*, and *Acidithiobacillus thiooxidans*) followed by Legionellales

and Rhodospirillales that included *Acidiphilium cryptum* (1,681 assigned reads). A high proportion of sequences related to enzymes involved in transposition and integration of mobile genetic elements (transposases) were mapped to the *A. cryptum* JF-5 genome. By using BLASTX against the NCBI-nr database and the MEGAN v4.0 software, 19,876 sequences were associated with KEGG pathways, specifically to metabolism of carbohydrates (2,623), amino acids (2,584), energy (1,920), and nucleotides (1,431). A total of 87,023 reads (30.9 %) were assigned to 25 COG categories and most of the sequences were related to replication, recombination, and repair (10,712 reads), suggesting that these systems could be important in this ecosystem where high UV radiation, acidic pH, and high water temperature may cause significant damage to DNA. Deep sea hydrothermal vent chimneys and hot spring microbial communities are enriched in genes involved in mismatch DNA repair and homologous recombination, perhaps due to the need for extensive DNA repair systems to cope with extreme conditions that could have potential deleterious effects on their genomes (Klatt et al. 2011; Xie et al. 2011). In this study we also identified sequences associated with *quorum sensing* and cellular communication in biofilms, structures that could form on the surfaces of these acidic hot springs and could be relevant for ecosystem functionality (Fig. 1).

Metagenomic Approach: Nitrogen and Sulfur Transformations

Pathways involved in nitrogen and sulfur metabolism could be important in acidic hot spring habitats where terminal electron acceptors other than O₂ may be relevant, such as nitrate, ferric iron, arsenate, thiosulfate, elemental S, sulfate, or CO₂. Genes related to the dissimilatory reduction of nitrate to nitrite (*nar* GHI genes), conversion of nitrite to N₂ (*nir* K, *nir* S, *nor* B, *nos* Z), and associated with ferredoxin-nitrite reductase (*nir* A) were found in the metagenome of EC hot spring (Fig. 2a). The presence of *nif*



Metagenome of Acidic Hot Spring Microbial Planktonic Community: Structural and Functional Insights, Fig. 2 Partial (a) nitrogen and (b) sulfur pathways identified by KEGG affiliation of the sequences from EC hot spring. Boxes indicate the KEGG characteristic identified and numbers in gray circles indicate the amount of sequence reads affiliated to the KEGG function (Jiménez et al. 2012)

Metagenome of Acidic Hot Spring Microbial Planktonic Community: Structural and Functional Insights, Fig. 2 Partial (a) nitrogen and (b) sulfur pathways identified by KEGG affiliation of the sequences from EC hot spring. Boxes indicate the

K genes (associated with sulfate-reducing *Thermodesulfovibrio* and sulfur-reducing bacteria *Desulfotobacterium*) also indicated that in addition to denitrification, nitrogen fixation could also be taking place in this acidic hot spring. Based on taxonomic affiliation, the dissimilatory nitrate reduction is most likely carried out by Proteobacteria-like organisms, while assimilatory reduction of nitrate was associated mostly with acidophilic micro-algae, Acidobacteria, Spartobacteria, and Alphaproteobacteria (Jiménez et al. 2012). Conversion of sulfate into adenylylsulfate and, further, to generate sulfite and H₂S were also predicted from sequence analysis of the EC metagenome. This included genes involved in conversion of adenylylsulfate to sulfite (*apr* AB; *cys* H), in sulfite reduction and H₂S formation (*cys* I), and in the oxidation of sulfite to sulfate (sulfite oxidase enzymes) (Meyer and Kuever 2008). These pathways indicate that the oxidation of H₂S and (or) SO₂ could be linked to the acidity of the environment (Jones et al. 2012).

PCR-Target Approach: Proteorhodopsin-Like Genes in Andean Acidic Hot Springs

These Andean mountain hot springs are subjected to a large amount of solar light, yet taxonomic surveys identified only few phototrophic bacteria (Bohórquez et al. 2012a; Jiménez et al. 2012). Thus a search was conducted to identify energy-harvesting bacterial proteorhodopsins (PRs) that could also contribute to productivity in these ecosystems (Bohórquez et al. 2012b). PRs are retinal-binding bacterial transmembrane proton pumps that can generate energy from light, which are therefore important in terms of carbon cycling and energy flux in various aquatic ecosystems (Fuhrman et al. 2008). PCR with degenerate primers designed to target an internal conserved region in the PR gene was used to identify putative PR sequences. Recovered sequences showed between 80 % and 100 % identity at

the amino acid level with previously reported PR sequences from both freshwater and marine samples. These sequences contained conserved residues indicative of proton-pumping activity and of pigments that absorb green light. They harbored diversity at the amino acid level and clustered into three groups, showing similarity with both freshwater and marine sequences. The presence of these genes indicated that PR phototrophy might play a role in these oligotrophic high-mountain aquatic habitats exposed to abundant sunlight by providing a possible advantage that could contribute to survival.

Summary

The sequence-based exploration of the metagenomic content in Andean hot springs goes beyond the identification of taxa using 16S rRNA gene analysis and provides insight into metabolic potential and ecosystem function. Taxonomic surveys of EC spring and other similar springs indicated overall predominance of Bacteria over Archaea, even in the most acidic waters. Certain bacterial taxa predominated, but there were also site-specific groups at each spring, indicating that the surveyed microbiomes were different. The functional annotation showed that the microbial community in EC spring contained pathways involved in nitrogen and sulfur metabolism, as well as extensive DNA repair systems, possibly to cope with UV radiation at such high altitudes. Processes involved in denitrification, nitrogen fixation, and sulfide oxidation were likely linked to the acidity of the environment. Finally, the presence of PR sequences in these communities suggests that these genes might play a role important for bacterial survival in these aquatic ecosystems.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)

- ▶ [Biological Treasure Metagenome](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [KEGG and GenomeNet, New Developments, Metagenomic Analysis](#)
- ▶ [Lateral Gene Transfer and Microbial Diversity](#)
- ▶ [Metagenomic Potential for Understanding Horizontal Gene Transfer](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)

References

- Aguilera A, Souza-Egipsy V, González-Toril E, Rendueles O, Amils R. Eukaryotic microbial diversity of phototrophic microbial mats in two Icelandic geothermal hot springs. *Int Microbiol.* 2010;13(1): 21–32.
- Bohórquez LC, Delgado-Serrano L, Lopez G, Osorio-Forero C, Klepac-Ceraj V, et al. In-depth characterization via complementing culture-independent approaches of the microbial community in an acidic hot spring of the Colombian Andes. *Microb Ecol.* 2012a;63:103–15.
- Bohórquez LC, Ruiz-Pérez CA, Zambrano MM. Proterhodopsin-like genes present in thermoacidophilic high-mountain microbial communities. *Appl Environ Microbiol.* 2012b;78(21):7813–7.
- Bouraoui H, Rebib H, Aissa MB, Touzel JP, O'donohue M, Manai M. *Paenibacillus marinum* sp. nov., a thermophilic xylanolytic bacterium isolated from a marine hot spring in Tunisia. *J Basic Microbiol.* 2013. doi:10.1002/jobm.201200275. [Epub ahead of print].
- Fuhrman JA, Schwalbach MS, Stingl U. Proterhodopsins: an array of physiological roles? *Nat Rev Microbiol.* 2008;6:488–94.
- Jiménez DJ, Andreote FD, Chaves D, Montaña JS, Osorio-Forero C, et al. Structural and functional insights from the metagenome of an acidic hot spring microbial planktonic community in the Colombian Andes. *PLoS ONE.* 2012;7(12):e52069.
- Jones B, Renaut R. Hot springs and geysers. In: Reitner J, Thiel V, editors. *Encyclopedia of geobiology.* Berlin: Springer; 2011. doi:10.1007/Springer-Reference_187284 2012-09-10 14:32:43 UTC. Springer Reference (www.springerreference.com).
- Jones DS, Albrecht HL, Dawson KS, Schaperdorth I, Freeman KH, et al. Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. *ISME J.* 2012;6:158–170.
- Klatt CG, Wood JM, Rusch DB, Bateson MM, Hamamura N, et al. Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J.* 2011;5:1262–78.
- Liu Z, Klatt CG, Wood JM, Rusch DB, Ludwig M, et al. Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J.* 2011;5:1279–90.
- López-López O, Cerdán ME, González-Siso MI. Hot spring metagenomics. *Life.* 2013;2:308–20.
- Mathur J, Bizzoco RW, Ellis DG, Lipson DA, Poole AW, et al. Effects of abiotic factors on the phylogenetic diversity of bacterial communities in acidic thermal springs. *Appl Environ Microbiol.* 2007;73(8): 2612–23.
- Meyer B, Kuever J. Homology modeling of dissimilatory APS reductases (AprBA) of sulfur-33 oxidizing and sulfate-reducing prokaryotes. *PLoS One.* 2008;3(1): e1514.
- Montaña JS, Jiménez DJ, Hernandez M, Angel T, Baena S. Taxonomic and functional assignment of cloned sequences from high Andean forest soil metagenome. *A Van Leeuw J Microb.* 2012;101:205–15.
- Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J. Biodiversity hotspots for conservation priorities. *Nature.* 2000;403:853–8.
- Norris PR. Acidophiles. In: Wiley J and Sons, editors. *Encyclopedia of life sciences.* 2001. p. 1-6. doi:10.1038/npg.els.000033. <http://els.net>. Accessed 11 Nov 2011.
- Pentecost A, Jones B, Renaut RW. What is a hot spring? *Can J Earth Sci.* 2003;40:1443–6.
- Rzonca B, Schulze-Makuch D. Correlation between microbiological and chemical parameters of some hydrothermal springs in New Mexico, USA. *J Hidrol.* 2003;280:272–84.
- Siering PL, Clarke JM, Wilson MS. Geochemical and biological diversity of acidic, hot springs in Lassen volcanic National Park. *Geomicrobiol J.* 2006;23(2): 129–41.
- Stout LM, Blake RE, Greenwood JP, Martini AM, Rose EC. Microbial diversity of boron-rich volcanic hot springs of St. Lucia, Lesser Antilles. *FEMS Microbiol Ecol.* 2009;70(3):402–12.
- Tirawongsaroj P, Sriprang R, Harnpicharnchai P, Thongaram T, Champreda V, et al. Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. *J Biotechnol.* 2008;133:42–9.
- Wang S, Hou W, Dong H, Jiang H, Huang L, et al. Control of temperature on microbial community structure in hot springs of the Tibetan Plateau. *PLoS ONE.* 2013;8(5):e62901.
- Wemheuer B, Taube R, Akyol P, Wemheuer F, Daniel R. Microbial diversity and biochemical potential encoded by thermal spring metagenomes derived from the Kamchatka Peninsula. *Archaea.* 2013: (136714).
- Xie W, Wang F, Guo L, Chen Z, Sievert SM, et al. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J.* 2011;5:414–26.

Metagenomes: 23S Sequences

23S rRNA Genes in Metagenomes

Pelin Yilmaz¹ and Frank Oliver Glöckner^{1,2}

¹Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

²Jacobs University Bremen gGmbH, Bremen, Germany

Synonyms

Environmental genomics; Large subunit rRNA; Metagenomes; Metagenomics; 23S ribosomal RNA gene; 23S rRNA

Definition

As an evolutionary marker, 23S ribosomal RNA (rRNA) offers more diagnostic sequence stretches and greater sequence variation than 16S rRNA. The main drawback of using 23S rRNA as a phylogenetic marker is that it is still not as widely used. In a survey of 23S rRNA gene sequences found in metagenomic datasets, the Global Ocean Sampling (GOS) metagenome revealed that 23S rRNA gene sequences are twice as abundant as 16S rRNA gene fragments, with 23S rRNA gene fragments being generally about 100 bp longer.

Introduction

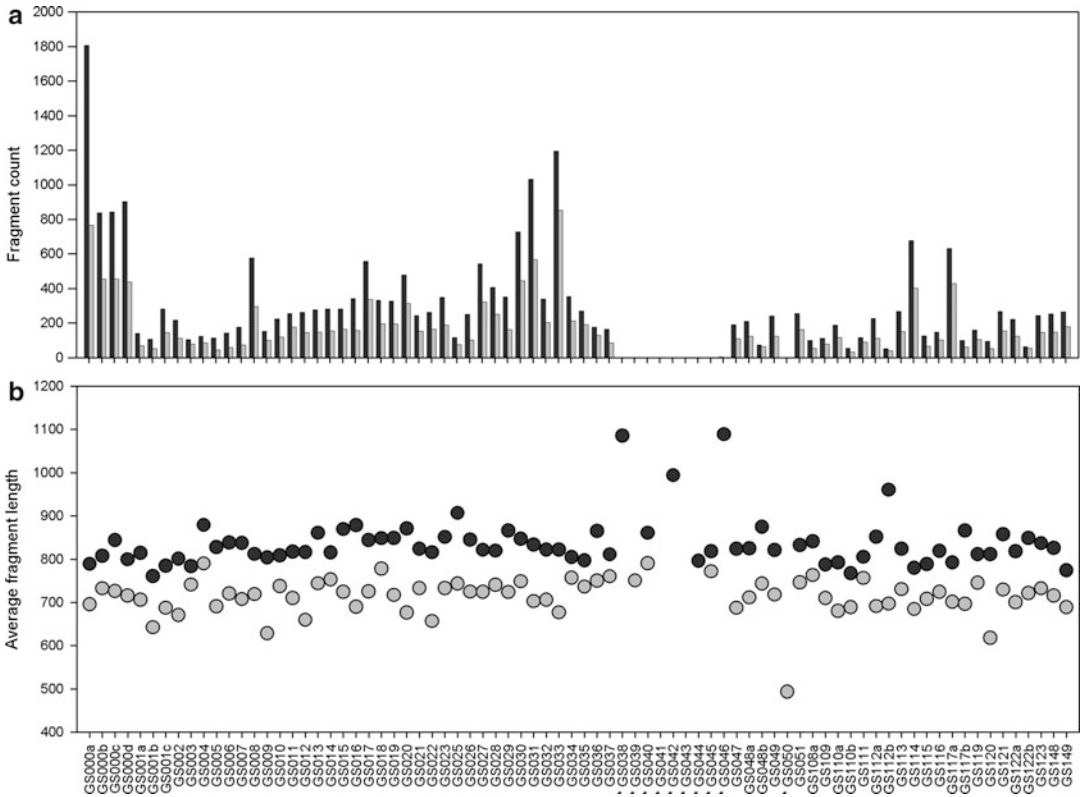
The distribution of 23S rRNA gene sequences in the GOS and other metagenomes remains unexplored. Although the 16S rRNA gene has been established as the standard molecule for analyzing the taxonomic diversity in metagenomes, using the 23S rRNA gene as a phylogenetic marker offers advantages over using the 16S rRNA gene. With an average length

of 2,900 bases, it is almost twice as long as the 16S rRNA and, therefore, is theoretically a more informative phylogenetic marker than the 16S rRNA gene (Ludwig and Schleifer 1994; Ludwig et al. 1995; Ludwig and Klenk 2001). Both the 23S and 16S rRNA molecules share the same properties in terms of molecule ubiquity, as well as sequence and structure conservation. Furthermore, phylogenetic trees based on 16S rRNA and on 23S rRNA genes have comparable topologies (Rijk et al. 1995; Ludwig and Schleifer 1999).

A disadvantage of the 23S rRNA gene is the relatively low number of sequences available in the public databases as compared to 16S rRNA genes. Currently (May 2014), only 446,998 23S/28S sequences are publicly available, compared to 4,346,367 16S/18S sequences (Quast et al. 2013). Furthermore, the low number of 23S/28S rRNA sequences (29,397) longer than 1,900 bases (full length) limits the assessment of taxonomic diversity due to reduced resolution in taxonomic assignments. The lower number of available 23S rRNA gene sequences can historically be explained by the technical difficulty and higher cost of sequencing the larger molecule with Sanger sequencing technology. However, with new technologies and constantly decreasing sequencing costs, these difficulties are becoming less pronounced.

Summary of rRNA Gene Fragment Retrieval

The 23S/28S rRNA gene is twice as long as the 16S/18S rRNA gene; hence, the probability of retrieving a 23S/28S rRNA gene fragment should be proportionately higher. Ratios of approximately 2:1 of identified 23S/28S rRNA over 16S/18S rRNA observed at different sites in the GOS metagenome study support this expectation – GS000d (904 23S/28S vs. 438 16S/18S), GS029 (351 23S/28S vs. 162 16S/18S), or GS112a (227 23S/28S vs. 113 16S/18S) (Fig. 1a). This twofold difference is also reflected by the average number of fragments retrieved per site, which is 301 for 23S/28S rRNA and 177 for



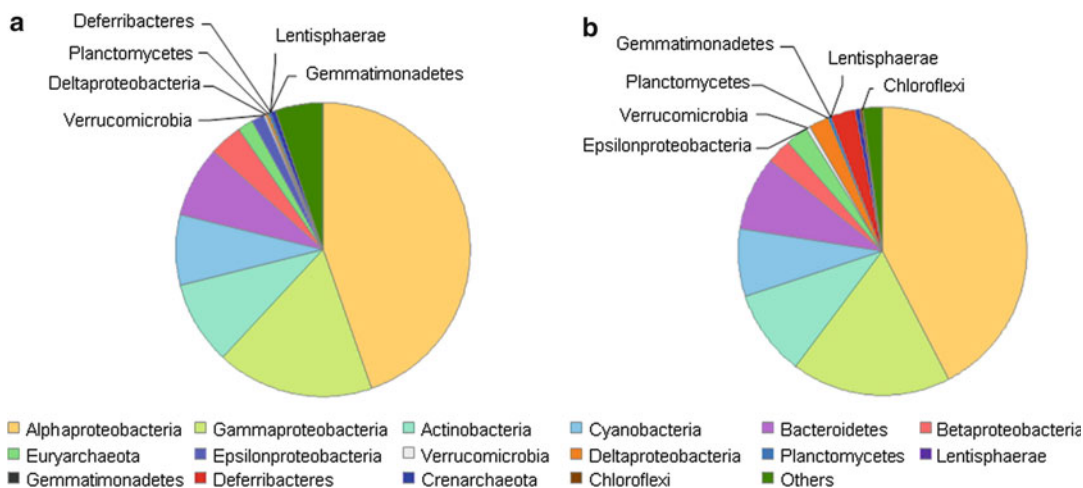
Metagenomes: 23S Sequences, Fig. 1 (a) Comparison of number of 23S/28S (dark gray bars) and 16S/18S (light gray bars) rRNA fragments retrieved from each GOS sample dataset. (b) Average length of 23S/28S (dark gray circles) and 16S/18S (light gray circles)

rRNA fragments from each GOS sample dataset in terms of number of aligned bases within the rRNA gene boundaries, excluding any fragment (23S/28S or 16S/18S) that contained less than 100 aligned bases. Sites marked with an “*” indicate that less than five fragments were retrieved

16S/18S rRNA. Furthermore, 23S/28S rRNA gene fragments are considerably longer than 16S/18S gene fragments (Fig. 1b). Where an average 23S/28S rRNA fragment has 836 aligned bases within the rRNA gene boundaries, a 16S/18S rRNA fragment has 713 aligned bases. More abundant and longer rRNA gene fragments may provide additional information in assessing taxonomic diversity, both with phylogeny and operational taxonomic unit-based methods, as well as increasing the chances to affiliate other gene fragments with specific lineages. Both 23S/28S and 16S/18S rRNA fragments are randomly distributed over the rRNA gene regions, meaning that no specific sequence region is over- or underrepresented.

Taxonomic Diversity Based on 23S and 16S rRNA Genes

Percentages of both 23S and 16S rRNA fragments associated with major marine bacterial and archaeal taxa show good agreement with each other (Fig. 2, b). Specifically, based on 23S rRNA assignments, 43 % of the retrieved rRNA fragments are associated with *Alphaproteobacteria*, followed by 17 % *Gammaproteobacteria*, 9 % *Actinobacteria*, 8 % *Cyanobacteria*, 8 % *Bacteroidetes*, 3 % *Betaproteobacteria*, 2 % *Euryarchaeota*, and 0.4 % *Crenarchaeota* (Fig. 2a). However, less agreement in the assignment of 23S rRNA and 16S rRNA fragments is observed with less



Metagenomes: 23S Sequences, Fig. 2 Percentage of 23S (a) and 16S (b) rRNA fragments associated with major marine bacterial and archaeal taxa among all GOS

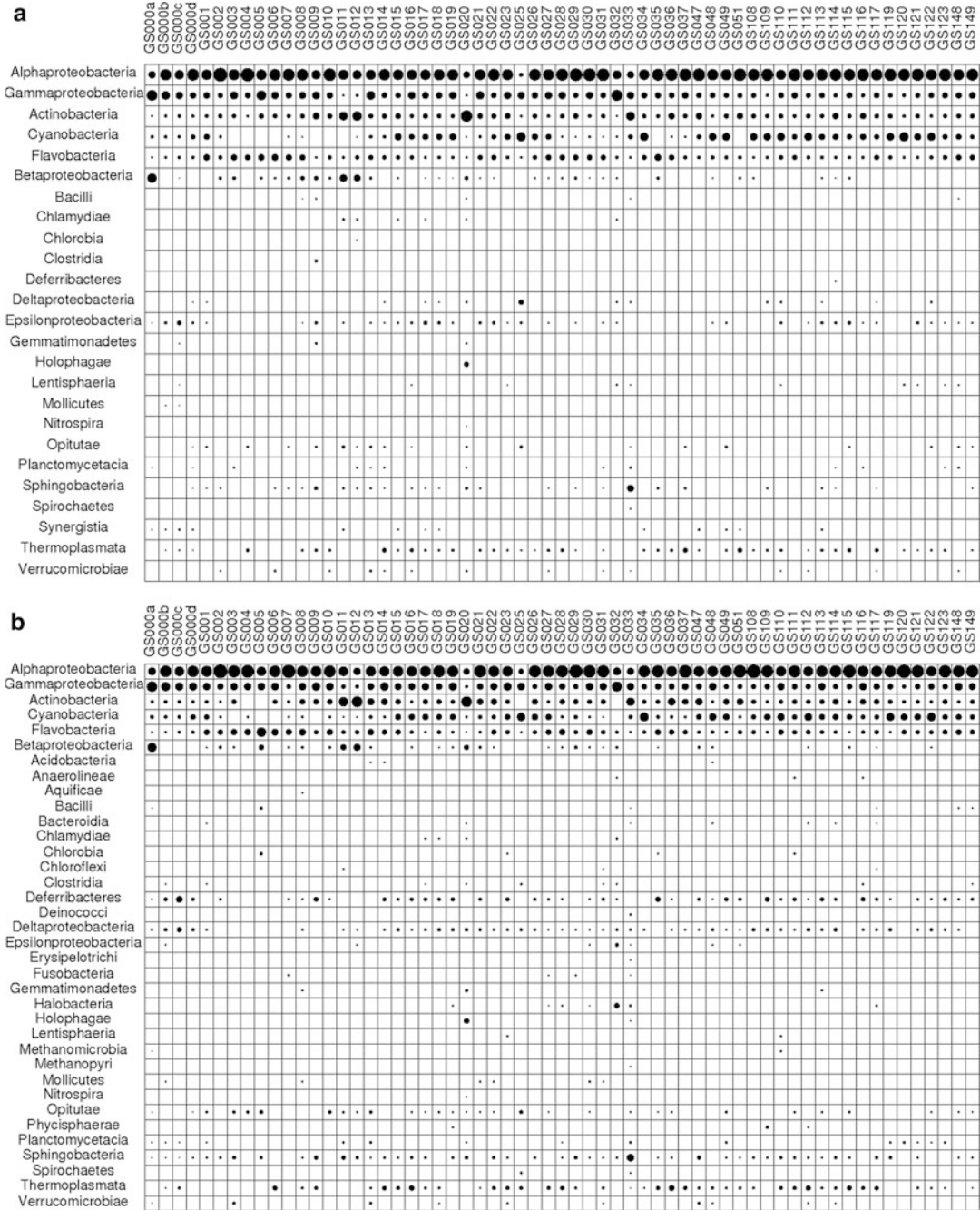
sample datasets, except GS038–GS046 and GS050. Percentages were calculated based on absolute numbers of fragments associated with a given taxa

abundant marine taxa. For example, *Chloroflexi*- and *Deferribacteres*-associated fragments are not observed in the 23S rRNA gene-based classification, which may be ascribed to the lack of annotated clades for these taxa. In such cases, 16S rRNA gene-based classifications appear to provide better estimations.

Similar trends are also observed in sample-by-sample distribution of taxa at the “class” level for both 23S and 16S rRNA-based assignments, as compared to the previous overall assessment (Fig. 3a, b). *Alphaproteobacteria*, *Gammaproteobacteria*, *Actinobacteria*, *Cyanobacteria*, *Flavobacteria*, and *Betaproteobacteria* are the most abundant taxa in the majority of sample datasets. However, differences are observed in the occurrence or relative abundance of minor groups, such as *Planctomycetacia* or *Aquificae*. In certain cases, 23S rRNA-based assessments predict higher relative abundances or occurrence in sample datasets for other taxa. Up to 12-fold more *Epsilonproteobacteria*-associated 23S rRNA fragments are found in sample dataset GS000b compared to 16S rRNA fragments. Additionally, *Lentisphaeria*, which appears to be present in ten sites according to 23S rRNA classifications, are observed only at two sites according to 16S rRNA gene classifications.

The former case, where 16S rRNA-based assignments estimated more taxa in more sample datasets, demonstrates the current drawback of 23S rRNA-based classification (i.e., its lack of resolution due to insufficient full-length reference sequences). On the other hand, the latter observations demonstrate that when reference sequences are present for a taxon, the higher number of 23S rRNA fragments retrieved can capture what is missed with 16S rRNA fragments.

Investigating relative abundances at lower taxonomic levels can shed light on more prominent habitat-specific diversity patterns. However, with the current size and content of LSU rRNA reference databases, the 23S rRNA has a distinct disadvantage in achieving this. As summarized in Table 1, the percentage of 23S rRNA gene fragments that can be classified to a certain taxa is comparable to the 16S rRNA gene-based classification at domain, phylum or class levels. A decrease in percentage of classified 23S rRNA fragments was observed at lower levels, from 95 % at the class level down to even 17 % at the genus level. This can be explained by the 23,197 sequences of taxonomically classified cultured organisms in the SILVA SSU Ref dataset (release 102) versus only 3,602 sequences in the LSU Ref dataset of the same release.



Metagenomes: 23S Sequences, Fig. 3 The relative abundance of 23S (a) and 16S (b) rRNA fragments associated with different taxa (rows) at each GOS sample dataset (columns). Presence of a spot indicates the presence of fragments associated with a given taxa, and the area of a spot represents the relative abundance. Relative abundances are based on absolute counts of all fragments from a given site associated with a certain taxa, which are

then normalized according to the total fragment counts from that site. Abundances are not normalized with respect to single copy genes, and since rRNA operons can occur multiple times in a genome, the numbers do not represent cell abundances. The taxa shown here are on the “class” level, except *Cyanobacteria*, which is at the “phylum” level

Metagenomes: 23S Sequences, Table 1 Percentage of 23S and 16S rRNA gene fragments that can be classified up to domain, phylum, class, order, family, and genus levels. Total number of fragments classified are 20,036 and 12,491 for 23S and 16S rRNA, respectively, excluding *Eukarya* and fragments with less than 300 aligned bases for LSU and less than 100 aligned bases for SSU

	23S rRNA gene fragments (%)	16S rRNA gene fragments (%)
Domain	99.9	100.0
Phylum	96.6	100.0
Class	94.4	99.1
Order	78.8	96.3
Family	35.4	80.0
Genus	16.6	31.2

Specificity of Common 23S rRNA Primers and Probes

Including the 23S rRNA gene sequences identified in the GOS metagenome dataset in the SILVA LSU Parc dataset increased its size by 12 % (SILVA release 102). Furthermore, they have not undergone PCR amplification and hence provide a unique opportunity for testing the coverage of previously described universal amplification primers, as well as widely used class-specific probes.

The most recently developed primer sets (129f, 189f, 457r, 2490r) (Hunt et al. 2006), as well as primer 2241r (Lane 1991), show reasonable group coverage for the 23S rRNA gene sequences identified in the GOS dataset with an average of 85 % (Table 2), and the results are comparable to those obtained from matching the primers against the SILVA LSU Parc dataset (release 102) with a difference of only ± 2 %. The reference dataset used by Hunt and colleagues is with 2,176 sequences smaller than both the LSU Parc (average of 11,000 target group sequences) and the GOS 23S (average of 5,400 target group sequences) datasets used in this study. However, the authors have included environmental shotgun sequences from the Sargasso Sea pilot study (Venter et al. 2004) in their dataset, which would account for the comprehensiveness of these primers also in the GOS 23S dataset.

Contrary to these results, the primers developed for the amplification of variable regions of bacterial 23S rRNA sequences (11a–97ar) (Van Camp et al. 1993) show very poor group coverage in the GOS 23S dataset sequences, with generally less than 50 % coverage of the target group. 90 % group coverage is only observed for 69ar (Table 2). Although the primer binding sites were highly conserved, this is counteracted by the very small dataset that these primers were based on. Surprisingly, primers 53a to 97ar are observed to have higher group coverage within the GOS 23S rRNA sequences than within LSU Parc.

The two archaeal primers (LSU190-F and LSU2445a-R) (DeLong et al. 1999) show very low group coverage in the GOS 23S dataset (Table 2), with 14 % and 5 %, respectively. Nevertheless, while the percentages are higher in the LSU Parc, they do not exceed 50 %.

For the BET42a probe (Manz et al. 1992), 79 % group coverage is observed. This, as well as the number of outgroup hits within the GOS 23S dataset, is close to that reported by a previous evaluation (Amann and Fuchs 2008) (Table 2). Group coverage within LSU Parc (87 %) is in accordance with Amann and Fuchs (Amann and Fuchs 2008) (Table 2), although considerably more outgroup hits, 348 in LSU Parc versus 62, are observed.

The GAM42a probe coverage in the GOS 23S dataset (Table 2) is almost half (42 %) of the value reported previously (76 %) (Amann and Fuchs 2008) and the corresponding evaluation of the LSU Parc (78 %) dataset. Since the mismatches could result from sequencing errors, the alignments of sequences with mismatches to the probe GAM42a were manually inspected. A few cases were likely to be sequencing errors and were mainly observed in fragments obtained from ends of sequencing reads. The majority of the mismatches revealed consistent, class-specific mismatches. These mismatches are up to four bases and are found mainly between *E. coli* positions 1,030–1,040. Although this evaluation of the GAM42a probe was based on a single environment, the surface ocean, limitations and anomalous results with the GAM42a

Metagenomes: 23S Sequences, Table 2 Specificities of selected primers and probes, evaluated on the 23S/28S rRNA gene fragments retrieved from the GOS metagenomes having more than 300 aligned bases within the rRNA gene boundaries and on the SILVA Parc release 102 LSU dataset. Outgroup hits are the sum of both *Archaea* and *Eukarya* in case of bacterial primers, both *Bacteria* and *Eukarya* in case of archaeal primers, only *Eukarya* in case of bacterial and archaeal primers, and non-*Betaproteobacteria* and non-*Gammaproteobacteria* for BET42a and GAM42a probes

Primer/ probe	Target group	GOS 23S/28S			LSU Parc		
		Size of target group	Group coverage (%)	Outgroup hits	Size of target group	Group coverage (%)	Outgroup hits
129f ^a	<i>Bacteria</i>	4,853	74 %	0	10,640	82 %	4
189f ^a	<i>Bacteria</i>	5,285	87 %	0	11,508	87 %	0
457r ^a	<i>Bacteria</i>	5,551	86 %	4	11,177	83 %	279
2241r ^b	<i>Bacteria</i>	5,832	84 %	10	11,457	86 %	3,967
2490r ^a	<i>Bacteria</i>	5,734	94 %	0	10,821	98 %	0
11a ^c	<i>Bacteria</i>	5,256	20 %	0	11,478	39 %	0
23ar ^c	<i>Bacteria</i>	5,619	23 %	0	10,526	49 %	4
43a ^c	<i>Bacteria</i>	5,633	6 %	0	10,999	44 %	0
53a ^c	<i>Bacteria</i>	5,320	3 %	0	10,594	1 %	0
62ar ^c	<i>Bacteria</i>	5,540	8 %	0	11,455	5 %	0
69ar ^c	<i>Bacteria</i>	5,731	90 %	0	11,443	87 %	0
93a ^c	<i>Bacteria</i>	5,737	62 %	0	10,322	55 %	0
93ar ^c	<i>Bacteria</i>	5,731	63 %	0	10,327	56 %	2
97ar ^c	<i>Bacteria</i>	4,969	55 %	0	9,165	29 %	38
LSU190-F ^d	<i>Bacteria</i> and <i>Archaea</i>	5,348	14 %	0	11,741	24 %	0
						28 %	
LSU2445a- R ^d	<i>Archaea</i>	142	5 %	0	262	28 %	0
BET42a ^e	<i>Betaproteobacteria</i>	209	79 %	63	570	87 %	348
GAM42a ^e	<i>Gammaproteobacteria</i>	980	42 %	1	2,877	78 %	10

References: ^aHunt et al. 2006; ^bLane 1991; ^cVan Camp et al. 1993; ^dDeLong et al. 1999; ^eManz et al. 1992

probe have been reported previously for other environments as well, which were found to be mainly due to polymorphisms at *E. coli* position 1,033 (Yeates et al. 2003; Barr et al. 2010). Our observation confirms these reports, by adding additional polymorphisms before and after this position. Consequently, the limitations of the GAM42a probe might be more severe than previously thought, and therefore, we recommend the design and testing of novel *Gammaproteobacteria* probes.

Summary

This comparative overview of 16S and 23S rRNA fragments retrieved from the GOS metagenomes exemplifies the possibility and

power of using 23S rRNA genes. High-quality taxonomic classification for biodiversity analysis, as well as primer and probe design, depends on the size and extent of the reference dataset used. The advantage of using the larger 23S rRNA genes for biodiversity analysis, especially for the marine system, has been shown previously (Peplies et al. 2004). Additionally, a recent study assessing the diversity of paralogous 23S rRNA genes has shown that significant sequence diversification was observed in 184 species, further supporting the suitability of this molecule for taxonomy (Pei et al. 2009). Although an obvious limitation faced during this study was the small size of the 23S rRNA gene reference datasets, this is likely to be overcome in the near future with the contribution of (meta-)genomic sequences from

mega-sequencing projects, such as the Human Microbiome Project, the TerraGenome, the Tara Oceans, or the Genomic Encyclopedia of *Bacteria* and *Archaea*. Moreover, studies assessing the characteristics and sequence diversity of 23S rRNA genes in bacterial and archaeal genomes, in combination with efforts to design, test and, reevaluate universal and group-specific primers and probes, can renew the interest and utilization of this molecule. The application of continually advancing, cheaper sequencing technologies to the undiscovered fraction of the 23S rRNA gene sequences can result in a higher appreciation of this valuable phylogenetic marker.

References

- Amann R, Fuchs BM. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nat Rev Microbiol*. 2008;6(5):339–48.
- Barr JJ, Blackall LL, et al. Further limitations of phylogenetic group-specific probes used for detection of bacteria in environmental samples. *ISME J*. 2010;4:959–61.
- DeLong E, Taylor L, et al. Visualization and enumeration of marine planktonic *Archaea* and *Bacteria* by using polyribonucleotide probes and fluorescent *in situ* hybridization. *Appl Environ Microbiol*. 1999;65(12):5554–63.
- Hunt DE, Klepac-Ceraj V, et al. Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl Environ Microbiol*. 2006;72(3):2221–5.
- Lane DJ. 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M, editors. *Nucleic acid techniques in bacterial systematics*. Chichester/New York: Wiley; 1991. p. 115–75.
- Ludwig W, Klenk HP. A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, editors. *The Archaea and the deeply branching and phototrophic Bacteria*. New York: Springer; 2001. p. 49–65.
- Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol Rev*. 1994;15(2–3):155–73.
- Ludwig W, Schleifer K. Phylogeny of *Bacteria* beyond the 16S rRNA standard. *ASM News*. 1999;65(11):752–7.
- Ludwig W, Rossello-Mora R, et al. Comparative sequence analysis of 23S rRNA from *Proteobacteria*. *Syst Appl Microbiol*. 1995;18:164–88.
- Manz W, Amann R, et al. Phylogenetic oligodeoxynucleotide probes for the major subclasses of *Proteobacteria*: problems and solutions. *Syst Appl Microbiol*. 1992;15(4):593–600.
- Pei A, Nossa CW, et al. Diversity of 23S rRNA genes within individual prokaryotic genomes. *PLoS ONE*. 2009;4(5):e5437.
- Peplies J, Glöckner FO, et al. Comparative sequence analysis and oligonucleotide probe design based on 23S rRNA genes of *Alphaproteobacteria* from North Sea bacterioplankton. *Syst Appl Microbiol*. 2004;27(5):573–80.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acid Res*. 2013;41:D590–D596.
- Rijk P, Peer Y, et al. Evolution according to large ribosomal subunit RNA. *J Mol Evol*. 1995;41(3):366–75.
- Van Camp G, Chapelle S, et al. Amplification and sequencing of variable regions in bacterial 23S ribosomal RNA genes with conserved primer sequences. *Curr Microbiol*. 1993;27(3):147–51.
- Venter JC, Remington K, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304(5667):66–74.
- Yeates C, Saunders AM, et al. Limitations of the widely used GAM42a and BET42a probes targeting bacteria in the *Gamma*proteobacteria radiation. *Microbiology*. 2003;149(5):1239–47.

Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome

Brian V. Jones¹ and C. G. M. Gahan²

¹Centre Biomedical and Health Science Research, University of Brighton, School of Pharmacy and Biomolecular Sciences, Brighton, East Sussex, UK

²Department of Microbiology, School of Pharmacy & Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland

Definitions

Metagenome/metagenomics: The collective genomes of all members of a particular microbial community may be referred to as the metagenome (or a genome of many).

Metagenomics refers to methods which seek to understand the composition, development, and function of microbial ecosystems through analysis of the community metagenome.

Function-driven metagenomics: A metagenomic approach in which emphasis is placed on the recovery of genes encoding a defined function of interest, through assays based on heterologous gene expression. Typically metagenomic DNA is used to generate genetic libraries in a surrogate host species that may be easily manipulated in the laboratory. Each clone in the library (analogous to books in a conventional library) represents a fragment of metagenomic DNA from a member of the microbial community under study. Libraries are then subsequently screened to identify clones encoding and expressing activities of interest.

Large-insert library/genetic library: Due to the complexity of microbial communities, genetic libraries constructed for function-driven metagenomic analysis often seek to clone large fragments of metagenomic DNA (typically ranging from 40 to 200 kb in size, depending on the specifics of the cloning system used). The term “insert” refers to the metagenomic DNA fragments which are ligated, or “inserted,” into a plasmid vector that maintains them in the surrogate host bacterium. Insert sizes of ~40 kb and over are usually referred to as “large inserts,” giving rise to the term “large-insert library.”

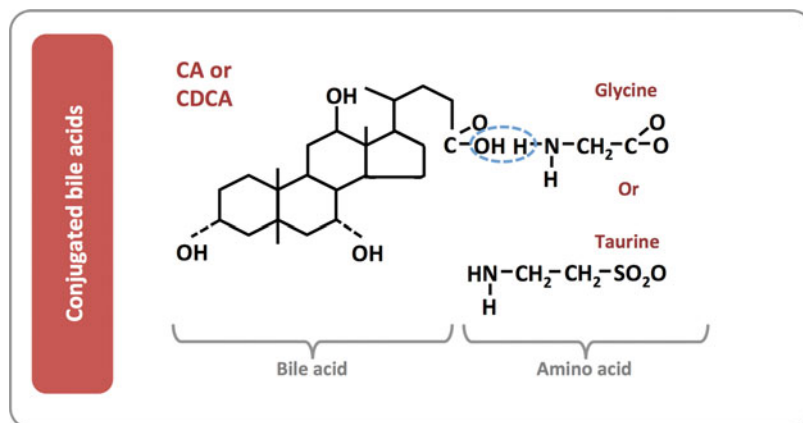
Sequence-driven metagenomics: A metagenomic approach in which the emphasis is placed on the generation and analysis of nucleotide sequence data from metagenomic DNA. Typically sequence-based approaches are utilized to provide a broad overview of the population structure and predicted functions undertaken by a microbial community.

Heterologous gene expression: Refers to the expression of genes in an organism from which they did not originate. For function-driven metagenomics, this generally refers to the expression of genes encoded by cloned fragments of metagenomic in the surrogate host species used to construct genetic libraries (typically *Escherichia coli*).

Bile Acids and Microbial Bile Acid Metabolism

Bile acids (BA) are cholesterol derivatives synthesized in the liver and linked with either glycine or taurine to form conjugated bile acids (CBA) (Ridlon et al. 2006; Begley et al. 2005a, b; Fig. 1). The dominant CBA in humans are glycine conjugates of cholic acid and chenodeoxycholic acid, with CBA forming a major component of bile stored in the gall bladder (Ridlon et al. 2006). In response to food intake, bile is secreted into the lumen of the intestine where CBA facilitate the digestion of dietary fat, promoting the emulsification of lipids and their subsequent absorption across the intestinal epithelium (Ridlon et al. 2006; Begley et al. 2005a). However, the functions of bile acids are not limited to digestion, and BA are also important signaling molecules that contribute to the regulation of diverse metabolic processes (Thomas et al. 2008; Fig. 2). These include regulation of mucosal immune responses in the intestine, as well as aspects of energy homeostasis and fat storage (Thomas et al. 2008; Inagaki et al. 2006; Houten et al. 2006; Jones 2011; Watanabe et al. 2006; Fig. 2). As such, BA are now no longer viewed as purely digestive secretions but also as metabolic integrators and key regulators of intestinal homeostasis (Thomas et al. 2008; Hofmann and Eckmann 2006; Jones 2011).

The regulatory functions of bile acids are believed to act through two main receptors, the nuclear receptor FXR α and the membrane receptor TGR5, for which bile acids are the natural ligands (Thomas et al. 2008). These receptors are highly expressed in the liver and intestinal tissues but also in numerous extraintestinal tissues (Thomas et al. 2008). Although the majority of bile acids are efficiently reclaimed from the intestine and returned directly to the liver for reuse (referred to as enterohepatic circulation), a portion enter the systemic circulation and signal other organs through these receptors, coordinating cholesterol, triglyceride and glucose metabolism, as well as fat storage (Thomas et al. 2008; Fig. 2).



Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 1 Structure of dominant conjugated bile acids in humans (Modified from Begley et al. 2005a). Major bile acid species in the human bile acid pool are conjugated forms linked to either glycine or taurine via amide bonds, with glyco-conjugates dominant in humans (Ridlon et al. 2006). The predominant bile acid species are cholic acid (CA) and chenodeoxycholic acid (CDCA), which are generated de

novo in the liver from cholesterol and referred to as primary bile acids (Ridlon et al. 2006; Thomas et al. 2008). De-conjugated CA and CDCA, as well as derivatives of these primary BA formed in the intestine, are recovered and returned to the liver where they are conjugated and re-assimilated into the bile acid pool. For comprehensive reviews, see Ridlon et al. (2006) and Begley et al. (2005a)

CBA have also been implicated in the control of microbial growth in the small intestine via toxic effects on colonizing bacteria (Begley et al. 2005a; Ridlon et al. 2006). This antimicrobial effect is thought to repress bacterial growth in the small intestine and prevent microbes proliferating to levels which are harmful to the human host. Local mucosal immune responses in the intestine are also regulated by bile acids (through FXR α) and implicated in microbial population control in this compartment (Inagaki et al. 2006). It is most likely that bile acid mediated mucosal immune regulation works in synergy with the direct effects of bile acids on resident microbes, to prevent bacterial overgrowth in the small intestine and associated deleterious effects on host health (Inagaki et al. 2006; Hofmann and Eckmann 2006; Begley et al. 2005a; Fig. 2).

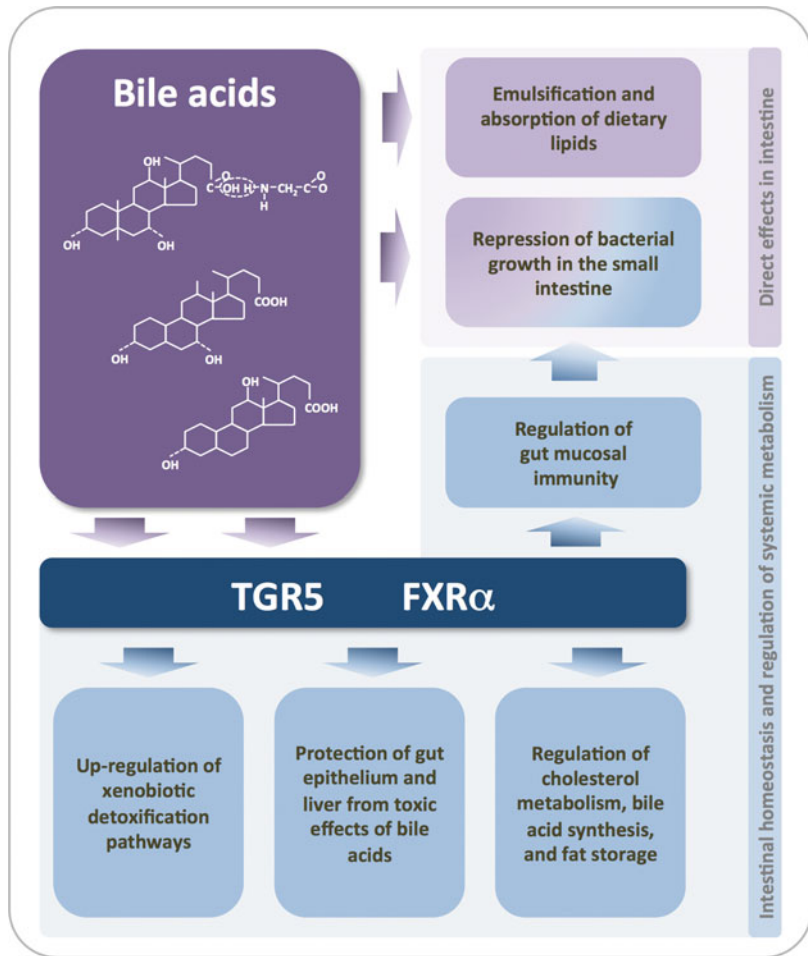
However, once secreted into the intestinal lumen, CBA are subject to extensive biotransformation by indigenous gut microbes, leading to the formation of a range of secondary and tertiary products (Ridlon et al. 2006; Begley et al. 2005a;

Jones et al. 2008; Fig. 3). These modified bile acids display altered binding characteristics for bile acid receptors, with microbial products of bile acid metabolism among the most potent agonists (Thomas et al. 2008). This highlights the potential for microbes resident in the human gut microbiome to influence wider aspects of host metabolism and phenotype, through interaction with bile acid signaling pathways (Jones et al. 2008; Thomas et al. 2008; Jones 2011; Ogilvie and Jones 2012). Congruent with this hypothesis is the accumulating body of evidence implicating microbial bile acid metabolism as the basis of a long-standing dialogue between the human host and its gut microbiome (Jones 2011; Inagaki et al. 2006; Gadaleta et al. 2011; Maran et al. 2009; Modica et al. 2008; Duboc et al. 2013; Jones et al. 2008). As such, there is increasing interest in understanding the role of this activity in human health and disease processes, with this function of the gut microbiome likely to be a viable target for disease prevention through manipulation, or augmentation of the intestinal microbial ecosystem.

Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome,

Fig. 2 Overview of physiological functions undertaken by bile acids.

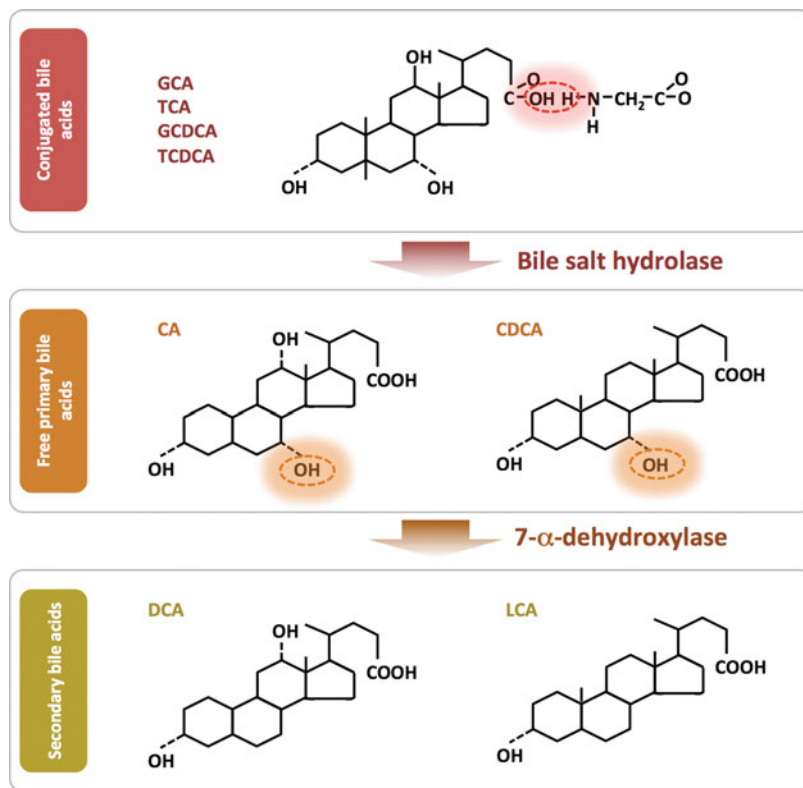
Boxes shaded violet summarize the direct functions of bile acids in the small intestine, attributed to their physical properties. Boxes shaded blue summarize regulatory functions of bile acids, through interaction with the main bile acid receptors TGR5 and FXR α . For comprehensive reviews of bile acid signaling, see Thomas et al. (2008)



Overview of Bile Salt Hydrolases: Biochemistry, Structure, and Function

Bile salt hydrolases (BSH; EC 3.5.1.24) (also designated as choloylglycine hydrolases or conjugated bile acid hydrolases) are members of the N-terminal nucleophilic (Ntn) hydrolase superfamily of proteins and catalyze the hydrolysis of conjugated bile acids, linked with the amino acids taurine or glycine (tauro-CBA, glyco-CBA), to liberate free primary bile acids and amino acids (Fig. 3; Begley et al. 2006; Kumar et al. 2006). The wider enzyme superfamily also contains the penicillin V acylase (PVA; EC 3.5.1.11) enzyme family, and BSH and PVA enzymes share significant homology and catalyze hydrolysis of the same type of chemical bond.

While the main substrates for BSH and PVA enzymes (conjugated bile acids and penicillins, respectively) vary considerably in structure, PVA has been shown to exhibit some moderate activity against bile acids and some BSH enzymes demonstrate mild activity against penicillin V (Kumar et al. 2006). This suggests that each enzyme group has preferential activity against a specific substrate but that some overlap in activities also occurs (Kumar et al. 2006). The sequence homology between these enzyme families has led to mis-annotation of PVA in some bacterial genomes, for example, in the initial genome annotation of *Listeria monocytogenes* (Begley et al. 2005b) and *Lactobacillus plantarum* WCFS1 (Lambert et al. 2008a). This highlights a requirement for functional enzymatic



Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 3 Major bile acid transformations undertaken by the human gut microbiota (Modified from Jones 2011). Bile salt hydrolase (*BSH*) catalyzes the initial de-conjugation of CBA to liberate free primary bile acids and amino acids. Free primary bile acids are then available to further modification by the gut microbiome and converted to secondary forms. A multistep 7- α dehydroxylation pathway is responsible for generation of key secondary BA species

which accumulate in the bile acid pool. *GCA*, *TCA*, glyco- and tauro-conjugated cholic acid, respectively; *GCDCA*, *TCDCa*, glyco- and tauro-conjugated chenodeoxycholic acid, respectively; *CA*, *CDCA*, free primary bile acids cholic acid and chenodeoxycholic acid, respectively; *DCA*, *LCA*, free secondary BA deoxycholic acid and lithocholic acid, respectively. For comprehensive reviews of microbial bile acid transformations, see Ridlon et al. (2006) and Begley et al. (2005a)

analysis in order to determine substrate preferences and to guide annotation (Jones et al. 2008; Lambert et al. 2008b).

The crystal structure has been solved for a number of BSH (Kumar et al. 2006; Rossocha et al. 2005) and PVA (Suresh et al. 1999) enzymes and demonstrates a conservation in overall structure suggestive of shared mechanisms of action and an evolutionary relationship between BSH and PVA (Kumar et al. 2006). Detailed analysis of the structure of BSH and PVA enzymes indicates that there is a significant difference in the organization of

specific loops near the active site in each case which may explain differences in substrate specificity (Kumar et al. 2006).

Structural and functional analysis of BSH enzymes from different bacteria has revealed the presence of conserved amino acids that are thought to be essential for bile hydrolysis. In particular the thiol group of the Cys-1 amino acid has been shown to be essential for catalytic activity (Kim et al. 2004; Lodola et al. 2012). In addition a number of amino acids including Asp-20, Tyr-82, Asn-175, and Arg-228 are highly conserved across numerous BSH enzymes

(Begley et al. 2006) and have recently been shown to be essential for catalytic activity mainly through electrostatic interactions with the Cys-1 sulfur atom (Lodola et al. 2012)

(Begley et al. 2006). Despite high levels of amino acid conservation, different BSH enzymes display subtle differences in their preferred bile substrates with some enzymes exhibiting hydrolysis of glyco- and tauro-conjugated bile acids and others demonstrating specific hydrolysis of tauro-conjugated bile acids (Jones et al. 2008). BSH enzymes with specificity for tauro-conjugated bile acids are highly represented among the Bacteroidetes and form a separate phylogenetic group relative to other BSH enzymes but have not been characterized in detail (Jones et al. 2008). Further biochemical analysis of a variety of BSH enzymes is warranted to determine the structural variances that give rise to these subtle differences in bile acid substrate range.

Metagenomic Analysis of Bile Salt Hydrolases (BSHs)

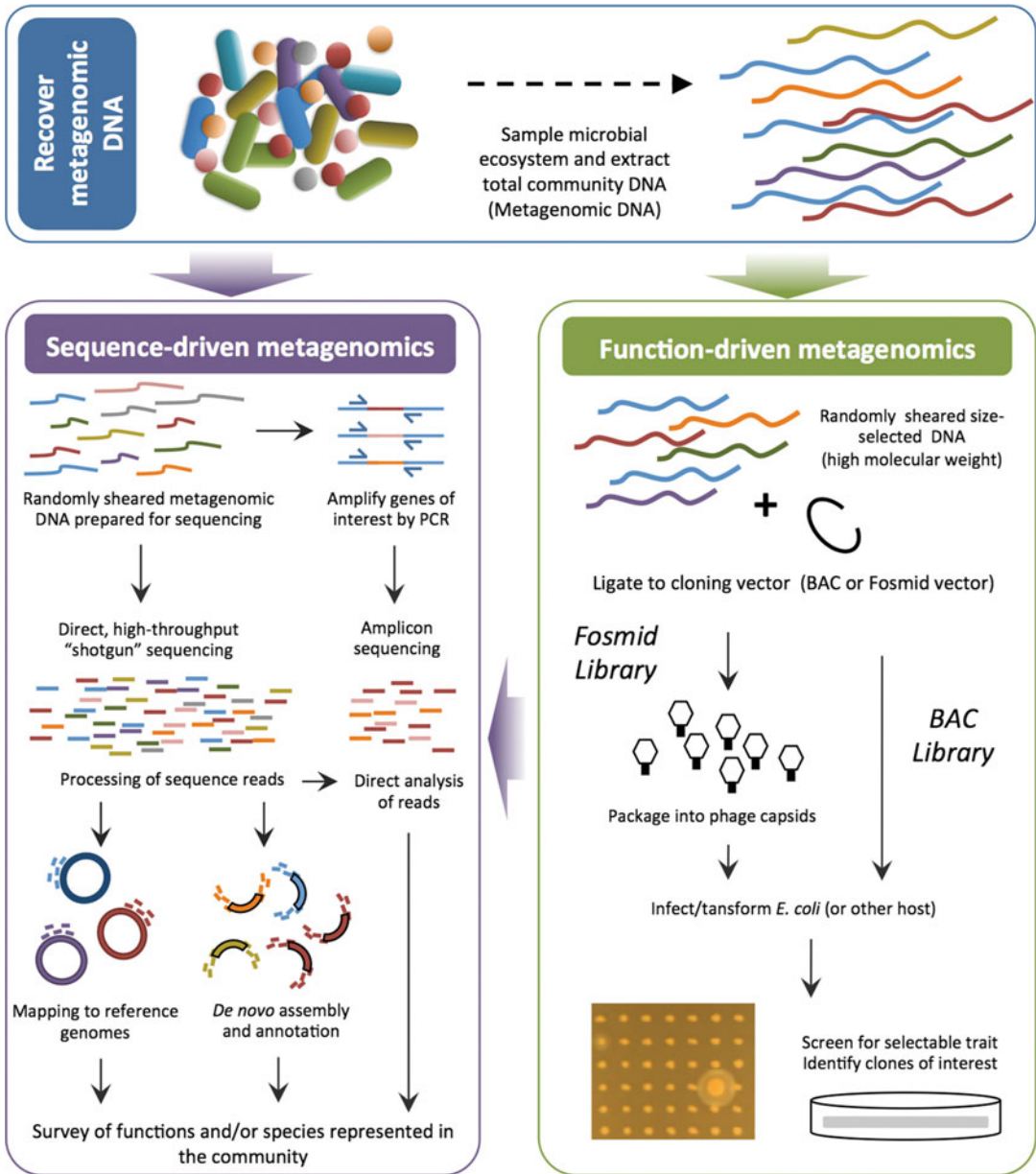
As the human gut microbiota is composed predominantly of microbes which are yet to be grown in the laboratory, a range of culture-independent approaches have been developed and applied to study this and other microbial communities (Handelsman 2004; Jones and Marchesi 2007; Qin et al. 2010; Kurokawa et al. 2007; Gill et al. 2006). Metagenomic approaches constitute a particularly powerful branch of the culture-independent techniques available for characterization of microbial ecosystems, in which the collective genomes of all species comprising a community are considered as a single, community-wide, genetic unit (the metagenome) (Handelsman 2004). Access and analysis is guided by this basic principle, and metagenomic approaches are rooted in the extraction of total, mixed community DNA (metagenomic DNA) without any prior cultivation (Handelsman 2004). Recovered community DNA is then either subject to direct analysis using high-throughput

sequencing (shotgun metagenomics or sequence-based metagenomics) or used to construct large-insert genetic libraries for function-based screening (function-driven metagenomics) (Handelsman 2004) (Fig. 4).

The resulting data not only affords access to census-type information describing the composition of a community (who is there?) but also permits access to the broader functional content encoded by microbial ecosystems (what are they doing?) (Handelsman 2004; Jones et al. 2008). Recently both function-driven and sequence-based metagenomic approaches have been applied to analyze BSH activity in the gut microbiome and provide good examples of the capacity for metagenomics to generate novel functional insights into a microbial community and, in the case of the human microbiome, to understand its influence on host health (Jones et al. 2008; Ogilvie and Jones 2012).

Function-Driven Metagenomic Analysis of Bile Salt Hydrolases: Due to the relative paucity of information regarding the genes underpinning bile acid metabolism in the gut microbiome, initial community-wide studies of this activity utilized a function-driven metagenomic approach, to assess the diversity and phylogenetic distribution of BSH activity in this ecosystem (Jones et al. 2008; Fig. 4). The reliance on heterologous gene expression in the surrogate host (typically *E. coli*) and the requirement for a phenotypic screen for the trait of interest are clear limitations of the function-based strategy but are offset by unique benefits of this approach over other metagenomic techniques (Handelsman 2004).

A major advantage of the function-driven approach is that no prior knowledge or sequence data for the genes underpinning an activity is required, which not only allows the application of metagenomics to poorly studied microbial functions in a community (such as bile acid metabolism) but also permits the recovery of novel, unrelated enzyme classes catalyzing a particular reaction (Jones et al. 2008; Handelsman 2004). Furthermore, a clear confirmation of activity among the genes identified is intrinsic to the function-driven approach. This is



Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 4 Overview of metagenomic approaches to study microbial ecosystems. *Recovery of metagenomic DNA* (Modified from Ogilvie et al. 2012): Metagenomic approaches begin with sampling the microbial ecosystem and extracting DNA from the mixed community as a whole, without any prior cultivation. This metagenomic DNA may then be subjected to one or more strategies to access the functional content of the ecosystem under study and/or explore the population structure and identify species present. *Sequence-driven metagenomics*: Metagenomic DNA

may be subject directly to high-throughput sequencing (shotgun metagenomics) or first used as a template for PCR reactions intended to amplify key genes of interest. The latter is most typically used to amplify phylogenetic anchors, such as genes for 16S ribosomal RNA, which permit a census of the species present in a community. Sequences generated directly in the shotgun approach can subsequently be compared with well-characterized microbial genomes and/or assembled into large contigs and genes predicted, in order to assess the functions encoded by community members (with information on population structure also captured in this strategy where relevant gene

of major benefit in the analysis of enzymes such as BSH, which share a considerable degree of sequence homology with closely related enzymes in the wider Ntn_CGH-like (COG3049) family of proteins (Jones et al. 2008; Kumar et al. 2006). In particular BSH are closely related to penicillin V amidases, from which they are believed to have evolved, and comparison of sequence data alone is often insufficient for the accurate prediction of function in these enzymes (Kumar et al. 2006; Jones et al. 2008).

The function-driven approach employed to survey BSH activity in the gut microbiome was based on screening large-insert genetic libraries (constructed from metagenomic DNA derived from stool samples), using a simple plate-based assay to identify clones able to de-conjugate CBA (Fig. 5 Library construction and Screen). The basis of this screen is the complementation of the BSH-deficient *E. coli* host used to construct libraries and the subsequent de-conjugation of CBA incorporated into the bacterial growth media used for screening (Jones et al. 2008; Dashkevicz and Feighner 1989). Once liberated, free bile acids are no longer soluble and precipitate to form a halo around BSH-positive clones, allowing those harboring active BSH to be easily identified and recovered for further analysis (Jones et al. 2008; Fig. 5). Characterization of BSHs recovered from the human gut metagenomic library through function-based screening provided the basis to subsequently examine the distribution and evolution of this activity among

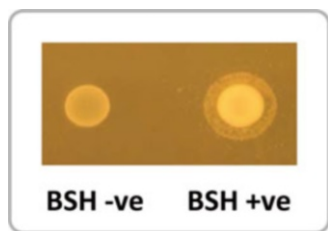
members of the gut microbiome, the conservation of this function between distinct human microbiomes, and the role of this activity in gut-associated bacteria (Jones et al. 2008; Ogilvie and Jones 2012; Fig. 6).

Distribution of BSH Activity Among Members of the Human Gut Microbiome: Sequence data obtained from metagenomic clones encoding BSH activity was used to predict the phylogenetic origin of the BSHs obtained and determine which members of the gut microbiome encode this function (Jones et al. 2008). Although the taxonomic resolution afforded by this analysis was limited by a lack of conserved phylogenetic anchors in many metagenomic clones (such as 16S rRNA genes) and the limited availability of genome sequences from gut-associated bacterial species at the time of analysis (against which recovered BSH sequences could be compared), this survey nevertheless revealed a broad distribution of BSH activity within the gut microbiome (Jones et al. 2008).

All major bacterial phyla comprising the human gut microbiome (Bacteroidetes, Firmicutes, Actinobacteria) were shown to encode this function, highlighting the high level of redundancy and general stability of BSH activity within the community (Jones et al. 2008). Furthermore, BSH activity was also identified in the archaeal species *Methanobrevibacter smithii*, which commonly forms a part of the human gut microbiome (Jones et al. 2008). These observations further expanded the representation of BSH among community members and revealed this function to be

Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 4 (continued) such as 16S rRNA genes are identified). **Function-driven metagenomics:** these approaches rely on the construction of large-insert genetic libraries and the heterologous expression of cloned genes in the surrogate host species (as used to explore BSH activity in the human gut microbiome; Jones et al. 2008). Although the requirement for genes originating in diverse and distantly related species to express functional proteins in the library host is a limitation of this method, unlike sequence-driven approaches, there is no requirement for prior information or well-characterized sequences

from genes of interest. This is a major advantage of the function-driven approach which facilitates the identification of novel enzyme classes and is well suited to explore activities for which few initial examples of well-characterized genes or proteins exist. However, a second major caveat of the function-driven approach is that a suitable high-throughput screen for the activity of interest must also be available (see Fig. 5). Fosmid (vectors based on the *E. coli* F-plasmid) and BACs (*bacterial artificial chromosomes*) represent the most commonly used systems for construction of large-insert metagenomic libraries



Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 5 Example of function-based screening for bile salt hydrolase activity (Modified from Jones et al. 2008). High-throughput function-driven metagenomic analysis of BSH activity in the human gut microbiome utilized a simple plate-based screen to identify clones encoding this activity (Dashkevicz and Feighner 1989). De-conjugation of CBA incorporated in the media results in precipitation of free bile acids and the formation of a distinct halo around BSH + clones. The image shows the phenotype of the surrogate, BSH-deficient *E. coli* host, and a corresponding BSH-positive metagenomic clone on the bile agar media

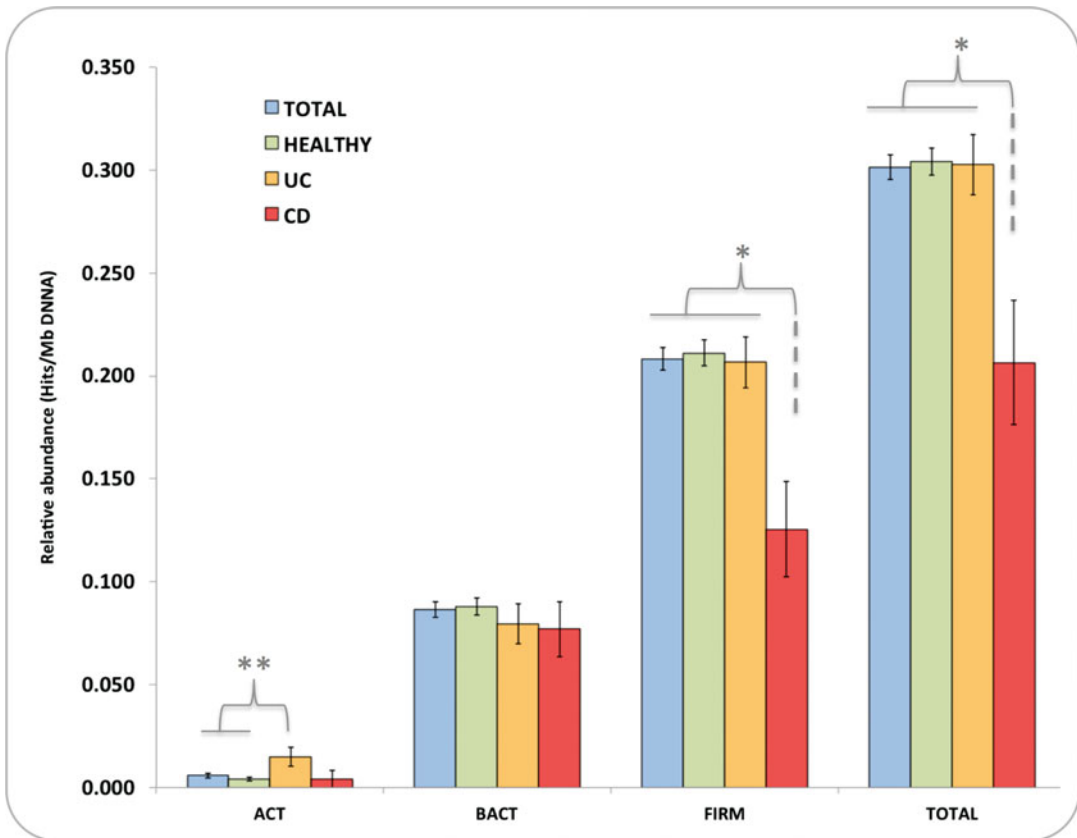
present in two domains of life (bacteria and archaea) in the gut microbiome (Jones et al. 2008).

In addition, the expression in *E. coli* of BSH genes predicted to originate from a wide range of bacterial species, as well as archaea, highlights the coverage afforded by the function-driven approach in this case (Jones et al. 2008). Despite this strategy being limited by the ability of the surrogate host to express the trait of interest, and genes derived from a wide range of often distantly related microbes, the function-driven survey of BSH demonstrates the clear potential for genes of diverse phylogenetic origin to be obtained by this method (Jones et al. 2008). Continued improvements in the range of hosts and vector systems available will further enhance the utility of this approach and expand the role of this strategy in the analysis of microbial communities.

Insights into BSH Evolution and Its Role in Gut Bacteria: The recovery of novel BSH sequences from the gut microbial ecosystem with confirmed function also allowed a deeper insight into the evolution and role of this activity in the gut microbiome (Jones et al. 2008). To understand the evolution of this activity within the gut community, sequences from novel BSH

obtained by function-driven metagenomics were compared with a large collection of related sequences, from gut-associated and non-gut-associated species, belonging to the wider Ntn_CGH family of which BSH are members. Clustering of these enzymes based on similarity of amino acid sequences revealed those derived from gut-associated microbes generally grouped together, despite originating from very different species (Jones et al. 2008). When the substrate range of enzymes with proven function was mapped against the observed groupings, a clear shift toward BSH activity was also evident among sequences that originated from gut microbes (Jones et al. 2008). Subsequently, murine experiments designed to test the contribution of BSH to bacterial survival in the gut clearly demonstrated the role of BSH in facilitating colonization of this habitat by mitigating the toxic effects of bile acids in the intestine (Jones et al. 2008).

Collectively, the results of experiments in murine models, together with trends observed in comparisons of functional BSH with related sequences, indicated BSH activity to be a common microbial adaptation to the gut environment, with selective pressure from conjugated bile acids likely to have driven the divergence of members of the Ntn_CGH family of proteins in gut bacteria toward BSH activity (Jones et al. 2008). Overall, this points to CBA as a key selective pressure in the gut habitat, and the development of a common mechanism for dealing with this stress in a diverse cross section of the community is congruent with the concept of host-level selection on functions of the gut microbiome (Ley et al. 2006). Bacteria face many challenges when colonizing and persisting in the mammalian intestinal tract, but the solutions developed for mitigating these barriers to survival must also be acceptable to the higher host organism and facilitate bacterial colonization without negative impact on fitness of the host (Jones et al. 2008; Ley et al. 2006). Therefore, the human host is believed to exert a selective pressure on functions and activities undertaken by the gut microbiome as a whole, and analysis of



Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome, Fig. 6 Relative abundance of bile salt hydrolases in the gut microbiome in health and disease (From Ogilvie and Jones 2012). Human gut microbiomes from the MetaHIT dataset were surveyed using sequence from BSH with proven function to identify homologues to these genes (minimum of 35 % amino acid identity ≥ 50 aa or more and $1e^{-5}$ or lower) in the 124 individual gut microbiomes represented in this dataset (Qin et al. 2010). Identified BSH sequences were subsequently affiliated to different bacterial divisions based on

sequence similarities and used to calculate the relative abundance of BSHs for major phylogenetic divisions in each gut microbiome (expressed as Hits/Mb DNA). *ACT* Actinobacteria; *BACT* Bacteroidetes; *FIRM* Firmicutes; *TOTAL* BSH relative abundance in MetaHit dataset as a whole irrespective of phylogenetic affiliations. *Healthy* healthy individuals only ($n = 99$), *UC* individuals with ulcerative colitis only ($n = 21$), *CD* individuals with Crohn's disease only ($n = 4$). Error bars indicate standard error of the mean. Level of significance: * $P < 0.01$; ** $P < 0.00$

microbial BSHs suggests that these may be an example of a mutually acceptable arrangement between the host and its microbiome (Jones et al. 2008).

BSH Activity as a Conserved Feature of the Human Gut Microbiome: Although the initial application of function-driven screens provided much fundamental insight into bile acid metabolism by the gut microbiome, these studies also fill an additional role in generating baseline

sequence data from genes with proven functions or activities. Such data in itself constitutes a useful and valuable resource for numerous other applications, including the accurate annotation and interpretation of shotgun metagenomes (and complete bacterial genomes), opening the way for larger-scale sequence-based surveys of key functions within microbial ecosystems. This is exemplified by the use of BSH recovered through function-driven metagenomics to

subsequently interrogate a range of sequence-based shotgun metagenomes, in order to examine the representation of this activity among distinct gut communities and other microbial ecosystems (Ogilvie and Jones 2012; Jones et al. 2008).

This approach was first applied to survey 15 human gut metagenomes and several non-gut metagenomes from a range of habitats (Jones et al. 2008). Comparison of the relative abundance of genes with homology to functional BSHs in human gut microbiomes with non-gut habitats revealed an enrichment of putative BSHs in the human gut microbiome (Jones et al. 2008). This is in keeping with the concept of CBA as an important habitat-associated selective pressure for gut microbes (absent in non-gut environments) and BSH as a conserved microbial adaptation to life in the mammalian intestinal tract (Jones et al. 2008).

When relative abundance of BSH homologues was compared between individual gut microbiomes, the potential for interindividual variation in abundance and types of BSH was also highlighted (Jones et al. 2008). Because BSH catalyzes the initial rate limiting step in the wider pathway of microbial bile acid metabolism facilitated by the gut microbiome (Fig. 3), variation in overall levels of BSH should be good predictors of the capacity for bile acid modification in a given microbiome (Jones et al. 2008; Ogilvie and Jones 2012). Furthermore, previous characterization of BSH types originating from the main phylogenetic groups in the human gut microbiome revealed differences in substrate range of enzymes encoded by different phyla, highlighting the potential for shifts in community structure to also alter aspects of bile acid metabolism by altering the prevailing bile acid modifications undertaken by gut microbes (Jones et al. 2008).

Metagenomic Analysis of Bile Salt Hydrolases in Health and Disease

Due to the role of bile acids in regulating metabolism and mucosal immune responses and the potential for the gut microbiome to influence

this signaling network through bile acid transformations, alterations in capacity for bile acid metabolism in the human gut microbiome may play a role in the pathogenesis of numerous diseases (Jones et al. 2008; Ogilvie and Jones 2012; Jones 2011). For example, the products of microbial bile acid metabolism have been linked to the initiation and pathogenesis of colorectal cancer (CRC) through several mechanisms, including the direct carcinogenicity of some BA (Bernstein et al. 2005; Hill 1990; O'Keefe 2008; Debruyne et al. 2001).

Recent observations also implicate the perturbation of bile acid signaling as a potential mechanism contributing to the pathogenesis of CRC and other inflammatory bowel diseases, with the dedicated bile acid receptor FXR α demonstrated to be protective against both CRC and Crohn's disease in murine models (Gadaleta et al. 2011; Modica et al. 2008; Duboc et al. 2013; Maran et al. 2009). Since activation of this receptor is implicated in the downregulation of mucosal immune responses and protection against autoimmune damage and induction of antiapoptotic pathways in the human gut (Gadaleta et al. 2011; Duboc et al. 2013), alterations to microbial bile acid metabolism leading to changes in the balance of BA species available for receptor binding have clear implications for disease initiation and progression.

The initial function-driven metagenomic analysis of BSH activity in the gut microbiome also provided the basic information to explore these theories further and to begin to explore the association between microbial bile acid metabolism and intestinal diseases (Jones et al. 2008; Ogilvie and Jones 2012). This is exemplified by the application of gut-derived BSH sequences (with proven activity) to explore changes in the BSH profile in the microbiomes of individuals with inflammatory bowel disease (Ogilvie and Jones 2012). Surveys of whole community shotgun metagenomes for genes homologous to functional BSH sequences revealed a distinct reduction in the relative abundance of BSH homologues in the gut microbiomes of individuals with Crohn's disease (CD), primarily within BSHs affiliated with

the Firmicutes division (Ogilvie and Jones 2012). These changes are in keeping with the well-documented dysbiosis and shift in community structure characteristic of CD (where the diversity of Firmicutes is markedly reduced) (Manichanh et al. 2006; Qin et al. 2010) and the role of FXR α signaling in regulation of mucosal immune responses (Gadaleta et al. 2011). These metagenomic-based predictions of changes in functional capacity of the CD gut microbiome related to bile acid metabolism have since been validated and a reduction in capacity for bile acid modification demonstrated in active disease (Duboc et al. 2013). The apparent deficiency of this function in the CD gut microbiome now raises the potential for targeting bile acid metabolism in the gut microbiota as a marker for disease risk or therapeutic intervention.

Summary

The analysis of bile acid metabolism in the human gut microbiome has benefited greatly from the application of metagenomics and provides an excellent example of how these powerful community-level approaches can rapidly provide significant insight into the functioning and development of microbial ecosystems. In the case of the human gut microbiome, and other host-associated microbial consortia, metagenomic approaches can also generate new understanding of how bacteria interact with and impact upon their higher host organisms.

In the case of bile acid metabolism by the gut microbiome, the deployment of metagenomics to explore this aspect of the indigenous intestinal microbiota has rapidly enhanced our understanding of this activity, its effect on human health, and its function within the gut microbiome. Our knowledge of bile acid metabolism by the gut microbiome has now been elevated to a point where tangible hypotheses regarding impacts on host health can be formulated and tested. Although much remains to be done and our understanding is far from complete, metagenomics will undoubtedly continue to play a key role in ongoing studies

and has already yielded new targets for disease diagnosis, prophylaxis, or treatment which can now be explored further.

References

- Begley M et al. The interaction between bacteria and bile. *FEMS Microbiol Rev.* 2005a;29:625–91.
- Begley M et al. Contribution of three bile-associated loci, *bsh*, *pva*, and *btlB*, to gastrointestinal persistence and bile tolerance of *Listeria monocytogenes*. *Infect Immun.* 2005b;73:894–904.
- Begley M et al. Bile salt hydrolase activity in probiotics. *Appl Environ Microbiol.* 2006;72:1729–38.
- Bernstein H et al. Bile acids as carcinogens in human gastrointestinal cancers. *Mutat Res.* 2005;589:47–65.
- Dashkevich MP, Feighner SD. Development of a differential medium for bile salt hydrolase-active *Lactobacillus* spp. *Appl Environ Microbiol.* 1989;55:11–6.
- Debruyne PR et al. *Mutat Res.* 2001;480–81:359–69.
- Duboc H et al. Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. *Gut.* 2013;62:531–9.
- Gadaleta RM et al. Farnesoid X receptor activation inhibits inflammation and preserves the intestinal barrier in inflammatory bowel disease. *Gut.* 2011;60:463–72.
- Gill SR et al. Metagenomic analysis of the human distal gut microbiome. *Science.* 2006;312:1355–9.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev.* 2004;68:669–85.
- Hill MJ. Bile flow and colon cancer. *Mutat Res.* 1990;238:313–20.
- Hofmann AF, Eckmann L. How bile acids confer gut mucosal protection against bacteria. *Proc Natl Acad Sci U S A.* 2006;103:4333–4.
- Houten SM et al. Endocrine functions of bile acids. *EMBO J.* 2006;25:1419–25.
- Inagaki T et al. Regulation of antibacterial defense in the small intestine by the nuclear bile acid receptor. *Proc Natl Acad Sci U S A.* 2006;103:3920–5.
- Jones BV. Bacterial bile acid modification and potential pharmaceutical applications. *J Appl Ther Res.* 2011;8:94–100.
- Jones BV, Marchesi JR. Accessing the mobile metagenome of the human gut microbiota. *Mol Biosyst.* 2007;3:749–58.
- Jones BV et al. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc Natl Acad Sci U S A.* 2008;105:13580–5.
- Kim GB et al. Cloning and characterization of the bile salt hydrolase genes (*bsh*) from *Bifidobacterium bifidum* strains. *Appl Environ Microbiol.* 2004;70:5603–12.

- Kumar RS et al. Structural and functional analysis of a conjugated bile salt hydrolase from *Bifidobacterium longum* reveals an evolutionary relationship with penicillin V acylase. *J Biol Chem.* 2006;281:32516–25.
- Kurokawa K et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 2007;14:169–81.
- Lambert JM et al. Functional analysis of four bile salt hydrolase and penicillin acylase family members in *Lactobacillus plantarum* WCFS1. *Appl Environ Microbiol.* 2008a;74:4719–26.
- Lambert JM et al. Improved annotation of conjugated bile acid hydrolase superfamily members in Gram-positive bacteria. *Microbiology.* 2008b;154:2492–500.
- Ley RE et al. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell.* 2006;124:837–48.
- Lodola A et al. A catalytic mechanism for cysteine N-terminal nucleophile hydrolases, as revealed by free energy simulations. *PLoS ONE.* 2012;7:e32397. doi:10.1371/journal.pone.0032397.
- Manichanh C et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut.* 2006;55:205–11.
- Maran RRM et al. Farnesoid X receptor deficiency in mice leads to increased intestinal epithelial cell proliferation and tumor development. *J Pharmacol Exp Ther.* 2009;328:469–77.
- Modica S et al. Nuclear bile acid receptor FXR protects against intestinal tumorigenesis. *Cancer Res.* 2008;68:9589–94.
- O'Keefe SJD. Nutrition and colonic health: the critical role of the microbiota. *Curr Opin Gastroenterol.* 2008;24:51–8.
- Ogilvie LA, Jones BV. Dysbiosis modulates capacity for bile acid metabolism in the gut microbiomes of patients with IBD: a mechanism or marker for disease? *Gut.* 2012. doi:10.1136/gutjnl-2012-302137.
- Ogilvie LA et al. Evolutionary, ecological and biotechnological perspectives on plasmids resident in the human gut mobile metagenome. *Bioeng Bugs.* 2012;3:1–19.
- Qin J et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:59–65.
- Ridlon JM et al. Bile salt biotransformations by human intestinal bacteria. *J Lipid Res.* 2006;47:241–59.
- Rossocha M et al. Conjugated bile acid hydrolase is a tetrameric N-terminal thiol hydrolase with specific recognition of its cholyl but not of its tauryl product. *Biochemistry.* 2005;44:5739–48.
- Suresh CG et al. Penicillin V acylase crystal structure reveals new Ntn-hydrolase family members. *Nat Struct Biol.* 1999;6:414–6.
- Thomas C et al. Targeting bile-acid signalling for metabolic diseases. *Nat Rev Drug Discov.* 2008;7:678–93.
- Watanabe M et al. Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation. *Nature.* 2006;439:484–9.

Metagenomic by RAPD Profiling

Jaime Henrique Amorim, João Carlos Teixeira Dias and Rachel Rezende
Universidade Estadual de Santa Cruz,
Laboratório de Biotecnologia Microbiana,
Ilhéus, BA, Brazil

Detailed description and study of taxa, metabolic pathways, protein/peptide interactions, and molecular relationships in microenvironments bring out great interest due to the possibility of yielding new molecules with important applicability and new knowledge about the microenvironment dynamics. However, such advances are not possible by culturing dependent techniques, due to lack of knowledge of culturing conditions to unknown microorganisms (Yun et al 2004; Riesenfeld et al. 2004). Metagenomic approaches have been pointed as a way to further access data contained in these ecosystems (Johnson and Slatkin 2006; McHardy and Rigotsos 2007). This technology allows access to taxonomical and metabolic data (Streit and Schmitz 2004; Roh et al. 2006; McHardy and Rigotsos 2007) independently of culturing proceedings. Nevertheless, a main drawback using metagenomic approaches is that most of them are preceded by conventional DNA extraction methods that prejudice taxonomical representativeness and difficulties cloning steps by interference substances. By biasing taxonomical representativeness, such methods also limit the mining of new molecules. In addition, such metagenomic conventional approaches based on cloning of polymerase chain reaction (PCR) products are unfeasible if the aim is to access taxonomical and metabolic diversity at the same time (Schloss and Handelsman 2003). Another aspect that limits the study of metagenomic content is the use of bioinformatics methods (Rondon et al. 2000; Roh et al. 2006; Huson et al. 2007) that depend on sequences of the same gene to compute taxonomical or metabolic profile of the environment (Rondon et al. 2000; Roh et al. 2006). Thus,

once more, they do not allow understanding of taxonomical and metabolic content at the same time (Rondon et al. 2000; Roh et al. 2006).

The random amplified polymorphic DNA (RAPD) is an approach that allows the study of genetic diversity and population structure of bacteria (Baker and Banfield 2003; Akbar et al. 2005). In metagenomic studies, it has been exploited in its conventional form, through analysis of the polymorphic amplified DNA segments in electrophoretic devices (Helton and Wommac 2009; Patel and Behera 2011). However, we recently reported a new and interesting application for RAPD in metagenomics by coupling it with an innovative metagenomic DNA extraction method (Amorim et al. 2012). By cloning RAPD instead of PCR products, we were able to access taxonomical and metabolic content at the same time. This advantage is due to the capacity of RAPD primers to anneal in a more broad number of DNA segments, yielding amplified DNA with different sizes and from different gene families. Randomly amplifying metagenomic DNA segments may result in at least three great fields of investigation: (i) it is possible to infer the taxonomical diversity of a specific environment if a suitable bioinformatic approach is available (Huson et al. 2007; Amorim et al. 2012); (ii) it is possible to take advantage on the variety of gene families amplified and mine new genes and molecules; (iii) if DNA fragments with different sizes and from different gene families are amplified, it is possible to infer the metabolic network in a specific environment. All of these possibilities may significantly improve and expand the use of RAPD in the study of genetic diversity and population structure of microorganisms.

To study the environmental taxonomical and genetic diversity using RAPD in the context of metagenomics, it is necessary to clone such amplified DNA and then study their sequences. Another possibility is to determine their sequences directly on pyrosequencing devices. However, it is important to determine a size cut-off of sequences that will be used to compute the taxonomical diversity of the environment, in order to avoid inconclusive sequences regarding

taxonomical information (Huson et al. 2007). In addition, it is also necessary to use a suitable algorithm that is able to compute such taxonomical content based on sequences from different gene families (Amorim et al. 2012). The advantage of studying environmental diversity with this approach is the possibility to amplify not only bacterial but also viral, fungi, and other eukaryotic sequences at the same time due to the no specificity of RAPD primers. Thus, this approach makes possible to infer the whole taxonomical diversity of a specific environment.

Randomly amplifying sequences from a variety of gene families may yield new molecules that may have important applications. Again, the nonspecificity of RAPD primers works in benefit of the diversity of amplified environmental DNA. However, it is necessary to determine again a size cutoff, in order to maximize the probability of cloning a complete viable gene. In addition, the representativeness of metagenomic content must be considered in order to also maximize the probability of new genes and molecule mining. Such requirement is due to some DNA extraction methods that precede metagenomic approaches and have the characteristic of restricting or limiting the metagenomic representativeness regarding the taxonomical and metabolic diversity of a specific environment (Amorim et al. 2008, 2012). The advantage of mining new substances with this approach is the possibility to search for molecules with different biological functions at the same time. In addition, it is possible to couple new molecule mining to genetic diversity profiling by using RAPD.

As a new and interesting applicability, it was shown that the use of RAPD in metagenomics may turn possible to infer the metabolic network of a specific environment. Once that genes involved with related metabolic pathways are amplified, it is possible to use suitable algorithms to study its relationships in the same taxon or between different taxons. For all possibilities discussed here, RAPD seems to be a simple but robust tool to significantly improve the metagenomic research.

References

- Akbar T, Akhtar K, Ghauri MA, Anwar MA, Rehman M, Rehman M, Zafar Y, Khalid AM. Relationship among acidophilic bacteria from diverse environments as determined by randomly amplified polymorphic DNA analysis (RAPD). *World J. Microbiol. Biotech.* 2005;21:645–8.
- Amorim JH, Macena TNS, Lacerda-Junior GV, Rezende RP, Dias JCT, Cascardo JCM. An improved extraction protocol for metagenomic DNA from a soil of the Brazilian Atlantic Rainforest. *Genet Mol Res.* 2008;4:1226–32.
- Amorim JH, Vidal RO, Lacerda-Junior GV, Dias JCT, Brendel M, Rezende RP, Cascardo JCM. A simple boiling-based DNA extraction for RAPD profiling of landfarm soil to provide representative metagenomic content. *Genet. Mol. Res.* 2012;11:182–9.
- Baker BJ, Banfield JF. Microbial communities in acid mine drainage. *FEMS Microbiol Ecol.* 2003;44(2): 139–52.
- Helton RR, Wommac KE. Seasonal dynamics and metagenomic characterization of estuarine viriobenthos assemblages by randomly amplified polymorphic DNA PCR. *Appl Environ Microbiol.* 2009;75(8):2259–65.
- Huson DH, Auch AF, Stephan JQ, Schuster C. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17:377–86.
- Johnson PLF, Slatkin M. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res.* 2006;112:1320–7.
- McHardy AC, Rigoutsos I. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol.* 2007;10:499–503.
- Patel AK, Behera N. Genetic diversity of coal mine spoil by metagenomes using random amplified polymorphic DNA (RAPD) marker. *Indian J Biotechnol.* 2011;10:90–6.
- Riesenfeld CS, Schloss PD, Handelsman J. METAGENOMICS: genomic analysis of microbial communities. *Annu Rev Genet.* 2004;38:525–52.
- Roh C, Villatte F, Kim B, Schmid RD. Comparative study of methods for extraction and purification of environmental DNA from soil and sludge samples. *Appl Biochem Biotechnol.* 2006;134:97–112.
- Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loicono KA, Handelsman J, Goodman RM. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol.* 2000;115:2541–7.
- Schloss P, Handelsman J. Biotechnological prospects from metagenomics. *Curr Opin Biotechnol.* 2003;14:303–10.
- Streit WR, Schmitz RA. Metagenomics – the key to the uncultured microbes. *Curr Opin Microbiol.* 2004;7: 492–8.
- Yun J, Kang S, Park S, Yoon H, Kim M, Heu S, Ryu S. Characterization of a novel amylolytic enzyme encoded by a gene from a soil-derived metagenomic library. *Appl Environ Microbiol.* 2004;11:7229–35.

Metagenomic Potential for Understanding Horizontal Gene Transfer

Luigi Grassi¹, Jacopo Grilli² and Marco Cosentino Lagomarsino^{3,4}

¹Physics Department, Sapienza University of Rome, Rome, Italy

²Dipartimento di Fisica “G. Galilei”, CNISM and INFN, Università di Padova, Padova, Italy

³Computational and Quantitative Biology, University Pierre et Marie Curie, Paris, France

⁴CNRS, Paris, France

Definition

Horizontal gene transfer (HGT) describes the biological phenomenon by which an organism acquires genes from organisms belonging to other species, genera, or taxa. Its name reflects the fact that the transfer of genetic information between organisms that are not necessarily related is different from the “vertical” transmission of genes from parent to offspring. Early reports (Smith et al. 1992) interpreted HGT as a rare event, unable to significantly influence the global composition of target genomes. This first impression was rapidly subverted by the advent of genomic sequencing technologies. For example, the comparison of the genomes of *Escherichia coli* and *Haemophilus influenzae*, two bacteria belonging to the same evolutionary lineage, shows a significant difference in their gene content (Tatusov et al. 1996). This difference, which is not at all justifiable only in terms of vertical descent, gave a first indication of the massive role played by HGT in the evolution of prokaryotic genomes. Subsequent evidence from multiple genomes indicates that HGT acts

pervasively on prokaryotic genomes (Woese 2000). For example, a detailed analysis made by Dagan et al. (Dagan and Martin 2007) considered 57670 gene families across 190 sequenced genomes demonstrating that at least two-thirds and possibly all of them have been affected by HGT at some time in their evolutionary past.

Introduction

The pervasive HGT occurrence reveals the importance of this process in forging the extant prokaryotic genomes (Ochman et al. 2005). The consequent question is how and to what extent the transferred genes innovate a genome over the course of evolution and how they are incorporated into a genome's existing biochemical and regulatory networks. This issue requires the study of multiple genomes and naturally overlaps with metagenomics, because HGT is affected by environmental, ecological, and population factors acting at the level of communities of coexisting species. In brief, the understanding of HGT passes through the knowledge of its consequences on genomes, populations, and ecosystems.

HGT Impact on the Evolution of Genomes

A genome affected by HGT can acquire two typologies of genes: genes homologous to existing ones and genes that are not (Ochman et al. 2000). Both types of HGT influence the evolution of a lineage but do so in very different manners and contexts. The first mechanism is favored when the phylogenetic distance between donor and acceptor is small (Andam and Gogarten 2011); these transfers may occur via homologous exchange, whose probability increases with genetic similarity (Vulic et al. 1997). The second type of HGT involves the acquisition of new genes, with a sporadic phylogenetic distribution. Such transfers might supply genes that confer novel phenotypic properties and result in the rapid adaptation of a bacterial species. Both

types of transfers can leave traces in the metabolism of the acquiring genomes. Several studies found that the majority of changes to the metabolic network of *Escherichia coli* in the past 100 million years are due to HGT (Pál 2005; Lercher and Pal 2008). Interestingly it appears that horizontally transferred genes are integrated at the periphery of the network, whereas central parts remain evolutionarily stable. This is also supported by the modular nature of prokaryotic genes. Indeed, metabolically related genes (e.g., genes coding proteins in physiologically coupled reactions) are often transferred together as operons. Thus, HGT appears to be the main force able to expand bacterial metabolic networks by enlarging their periphery in response to changing environments (Lercher and Pal 2008; Pang and Maslov 2011). Necessarily, this carries consequences on the addition of new genes regulating the metabolic pathways, defining some observed quantitative features of genome composition (Grilli et al. 2012; Koonin 2011). All the above studies evaluate the contribution of HGT to the evolution of prokaryotic genomes using the tools of comparative genomics. Quite interestingly, the effects and consequences of HGTs can also be evaluated with direct experiments. For example a single-cell analysis was performed in *Escherichia coli* in 2008 (Babic et al. 2008). This study proves the high efficiency (up to 96 % of recipients) of recombination and integration of transferred DNA. In another study, Babic et al. monitored in real time, through fluorescence microscopy, the sequential conjugation events of an integrative and conjugative element encoding for a green fluorescent protein (GFP) (Babic et al. 2011). A recent study investigated the novelty of protein domains acquired through HGT in Proteobacteria, focusing on their specific features (Grassi et al. 2012). The results indicate that protein domains subject to HGT have a transferability proportional to their total frequency in the pool of considered genomes, and at the same time, HGTs of exogenous protein families are found less frequently for larger genomes. Based on these observations, one can conclude that HGTs behave as if they were drawn

randomly from a cross-genomic community gene pool, much like gene duplicates are drawn from a genomic gene pool. Similar conclusions were drawn from a recent comparative study of *Escherichia coli* and *Salmonella enterica* genomes (Karberg et al. 2011). These results indicate a role of a common gene pool in determining the genes available for horizontal transfer and link the problem to the structure of past and existing bacterial communities and ecosystems.

Advantage of Metagenomics in the Study of HGT

The use of single (often cultured) bacterial species for the study of HGT has several limitations. Firstly the great majority of bacteria (more than 99 %) cannot be cultured in the laboratory. Furthermore, such a “single-species” approach gives, by definition, an organism-centric view of the phenomenon. This implies a limited understanding of microbial physiology, genetics, and community ecology. Many recent shotgun sequencing projects characterized the genome content of whole microorganism communities (Riesenfeld et al. 2004). For example, the “Sorcerer II” Global Ocean Sampling expedition was designed with the precise aim of giving a global snapshot of the marine microbiological world (Rusch et al. 2007). The results of this important project traced an impressive distance between marine microorganisms and cultivated ones. Very few metagenomic sequences were found to be similar to the ones of annotated genomes. A subsequent analysis of these data indicates that the abundant and cosmopolitan picoplanktonic prokaryotes tend to have smaller genomes (Yooseph et al. 2010). Such condition is probably associated to a slow growth lifestyle and with the relative inability to sense or rapidly acclimate to energy-rich conditions. By contrast, the microbial taxa display the ability of growing slowly and surviving in energy-limited environments, while growing rapidly in energy-rich environments. One other focus of interest for metagenomics is the exploration of the human microbiome. This large project has the final goal

of characterizing associations between human microbiome and health of an individual (Nelson et al. 2010), i.e., the ecological influence that microorganisms have on humans.

Metagenomic data make it difficult to formalize the traditional concept of species (and consequently of recombination and HGT among prokaryotes). Nevertheless, there are interesting findings that reveal the crucial role played by HGT in microbial communities. For example, Hehemann et al. (2010) point out that metagenomic samples derived by feces of Japanese individuals are enriched in carbohydrate-active enzymes (e.g., porphyranases and agarases), while the same enzymes are absent in metagenomic samples derived by North American individuals. Interestingly, gut bacteria from Japanese individuals have acquired these enzymes through HGT. This finding confirms the observation that HGT events among bacteria from different environments can occur also inside the human intestine (Lester et al. 2006). Another recent study of Smillie et al. (2011) uses metagenomics to describe the forces governing HGT. The authors identified, through a heuristic method, recent HGTs among thousands of microbial genomes. Roughly one-quarter of the identified transfers includes at least one predicted mobile element, confirming the importance of such elements in facilitating gene exchange. However, the most interesting finding of this study is that bacteria isolated from human body are 25 times more likely to share HGT genes than bacteria living in different environments (aquatic, terrestrial, and nonhuman host associated). This phenomenon is even more striking considering human isolates derived by the same body site, with a rate of transfer increased by a factor of two. The authors also studied this high transferability in human microbiome separating bacteria by their ribosomal 16S distance, reporting that even most divergent bacteria, separated by billions of years of evolution, but sharing the same ecological niche, are affected by more HGT than the most closely related isolates living in different niches.

The above findings indicate that ecological factors are relevant for driving HGT in the

human microbiome and thus play a role in its evolution and genomic composition. A global understanding of these aspects is a future challenge for metagenomics, which could expand our fundamental understanding of evolution, with implications for biotechnology and health.

Summary

Horizontal gene transfer (HGT) is a widespread phenomenon in prokaryotes. Its pervasive modality of action enormously influences the receiving genomes. In light of this HGT appears among the main forces able to expand bacterial metabolic networks in response to changing environments. Metagenomics opens up new perspective to the study of HGT by giving the possibility to uncover the ecological factors relevant for driving HGT. Interestingly it can directly investigate complex ecosystems as marine microbiological world and the human microbiome.

Cross-References

- ▶ [Integrins as Repositories of Genetic Novelty](#)
- ▶ [Lateral Gene Transfer and Microbial Diversity](#)
- ▶ [Metagenome of acidic hot spring microbial planktonic community: Structural and functional insights](#)

References

- Andam CP, Gogarten JP. Biased gene transfer and its implications for the concept of lineage. *Biol Direct*. 2011;6:47.
- Babic A, Lindner AB, Vulic M, Stewart EJ, Radman M. Direct visualization of horizontal gene transfer. *Science*. 2008;319:1533–6.
- Babic A, Berkmen MB, Lee CA, Grossman AD. Efficient gene transfer in bacterial cell chains. *MBio*. 2011;2(2):e00027.
- Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A*. 2007;104:870–5.
- Grassi L, Caselle M, Lercher MJ, Lagomarsino MC. Horizontal gene transfers as metagenomic gene duplications. *Mol Biosyst*. 2012;8:790–5.
- Grilli J, Bassetti B, Maslov S, Lagomarsino MC. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucleic Acids Res*. 2012;40:530–40.
- Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*. 2010;464:908–12.
- Karberg KA, Olsen GJ, Davis JJ. Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome. *Proc Natl Acad Sci U S A*. 2011;108:20154–9.
- Koonin EV. Are there laws of genome evolution? *PLoS Comput Biol*. 2011;7:e1002173.
- Lercher MJ, Pal C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol*. 2008;25:559–67.
- Lester CH, Frimodt-Moller N, Sorensen TL, Monnet DL, Hammerum AM. In vivo transfer of the vanA resistance gene from an *Enterococcus faecium* isolate of animal origin to an *E. faecium* isolate of human origin in the intestines of human volunteers. *Antimicrob Agents Chemother*. 2006;50:596–9.
- Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, et al. A catalog of reference genomes from the human microbiome. *Science*. 2010;328:994–9.
- Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000;405:299–304.
- Ochman H, Lerat E, Daubin V. Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A*. 2005;102 Suppl 1:6595–9.
- Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*. 2005;37:1372–5.
- Pang TY, Maslov S. A toolbox model of evolution of metabolic pathways on networks of arbitrary topology. *PLoS Comput Biol*. 2011;7:e1001137.
- Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*. 2004;38:525–52.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*. 2007;5:e77.
- Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*. 2011;480:241–4.
- Smith MW, Feng DF, Doolittle RF. Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci*. 1992;17:489–93.
- Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, et al. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr Biol*. 1996;6:279–91.

- Vulic M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A*. 1997;94:9763–7.
- Woese CR. Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A*. 2000;97:8392–6.
- Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M, et al. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*. 2010;468:60–6.

Metagenomic Research: Methods and Ecological Applications

Navneet Batra¹, Sonu Bhatia¹, Arvind Behal¹, Jagtar Singh² and Amit Joshi³

¹Department of Biotechnology, GGSDS College, Chandigarh, India

²Department of Biotechnology, Panjab University, Chandigarh, India

³Department of Biotechnology & Bioinformatics, SGS College, Chandigarh, India

Synonyms

Community genomics; Ecological genomics; Environmental genomics

Definition

The aim of metagenomics is to investigate enormous diversity of taxonomically and phylogenetically relevant genes, individual catabolic genes, and whole operons by explicating the genomes of uncultured microbes. The concept of metagenomics was introduced by Handelsman which involves the extraction of genomic DNA from the microbial community inhabiting the environment. Either this DNA is cloned as libraries for functional screening, or PCR-based enrichment is performed with respect to gene of interest. Generally, DNA is considered as the most appropriate method for assessing environmental microbial community's structure as any

selection biasness is not involved. High-throughput methods can be employed for direct sequencing of the metagenome. The functional approach is used to explore genes that encode novel enzymes or drugs, but advancements are needed for function-based metagenomics by employing high-throughput screenings.

Introduction

Enormous genetic and biological pool of microbial diversity is present on the Earth. It accounts for $4-6 \times 10^{30}$ prokaryotic cells containing 10^6-10^8 distinct genospecies. Various studies based on molecular approaches prove that approximately 1 % of the total vast microbial population is culturable under cultivation conditions and in media of restricted and optimized range. Cultivation techniques pose several difficulties and limitations. To encounter this problem, various DNA-based molecular methods have been developed. Culture-independent methods were firstly applied to environmental system like hot springs in Yellow Stone National Park. Further technical developments in this field have led in to an era of metagenomics (Singh et al. 2009).

Conserved rRNA gene sequences are used in sequence-driven strategies of genomics to explore microbial diversity. 16S rRNA gene has been established as the standard molecule for taxonomic diversity analysis in metagenomics. Recently 23S rRNA has also been considered as it offers advantages over 16S rRNA with size almost twice as long as that of 16S rRNA. Thus, it can prove to be a more informative phylogenetic marker in comparison to 16S rRNA gene (Yilmaz et al. 2011); however, 23S rRNA faces a drawback as lower number of reference sequences of this marker is present in public databases. All metagenomic output is collected and shared across public databases and retrieved using bioinformatic tools capable of dealing with extensive generated data. Genomic Standards Consortium (GSC) was established in 2005 with its main intention to promote and share the information about the resources required in the

development of better and improved mechanisms of metadata capture and exchange (Mocali and Benedetti 2010).

Methodology

All metagenomic approaches are mainly based on the technique of isolation and examination of DNA extracts directly from naturally occurring microbial populations. In functional metagenomic studies (expression dependent), the examination of DNA libraries is done by high-throughput assays to identify clones that have specific desired phenotype. However, homology-based approach constitutes probing of the library to identify clones containing conserved sequences. Metagenomics includes the following major steps.

Sampling

Most of the studies illustrate that estimate of microbial diversity increases with the areas sampled. While beginning with metagenomic analysis of microorganisms, sampling is done using well-established protocols to provide the best representative sample of the desired site (Hirsch et al. 2010). Microbial activity and growth patterns in soil are influenced by physical, chemical, and biological properties. Soil substratum and geographical location affect the phylogenetic composition of the microbial community. Selection of sampling site and sample method are important considerations (Kakirde et al. 2010). The number and diversity of microorganisms that is to be sampled are affected by the depth of the soil at which sampling is performed. Multiple spatial scales are used for sampling at different intervals, to demonstrate spatial heterogeneity of soil microbial communities in an agriculture soil. It was suggested that geo-statistical analysis can be used to describe spatial distribution of the microbes present at the subsurface of soil along with power analysis for the assessment of the required sample size (Mocali and Benedetti 2010). Sampling variability can be significantly reduced by using such a regime. Sampling from aquatic systems based on marine habitat

is comprehensively been undertaken by Global Ocean Sampling (GOS) expedition. The Atlantic, Pacific, and Indian oceans are covered as part of their metagenomic studies (Yilmaz et al. 2011).

Techniques for environmental sampling are dependent on the purpose of the study, the habitat to be sampled, and the desired downstream analysis (Lewin et al. 2012). With the advent of an era of new high-throughput methods, the number of samples accessed can be greatly increased. Sampling has been reported from different environment for metagenomic analysis including soils, surface water of the sea, deep sea sediments, various organs of animals and humans, compost, sludge, acid mine site, Arctic sediments, etc. (Singh et al. 2009; Kakirde et al. 2010; Xing et al. 2012). Oil samples were obtained from a production well at the onshore Potiguar Basin, Brazil, with an in situ temperature of 42.2 °C and depth of 535.5–540.5 m for the purpose of screening for hydrocarbon biodegraders in a metagenomic clone library (Vasconcellos et al. 2010).

Extraction Methods

Quality of extracted DNA samples should be high for construction of metagenomic library. This extraction and purification of nucleic acids should be performed critically. Methodology of DNA extraction is based according to the size of the target genes and on screening strategies. Metagenome extraction is the arbitration between vigorous extraction that is done for the representation of all microbial genome and lowering DNA shearing with simultaneous co-extraction of sample contaminants (Cowan et al. 2005). In direct extraction methodology, the samples are processed without the cultivation of the microbes and involve the use of detergents and enzymes. The samples are further treated with phenol or chloroform. It has been argued that this method of extraction is biased; for example, ammonia-oxidizing bacteria and methane-oxidizing bacteria are not easily displaced from soil particles, when compared to the other bacterium inhabiting the soil, and also actinomycete spores may be underrepresented (Hirsch et al. 2010). Physical

means of separation of microbes with lysis-based extraction is employed in indirect extraction approach. Cell lysis can be performed using methods like sonication, grinding, freezing-thawing, and solubilization of cell membranes and cell walls by detergents or by employing enzymatic means (Singh et al. 2009). Bead-beating method was used for microbial lysis on soil samples collected from barren regions of Gujrat (India) to index microbial population and community structure in saline-alkaline soil using gene target metagenomics (Keshri et al. 2013). Direct DNA extraction methods show higher recovery rate of DNA (10–100 times) as compared to indirect methods but length of DNA fragments is larger in case of indirect methods. Impurity content is more in DNA extracted using direct methods as compared to indirect methods (Xing et al. 2012).

The extraction method is selected on the basis of desired applications. Delmont et al. (2011) compared direct and indirect soil extraction approaches, and they concluded that there was a more than 40 % decrease in Eukarya sequences when using indirect DNA extraction as compared to direct method. Archaeal and bacterial sequences also increased in indirect approaches. Another concern in extraction process is the presence of various contaminants in the soil, for example, humic acid, polyphenols, polysaccharides, and nucleases, which can prove inhibitory to different applications including PCR and metagenomic library construction. Single-DNA extraction methods can underestimate the total number of bacterial ribotypes present in marine sediments (Singh et al. 2009; Kakirde et al. 2010). To obtain the significant amount of DNA, large quantities of material is required. RNA recovery from environments is quite similar to that of DNA isolation. To decrease physical degradation and RNase activity, samples are specifically processed. Harvesting of samples is followed by freezing it at -80°C . Sulfate salt solution can be used to coprecipitate cellular RNA with proteins. cDNA metagenomic libraries can be constructed to identify functional eukaryotic genes using RNA extraction approach (Cowan et al. 2005).

Enrichment of Sample

Whereas non-enrichment methods have a capacity to maintain high diversity of microbial communities, to increase the specificity of a sample's genomic DNA, enrichment is performed. Screening of sequence-based novel genes is benefited by enrichment (Xing et al. 2012). Such methods have power to select particular community based on its function. The loss of diversity can be moderated by alteration of the degree of the selection pressure applied. Active biomolecule of the microorganisms can be targeted using genome enrichment strategies. Enrichment of the target population can be achieved by the use of selective media due to its capability to utilize specific substrate. Novel techniques are employed to enrich specific microbial community such as 5-bromo-2-deoxyuridine (BrdU) labeling that can be done on actively growing microbes, followed by separation of labeled nucleic acids by density gradient centrifugation and immuno-capture techniques. Growth of specific substrate utilizing microbial community can be enhanced by addition of substrates along with BrdU (Singh et al. 2009).

SSH (suppressive subtractive hybridization) technique is also used for specific gene enrichment and identifying genetic differences between microorganisms. Samples are ligated with adaptors and such fragments are selected on the basis of subtractive hybridization. The effect of specific pollutants on the community DNA can be determined with this enrichment technique by making a comparison with reference metagenome in the absence of that pollutant (Cowan et al. 2005). Another enrichment method is stable isotope probing (SIP) that can be used to target metagenomics to specific populations. SIP involves stable isotope-labeled substrate and separation of heavier nucleic acids (DNA/RNA) by density gradient centrifugation. Metagenome expression profile can be compared in response to specific substrates or xenobiotic compounds in method based on differential expression analysis (DEA), which can identify genes which are upregulated for specific activity (Cowan et al. 2005). In addition, microarray method, phage display expression system, and multiple

displacement amplification and differential display are other methods for enrichment of genomic DNA (Singh et al. 2009). Aerobic and anaerobic microbial enrichments can also be performed as done in a study involved in screening for hydrocarbon degraders. These cultures were grown in Schott bottles containing 500 ml Widdel B mineral medium supplemented with n-hexadecane as carbon source. This resulted in anaerobic enrichment of sample (Vasconcellos et al. 2010). WGA (whole genome amplification) approach involves the use of short random primers to replicate DNA and is employed when limited-sized sample (microsamples) is to be processed (Hirsch et al. 2010).

Construction of Metagenomic DNA Libraries

Construction of a metagenomic library depends on appropriate vector. Quality of extracted DNA and associated research goals plays an important role in vector selection. Plasmids, cosmids, bacterial artificial chromosomes, and fosmids are extensively used vectors. The choice of vector is influenced by the size of the insert fragment, copy number of vector required, host used, and screening methods (Xing et al. 2012). Cosmid DNA libraries are constructed with an insert size ranging between 25 and 35 Kb. BAC libraries can permit the size up to 200 Kb and fosmid libraries with inserts of 40 Kb of foreign DNA (Streit and Schmitz 2004). pCR 2.1 vector was used for cloning, and plasmids were further screened for insert size by PCR-based amplification in the study of community structure in saline-alkaline soil (Keshri et al. 2013). Molecular classification of gliomas was done using P-1-derived artificial chromosome (PAC). Large-sized human genomic DNA is best cloned in YAC or BAC (Xing et al. 2012). Entire metabolic pathways can be recovered by cloning large fragments of metagenomic DNA in vectors. Host selection is another important criteria considered for efficient cloning. Host strain should be selected on the basis of efficiency of the conversion process, gene expression, plasmid stability in the host

cells, and screening of the target genes. Commonly used host strains are *E. coli*, *Streptomyces sp.*, *Pseudomonas sp.* and *Rhizobium sp.* Highly sheared DNA poses a major problem in library generation, because ligatable sticky ends cannot be formed out of highly sheared DNA. Blunt-end ligation can overcome this problem to some extent (Singh et al. 2009).

Integrated approach of stable isotope probing (SIP) and metagenomics has increased the frequency of clones containing target genes which are desirable. In one of the study on methane-utilizing bacteria in a forest soil, the sample was labeled with CH_4 and “heavy” DNA was used to construct a bacterial artificial chromosome library. 2,300 clones had to be screened in order to obtain pmoCAB operon encoding subunits of methane monooxygenase, whereas in non-SIP study 250,000 fosmid clones were screened to find pmoCAB operon (Uhlik et al. 2013).

Screening of Clones from Metagenomic Libraries

After obtaining the metagenomic library, screening of clones is done. Function-based screening also known as biological activity screening selects positive clones that express desirable characteristics. Specific phenotypes of the individual clones can be directly detected by using functional assays, by adding chemical-based dyes or chromophore-conjugated enzymes. New antibiotic resistance gene determinants (ARGD) can be investigated by functional analysis. A novel chloramphenicol-florfenicol-resistant gene was discovered by screening Alaska soil metagenomic clone library (Monier et al. 2011).

Banik and Brady (2010) reviewed the metagenomic approaches toward discovery of antimicrobials. In a work performed by Schmitz and coworkers, bacteriophage DNA was isolated from bat, guano, and earthworm guts; its functional screening led to the discovery of three new lysins capable of inhibiting *Bacillus anthracis* proliferation. Another function-based approach is the use of host strains requiring heterologous complementation by foreign genes for growth

under selective conditions. The recombinant clones that contain target gene and produce corresponding gene product in active form show optimum growth. This functional complementation was used to isolate lysine racemase (Lyr) gene from soil metagenome; in this *E. coli* BCRC 51,734 cells were used as the host and D-lysine as selection agent (Chen et al. 2009). The above approach faces certain problems including that of inaccurate transcription of target genes and assemblage problems of the corresponding enzymes. There is a scope of improvement in screening efficiency by enrichment of target microbes or use of screening sensitive substrate (Streit and Schmitz 2004).

Sequence-driven screening methods comprise of primers and probes of known conserved sequence that include phylogenetic or functional genes. Target clone is identified by PCR-based amplification or hybridization. PCR amplification of 30 genes encoding novel patellamide-like precursor peptide from *Prochloron sp.* symbionts living in consortia with marine sponges was reported by Schmidt and coworkers (Banik and Brady 2010). Fifteen new variants of the gene encoding precursor to the microviridin peptide were identified by Ziemert and coauthors in a PCR-based methodology. Homology-based screening is carried out mostly by using degenerate PCR primers, RT-PCR, DNA microarrays, integron, and affinity capture methods of sequence-based screening, as reported in literature (Xing et al. 2012). Relatively a new method for genetic screening is substrate-induced gene expression screening (SIGEX). These metabolism-related genes are selectively expressed in the presence of certain substrates. Chromatography-based screening techniques known as compound configuration screening are also reported. Clones are screened on their capability to produce new structural compounds depicting different chromatographic peaks relative to the host cells. Microarray-based GeoChip technology has been developed to access genetic and functional diversity of microbial community. Reactome array is a new sensitive metabolite array which offers functional analysis of metabolic phenotypes (Streit and Schmitz 2004).

Gene of interest can also be identified by random sequencing. Phylogeny can be linked with the functional gene by performing phylogenetic analysis with flanking DNA.

Metagenomic Sequencing

Gradual change has been experienced in the area of sequencing. Classical Sanger's sequencing technology is being proceeded by next-generation sequencing (NGS). Sanger method is preferred for its low error rate, long read length (>700 bp), and large insert sizes, but it has a drawback of being a labor-intensive process. Array-based sequencing and in vitro amplification of target DNA fragments constitute the second-generation DNA sequencing. Such technology is implemented in 454 Genome Sequencer, Illumina Genome Analyzer, and SOLiD platform (Xing et al. 2012). These next-generation approaches have the capacity for abundant parallel sequencing of samples. Pyrosequencing allows sequencing of 100–200 bp of single-stranded DNA and employs luciferase-based real-time monitoring of pyrophosphate release (Guazzaroni et al. 2009) and has high accuracy rates comparable to Sanger's sequencing.

Metagenomics employs two approaches: firstly, system-based approach, where complete sample of DNA is processed and analyzed. MG-RAST (Metagenomic Rapid Annotation Using Subsystem Technology) characterizes HTS pyrosequencing run (Larsen et al. 2012). Secondly, species identification-based approach involves the probability of potentially missing certain taxa in the process of PCR-based amplification of specific regions. One of the efficient methods of high-throughput analysis (HTS) of genes is based on microarrays; differential gene expression quantification of environmental bacterial diversity can be monitored (Cowan et al. 2005). Second-generation sequence technologies help in obtaining more information from complex microbial communities (Logares et al. 2012). Open reading frames and operons can be identified by analysis of longer contiguous sequences. Colony hybridization and pyrosequencing when combined with

metagenomic approach helped in gaining information about genetic organization and diversity of specific operon.

Addition of sample specific oligonucleotides barcode to PCR primers had an advantage of sequencing a number of samples simultaneously at a relatively reduced cost, also known as barcoding or multiplexing (Willner and Hugenholtz 2013). Third-generation sequencing is evolving fast. The first such technology became available was PacBio RS from Pacific Biosciences. This immobilized polymerase performs sequencing, and four differently colored nucleotides are detected in real time (Logares et al. 2012). Another innovative sequencing platform known as Ion Torrent is based on the principle that DNA polymerization releases protons which can help in the detection of nucleotide incorporation. Read length >100 bp can be obtained in the above technology. DNA nanoballs can be sequenced in a technology offered by Complete Genomics (Thomas et al. 2012).

Assembly, Binning, and Annotation

Recovering and characterization of genome of cultured organisms requires assembly of short-read fragments into longer genome contigs. Reference-based assembly method is applied, if closely related reference genomes are available. Large computational resources are required for de novo assembly (Thomas et al. 2012). A process based on sequence comparison of unknown DNA with reference databases, known as binning, helps to sort DNA sequences into groups representing genomes from closely related organisms. Metagenome sequence data is generally annotated by feature prediction and functional annotation. Feature prediction labels the sequences as gene, and functional annotation assigns taxonomic neighbors and putative gene function.

Data Handling and Statistical Analysis

Statistical approach aids metagenomics to link functional and phylogenetic information to the

biological, physical, and chemical parameters that fully characterize a microbial community. Multivariate statistical analysis is provided by various tools like Primer-E package. This package helps in the generation of multidimensional scaling (MDS) plots, analysis of similarities (ANOSIM), and species identification (SIMPER) (Thomas et al. 2012). A wide variety of bioinformatic tools and databases are available for metagenomic studies (Table 1).

Ecological Inferences

Community Studies

The ecological role of the microorganisms can be highlighted by conducting a genome-wide analysis. The ecosystem is highly dynamic in structure, and by employing shotgun metagenomics, direct sequencing of community DNA can be achieved. Metagenomics generate environmental microbial community data that helps in the investigation of microbial environmental interactome (MEI) (Larsen et al. 2012). PCR-based methods such as amplified ribosomal DNA restriction analysis (RISA), denaturing gradient gel electrophoresis (DGGE), and terminal restriction fragment length polymorphism (T-RFLP) have been used for the characterization of community microorganisms.

These techniques were applied to study the bacterial response in a pesticide contaminated soil (Imfeld and Vuilleumier 2012). Subsurface oil reservoirs with high pressure, salt, heavy metals, and organic solvent concentration have been analyzed by metagenomics. In another study, permafrost samples from the Canadian High Arctic and Alaska were investigated, in order to understand its potential linkage to global warming (Lewin et al. 2012). Microbial niche study was conducted on flowing acid mine drainage to determine the industrial community structure of a natural acidophilic biofilm growing on it (Streit and Schmitz 2004). ECOMIC-RMQS project is a French initiative to characterize soil microbial communities. Innovative studies and methodologies can determine organism's possible habitat in

Metagenomic Research: Methods and Ecological Applications, Table 1 Bioinformatic tools and databases commonly used in metagenomic studies

Name	Description	Website
ARB	Tools for sequence database handling and data analysis	www.arb-home.de
CAMERA	Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis	http://camera.calit2.net
CARMA	Characterizing short-read metagenomes	www.cebitec.uni-bielefeld.de/brf/carma/
COG	Clusters of Orthologous Groups	http://www.ncbi.nlm.nih.gov/COG/
DDBJ	DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/
DOTUR	Defining Operational Taxonomic Units and Estimating Species Richness	http://www.plantpath.wisc.edu/fac/joh/dotur.html
EMBL	European Molecular Biology Laboratory	www.embl.de/services/bioinformatics/index.php
GAAS	Genome relative Abundance and Average Size	http://sourceforge.net/projects/gaas/
GenBank	Genetic sequence database	www.ncbi.nlm.nih.gov/Genbank/metagenome.html
GOLD	Genomes Online Database	www.genomesonline.org
GSC	Genomic Standards Consortium	www.gensc.org
INSDC	International Nucleotide Sequence Database Collaboration	http://www.insdc.org/
IMG/M	Integrated Microbial Genomes	http://img.jgi.doe.gov/
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
LefSe	LDA Effect Size	http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload
MEGAN	MEtaGenome ANalyzer	www-ab.informatik.uni-tuebingen.de/software/megan
Megx.net	Marine Ecological GenomiX	www.megx.net
MetaPhlAn	Metagenomic Phylogenetic Analysis	http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload
GraPhlAn	Graphical Phylogenetic Analysis	http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload
METAREP	JCVI Metagenomics Reports	http://jvci.org/metarep/
PyNASt	Python Nearest Alignment Space Termination	www.qiime.org/pynast/
Naive Bayes classifier	Probabilistic classifier	http://www.statsoft.com/textbook/naive-bayes-classifier/
MG-RAST	Metagenomic RAST	http://metagenomics.nmpdr.org
PHACCS	Phage Communities from Contig Spectra	http://sourceforge.net/projects/phaccs/
RefSeq	Reference Sequence	http://www.ncbi.nlm.nih.gov/refseq/
ShotgunFunctionalizeR	R-package for functional comparison of metagenomes	http://shotgun.zool.gu.se
SILVA	Comprehensive online ribosomal RNA sequence database	www.arb-silva.de
SINA	Bioinformatic tools for sequence alignment	www.arb-silva.de
SmashCommunity	Stand-alone metagenomic annotation and analysis pipeline	http://www.bork.embl.de/software/smash/
Sort-ITEMS	Sequence orthology-based approach for improved taxonomic estimation of metagenomic sequences	http://metagenomics.atc.tcs.com/binning/SORT-ITEMS/
STAMP	Statistical Analysis of Metagenomic Profiles	http://kiwi.cs.dal.ca/Software/STAMP

(continued)

Metagenomic Research: Methods and Ecological Applications, Table 1 (continued)

Name	Description	Website
TACOA	Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach	http://www.cebitec.uni-bielefeld.de/brf/tacoa/tacoa.html
TETRA	Fragment assignment by intrinsic tetranucleotide frequencies	www.megx.net/tetra
Treephyler	Fast taxonomic profiling of metagenomes	http://www.gobics.de/fabian/treephyler.php
Fast UniFrac	Comparison of microbial communities	http://bmf2.colorado.edu/fastunifrac
XplorSeq	Mac OSX software for sequence analysis	www.phyloware.com/Phyloware/XplorSeq.html
Xipe	Statistical comparison program	http://edwards.sdsu.edu/cgi-bin/xipe.cgi

multidimensional space. Microbial assemblage prediction (MAP), a bioclimatic tool, helps in modeling relative abundance of microbial taxa as a function of environmental parameters (Larsen et al. 2012). SIP-metagenomics approach can be employed in the identification of microbial species degrading xenobiotic compounds.

Bioprospecting

New thermostable and thermolabile biocatalyst can be discovered from extreme ecological communities. High-temperature metagenomes of virus recently gave a new thermostable DNA polymerase with reverse transcriptase activity for RT-PCR (Lewin et al. 2012). Variety of enzymes has been isolated using metagenomics (Table 2). By using metagenomics, scientists were able to identify many genes playing a role in various processes including cell cycle, metabolism, DNA repair, transcriptional regulation, etc. (Sharma et al. 2005).

A novel cold-active xylanase gene was isolated from the community DNA of goat's rumen contents. Human gut metagenomic library was subjected to high-throughput screening; 310 clones were isolated showing various enzyme activities (Xing et al. 2012). Novel antibiotics were successfully achieved through metagenomics, e.g., indirubin, deoxyviolacein, and violacein (Ghazanfar et al. 2010).

Metagenomic Research: Methods and Ecological Applications, Table 2 Enzymes/biocatalysts isolated using metagenomic approaches

Name of enzymes		
Agarase	DNA polymerase	Nitrile hydratase
Alcohol oxidoreductase	Endoglucanase	Nuclease
Alkane hydroxylase	Exoglucanase	Pectinase
Amidase	Esterase	Phytase
Amylase	b-Glucosidase	Polyketide synthase
Cellulase	Glucoamylase	Protease
Chitinase	Laccase	Rhamnosidase
Decarboxylase	Lipase	Single-stranded DNA ligase
4-Hydroxybutyrate dehydrogenase	Mannanase	Xylanase
Dehydratase	Nitrilase	b-Lactamase

Compiled from Cowan et al. 2005; Singh et al. 2009; Ghazanfar et al. 2010; Xing et al. 2012

Clinical Metagenomics

In an initiative by the National Institute of Health (NIH), the Human Microbiome Project is being undertaken to characterize microbial community present at various sites in the human body. The main objective of this project is to study these microbes in healthy and diseased state of the human body. Nelson et al. (2010) sequenced 178 microbial genomes present at multiple body sites, and further novel predicted polypeptides

were identified. Crohn's disease patient's gut metagenome revealed a characteristic disease-associated microbiota. By using metagenomics, healthy human virome can be characterized and infectious diseases with unknown etiology can be diagnosed. Novel viruses such as cardiovirus and klassevirus have been reported in viral metagenome of human fecal samples. For fungal diversity analysis, nuclear ribosomal internal transcribed spacer region (ITS) is employed. Mycobiome was prepared from oral rinse samples to study fungal species diversity. High-throughput sequencing of fungal metagenome was applied to the samples from patients with cystic fibrosis for identifying new species. Community profiling using HTS provides new insights in the area of clinical microbiology (Willner and Hugenholz 2013).

Summary

Metagenomic approach is a repertoire of huge genetic information. DNA/RNA from numerous ecosystems are sampled, extracted, and processed. Functional- and sequence-based screening of metagenomic libraries has helped in establishing phylogenetic relationships among the communities. It has opened a new era of discovery of novel genes and microbial interaction-based studies. Innovative metagenomic sequencing efforts will be essential to resolve the complexity involved in various microbiomes. It is important to share and critically apply outcomes of metagenomic research. With the advent in the areas of metagenome library construction, screening methodology, and enhanced gene expression, metagenomics can evolve as a significant technology in microbial diversity analysis.

Cross-Reference

- ▶ [A 123 of Metagenomics](#)
- ▶ [Biological Treasure Metagenome](#)
- ▶ [Microbial Ecology in the Age of Metagenomics: An Introduction](#)
- ▶ [Next-Generation Sequencing for Metagenomic Data: Assembling and Binning](#)

References

- Banik JJ, Brady SF. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr Opin Microbiol.* 2010;13:603–9.
- Chen IC, Lin WD, Hsu SK, et al. Isolation and characterization of a novel lysine racemase from a soil metagenomic library. *Appl Environ Microbiol.* 2009;75:5161–6.
- Cowan D, Meyer Q, Stafford W, et al. Metagenomic gene discovery: past, present and future. *Trends Biotechnol.* 2005;23:321–9.
- Delmont TO, Robe P, Clark I, et al. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods.* 2011;86:397–400.
- Ghazanfar S, Azim A, Ghazanfar MA, et al. Metagenomics and its application in soil microbial community studies: biotechnological prospects. *J Anim Plant Sci.* 2010;6:611–22.
- Guazzaroni ME, Beloqui A, Golyshin PN, et al. Metagenomics as a new technological tool to gain scientific knowledge. *World J Microbiol Biotechnol.* 2009;25:945–54.
- Hirsch PR, Mauchline TH, Clark IM. Culture-independent molecular techniques for soil microbial ecology. *Soil Biol Biochem.* 2010;42:878–87.
- Imfeld G, Vuilleumier S. Measuring the effects of pesticides on bacterial communities in soil: a critical review. *Eur J Soil Biol.* 2012;49:22–30.
- Kakirde KS, Parsley LC, Liles MR. Size does matter: application-driven approaches for soil metagenomics. *Soil Biol Biochem.* 2010;42:1911–23.
- Keshri J, Mishra A, Jha B. Microbial population index and community structure in saline-alkaline soil using gene targeted metagenomics. *Microbiol Res.* 2013;168:165–73.
- Larsen P, Hamada Y, Gilberta J. Modeling microbial communities: current, developing, and future technologies for predicting microbial community interaction. *J Biotechnol.* 2012;160:17–24.
- Lewin A, Wentzel A, Valla S. Metagenomics of microbial life in extreme temperature environments. *Curr Opin Biotechnol.* 2012;24:1–10.
- Logares R, Haverkamp THA, Kumar S, et al. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J Microbiol Methods.* 2012;91:106–13.
- Mocali S, Benedetti A. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Res Microbiol.* 2010;161:497–505.
- Monier JM, Demaneche S, Delmont TO, et al. Metagenomic exploration of antibiotic resistance in soil. *Curr Opin Microbiol.* 2011;14:229–35.
- Nelson KE, Weinstock GM, Highlander SK, et al. A catalog of reference genomes from the human microbiome. *Science.* 2010;328:994–9.

- Sharma R, Ranjan R, Kapardar RK, et al. Unculturable bacterial diversity: an untapped resource. *Curr Sci.* 2005;89:72–7.
- Singh J, Behal A, Singla N, et al. Metagenomics: concept, methodology, ecological inference and recent advances. *Biotechnol J.* 2009;4:480–94.
- Streit WR, Schmitz RA. Metagenomics - the key to the uncultured microbes. *Curr Opin Microbiol.* 2004;7:492–8.
- Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microbiol Inform Exp.* 2012;2:3.
- Uhlik O, Leewis MC, Strejcek M, et al. Stable isotope probing in the metagenomics era: a bridge towards improved bioremediation. *Biotechnol Adv.* 2013;31:154–65.
- Vasconcellos SP, Angolini CFF, García INS, et al. Screening for hydrocarbon biodegraders in a metagenomic clone library derived from Brazilian petroleum reservoirs. *Org Geochem.* 2010;41:675–81.
- Willner D, Hugenholtz P. Metagenomics and community profiling: culture-independent techniques in the clinical laboratory. *Clin Microbiol Newsl.* 2013;35:1–9.
- Xing MN, Zhang XZ, Huang H. Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. *Biotechnol Adv.* 2012;30:920–9.
- Yilmaz P, Kottmanna R, Pruesse E, et al. Analysis of 23S rRNA genes in metagenomes – a case study from the global ocean sampling expedition. *Syst Appl Microbiol.* 2011;34:462–9.

Metagenomics Potential for Bioremediation

Terrence H. Bell¹, Charles W. Greer² and Etienne Yergeau²

¹Department of Natural Resource Sciences, McGill University, Sainte-Anne-de-Bellevue, QC, Canada

²National Research Council Canada, Montreal, QC, Canada

Synonyms

Metagenomics of polluted substrates/environments

Definition

Bioremediation refers to the detoxification of environments through the activities of living

organisms. In many environments, microorganisms are the main agents of bioremediation, as they adapt their existing biochemical pathways to the degradation or conversion of pollutants. Human intervention can often improve the ability of microorganisms to rapidly remediate contaminants, but how treatments affect species diversity and gene allocation in complex microbial communities is not well characterized. The metagenome of a contaminated environment includes all DNA contained within it; however, a variety of screening methods can be used in bioremediation studies to simplify the collection and analysis of targeted genomic information.

Introduction

Pollution is a ubiquitous global concern, as many natural and synthetic compounds have been introduced into environments in which they are posing hazards to the health of humans and ecosystems. Bioremediation is the degradation, conversion, or stabilization of these compounds by organisms, generally performed by microorganisms and plants. When the organisms that are native to a contaminated site effectively remove contaminants without intervention, the toxicity at the site may simply be monitored as the pollutant is reduced or converted to a less toxic form. In many cases, however, intervention can increase the rate of bioremediation. The addition of stimulating amendments on site (e.g., nutrients, organic matter) and the relocation of contaminated material to off-site treatment facilities are the most common approaches to encouraging remediation.

Often it is microorganisms that play the most significant role in bioremediation. High-resolution genetic information is required to understand how contaminants and treatments affect the complex microbial communities that exist in natural environments. Some taxonomic groups have been linked to the presence of various pollutants, but many of the taxa and enzymes that can potentially participate in bioremediation remain unknown. Thousands of microbial species may exist in a single gram of soil, so when pollutants are similar in composition to compounds

that occur naturally in the environment, a large number of species are able to compete to use the pollutant as a source of carbon, nutrients, or energy. At the other extreme, when the introduced pollutant is complex or synthetic in origin, there may be no local strains that are immediately capable of metabolizing it or reducing its toxicity.

A number of bioremediating microorganisms have been isolated from contaminated sites, but it is now generally understood that the information obtained from these isolates is insufficient to understand the workings of complex microbial communities. More complete genetic information from natural environments is required to understand how contamination affects microbial communities on the whole, and whether there is the potential for further optimization of bioremediation. The large-scale, culture-independent studies that are required to meet this end are now possible with the advent of new high-throughput sequencing technologies.

Aspirations for Metagenomics in Bioremediation

Understanding the differences between a contaminated environment and its uncontaminated equivalent is a major topic of study in bioremediation research, as it can help in determining how much of the natural function of the system has been altered by contamination. Metagenomic data can provide information about taxonomic and enzymatic diversity both pre- and post-contamination, which will allow the mining of potentially active genes and organisms. Accumulating metagenomes from a variety of contaminated and uncontaminated equivalent environments will make it possible to link changes in contaminant composition and concentration to specific genes and taxa. In addition, such studies will answer questions about the microbial ecology of the contaminated system, specifically how microorganisms respond to the disturbance created by the contaminant. Adjustments of nutrients, carbon sources, pH, temperature, oxygen, and water content are frequently

parts of treatment scenarios applied to contaminated sites, so metagenomic studies of bioremediation will also provide information on how microbial communities respond to changes in a variety of environmental factors. To date, only a handful of such studies have been conducted (Table 1).

Types of Metagenomic Studies Used in Bioremediation

Strictly speaking, metagenomics involves the entirety of genetic information contained within a sample. More efficient sequencing now makes it possible to produce this data, but the effort required to thoroughly analyze such huge datasets is a limiting step in metagenomic studies. Even when full metagenomes are sequenced, analysis of the data will often focus on specific genes of interest. There is also a trade-off between the number of samples analyzed and the depth of sequencing possible. While it is tempting to completely sequence and annotate single samples, it is difficult to know how representative this sample is of an entire environment or in the case of composite samples, the variability that exists within the environment.

As a compromise, many studies of contaminated sites have used what has been referred to as gene-targeted metagenomics (Iwai et al. 2010), in which specific gene regions are amplified and then sequenced using high-throughput technologies. This has been used in bioremediation studies to look at specific functional genes (Bell et al. 2011; Iwai et al. 2010) as well as 16S rRNA gene diversity (e.g., Bell et al. 2011; Gihring et al. 2011). The limitations of gene-targeted metagenomics are that (1) genetic information that is not immediately of interest cannot be explored in the future, (2) novel genes that cannot be amplified by the selected primers are excluded from the analysis, and (3) information about the relative occurrence of the targeted genes within the sample will be lost.

Several recent reports have incorporated some type of metagenomics into the study of the

Metagenomics Potential for Bioremediation, Table 1 Studies that have used metagenomics to study microbial populations in contaminated substrates

Substrate	Contaminant	Treatment	Gene groups examined	Key finding	Sequencing type	References
Whole genome sequencing						
Groundwater	Heavy metals, nitrate, organic solvents	None	16S rRNA, metabolism, stress response	Significant loss of species and metabolic diversity following more than 50 years of contamination	PRISM 3730 capillary DNA sequencer	Hemme et al. 2010
Soil	Diesel	Monoammonium phosphate and aeration	16S rRNA, alkyl group hydroxylases, extradiol dioxygenase, intradiol dioxygenase, gentisate/homogentisate dioxygenase	Shift from <i>Gammaproteobacteria</i> to <i>Alphaproteobacteria</i> and <i>Actinobacteria</i> after 1 year of remediation	Roche/454 GS FLX Titanium	Yergeau et al. 2012
Gene-targeted sequencing						
Soil	JP-8 jet fuel	Monoammonium phosphate	16S rRNA, <i>alkB</i>	<i>Alphaproteobacteria</i> in contaminated soils were more effective at incorporating added nitrogen than were other bacterial taxa	Roche/454 GS FLX Titanium	Bell et al. 2011
Rhizosphere soil	PCB	None	Toluene/biphenyl dioxygenases	Unexpected gene diversity, including 25 novel clusters	Roche/454 FLX	Iwai et al. 2010
Subsurface sediment	Uranium (VI)	Ethanol injection	16S rRNA	Identified indicator taxa specific to various hydrochemical conditions and those that responded to treatment	Roche/454 FLX	Cardenas et al. 2010
Mangrove sediment	MF380 heavy fuel oil	None	16S rRNA	Wide diversity in both contaminated and uncontaminated sediment, with indicator taxa detected for each	Roche/454 FLX	dos Santos et al. 2011
Groundwater	Uranium, sulfate, nitrate	Emulsified vegetable oil	16S rRNA	Very narrow group of microorganisms that were stimulated by the treatment and/or involved in remediation	Roche/454 FLX	Gihring et al. 2011
Liquid media	Synthetic aromatic alkanolic acids	Added individual alkanolic acids	16S rRNA	Microbial community was unique to the contaminant added, which varied in alkyl side branching	Roche/454	Johnson et al. 2011

(continued)

Metagenomics Potential for Bioremediation, Table 1 (continued)

Substrate	Contaminant	Treatment	Gene groups examined	Key finding	Sequencing type	References
<i>Functional screening</i>						
Soil	Aliphatic and aromatic hydrocarbons	Air sparging	Extradiol dioxygenase	High diversity of extradiol dioxygenase genes in contaminated soil; one extradiol dioxygenase gene found per 3.6 Mb of DNA	ABI PRISM 3100 Genetic Analyzer	Brennerova et al. 2009
Activated sludge	Various aromatic compounds	None	Extradiol dioxygenase	Identified novel arrangements of the extradiol dioxygenase degradation pathway on plasmid-like DNA	ABI 3730xl DNA Analyzer	Suenaga et al. 2009

microorganisms living in contaminated environments. Since the labor required to process data is beginning to outweigh the cost of sequencing as the limiting step in metagenomic analyses, a variety of screening methods have been used in bioremediation studies to optimize the output of information (Fig. 1). The various approaches to metagenomics that have been taken in bioremediation research are outlined below.

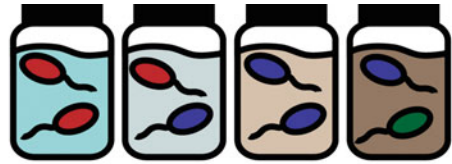
Multiplexed 16S rRNA Gene Sequencing

Because of its potential to quickly assign taxonomy to large numbers of microorganisms, 16S rRNA gene sequencing has gone through several waves of popularity in microbial ecology. Comparisons of the 16S rRNA gene profiles of environmental samples have taken off again with the advent of high-throughput sequencing (Tringe and Hugenholtz 2008) and are currently more popular than any other type of metagenomic study. One reason is that a large number of 16S rRNA gene entries exist in NCBI and EMBL, as do curated 16S rRNA gene databases such as the Ribosomal Database Project (<http://rdp.cme.msu.edu/>) and Green Genes (<http://greengenes.lbl.gov/>). As a result, profiles of community diversity can be conducted with only a cursory understanding of bioinformatics. While early techniques such as T-RFLP and DGGE gave some indication

of the variation between samples, they only described small portions of microbial communities. Even clone library studies rarely sampled more than a few hundred clones, whereas multiplexed next-generation sequencing easily provides several thousand sequences per sample.

Since studies into bioremediation generally aim to identify effective pathways for converting or tolerating contaminants, how relevant is taxonomy? There is still debate surrounding how much functional redundancy exists between microbial species and how prevalent horizontal gene transfer (HGT) is within microbial communities, yet a recent metagenomic study shows that distinct bacterial species likely do exist (Caro-Quintero and Konstantinidis 2012). A number of 16S rRNA gene surveys have been conducted in contaminated environments and have been used to assess how microbial communities vary in relation to uncontaminated reference environments or how a community changes in a contaminated environment over time. In several of these studies, 16S rRNA gene-targeted metagenomics has identified indicator species that are specific to certain contaminants and environmental conditions (Cardenas et al. 2010; dos Santos et al. 2011). Similar multiplexed studies may be used to identify indicator species across multiple environments at similar stages of

Contaminated substrate under various bioremediation treatments



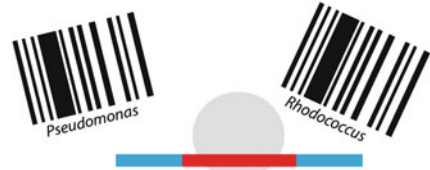
1. Multiplexed 16S rRNA Sequencing

Advantages

- Can process many samples at low cost

Disadvantages

- Horizontal gene transfer can make it difficult to pin functions to specific taxonomic groups
- Primers bias against certain groups



2. Multiplexed Gene-Targeted Sequencing

Advantages

- Can process many samples at low cost

Disadvantages

- Primer bias against unknown sequences



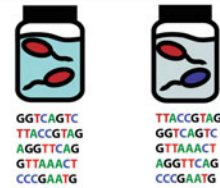
3. Screen Substrates for Bioremediation Capacity

Advantages

- Eliminates samples that are less effective at remediation
- Allows greater sequencing depth for money available

Disadvantages

- Omits potentially interesting information from less effective substrates



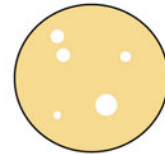
4. Screen DNA in Plasmids for Function

Advantages

- Eliminates large stretches of DNA that are not involved in a specific process

Disadvantages

- Requires that an entire pathway exist in one plasmid
- Some gene products may be toxic to the host



5. Mixed Culturing *in vitro*

Advantages

- Population usually enriched in effective bioremediating microorganisms

Disadvantages

- Does not represent natural system
- Limited information on true ecology



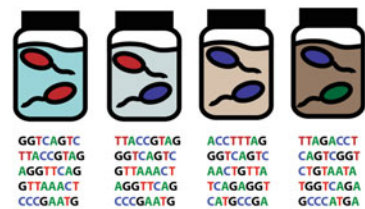
6. Full Metagenome Analysis of All Samples

Advantages

- Most information possible
- Can perform any number of post hoc analyses
- Data available for future research

Disadvantages

- High cost
- Comprehensive databases not available for many genes
- A lot of data is ignored in the immediate study



Metagenomics Potential for Bioremediation, Fig. 1 Methods for integrating metagenomics into bioremediation studies

contamination, and these indicator species could theoretically be used to assess the state of other contaminated sites.

The major advantage of the high-throughput sequencing approach when compared with earlier 16S rRNA gene profiling techniques is the depth of coverage. In mangrove sediment contaminated with heavy fuel, little change was seen at the phylum level following contamination, while large shifts were observed at finer taxonomic levels (dos Santos et al. 2011), an effect that may not have been visible using coarser profiling methods. Similarly, 16S rRNA gene pyrosequencing showed that a very narrow group of taxa were stimulated by emulsified oil injection in a uranium-contaminated aquifer (Gihring et al. 2011). With less sequencing coverage, it would be impossible to determine whether these were the only taxa stimulated or simply the most dominant members of the community.

Multiplexed Functional Gene Sequencing

In many bioremediation studies, specific catabolic, reducing, or oxidizing genes are the subjects of interest. In such cases, it may be desirable to simply amplify and sequence these targeted genes. As with 16S rRNA gene sequencing, many samples can be processed by multiplex sequencing for a limited cost. Degenerate primers have been used to amplify alkane monooxygenase genes from hydrocarbon-contaminated Arctic soil, and sequencing showed that those related to Alphaproteobacteria responded most positively to amendment with monoammonium phosphate (Bell et al. 2011). Amplicons were also obtained from a PCB-contaminated soil using degenerate primers targeting toluene/biphenyl dioxygenase genes, and sequencing identified a variety of novel dioxygenase gene clusters (Iwai et al. 2010). In terms of gene discovery, the major drawback of this approach is that gene identification depends on novel genes having significant homology at the primer-targeted regions. Even when the targeted genes are known, the chosen primers will bias the relative gene abundance within each sample. Amplicon size must also be

considered, since many sequencing technologies have a maximum read length, although with time, this is becoming less of a concern.

Functional Screening

Since bioremediation is generally focused on which microbial communities most effectively degrade pollutants, it can potentially be straightforward to functionally screen for samples of interest. A study of contaminated Arctic soils compared the hydrocarbon-degrading efficiency of various soils in response to different in situ and ex situ treatments, with degradation occurring significantly more effectively in one location. Subsequently, a metagenomic analysis was conducted throughout a year-long time course on the soil that most rapidly degraded the contaminating hydrocarbons, along with an uncontaminated reference soil (Yergeau et al. 2012). Metagenomic studies that are conducted in vitro also involve an aspect of selection, as only microorganisms that are capable of growing in mixed culture prevail. Mixed culture studies are common, as they often evaluate the potential for bioremediation in treatment facilities. Metagenomics is starting to be applied to such studies, as in one case in which it was determined that the amount of branching in synthetic aromatic alkanolic acids led to vastly different microbial communities (Johnson et al. 2011).

Prescreening of DNA can also be conducted on large genomic fragments that are contained within plasmids, such as fosmids or cosmids. By transforming these vectors into hosts such as *E. coli*, the DNA fragment can be screened for the ability to mineralize or tolerate a specific contaminant. This strategy permits the identification of genes that are involved in the catabolism of particular pollutants, or that permit host survival, provided the essential pathway can be contained in a single DNA fragment and can be expressed in the host. Sequencing is also more targeted using this approach, as the sequencing of housekeeping and rRNA genes is limited.

To search for genes capable of degrading catechol, metagenomic DNA from a hydrocarbon-contaminated soil was fragmented, cloned into fosmid vectors, transformed into *E. coli*, and plated with catechol as a carbon substrate. A high

diversity of extradiol dioxygenase genes was observed, as well as a surprisingly high density of one extradiol dioxygenase per 3.6 Mb of DNA screened (Brennerova et al. 2009). A similar approach identified novel extradiol dioxygenase genes, as well as previously unknown arrangements of catechol-degrading pathways (Suenaga et al. 2009). The drawbacks of this approach are that the entire genetic pathway must be contained within a single plasmid; that the host may be unable to survive in the presence of any toxic gene products, meaning that not all relevant genes will necessarily be identified; and that some genes may not be expressed if the chosen host is not closely related to the organism from which the DNA fragment originated.

Full Metagenome Analysis

Full metagenomic sequencing, when possible, provides the greatest amount of information. With this approach, any number of post hoc analyses can be conducted on a dataset. While much of the genetic information obtained from a given environment may lack appropriate comparators in existing gene banks, collecting full metagenomic information will allow future researchers the opportunity to analyze the dataset. At the moment, a number of database projects are ongoing in an attempt to collect and annotate metagenomic data, including some from contaminated sites (e.g., <http://www.hydrocarbonmetagenomics.com/>).

To date, only a handful of complete metagenomic studies have been conducted in contaminated environments. While 16S rRNA gene studies are useful in determining the relative microbial diversity of environments, the metabolic potential of a microbial community may not be strictly linked to its taxonomic profile. Thus, full metagenomic studies can be used to assess how diversity relates to functional potential. A metagenomic study of a diesel-contaminated Arctic soil showed that a shift in 16S rRNA gene sequences from Gammaproteobacteria to Alphaproteobacteria and Actinobacteria mostly correlated with a shift in hydroxylases and dioxygenases that were affiliated with those same organisms

(Yergeau et al. 2012), demonstrating that, in this case, there was significance to taxonomic affiliation. Similarly, most of the functional genes (stress response, metal resistance, etc.) identified in the metagenome of a heavy metal-contaminated groundwater community were traced to Gammaproteobacteria, the group that also dominated the 16S rRNA gene profile (Hemme et al. 2010).

Full metagenomes can also provide information on the relative abundance of genes of interest. PCR-based approaches introduce a primer bias prior to sequencing, whereas strict metagenomic analysis permits a more direct quantitative comparison. Within the contaminated groundwater metagenome, stress-response genes, such as those involved in DNA repair and heavy metal resistance, were more abundantly represented than would be expected in an uncontaminated community (Hemme et al. 2010). Most hydrocarbon-degrading genes were high in abundance in the contaminated Arctic soil metagenomes when compared with the uncontaminated reference soil, but extradiol aromatic ring-cleavage dioxygenase sequences decreased after a year of treatment, while other dioxygenases increased in abundance, and alkane hydroxylases remained constant throughout treatment (Yergeau et al. 2012). Caution should be exercised when using preparatory techniques such as whole genome amplification, since the quantitation of genes can be affected (Yergeau et al. 2010). Although the amount of DNA required for metagenomic sequencing is decreasing, whole genome amplification may still be necessary in very low biomass systems, as can be found in some highly contaminated environments.

Information Lacking from Bioremediation Literature

Genes Involved in Bioremediation

Key pathways involved in the bioremediation of major contaminants are known, but many novel enzymes and pathways are still being discovered. The lack of sequence conservation in some key gene families has made it difficult to determine their true diversity using PCR-based methods.

In the case of genes that code for enzymes that are involved in normal forms of metabolism or other housekeeping functions within the cell, this diversity may be extensive. Metagenomic studies across contaminated environments will help correlate gene groups with contaminants, and this may identify roles for pathways that had previously been considered unimportant in the conversion or tolerance of contaminants.

Microbial species that are not directly involved in bioremediation can also represent a sizeable proportion of a contaminated community. Soils contaminated with hydrocarbons have still provided homes for populations of nitrifying bacteria (Deni and Penninckx 1999) and cyanobacteria (Yergeau et al. 2012), while the stimulation of the microbial reductive chlorination of PCE and TCE by adding organic products tends to promote many microorganisms that are not involved in remediation (Strycharz et al. 2008). In addition, microorganisms that function in various nutrient cycles (e.g., nitrogen fixers) may be important to the functioning of the overall community. To date, it is not really known how much these other species affect functioning in contaminated environments or how bioremediation is affected if some processes are disrupted.

Extent of Horizontal Gene Transfer

It can be difficult to determine the taxonomic affiliation of plasmid-borne DNA, and certain key genes involved in bioremediation, such as naphthalene dioxygenases and alkane monooxygenases (Whyte et al. 1997), have been found on plasmids. Mobile genetic elements are known to be common in at least some natural environments, but it is not known how significant a role HGT plays in the adaptation of microbial communities to contamination.

In metagenomic studies, genes can be compared with the background DNA of the community metagenome, which can help in identifying the prevalence of HGT. Bioinformatic analysis of a metagenome under long-term contamination showed that roughly 12 transposons were present per Mb of DNA, which was similar to reference strains of *Xanthomonas*, the dominant

community member. In addition, large differences in % G+C and codon bias between putatively transposed genes suggested a very recent origin for acetone carboxylases, mercuric resistance operons, and *czcD* divalent cation transporters (Hemme et al. 2010). The persistence of HGT after 50 years of continued contaminant stress suggests that it may be very important to the survival of microorganisms in a contaminated environment.

Horizontal gene transfer was also suspected when a mismatch between the number of cytochrome P450 genes affiliated with *Rhodococcus* and the relative abundance of Actinobacteria was observed in the metagenome of diesel-contaminated Arctic soils (Yergeau et al. 2012). A number of the genes detected in this study can be plasmid-borne, so this may be a common response. Future metagenomic analyses pre- and post-contamination may show how quickly this process can shape the genetic structure of microbial communities. If HGT is determined to be a major force shaping newly contaminated environments, the metagenomic screening of mobile elements alone may be another method of eliminating large amounts of housekeeping and redundant genetic information.

Quantitation

As mentioned, metagenomes that have not been modified by processes such as whole genome amplification may permit actual quantification of gene abundances. Whereas techniques such as qPCR and PCR-based diversity studies are subject to amplification biases, the metagenome represents all of the genetic information that could be extracted from a sample. Most previous attempts to quantify microbial allocation of gene resources to important processes in contaminated sites have relied heavily on PCR methods.

Some early metagenomic studies have already shown the potential of quantitation. The relative genomic allocation to the degradation of various components of jet fuel, a complex contaminant, was observed in a contaminated soil community. It was also observed that known hydrocarbon-degrading genes represented a disproportionate amount of the total metagenome

(Brennerova et al. 2009). An overabundance of genes conferring resistance to heavy metals, nitrate, and organic solvents was observed in a heavy metal-contaminated aquifer (Hemme et al. 2010). Semiquantitative approaches have also been used to determine relative shifts in species abundance and nitrogen incorporation in contaminated environments (Bell et al. 2011; Cardenas et al. 2010), and future studies using full metagenome analysis would permit actual quantification.

The Future of Metagenomics in Bioremediation

Technologies that facilitate metagenomic research are advancing quickly, and many studies that had previously been outside the realm of consideration are becoming possible. Companies such as PacBio and Nanopore are producing sequencers that will allow Kb reads of DNA, which will make it possible to assemble continuous genomes in mixed communities. Even with current technologies this is becoming feasible, as the entire draft genome of a novel permafrost methanogen was assembled by end-to-end linking of 113 bp paired-end reads that were produced in a metagenomic study using Illumina GAII technology (Mackelprang et al. 2011).

The combination of various high-throughput techniques will enable comprehensive studies of microbial communities and shed light on the links between species diversity, gene density, gene expression, protein production, and chemical transformation in contaminated environments. Stable isotope probing (SIP) is a technique that involves adding heavy isotope-labeled compounds to a substrate and allowing microorganisms to consume it and incorporate the labeled atoms into cellular components such as DNA, RNA, and phospholipids. In the case of DNA-SIP, all DNA from a treated sample is extracted and then centrifuged in CsCl gradients to separate the “heavy” (labeled) from the “light” (unlabeled) DNA. This technique has great potential in terms of identifying functionally active microbes, specifically those involved in

contaminant breakdown, and a recent review describes the potential power of combining SIP with metagenomics (Chen and Murrell 2010). SIP-metagenomic analyses of contaminated substrates allow the genes and species that actively respond to pollutants to be separated from the huge amount of background genetic information that may remain from the initial, uncontaminated soil. The link between taxonomic affiliation and community function is already being explored through the combination of SIP and high-throughput sequencing (Bell et al. 2011), while advances in RNA-SIP will provide a comprehensive picture of how the addition of substrates, whether contaminants or amendments, directly affects transcription. At the moment, the CsCl gradients that are required to separate labeled and unlabeled nucleic acids are extremely cumbersome and limit the number of samples that can be processed within a given study.

However, a novel proteomic-SIP technique, using 2-dimensional liquid chromatography-tandem mass spectrometry (2D-LC-MS/MS), was able to examine the isotopic ratios of roughly 100,000 spectra while simultaneously searching a database of 31,966 protein sequences in under 24 h (Pan et al. 2011). The computing power required to conduct the analysis was enormous, but as with all high-throughput processing, this can be expected to change rapidly with time. The potential for applying the proteomic-SIP technique in bioremediation studies is enormous, as even small numbers of proteins produced by rare microorganisms can be tracked (Pan et al. 2011). This will be especially useful in examining bioremediation pathways that involve syntrophic interactions, or those involved in the processing of slowly degraded contaminants, in which nutrient flux and subsequent protein production are bound to be low.

In contaminated environments, metagenomics has been used to compare polluted substrates with uncontaminated reference substrates (e.g., Yergeau et al. 2012) and has also been used to directly measure species composition within the same matrix before and after contamination (dos Santos et al. 2011). These types of comparative studies are geared at understanding what genetic

information distinguishes a contaminated environment from similar pristine systems. One of the next major efforts in metagenomics is likely to be the identification of a core microbiome (Shade and Handelsman 2012). In other words, what genes and species are common across an environment and across multiple environments. With a more comprehensive idea of what core microbiomes exist, environments may be aligned by their conserved regions, such as sequences are now, and the true variability between environments can then be assessed. In the context of bioremediation, it will be important to understand whether there are critical genes and organisms that must respond positively to the introduction of a contaminant in order to achieve successful remediation. Genes promoted outside of this common core must then be the result of other environmental or stochastic processes.

Many current genomic studies focus on snapshots of genetic information in environmental samples, but the high growth rate of microorganisms means that many microbial communities are undergoing constant and rapid evolution. This suggests that longer-term metagenomic studies should be a focal point of future research. The metagenomic study by Hemme et al. (2010) of metal-contaminated groundwater showed that 50 years of pollutant stress had reduced species and metabolic diversity to a minimal level of complexity. While all necessary metabolic pathways were found, more than ten times fewer OTUs, with a similar loss in metabolic complexity, were present than were observed at an adjacent background site. Monitoring how evolution selects genes in contaminated environments over the long term will undoubtedly assist in the understanding and treatment of chronically contaminated sites, although the interpretation of large amounts of data will first require a solution to the human-processing bottleneck.

Summary

A variety of metagenomic approaches are available to bioremediation researchers. The choice of technique will depend heavily on the question that

is being asked, as well as the resources that are available. While full metagenomic studies provide the greatest amount of data per sample, surveying for indicator species or gene diversity across a wide range of samples may be more appropriate in many cases. These methods may change quickly as technology continues to improve, but ultimately, the best approaches will be those that answer questions about how to most efficiently improve the bioremediation of contaminated sites.

References

- Bell TH, Yergeau E, Martineau C, et al. Identification of nitrogen-incorporating bacteria in petroleum-contaminated Arctic soils by using [(15)N]DNA-based stable isotope probing and pyrosequencing. *Appl Environ Microb*. 2011;77:4163–71.
- Brennerova MV, Josefiova J, Brenner V, et al. Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ Microbiol*. 2009;11:2216–27.
- Cardenas E, Wu WM, Leigh MB, et al. Significant association between sulfate-reducing bacteria and uranium-reducing microbial communities as revealed by a combined massively parallel sequencing-indicator species approach. *Appl Environ Microb*. 2010;76:6778–86.
- Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. *Environ Microbiol*. 2012;14:347–55.
- Chen Y, Murrell JC. When metagenomics meets stable-isotope probing: progress and perspectives. *Trends Microbiol*. 2010;18:157–63.
- Deni J, Penninckx MJ. Nitrification and autotrophic nitrifying bacteria in a hydrocarbon-polluted soil. *Appl Environ Microb*. 1999;65:4008–13.
- dos Santos HF, Cury JC, do Carmo FL, et al. Mangrove bacterial diversity and the impact of oil contamination revealed by pyrosequencing: bacterial proxies for oil pollution. *PLoS One*. 2011;6:e16943.
- Gihring TM, Zhang GX, Brandt CC, et al. A limited microbial consortium is responsible for extended bioreduction of uranium in a contaminated aquifer. *Appl Environ Microb*. 2011;77:5955–65.
- Hemme CL, Deng Y, Gentry TJ, et al. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J*. 2010;4:660–72.
- Iwai S, Chai BL, Sul WJ, et al. Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. *ISME J*. 2010;4:279–85.

- Johnson RJ, Smith BE, Sutton PA, et al. Microbial biodegradation of aromatic alkanolic naphthenic acids is affected by the degree of alkyl side chain branching. *ISME J.* 2011;5:486–96.
- Mackelprang R, Waldrop MP, DeAngelis KM, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature.* 2011;480:368–71.
- Pan CL, Fischer CR, Hyatt D, et al. Quantitative tracking of isotope flows in proteomes of microbial communities. *Mol Cell Proteomics.* 2011; 10: M110.006049.
- Shade A, Handelsman J. Beyond the Venn diagram: the hunt for a core microbiome. *Environ Microbiol.* 2012;14:4–12.
- Strycharz SM, Woodard TL, Johnson JP, et al. Graphite electrode as a sole electron donor for reductive dechlorination of tetrachlorethene by *Geobacter lovleyi*. *Appl Environ Microb.* 2008;74:5943–7.
- Suenaga H, Koyama Y, Miyakoshi M, et al. Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. *ISME J.* 2009;3:1335–48.
- Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol.* 2008;11:442–6.
- Whyte LG, Bourbonnière L, Greer CW. Biodegradation of petroleum hydrocarbons by psychrotrophic *Pseudomonas* strains possessing both alkane (*alk*) and naphthalene (*nah*) catabolic pathways. *Appl Environ Microb.* 1997;63:3719–23.
- Yergeau E, Hogues H, Whyte LG, et al. The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. *ISME J.* 2010;4:1206–14.
- Yergeau E, Sanschagrin S, Beaumier D, et al. Metagenomic analysis of the bioremediation of diesel-contaminated Canadian high Arctic soils. *PLoS One.* 2012;7:e30058.

Metagenomics, Metadata, and Meta-analysis

Jack Gilbert
Department of Ecology & Evolution,
University of Chicago, Chicago, IL, USA

Synonyms

Comparative analysis; Contextual data; Environmental data; Network analysis; Shotgun metagenomics

Definition

The analytical approach of identifying emergent patterns in ecological properties of microbial communities by sequencing community structure and function and defining the physical, chemical, and biological parameters of the ecosystem.

Metagenomics is the study of all genetic material from all organisms in a defined sample (Handelsman et al. 1998). However, it is defined: metagenomics is just a term used to describe a selection of tools and techniques that enable us to uncover the DNA from the organisms in an environment (which can comprise any ecosystem, from soil to human intestinal tract). Metadata (also known as contextual data) refers directly to information regarding the original sample, the extraction and handling of the DNA, and the sequencing platform and data processing information (Field et al. 2011; Yilmaz et al. 2011). Without such metadata, metagenomic sequence data would be redundant for anything other than basic gene discovery. Meta-analysis, which is the process of performing comparative investigation of features between datasets, is greatly enhanced by the combination of metagenomic data and metadata (Knight et al. 2012).

Metagenomics

Our microbial planet is more than 1×10^{30} microbial cells (Whitman et al. 1998), a billion more cells than stars in the known universe (Gilbert 2010). This dominance of biomass is encapsulated nicely by a quotation accredited to Julian Davies, “Once the diversity of the microbial world is catalogued, it will make astronomy look like a pitiful science” (Gewin 2006). Microbial life comprises the main functional drivers of our planet’s ecosystems (Falkowski et al. 2008), yet their diversity and ecological networks remain largely unknown. In the last 15 years, metagenomics has provided a tool to explore the vast unseen majority with a greater resolution and depth of field than culturing has yet provided (Hugenholtz and Kyrpides 2009). The explosion

in 2004 of direct sequencing approaches, which provided a different route to market compared to clone-dependent sequencing, has accelerated the implementation and data generation capability of this technique. Existing studies have been well reviewed in terms of the impact on community ecology interpretation and novel biochemical process identification (Gilbert and Dupont 2011).

Metadata

The ensuing data bonanza (Field et al. 2011) has driven the need for more robust and comprehensive standards for recording and sharing information about why, how, and from where the sequencing data was generated. One person's metadata is another person's primary data, and so the community outreach to determine the consensus for recording different data types and information has been a mammoth effort. The Genomic Standards Consortium (Field et al. 2011) has risen to be one of the most prominent and successful standards communities. The central tenet of the Genomic Standards Consortium is to promote mechanisms that standardize the description of genomes, metagenomes, and amplicon sequences and the exchange and integration of these data and associated metadata (www.genesc.org). The GSC has created three minimal information checklists, which collectively are known as the Minimal Information about ANY sequence (MIxS) checklists. The three standards are the Minimal Information about a Genomic Sequence (MIGS; Field et al. 2008), the Minimal Information about a Metagenomic Sequence (MIMS), and the Minimal Information about a Marker Gene Sequence (MIMARKS) (Yilmaz et al. 2011). These information checklists and the ancillary environmental data sheets describe the types of information the community would like to see associated with the sequence data, and importantly provide a description for recording these data using a defined standard. This enables a level playing field for the provision and sharing of data between organizations and PIs, and the checklists

have been adopted by the International Nucleotide Sequence Database Collaboration (INSDC) and a considerable number of journals. The major proponent from the latter group is the GSC's own journal, *Standards in Genomic Science*, which requires a detailed but standard description of the associated metadata for genome and metagenome reports (Gilbert et al. 2010a; Nelson et al. 2009).

Meta-analysis

Meta-analysis is defined as the combination of results from different studies that have similar or related research hypotheses. While not strictly a meta-analysis, the use of comparative metagenomics to explore the principles of microbial ecology stems from the common analysis of data generated by different studies in different ecosystems to explore central hypotheses, usually related to the overall distribution of taxonomic functional attributes in the community. Initial efforts include comparative analysis of four metagenomic samples from soil and whale fall (Tringe et al. 2005), 87 viral and microbial metagenomic datasets from nine biomes (Dinsdale et al. 2008), metagenomic datasets from 86 viral and microbial communities (Willner et al. 2009), and more recently 77 metagenomes (Delmont et al. 2011). These studies have led to the conclusion that different environments have habitat-specific functional and taxonomic fingerprints that indicate environment-specific genomic adaptation. Of course this should be taken with a caveat that each comparative study has a small number of metagenomes in the analysis and that each metagenomic dataset only comprises a tiny fraction of the functional information present in any community. The latter point is made obvious by ultra-deep screening of microbial diversity, whereby even in marine coastal surface waters, the species richness can be astounding (>100,000 taxa per L of water; Caporaso et al. 2011).

Importantly, cross-sample comparisons should be performed in concert with dynamic

comparative analysis of the contextual environmental data. These physical, chemical, and biological data that describe the environment in which the microbial organisms under investigation were isolated are vital to interpreting the gradients of function and specific trends in gene persistence seen between samples and studies. Within one study, such as the Global Ocean Sampling (Rusch et al. 2007) or Western English Channel (Gilbert et al. 2010b), the link between environmental metadata and the functional or taxonomic sequence data can be implicit. However, in comparative studies, it is rare to be able to generate canonical correlations between specific functional gene abundances and different contextual metadata as different studies tend to measure different parameters differently. The Earth Microbiome Project (www.earthmicrobiome.org) is working to create not just comparable data on the basis of methodological standard protocols (e.g., DNA extraction, PCR, sequencing) but also by obtaining data with comparable contextual information, e.g., temperature measurements, latitude and longitude, ammonia concentrations, pH, etc. All these metadata are being collated into large-scale databases with the Genomic Standards Consortium's MIXS checklists as the data framework, and so they represent the community consensus for these records.

Summary

Metagenomics studies now need to be performed using the principles of scientific investigation and excellent statistical experimental design, using replication and adequate controls to determine if the perceived biological variation actually could be used to explore basic ecological principles. The only appropriate way to perform good meta-analysis for metagenomic studies is to utilize excellent metadata, and this comes back to the design of the experiment, long before any molecular analysis has even been suggested. It also must leverage multidisciplinary effort to obtain the right data to answer the relevant questions.

Cross-References

- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Biological Treasure Metagenome](#)
- ▶ [Challenge of Metagenome Assembly and Possible Standards](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)

References

- Caporaso JG, Field D, Paszkiewicz K, Knight R, Gilbert JA. Evidence for a persistent microbial community in the Western English Channel. *ISME J.* 2012;6:1089–1093.
- Delmont TO, et al. Metagenomic mining for microbiologists. *ISME J.* 2011;5(12):1837–43.
- Dinsdale EA, et al. Functional metagenomic profiling of nine biomes. *Nature.* 2008;452(7187):629–32.
- Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science.* 2008;320(5879):1034–9.
- Field D, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol.* 2008;26(5):541–7.
- Field D, et al. The genomic standards consortium. *PLoS Biol.* 2011;9(6):e1001088.
- Gewin V. Genomics: discovery in the dirt. *Nature.* 2006;439(7075):384–6.
- Gilbert JA. Beyond the infinite – tracking bacterial gene expression. *Microbiol Today.* 2010;37(2):82–5.
- Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci.* 2011;3:347–71.
- Gilbert JA, et al. Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Stand Genomic Sci.* 2010a;3(2):183–93.
- Gilbert JA, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One.* 2010b;5(11):e15545.
- Handelsman J, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5(10):R245–9.
- Hugenholtz P, Kyrpides NC. A changing of the guard. *Environ Microbiol.* 2009;11(3):551–3.
- Knight R, et al. Designing better metagenomic surveys: the role of experimental design and metadata capture in making useful metagenomic datasets for ecology and biotechnology. *Nat Biotechnol.* 2012;30(6):513–520.

- Nelson OW, Harrison SH, Garrity GM. Meeting report for SIGS1: first conference of the standards in genomic sciences eJournal. *Stand Genomic Sci.* 2009;1(1):72–6.
- Rusch DB, et al. The sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 2007;5(3):e77.
- Tringe SG, et al. Comparative metagenomics of microbial communities. *Science.* 2005;308(5721):554–7.
- Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA.* 1998;95(12):6578–83.
- Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol.* 2009;11(7):1752–66.
- Yilmaz P, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol.* 2011;29(5):415–20.

MetaRank: Ranking Microbial Taxonomic Units or Functional Groups for Comparative Analysis of Metagenomes

Tse-Yi Wang¹ and Huai-Kuang Tsai²

¹Department of Medical Research, Mackay Memorial Hospital, New Taipei City, Taiwan

²Institute of Information Science, Academia Sinica, Taipei, Taiwan

Definition

MetaRank is a rank conversion scheme for analyzing microbial communities based on the relative order of member (taxonomic unit or functional group) abundances rather than their estimated values (e.g., proportions). It leverages a series of statistical hypothesis tests to compare member abundances within microbial communities and determine their ranks, providing an alternative rank-based method for characterizing metagenomes.

Introduction

Metagenomics is a field that involves sampling, sequencing, and analyzing the genetic material of

microorganisms in microbial communities (Hugenholtz and Tyson 2008). A key question in metagenomics is whether and how changes in the microbial abundances of taxonomic units or functional groups relate to alterations of habitats (Hamady and Knight 2009). To characterize the relationship, it is important to compare microbial community compositions in different environments (Wooley et al. 2010).

Many statistical methods (e.g., *Metastats* (White et al. 2009), *ShotgunFunctionalizeR* (Kristiansson et al. 2009), *STAMP* (Parks and Beiko 2010)) have been developed for comparative metagenomics in attempt to identify differentially abundant features between microbial communities. Most of these methods employ statistical hypothesis tests to determine whether member abundances are equal in distinct communities and focus on the quantitative differences between microbial community compositions. They are highly dependent on the precision of estimated values in member abundances.

However, estimated abundances might deviate from the true abundances in habitats due to sampling biases and other systematic artifacts in metagenomic data processing (Ashelford et al. 2005; Brady and Salzberg 2009; Gomez-Alvarez et al. 2009; Mavromatis et al. 2007). Although systematic artifacts can be corrected through improvements in data processing techniques, sampling biases will remain unavoidable unless exhaustive data of the whole populations become available (Wooley and Ye 2010).

To reduce the effects of sampling biases, MetaRank performs a series of rank conversions for analyzing microbial communities based on the ranks of members rather than their estimated abundances. It leverages the fact that the ranks of highly abundant members are less affected by sampling biases because large values and, by extension, their relative order are robust against small deviations. It also utilizes statistical hypothesis testing to compare member abundances within communities and determine the ranks as follows: Highly abundant members are delegated to high ranks and any two members without statistically significantly different abundances are assigned the same rank.

Empirical tests on real datasets and synthetic samples (Kurokawa et al. 2007; Ley et al. 2006; Mavromatis et al. 2007) approve that MetaRank is able to downsize the effects of sampling biases and help to clarify the characteristics of metagenomes. The ranks converted by MetaRank have small normalized standard deviations, which clearly reveal the common traits within a set of metagenomes. The ranks also capably identify the discriminating features of microbial community compositions (Wang et al. 2011). In addition, it is noted that MetaRank as a rank-based approach has the same disadvantages of all nonparametric methods. There is a loss of information and the loss of ability to provide parametric statistics for inference. Therefore, MetaRank is a useful rank-based alternative for analyzing metagenomes that complements parametric methods.

Methods

Given a metagenomic sample of a microbial community, MetaRank first employs binomial tests to iteratively select highly abundant members within the community followed by multinomial tests to rank the selected members in each run.

Binomial Tests for Selecting Highly Abundant Members

For N members in a microbial community, let X_n represent the abundance of the n th member in the metagenomic sample and \hat{p}_n (i.e., X_n/S) be the sample proportion of the n th member, where $n = 1, 2, \dots, N$ and $S = X_1 + X_2 + \dots + X_N$. Under the assumption that all nucleic acids of microorganisms in habitats are equally likely to be sampled and sequenced in metagenomic experiments, the abundance X_n of the n th member in the sample is modeled as a binomial random variable:

$$X_n \sim \text{Binomial}(S, p_n),$$

where p_n is the unknown population proportion of the n th member in the habitat and estimated by the sample proportion \hat{p}_n .

To select highly abundant members with proportions that are significantly higher than the average proportion ($1/N$), MetaRank applies hypothesis tests, $H_o: p_n \leq 1/N$ vs. $H_a: p_n > 1/N$ for all $1 \leq n \leq N$. Since $X_n \sim \text{Binomial}(S, p_n)$ with mean $E(X_n) = Sp_n$ and variance $\text{Var}(X_n) = Sp_n(1 - p_n)$, the binomial distribution of the test statistic X_n under H_o is approximated by normal distribution with z -statistic Z_n :

$$\begin{aligned} Z_n &= \frac{X_n - E(X_n)}{\sqrt{\text{Var}(X_n)}} = \frac{X_n - \frac{S}{N}}{\sqrt{\frac{S}{N} \left(1 - \frac{1}{N}\right)}} \\ &= \frac{\hat{p}_n - \frac{1}{N}}{\sqrt{\frac{1}{SN} \left(1 - \frac{1}{N}\right)}} \sim N(0, 1) \end{aligned}$$

when sample size S is large enough such that $0 \leq E(X_n) - 3\sqrt{\text{Var}(X_n)} \leq S$. Otherwise, the exact binomial test is applied when S is small such that $E(X_n) - 3\sqrt{\text{Var}(X_n)} < 0$ or $S < E(X_n) + 3\sqrt{\text{Var}(X_n)}$.

The p -value for exact binomial test is calculated as follows:

$$P[X_n \geq x_n] = \sum_{k=x_n}^S \binom{S}{k} \frac{1}{N^k} \left(1 - \frac{1}{N}\right)^{S-k}$$

where x_n is the observed value of the test statistic X_n .

MetaRank considers members that reject the null hypothesis with statistical significance as highly abundant. For those that fail to reject the null hypothesis (assuming N' members remain), MetaRank temporarily sets them aside and continues to select members whose proportions are significantly larger than the average ($1/N'$) in the next iteration. When none of the remaining members reject the null hypothesis, MetaRank terminates the selection procedure and considers all remaining members as rare members.

Thus, in each iteration, the selected members (whose proportions are larger than the average) are higher than the remaining members (whose proportions are equal to or smaller than the average). Moreover, the members selected in distinct iterations are ranked in their selected order; more specifically, the members selected in first iteration are assigned a higher rank than the ones selected in the second iteration. At the end, the rare members are ranked the lowest in the community.

Multinomial Tests for Ranking Highly Abundant Members

Based on the above procedure, MetaRank ranks the abundances in the target community according to the following three rules. First, all rare members are assigned the same smallest rank. Second, the members selected in distinct iterations are ranked according to the order in which they were selected; thus, the members selected in the first iteration of the procedure are assigned higher ranks than all the others. Third, if two abundances (the i th and j th members) are selected in the same iteration, MetaRank determines their ranks ($R_i > R_j$, $R_i < R_j$ or $R_i = R_j$) by two hypothesis tests, $H_o: p_i \leq p_j$ vs. $H_a: p_i > p_j$ and $H'_o: p_j \leq p_i$ vs. $H'_a: p_j > p_i$. If H_o is rejected, $R_i > R_j$; conversely, if H'_o is rejected, $R_i < R_j$. However, if both H_o and H'_o are accepted, $R_i = R_j$.

Under the same assumption that all nucleic acids are equally likely to be sampled and sequenced, each abundance X_n is modeled as a binomial random variable; any two abundances X_i and X_j are jointly modeled by the multinomial distribution (i.e., the generalization of binomial distribution in multidimension):

$$(X_i, X_j) \sim \text{Multinomial}(S, p_i, p_j)$$

where p_i and p_j are the unknown population proportions of the i th and j th members in habitat and estimated by the sample proportions \widehat{p}_i and \widehat{p}_j . For large S such that $0 \leq E(X_i)p \pm 3\sqrt{\text{Var}(X_i)} \leq S$ and $0 \leq E(X_j)p \pm 3\sqrt{\text{Var}(X_j)} \leq S$, the z-statistics of the approximate tests are

$$Z_{ij} = \frac{X_i - X_j}{\sqrt{X_i + X_j}} \quad \text{and} \quad Z_{ji} = \frac{X_j - X_i}{\sqrt{X_j + X_i}}.$$

Otherwise, the exact multinomial tests are applied. The p -values are calculated as

$$P[X_i - X_j \geq x_i - x_j] = \sum_{h=x_i-x_j}^S \sum_{k=0}^{h-(x_i-x_j)} \frac{S!}{h!k!(S-h-k)!} \left(\frac{\widehat{p}_i + \widehat{p}_j}{2}\right)^{h+k} (1 - \widehat{p}_i + \widehat{p}_j)^{S-h-k}$$

and

$$P[X_j - X_i \geq x_j - x_i] = \sum_{h=0}^{S-(x_j-x_i)} \sum_{k=h+(x_j-x_i)}^S \frac{S!}{h!k!(S-h-k)!} \left(\frac{\widehat{p}_i + \widehat{p}_j}{2}\right)^{h+k} (1 - \widehat{p}_i + \widehat{p}_j)^{S-h-k}$$

where x_i and x_j are the observed values of X_i and X_j .

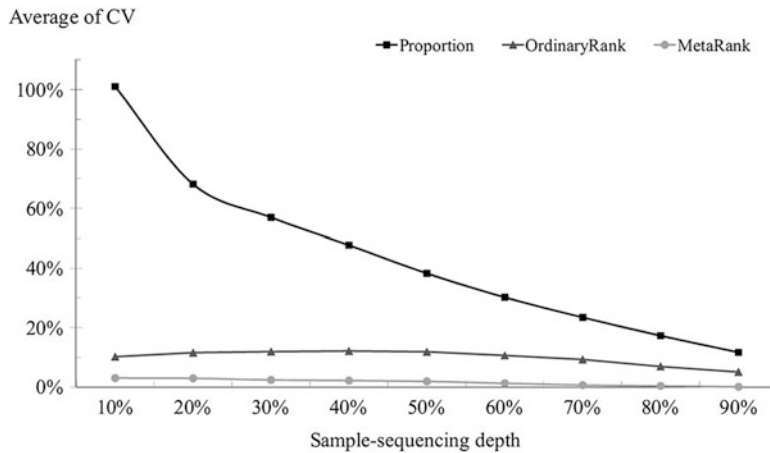
As a result, the sorted abundances $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(m)} \leq \dots \leq X_{(M)}$ are converted into ranks $1 \leq R_{(1)} \leq R_{(2)} \leq \dots \leq R_{(m)} \leq \dots \leq R_{(M)} \leq M$, where the subscript in parentheses (m) denotes the m th order in the community and M is the total number of members. For members whose abundances cannot be distinguished from each other by hypothesis testing, MetaRank converts them into their average order; i.e., for any m' , m'' such that $R_{(m')} < R_{(m'+1)} = R_{(m'+2)} = \dots = R_{(m''-1)} < R_{(m'')}$ (given $R_{(0)} = 0$ and $R_{(M+1)} = M + 1$), we have

$$R_{(m'+1)} = R_{(m'+2)} = \dots = R_{(m''-1)} = \frac{m' + m''}{2}$$

For example, the ranks of the rare members (assuming N'' members remain in the last iteration) are converted into $(N'' + 1)/2$.

Empirical Tests

To evaluate its utility in comparative analysis of microbiomes, MetaRank is applied to real



MetaRank: Ranking Microbial Taxonomic Units or Functional Groups for Comparative Analysis of Metagenomes, Fig. 1 The averages of CV, which is the normalized standard deviation, in the ranks converted by MetaRank, estimated proportions and ordinary ranks at

the phylum level of the 5,000 synthetic samples for each sample-sequencing depth $r \in \{10\%, 20\%, \dots, 90\%\}$. Under distinct sample-sequencing depth, the averages of CV in the ranks converted by MetaRank are smaller than the ones in the others

metagenomes and synthetic samples (Kurokawa et al. 2007; Ley et al. 2006; Mavromatis et al. 2007; Wang et al. 2011). In synthetic samples, it is shown that as compared with the estimated proportions or the ordinary ranks of straightforward sorted abundances, the ranks converted by MetaRank have smaller normalized standard deviation and are less affected by sampling biases. In real metagenomes, using MetaRank is able to clarify the common traits and detect the discriminating features of those microbiomes.

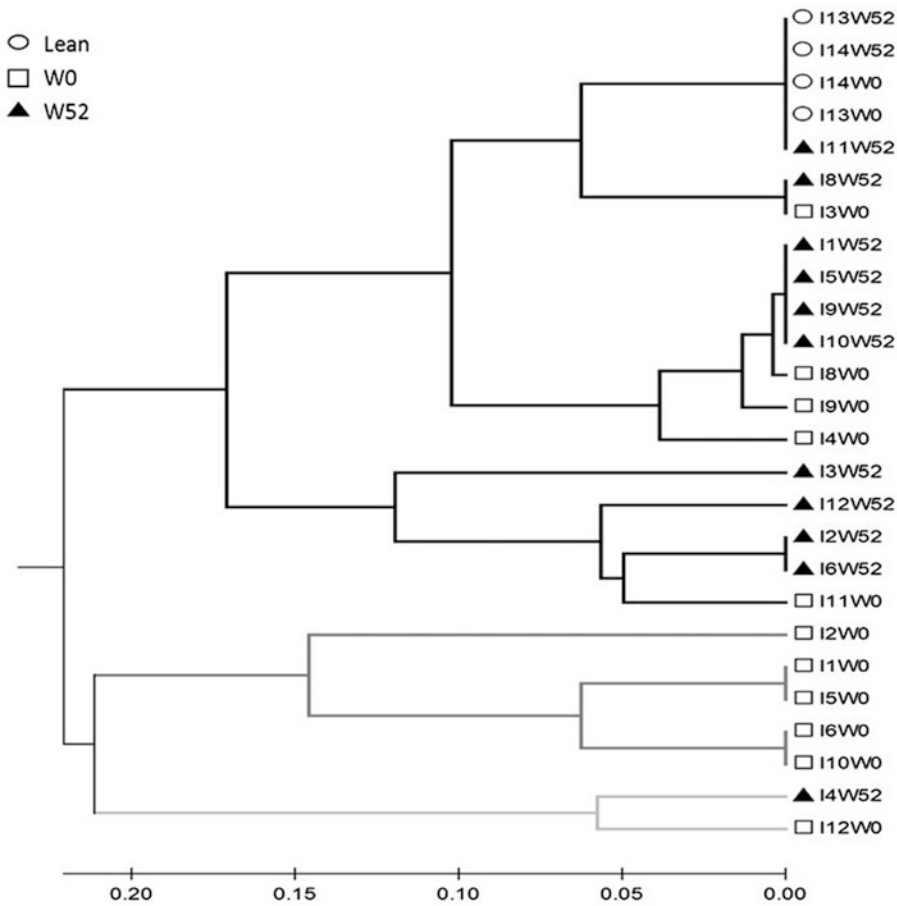
Simulation Analyses of Synthetic Samples

Synthetic samples are generated by randomly resampling reads from a pooled dataset of real metagenomes (Ley et al. 2006) for investigating the effects of sampling biases on the ranks converted by MetaRank, estimated proportions, and ordinary ranks (Wang et al. 2011). All the reads are pooled together as a synthetic library, and at the taxonomic level of phylum, five thousand synthetic samples are generated for each sample-sequencing depth $r \in \{10\%, 20\%, \dots, 90\%\}$. The effects of sampling biases are examined by the variability between the random synthetic samples, and the variability between the random samples is measured by the normalized standard deviation (CV; coefficient of variation) in

the ranks converted by MetaRank, estimated proportions, or ordinary ranks. As shown in Fig. 1, the normalized standard deviations in the ranks converted by MetaRank are smaller than the ones in the estimated proportions and the ordinary ranks. Similar observations are also found at the taxonomic levels of class, order, family, genus, and other simulated datasets in the Wang et al. (2011) study. The results confirm that MetaRank is able to reduce the effects of sampling biases.

Demonstration Studies in Real Metagenomes

In the real datasets from the human gut microbiomes (Ley et al. 2006; Kurokawa et al. 2007), MetaRank demonstrates its ability to clearly reveal the characteristics of metagenomes in comparative analyses (Wang et al. 2011). The first dataset contains samples from obese individuals and lean controls of human gut metagenomes in a one-year diet study. The second dataset includes infant and adult samples. In the first dataset, the obese samples are extracted from 12 obese individuals (I1, I2, ..., I12) at four distinct time points (week 0, 12, 26, and 52), and the lean controls are extracted from two lean individuals (I13 and I14) at two time point (week 0 and 52), all



MetaRank: Ranking Microbial Taxonomic Units or Functional Groups for Comparative Analysis of Metagenomes, Fig. 2 The hierarchical clustering results of the ranks converted by MetaRank at the phylum level in 12 obese individuals at week 0 (I1W0, I2W0, ..., I12W0) and 52 (I1W52, I2W52, ..., I12W52), including the four lean controls (I13W0, I14W0, I13W52, and I14W52), based on UPGMA. The hierarchical

agglomerative clustering (bottom-up clustering) initially treats each sample as a single cluster at the bottom and then successively agglomerates pairs of nearest clusters until all clusters have been merged into a single cluster at the top. Given a fix distance 0.2 (i.e., Pearson correlation 0.8), there are three main clusters, where the unweighted arithmetic mean of distances within clusters are smaller than 0.2

denoted by the convention, I_xW_y , where x represents the x th individual and y represents the time point. In the second dataset, four infant and nine adult samples were extracted from different individuals for COG-functional analysis.

When comparing metagenomes in the first dataset (Ley et al. 2006), using MetaRank is able to clarify the common traits of similar samples (Wang et al. 2011). The taxonomic abundances in the obese samples and the lean controls are converted into ranks by MetaRank, followed by hierarchical clustering with UPGMA

(Unweighted Pair Group Method with Arithmetic Mean). Figure 2 illustrates the result of the simple case that only consists of the samples at week 0 and 52 (before and after diet). As shown in Fig. 2, given a fix distance 0.2 (i.e., Pearson correlation 0.8), there are three main clusters, where the unweighted arithmetic mean of distances within clusters are smaller than 0.2. The four lean controls are closely grouped together in one cluster that contains some obese samples at week 0 and all the obese samples at week 52 except one (I4W52). More than half of the

obese samples at week 0 are in the other two clusters. The result shows that after dieting almost all the obese samples are clustered together with the four lean controls. Similar results are observed in the members of the biome at the taxonomic levels of class, order, family, and genus (Wang et al. 2011).

Additionally, MetaRank is able to detect rank-based differences and identify discriminating features between metagenomes in the second dataset (Kurokawa et al. 2007). The abundances of functional groups in the infant and adult samples are first converted into ranks by MetaRank. Then the *t*-test is applied to identify rank-based differences between the infant and adult samples. When compared with proportion differences detected by a parametric method (only *t*-test without MetaRank), it is found that MetaRank, a nonparametric approach, helped to identify additional functional groups as discriminating features (Wang et al. 2011). Since nonparametric and parametric methods are complementary to each other in statistics (one cannot replace the other), MetaRank is thus a useful rank-based approach complementary to parametric methods.

Summary

Most statistical methods for comparative analysis of microbial community compositions rely on estimated abundances of members. However, when processing metagenomic data, sampling biases and systematic artifacts cause noisy deviations that may result in estimated abundances differing from true abundances. MetaRank, which converts highly abundant members into higher ranks, is designed to cut the effects of noisy deviations. It leverages the fact that the ranks of highly abundant members are robust against small deviations. Empirical tests on synthetic samples and real metagenomes confirm that the ranks converted by MetaRank have small normalized standard deviations, facilitate the comparative analysis of metagenomes, and help to reveal the common characteristics or the discriminating features within a set of microbiomes. Therefore, MetaRank, as a

nonparametric approach, provides a useful rank-based alternative to analyzing microbial community compositions.

Cross-References

- ▶ [STAMP: Statistical Analysis of Metagenomic Profiles](#)

References

- Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol.* 2005;71:7724–36.
- Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods.* 2009;6:673–6.
- Gomez-Alvarez V, Teal TK, Schmidt TM. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* 2009;3:1314–7.
- Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* 2009;19:1141–52.
- Hugenholtz P, Tyson GW. Microbiology: metagenomics. *Nature.* 2008;455:481–3.
- Kristiansson E, Hugenholtz P, Dalevi D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics.* 2009;25:2737–8.
- Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 2007;14:169–81.
- Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. *Nature.* 2006;444:1022–3.
- Mavromatis K, Ivanova N, Barry K, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4:495–500.
- Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics.* 2010;26:715–21.
- Wang TY, Su CH, Tsai HK. MetaRank: a rank conversion scheme for comparative analysis of microbial community compositions. *Bioinformatics.* 2011;27:3341–7.
- White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol.* 2009;5:e1000352.
- Wooley JC, Ye Y. Metagenomics: facts and artifacts, and computational challenges. *J Comput Sci Technol.* 2010;25:71–81.
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol.* 2010;6:e1000667.

METAREP, Overview

Johannes Goll

Informatics Department, The J. Craig Venter
Institute, Rockville, MD, USA

With increasing scale and complexity of current metagenomic studies approaching terabase-volumes of sequence data, scalability of biological analysis software has become an essential requirement. Toward that end, we have developed JCVI Metagenomics Reports (METAREP), an open-source tool, which integrates the highly scalable search engine Solr/Lucene, R, and CAKEPHP into an extendible Web-based software to query, browse, compare, and share extremely large volumes of metagenomic annotations. The software allows flexible and simultaneous comparison of taxonomic and biological pathway and individual enzyme abundances across hundreds of samples. In this chapter, we provide an overview of this functionality, data format, import, installation, and customization. We present new features that have been released with version 1.4.0 including the implementation of two-way statistical tests to compare features of two datasets without replicates, protein sequence integration, and BLASTP homology search capabilities. The latest functionality can be tested on example data at JCVI's public METAREP instance, which is available at <http://www.jcvi.org/metarep> (via the "Try It" button). The open-source code of the software and developer information is accessible at the project's source code repository at <https://github.com/jcvi/METAREP>.

Introduction

Metagenomics describes a scientific approach in which DNA, extracted from microbes sampled from a certain environment, is used to reconstruct the genomic potential and interactions of whole microbial communities. This circumvents the problem that the majority of microbes cannot be

cultured outside their native habitat and thus cannot be investigated using a classic genome sequencing approach (Handelsman 2004). With increasing sequencing throughput of next-generation sequencing technologies, this approach has become commonplace and is being applied to the soils, oceans, agriculture, and human health. The goal is to understand how the microbe's genetic repertoire is used during nutrient cycle and energy production and, especially, what role it plays in human health and chronic disease.

As a consequence, the challenge for most microbiologists has shifted away from data generation to effective data storage and analysis methods (Stein 2010). An effective approach, to handle these immense data volumes, is to use workflows in combination with high-performance computer clusters or grids with hundreds of processors that execute homology searches in parallel for subsets of assembled or unassembled fragmented sequences (reads). Based on hits to reference sequences from completely sequenced genomes, the end data products are typically organismal as well as metabolic gene or read-based count profiles.

The Human Microbiome Project (HMP) (<http://nihroadmap.nih.gov>) is an excellent example of highlighting the scale of metagenomic projects currently taking place. The HMP is a very ambitious effort to characterize the microbial community associated with the human body. The jumpstart consortium consists of four sequencing centers: the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute (JCVI), and the Genome Center at Washington University. While it involves a range of activities including the extensive collection of metadata, generation of reference genomes, and marker gene studies, one of the most data-intensive phases of the project is the shotgun metagenomic survey of over 650 samples from 254 healthy individuals initially examining 15–18 body habitats. In future phases the HMP will compare this baseline data to clinical samples to examine the specific role the microbiome plays in disease and the maintenance of human health. At present over

METAREP, Overview, Table 1 Comparison of metagenomic software

Resource/ software	Year latest release	Maintaining institution	Free annotation services	Workflow	Open- source	Web site
IMG/M	2012	US Department of Energy	Yes	Annotation: COG, Pfam, TIGRFAM, InterPro, KEGG	No	http://img.jgi.doe.gov/cgi-bin/m/main.cgi
EBI portal metagenomics	2011	European Bioinformatics Institute	Yes	Annotation: InterPro, GO	Yes	https://www.ebi.ac.uk/metagenomics/
CLOVR	2011	University of Maryland School of Medicine	Yes	16 s, clustering, assembly, annotation: COG, RefSeq	Yes	http://clovr.org
Galaxy	2010	Penn State University	No	Only taxonomy/phylogeny, some community extensions {#15571}	Yes	http://galaxy.psu.edu
METAREP	2012	J. Craig Venter Institute	No	No inbuilt annotation workflow, users can upload existing annotations	Yes	http://www.jcvi.org/metarep/
CAMERA 2.0	2010	California Institute for Telecommunications and Information Technology	Yes	ORF finding, tRNA, rRNA finding, clustering, genome assembly, annotation: Pfam, TIGRFAM, COG	No	https://portal.camera.calit2.net/
MG-RAST	2008	Argonne National Lab cluster	Yes	SEED subsystem	Yes	http://metagenomics.anl.gov/

20,864 million reads of Illumina data have been produced from healthy individuals. The comparison of the sequence reads to protein databases alone is estimated to generate data exceeding 12 terabytes (Human Microbiome Project Consortium 2012). We believe that the HMP typifies the scope and complexity of metagenomic projects that will come. The collection, integration, sharing, and comparison of this data represent a characteristic example of the current metagenomic data analysis challenges. Toward this end we have developed METAREP, an open-source and thus adjustable software that enables exploratory data analysis for projects of this size and larger (Goll et al. 2010, 2012).

A variety of other free metagenomic annotation and analysis software is accessible to researchers (Table 1). Efforts that include annotation workflows and free compute resources are provided by the US Department of Energy (IMG/M (Markowitz et al. 2012)), the European Bioinformatics Institute, the Argonne National

Laboratory (MG-RAST (Meyer et al. 2008)), and the University of San Diego (CAMERA (Sun et al. 2011)). Efforts that require compute resources owned by the researchers (or rented via a cloud service) include CLOVR (Angiuoli et al. 2011), Galaxy (Goecks et al. 2010), and METAREP. The free annotation resources, however, are often tightly coupled to each center's specific infrastructure including its compute resources. Thus they cannot easily be installed and modified to satisfy custom needs including privacy concerns and advanced data access management. In contrast CLOVR, Galaxy, and METAREP are self-contained and can be run on other systems, and the source code can be adapted to handle project-specific needs. On the analysis side, most resources provide summary results that fit a certain workflow that are tailored toward answering a certain question. METAREP is an exception, as it supports generic exploratory data analysis for annotations from different workflows that can be

queried and filtered dynamically. For example, its functionality can be used to visualize how specific taxonomic or metabolic markers vary across samples. METAREP does not support a particular workflow but a generic annotation input format. As a consequence, it does not include annotation workflows. To bridge this gap, users can run a public annotation service or a custom local pipeline, format the data, and import it.

In the following sections we will describe how to import data, highlight features to analyze individual and multiple datasets, carry out BLAST searches, and customize the software.

Data Format and Import Process

The current METAREP tab-delimited format specification for 17 fields is shown in Table 2. Understanding this format is crucial for subsequent analysis. Following the outlined conventions will help users to leverage as much of the functionality as possible and understand what fields are supported. The format has been designed to accommodate common data types that are produced by many annotation workflows without being tied to a specific workflow. The disadvantage of this flexibility is that a custom parser needs to be written to format the output of a certain workflow according to this tab format before importing the data. However, in most cases, generating the METAREP tab-delimited format is trivial. In addition, METAREP provides data formatting functionality for two workflows: (1) the JCVI Prokaryotic Metagenomic Annotation Pipeline (JPMAP (Tanenbaum et al. 2010)) and (2) the HUMAnN metabolic reconstruction pipeline (Abubucker et al. 2012). The open-source code for formatting output from these two pipelines serves as a template for supporting other formats. The code base also includes a Perl utility script (`scripts/perl/metarep_loader.pl`) to import tab-delimited annotation files into METAREP projects (more details on how to use the import script can be found at <https://github.com/jcvi/METAREP/wiki/Installation-Guide-v-1.4.0>).

Except for the first two columns (`peptide_id` and `library_id`), which specify the unique ID of the respective annotation entry (gene/protein ID) and the library/dataset ID, respectively, columns are optional. This has the advantage that workflows that do not produce all of the data types are supported. The last two columns in Table 2 provide example values for each of the fields per pipeline. While the unique ID fields mentioned before store a single value, most of the other fields can store multiple values (as indicated in column 3). By convention, multiple values are double pipe separated. For example, information for a multi-enzymatic protein can be stored by setting the value of the `ec_id` field to “1.6.99.3||1.6.5.3”. By convention, the `ec_id` field stores the enzyme accessions according to IUBMB format. Higher-level enzymatic levels are encoded using dashes for all unspecified levels, e.g., 3.4.-.-. The `go_id` field stores accessions defined by the Gene Ontology (Ashburner et al. 2000) with accessions being prefixed using uppercase “GO:”. The `hmm_id` is a generic field for hidden Markov model-based assignments. It takes Pfam accessions (PF234) (Punta et al. 2012), TIGRFAM accessions (TIGR23423) (Haft et al. 2003), superfamily accessions (SSF345) (Madera et al. 2004), and combinations of the same (separated by double pipes).

The `blast_*` fields store information of BLAST (Altschul et al. 1990) alignments (but can hold alignment information from other alignment software). In particular, the `blast_tree` field stores organismal information in the form of the lowest taxon using the NCBI Taxonomy as the reference taxonomy. For example, to indicate that a certain annotation entry belongs to *Escherichia coli*, the `blast_tree` field can be set to NCBI taxon id “83333”. If multiple NCBI taxon IDs are provided, the lowest common ancestor will be determined during the data import process based on the NCBI taxon ID set provided by the user. The `blast_evalue`, `blast_pid` (proportion of identical amino acids), and `blast_cov` (proportion of coverage of query sequence) reflect alignment quality data types. The field values range from 0 to 1 and allow users to filter their data based on alignment quality (see searching and filtering). The `ko_id` field stores the KEGG Ortholog

METAREP, Overview, Table 2 Data format

Column	Field name	Multi-valued	Description	JPMAP	HUMANn
1	peptide_id	No	Unique entry ID	JCVI_PEP_1234123	ptr:453118
2	library_id	No	Dataset ID	SRS011061	SRS011061
3	com_name	Yes	Functional description	Sugar ABC transporter, periplasmic sugar-binding protein	LGMN, legumain, K01369 [legumain [EC:3.4.22.34]]
4	com_name_src	Yes	Functional description source	Uniref100_A23521	ptr:453118
			Description assignment		
5	go_id	Yes	Gene Ontology ID	GO:0009265	GO:0001509
6	go_src	Yes	Gene Ontology source	PF02511	K01369
			Assignment		
7	ec_id	Yes	Enzyme commission ID	2.1.1.148	3.4.22.34
8	ec_src	Yes	Enzyme commission source	PRIAM	ptr:453118
9	hmm_id	Yes	HMM ID	PF02511	NA
10	blast_tree	Yes	NCBI Taxonomy ID	246194	9598
11	blast_evalue	No	BLAST E-value	1.78E-20	Median
12	blast_pid	No	BLAST percent identity	0.93	Median
13	blast_cov	No	BLAST sequence coverage	0.82	N/A
14	Filter	Yes	Filter tag	Repeat	N/A
15	ko_id	Yes	KEGG Ortholog ID	N/A	K01369
16	ko_src	Yes	KEGG Ortholog source	N/A	ptr:453118
17	Weight	No	Weight to adjust abundance of assignments	1	43.23

accession (KO2134). Both the `ec_id` and `ko_id` fields are used to support two types of pathway analysis (see pathway analysis section). Pathway analysis based on the `ec_id` field allows analysis of 100, strictly metabolic, pathways. Pathway functionality based on `ko_id` is more comprehensive supporting 200 additional non-metabolic pathways such as transcription and translation. Depending on which field is populated, functionality is activated. Source fields (fields with a `_src` postfix) describe the origin of certain value. For example, an enzyme accession may have been assigned based on a certain TIGRFAM model or a reference gene/protein homology hit or other methods. The `ec_src` field can be used to track this information. Finally, the `weight` field allows users to assign weight to a certain entry to adjust the absolute and relative frequency of associated entry values. The field can be used for encoding abundance information such as the number of reads that support a certain gene/protein (in transcriptomic or assembly studies) or spectral counts in meta-transcriptomic studies. By default the `weight` field is set to 1.

When we subsequently refer to *annotation attributes*, we mean a selection of these fields that are used throughout the software to provide summary statistics and compare datasets. They refer to NCBI Taxonomy, Gene Ontology, Enzyme Classification, HMM, and KEGG/Metacyc pathways and KEGG Ortholog fields. A *feature* refers to a certain value that an *annotation attribute* can take. A feature-dataset matrix is a two-dimensional matrix with *features* of a certain *annotation attribute* as rows and datasets as columns. Cells represent the sum of weights of the respective feature-dataset combination (by default it reflects the number of genes/peptides with that specific feature).

Single Dataset Options

Dataset Summary Statistics

The *View Dataset Page* displays the imported annotations and provides high-level summaries of annotation attributes including detailed pathway summaries. The *Data Tab* shows the

imported data in tabular format. This is helpful to check if the data has been correctly imported. The *Summary Tab* provides an overview of overall annotation statistics including a high-level taxonomic breakdown. Subsequent tabs summarize statistics for a corresponding annotation attribute. For each, the top 20 ranked features with the absolute and relative counts are displayed. Users can adjust the number of top feature that is being displayed (up to 1,000 ranks) and download the data in tab-delimited format.

Dataset Search and Filter Options

The *Search Page* facilitates dynamic filtering of annotation and allows users to export matching entries and associated statistics. Once a query is executed, the page summarizes top 10 statistics for several annotation attributes in the form of lists and pie charts. The page also lists individual matching annotation entries so that users can confirm that the query correctly retrieved the desired results. The top 10 statistics, matching annotations, and underlying protein sequences (if configured, see configuration) can be exported. To search a dataset, users can enter a search term and select the field to search in from a drop-down box. Selections include ID-based and name-based searches. The former performs exact searches; the latter executes fuzzy name-based searches. For example, the user can enter 2.7.1.147 and select the Enzyme ID field from the drop-down box to search for exact matches. Alternatively, the user can carry out a fuzzy name-based search for “Glucokinase” which retrieves three matching enzymes: ADP-specific glucokinase (2.7.1.147), glucokinase (2.7.1.2), and phosphoglucokinase (2.7.1.10). For both search strategies, the selection triggers a query generation process that creates a query that is compatible with the Solr/Lucene query syntax (<http://wiki.apache.org/solr/SolrQuerySyntax>). The original search term is prefixed by the search field, and multiple terms can be logically combined using the AND, OR, and NOT keywords. In the ID-based example, the final query that will be generated is “`ec_id:2.7.1.147`”. For the name-based example, the final query represents a logical combination

(“OR”) of all individual matches, in this case “ec_id: 2.7.1.147 OR ec_id: 2.7.1.2 OR ec_id: 2.7.1.10”. The same principle is being applied to pathway name-based searches. A search for “starch and sucrose metabolism” using the name-based KEGG pathway name (EC) option searches for all enzymes in that pathway by generating the following query: “ec_id:1.1.1.22 OR ec_id:1.1.99.13 OR ec_id:2.4.1.1 OR ec_id:2.4.1.10 OR ...”. While the drop down helps to build queries, experienced users can enter the Solr/Lucene-formatted queries directly. This has the advantage of entering custom logical combinations of particular fields of interest (a complete list of fields and example queries are shown in Table 2). Note that if the value contains itself a colon (which is a special character of the Solr/Lucene language to separate field names from values), it needs to be preceded by a backward slash. For example, a search for “go_id:GO:0000160” instead of “go_id:GO\0000160” will return the desired results – fields that store complete hierarchies including the NCBI Taxonomy and the Gene Ontology. The former is encoded in the blast_tree field, which stores the whole taxonomic lineage (according to NCBI) for each entry in the form of NCBI taxon IDs. For example, a protein entry with a species assignment of “Escherichia coli” with NCBI taxon 562 has the following nine NCBI taxon IDs stored in the blast_tree field:

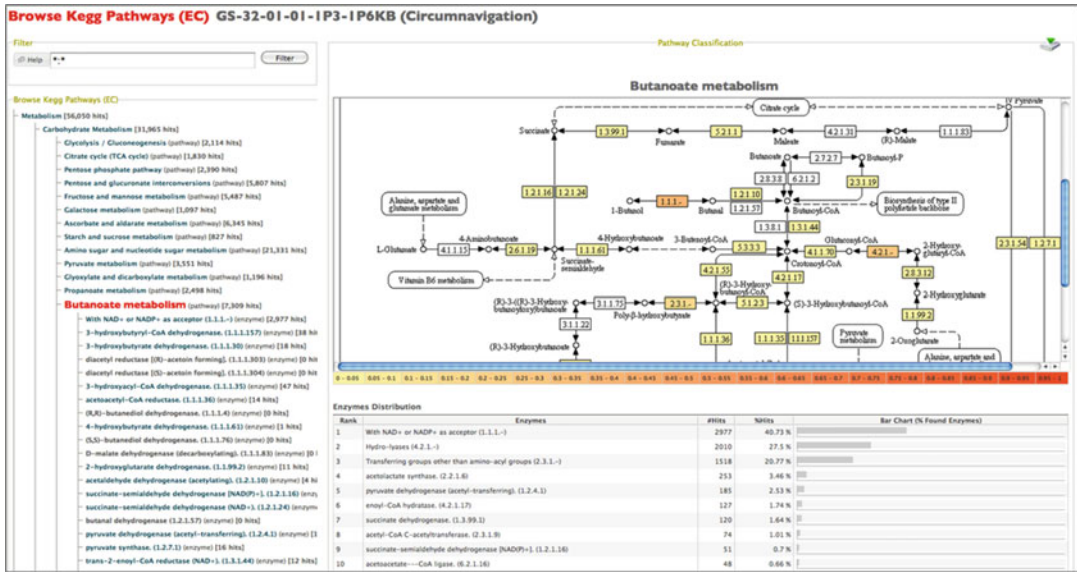
562 = Escherichia coli; 561 = Escherichia; 543 = Enterobacteriaceae; 91347 = Enterobacteriales; 1236 = Gammaproteobacteria; 1224 = Proteobacteria; 2 = Bacteria; 131567 = cellular_organisms; 1 = root;

This allows users to find the entry by searching for “blast_tree:562” (Escherichia coli) as well as “blast_tree:2” (Bacteria) or any other IDs that are part of that lineage. This can be very helpful for excluding or including proteins that were assigned to a certain taxonomic group. For example, “ec_id:2.7.1.2 AND blast_tree:2” can be used to filter the data for bacterial glucokinases. A search for “NOT blast_tree:9606” excludes entries that were assigned to “homo sapiens”. Another way of fuzzy searching (in addition to the name-based searches using the drop-down

menu) is to use Solr/Lucene wildcard characters. There are two supported wildcards: “?” and “*”. The “?” performs a single character wild card search. For example, to find common names like fliF, fliC, and fliS, one can search for “com_name_txt:fli?”. The “*” performs a multiple-character wild card search. For example, to search for all transferases 2.1.1.1, 2.1.1.2, 2.1.1.3, etc., one can enter “ec_id:2.*”. The quantitative alignment information (proportion of identical amino acids, proportion of covered query amino acids, E-value) range queries can be applied that identify entries that fall between a minimum and maximum. For example, to filter the data for a $1.0E-5 \leq E\text{-value} \leq 1.0E-20$, one can search the blast_value_exp (which stores the negative E-value exponent) for “blast_value_exp:[5 TO 20]”. To exclude the boundary values from the result list, the user can use “blast_value_exp:{5 TO 20}”. This is equivalent to $1.0E-5 < E\text{-value} \leq 1.0E-20$. When filtering for E-values, there is usually no defined lower bound. This can be reflected using a wild card character “*”. For example, “blast_value_exp:[5 TO *]” searches for all entries with an E-value $\leq 1.0E-5$. Finally, if the sequence store path has been defined (see section “[Installation and Configuration](#)”), the user can enter an amino acid sequence into the search box and select the *Search by Sequence* option with a certain minimum E-value. The software then executes a BLASTP search behind the scenes and returns the top-matching entry accessions (peptide_id field) concatenated by an OR and visualizes summary statistics for homologous proteins.

Drill into Datasets Using Hierarchical Datasets

The *Browse Dataset Pages* are available for several annotation hierarchies including NCBI Taxonomy, Gene Ontology, Enzyme Classification, and KEGG and Metacyc metabolic pathways. For KEGG two different pathway hierarchies can be selected: enzyme based and KO based. The difference is that the enzyme-based version uses enzyme assignments and maps them to EC-based KEGG pathways (a subset of KEGG



METAREP, Overview, Fig. 1 Screenshot of the METAREP Browse Pathway (EC) page

pathways that are mainly related to metabolism), while the KO-based version uses KEGG Orthologs to infer pathway membership and uses a more comprehensive set of pathways including non-metabolic processes such as translation and transcription. The number of hits is displayed for each node in the tree, and a user can click on a tree node and expand further. After clicking a node, a summary of that node is shown in the right panel featuring a pie chart calculated from its sub-nodes and top lists of functional and taxonomic assignments. Once the user has reached the pathway level, for the KEGG versions of the Browse Pathways pages, relative abundance of pathway members is visualized on top of pathway maps (Fig. 1).

Multi-dataset Analysis Options

Compare Feature Abundance Profiles Across Datasets

The *Compare Page* unifies a variety of descriptive, graphical, and statistical analysis options to compare annotation attributes of dozens of datasets. The page features three distinct panels, the *Dataset Select Panel*, the *Filter and Options*

Panel, and the *Results Panel* (Fig. 2). The right upper dataset select box in the *Dataset Select Panel* allows users to select datasets by dragging selected datasets to the left upper panel or by clicking on the plus symbol. The dataset selection can be narrowed down by entering keywords in the search textbox in the left upper panel. The *Filter and Options Panel* provides a textbox to enter a Lucene query (see section “[Dataset Search and Filter Options](#)” and Table 3). If applied, each dataset gets filtered and only annotation entries that match the query are being retained for the comparison. A typical example is to apply a more stringent BLAST E-value baseline.

Another example, highlighted in {REF}, is to filter all datasets for a certain enzymatic marker, e.g., pyruvate dehydrogenase complex (“ec_id:1.2.4.1 OR ec_id:2.3.1.12 OR ec_id:1.8.1.4” or the shorter version “ec_id:1.2.4.1 OR 2.3.1.12 OR 1.8.1.4”). A minimum count value can be entered into the *Min. Count Field* to filter out features whose minimum count across all datasets is equal or higher than the specified count. By default this field is set to 0 showing any features with at least one dataset having a count of one (features with zero counts across all datasets are discarded). The main compare



METAREP, Overview, Fig. 2 Screenshot/conceptual overview of the METAREP Compare Page. Current implementation of the METAREP Compare page (key options are highlighted in green panels). The page allows users to compare absolute and relative abundance of annotations categories across multiple datasets including taxonomic, pathway, enzyme, and GO classifications.

Visualization options include heatmap (shown), hierarchical clustering, multidimensional scaling, and Mosaic Plots. Advanced Compare options include statistical tests for pairwise dataset comparisons (Fisher’s Exact Test, Equality of Proportions Test) as well as for comparing two dataset populations (Wilcoxon Rank-Sum and a nonparametric *t*-test)

options can be selected from the drop down next to the *Min. Count Field* and are organized by the following categories: *Count Matrices*, *Statistical Tests (2 Datasets)*, *Statistical Tests (2 populations)*, and *Plot* options.

The *Results Panel* is automatically updated upon option selection displaying feature-dataset matrices or graphical representations of the same. A certain *annotation attribute* can be selected by clicking on the respective tab and exported using the disk with the green array key. In the following, we describe the option in more detail.

Count Matrices: Applicable if at Least One Dataset Was Selected

Absolute Count Matrix shows a numeric representation of a feature-dataset matrix with cells containing the number of counts for a feature-dataset combination.

Relative Count Matrix shows a numeric representation of a normalized feature-dataset matrix with cells containing the number of counts for a feature-dataset combination divided by the total dataset count. If a filter was entered, the cells represent the number of counts for a feature-dataset combination divided by the total count of the filtered dataset.

Heatmap Count Matrix shows a numeric representation of a row-normalized feature-dataset matrix with cells containing the relative counts per dataset divided by the sum of relative counts per row, i.e., across all datasets. Cells are color coded according to their row-normalized counts. The color scheme can be changed using a drop-down menu.

Statistical Tests (2 Datasets)

The following dataset tests are applicable if two datasets were selected. As input for the tests,

METAREP, Overview, Table 3 METAREP search fields

Field name	Description	Type/range	Example
Core annotation fields			
peptide_id	Peptide ID	text	peptide_id:1120333534885 Retrieve hit with the specified peptide id
com_name_txt	Common name (default field)	text	com_name_txt:phage All hits containing the word phage
com_name_src	Common name source	text	com_name_src:PF00204 All hits having names assigned based on this PFAM hit
go_id	Gene Ontology ID	text	go_id:GO:0000160 Hits with GO:0000160; use “\” before the colon
go_tree	Gene Ontology tree	Integer portion of ID	go_tree:160 Skip “GO:” prefix; all hits with GO:0000160 or lower including all hits with GO IDs that are lower (more specific) in the GO hierarchy
go_src	Gene Ontology source	text	go_src:PF00204 All hits that have GO terms assigned based this PFAM hit
ec_id	Enzyme ID	text	ec_id:5.99.1.3 All hits with Enzyme ID 5.99.1.3
ec_src	Enzyme source	text	ec_src:PF00204 All hits that have EC IDs assigned based on this PFAM hit
ko_id	KEGG Ortholog ID	text	ko_id: K01369
ko_src	KEGG source	text	ko_src: ptr\453118
hmm_id	HMM ID	text	hmm_id:PF00204 All hits that have a PF00204 HMM assignment
library_id	Library ID	text	library_id:GS-00a-01-01-2P5KB All hits that belong to library GS-00a-01-01-2P5KB (helpful to search for library entries within populations)
filter	Any filter tag (e.g., sequence duplicates)	text	filter:duplicate All hits with filter tagged with duplicate -filter:duplicate Exclude entries with filter tag duplicate
Alignment fields			
blast_species	Species	text	blast_species:Chlamydia* All Chlamydia species
blast_tree	Taxonomy	integer (NCBI Taxonomy ID)	blast_tree:2 All bacteria blast_tree:2 Exclude all bacteria
blast_evalue_exp	Negative E-value Exponent	positive integer	blast_evalue_exp:[20 TO *] All hits with BLAST E-value $\leq 10^{-20}$ blast_evalue_exp:[10 TO 20] All hits with $10^{-20} \leq E - \text{value} \leq 10^{-10}$
blast_pid	Percent identity	Float between 0 and 1	blast_pid:[0.9 TO *] All hits with BLAST percent identity $\geq 90\%$ blast_pid:[0.6 TO 0.8] All hits with $60\% \leq \text{percent identity} \leq 80\%$ All hits with $60\% \leq \text{percent identity} \leq 80\%$
blast_cov	Percent sequence coverage	Float between 0 and 1	blast_cov:[0.8 TO *] All hits with BLAST percent sequence coverage $\geq 80\%$ blast_cov:[0.2 TO 0.3] All hits with $20\% \leq \text{sequence coverage} \leq 30\%$

a 2×2 contingency table is generated separately for each feature with two dataset columns and rows representing observations for the presence and absence of the respective feature. As multiple features are simultaneously tested, Bonferroni-corrected p-values and FDR-based q-values are listed, which are recommended over the individual p-values.

Equality of Proportions Test tests whether the relative counts for two features is equal or not. It is equal to the chi-square test of independence in case of a 2×2 contingency table. It is a large sample approximation test, in which the normal distribution is being used to approximate the binomial distribution. Typically, a minimum cell count of five is recommended so that the large sample approximation holds reasonably well. The software accounts for this by automatically setting the *Min. Count Option* to five removing any features from the feature-dataset matrix that have counts lower than five. Results are sorted by ascending p-value. As multiple features are simultaneously tested, Bonferroni-corrected p-values and FDR-based q-values are listed. All three measures can be used to filter the data using the drop-down menu (q-values are being recommended). Result representation and filtering can be applied to any statistical tests described subsequently.

Fisher's Exact Test tests whether the relative counts for two features is equal or not. It is an exact test, in which the null distribution follows a hypergeometric distribution. Thus, it can be used for feature-dataset matrices that contain small cell counts. However, as it is computationally much more intense, execution takes much longer than for the Equality of Proportions Test.

Statistical Tests (2 Populations)

The following population tests are applicable if two populations were selected. A typical scenario for using these tests is to compare two groups of samples, for example, multiple samples taken from *healthy* and *diseased* individuals or from *unfarmed* and *farmed* land. The METAREP administrator has privileges to create populations from the collection of imported libraries via the *Project Page*. As for the two-way dataset tests,

multiple testing is taken into account by providing Bonferroni-corrected p-values and FDR-based q-values which are recommended over the individual p-values.

Wilcoxon Rank-Sum Test performs multiple two-sample nonparametric Wilcoxon rank-sum tests (also known as Mann-Whitney Test) in which each feature is being compared across two populations. It tests whether differences in the medians of the normalized counts for a certain feature are due to chance or not. The null hypothesis states that there is no difference between the dataset-normalized population medians of a feature. The alternative hypothesis states that there is a significant difference between the population medians.

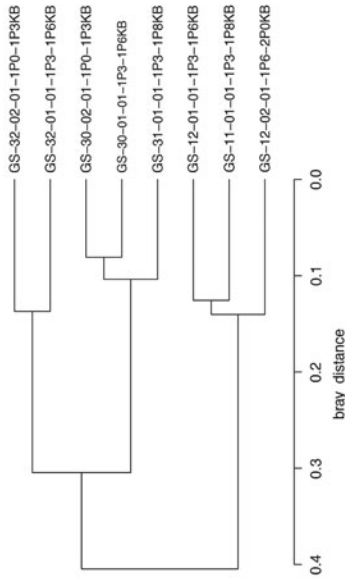
METASTATS is a modified nonparametric *t*-test for detecting differentially abundant features in metagenomic samples (White et al. 2009). The test can be used to compare features across two populations. The null hypothesis states that there is no difference between the dataset-normalized population means of a feature. The alternative hypothesis states that there is a significant difference between the population means. The null distribution approximated via randomization, and a *t*-statistic is being computed for each iteration (see the section “[Installation and Configuration](#)” on how to adjust the number of iterations). For low counts (less than 8), a Fisher's Exact Test is used instead of the non-parametric *t*-test.

Plots

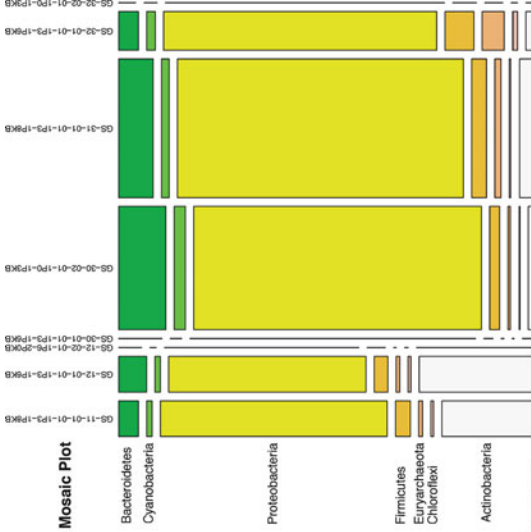
The following plot options are applicable if at least three datasets were selected:

Mosaic Plots draw groups of aligned rectangles, one for each dataset. Features are vertically stacked with the height of a feature (vertical axis) being proportional to the relative count (Fig. 3c). The width of the rectangle (horizontal axis) is proportional to the overall dataset size (compared to the other datasets). Thus, a Mosaic Plot provides a more comprehensive view than a Barplot as it provides a way of visualizing both, the relative feature contribution within datasets and the relative overall dataset size.

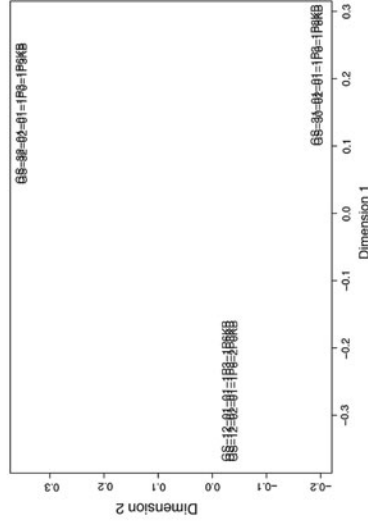
a Hierarchical Clustering Plot



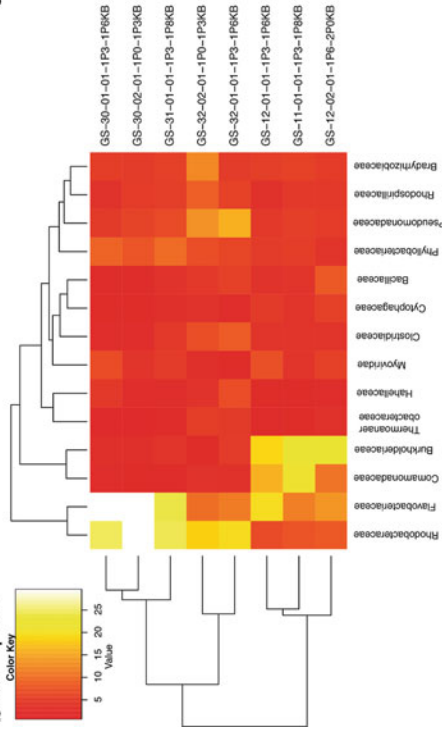
c Mosaic Plot



d MDS Plot



b Heatmap Plot



METAREP, Overview, Fig. 3 (continued)

Hierarchical Cluster Plots provide visual summaries of groups of dataset “clusters” that are similar with respect to their feature composition (Fig. 3a). The input to clustering is a normalized feature-dataset matrix. Here, for normalization, the total feature count across selected features is being used per dataset (note, this is different from the Relative Count Matrix normalization which uses the total dataset count). Distances (dissimilarities) between datasets in multidimensional space can be computed using the feature vectors. Users can choose from various distance metric options including Euclidean, Morisita-Horn, Bray-Curtis, and Jaccard that can be selected available via drop-down menu. After distances have been computed, datasets are clustered using an iterative procedure referred to as hierarchical clustering: initially, each dataset belongs to its own cluster. During each iteration, an optimal cluster pair is being aggregated into a higher-level cluster and distances are recomputed between the new and the remaining clusters. The process continues until there is one single cluster and a tree structure of successive clustering events, a dendrogram, is being drawn (Fig. 3a, b). Users can influence the process of recomputing distances based on several aggregation methods including single linkage (uses minimum distance between the new cluster members and an outside cluster), average linkage (uses average distance), and complete linkage (uses maximum distance). Centroid

uses the mean vector differences, while Ward’s minimum variance minimizes the overall within-cluster variance. For a review see Milligan et al. (1980). According to Milligan et al., the method with the best overall performance has been either average linkage or Ward’s minimum variance. A PDF of the dendrogram and computed distances can be downloaded via the *Results Panel’s* export option.

Heatmap Plots are similar to Hierarchical Clustering Plots in that they visualize datasets as well as feature differences using hierarchical clustering in the form of a dendrogram (Fig. 3b, shown on the right and top, respectively). The main difference is additional quantitative information in the form of a heatmap in which normalized feature-dataset counts are being color coded based on a color gradient. Users can change the base color of the gradient. Columns and rows are reordered to optimize the layout of the two dendrograms on each axis. Hierarchical clustering options include the *Distance Metric* as well as the *Cluster Method*. A PDF of the heatmap and both sets of computed distances can be downloaded via the *Results Panel’s* export option.

Multidimensional Scaling Plots apply non-metric multidimensional scaling to project differences between datasets onto a two-dimensional plane in which similar datasets are closer and less similar datasets area farther apart (Fig. 3d). Like for hierarchical clustering, a dissimilarity matrix based on

METAREP, Overview, Fig. 3 Compare Page plot options exemplified using a selection of eight Global Ocean Survey (GOS) samples. Plots A, B, and D show the same selection of datasets based on organismal composition on the family level (assigned based on the best reference hit using BLAST) with the Minimum Count Option set to 5. Plot C summarizes the same datasets for the phylum level. GS-11 and GS-12 were sampled from the Chesapeake Bay, Annapolis, MD, USA, and Delaware Bay, NJ, USA, respectively. GS-30, GS-31, and GS-G32 were sampled close to the Galapagos Islands (GS-30 off Roca Redonda, GS-31 Fernandina Island, and GS-32

mangrove on Isabella Island). For GS-11, GS-30, and GS-32, two samples were taken from the same location. The hierarchical clustering and heatmap dataset-based dendrograms and the MDS plot show that the replicated samples cluster together. The dendrogram shows that, although the mangrove samples are distinct from the rest of the Galapagos Islands, they are more close to each other when compared to the East Coast samples. The heatmap shows an increase of Rhodobacteraceae (*orange* to *white*) and a decrease in Comamonadaceae and Burkholderiaceae families (*orange* to *red*) when comparing these two groups based on the % abundance level

the normalized counts is used as input for the algorithm, which can be specified by the user. A PDF of the final heatmap and the computed distances can be downloaded via the *Results Panel's* export option.

Homology Searches

The *BLAST Sequence Page* provides functionality to screen multiple datasets for a protein sequence of interest (Fig. 4). Highly conserved single-copy marker genes, such as *dnaG*, for example, can be used to approximate the number of genomes in a dataset (Wu and Eisen 2008). The page uses the same “Select Datasets” panel as the *Compare Page*. BLAST options include the input sequence text area, BLAST Min. E-value, and a text field for entering a filter query. The *Result Panel* summarizes BLASTP alignment results filtered for homologous entries that match the filter query in different formats that can be selected by choosing one of three tabs. The Annotation Tab lists key alignment statistics along with annotations of homologous entries. The Alignment Tab displays the default BLASTP alignment output including textual representation of sequence alignments. The Tabular Tab tabulates the default tab-delimited BLASTP output (-m8 BLASTP option). Results for each of these tabs can be downloaded via the *Results Panel's* export option. To activate this option, protein sequences of each dataset need to be formatted using the BLAST utility program `formatdb` and organized in a sequence store on the Web server that runs the METAREP instance (see section “[Installation and Configuration](#)”).

Installation and Configuration

METAREP utilizes a variety of open-source software including R, Lucene/Solr, CAKEPHP, MySQL, Apache Http server, and SQLite that need to be downloaded and installed. Version 1.4.0 of METAREP can be downloaded at <https://github.com/jcvi/METAREP/zipball/1.4.0-beta>. For installation instruction please visit

<https://github.com/jcvi/METAREP/wiki/Installation-Guide-v-1.4.0>. For later versions, please visit the Project Page at <https://github.com/jcvi/METAREP/wiki>.

As part of the data import process, additional annotation attributes including NCBI Taxonomy lineage, GO assignments, and KEGG pathways are fetched from a SQLite database. The database can be updated using the `scripts/perl/metarep_update_database.pl` script. To update the KEGG attributes, the script needs to be pointed to a local snapshot of KEGG downloaded from the KEGG FTP site (license is required).

Once installation is completed, the instance can be configured modifying the “`app/config/metarep.php`” file. An important configuration that impacts performance and stability is the number of Solr/Lucene servers used for retrieving annotation information. While METAREP can be run in a setup with a single server (`SOLR_MASTER_HOST`), for best performance and stability, we recommend running a second server (slave), on another machine. The additional server can be configured using the `SOLR_SLAVE_HOST` variable. In theory, more than two slave servers can be defined, but METAREP currently supports only one slave server. A two-server setup can handle more concurrent traffic than a single server and thus can improve the average query response time (an important factor if many users are anticipated to access data simultaneously). The two Solr/Lucene servers will replicate data across the two different machines, and user traffic is balanced between the two servers using an internal load-balancing mechanisms implemented in the Web-logic component of the METAREP software. The slave server using Solr's inbuilt replication functionality will automatically replicate new index files that have been uploaded to the master server. If one server goes down (for maintenance, testing, malfunction, etc.), the other server can still handle user requests. The two-server system is thus also more fault tolerant and enables updates to the server without interfering with the user experience.

Select Datasets

4 items selected Remove all project datasets all datasets Add all

- library:MMETSP0022-20110809 *Micromonas sp. CCMP2099 is an abundant, pan-arcti...
- library:MMETSP0023-20110809 *Micromonas sp. CCMP2099 is an abundant, pan-arcti...
- library:MMETSP0024-20110809 *Micromonas sp. CCMP2099 is an abundant, pan-arcti...
- library:MMETSP0025-20110809 *Micromonas sp. CCMP2099 is an abundant, pan-arcti...
- library:MMETSP0032-20110527 Representative of prasinophyte clade
- population: Micromonas-CCMP2099-Control Micromonas-CCMP2099-+
- population: Micromonas-CCMP2099-Starvation Micromonas-
- library:MMETSP0032-20110809 Representative of prasinophyte clade
- library:MMETSP0034-20110809 Representative of prasinophyte clade

Enter Sequence

>PFAM00011.1
 DNSYITMAASGIGQELMAETRASQIVAEARIGRGRNQARVEAQQIDSTRKQDE
 EENWALGGGEGENAAALQAEINWQWGLGRNYSQAVDVAIRCCVSEVETA
 RYVATQKMGIXIVYRKMSEVIX

Options

Help Min. E-Value BlastP Update

Result Panel

Annotation Alignment Tabular zoom in zoom out

Peptide ID	Common Name	Blast Species	EC ID	HMM	KEGG Ortholog	Dataset	% Identity	E-Value	Bit Score
MMETSP0022-20110809 6152_1	hypothetical protein	Selaginella moellendorffii	3.6.3.14	PF03179 TIGR01147	K02152	MMETSP0022-20110809	36.00	1e-11	66.2
MMETSP0024-20110809 5198_1	hypothetical protein	Selaginella moellendorffii	3.6.3.14	PF03179 TIGR01147	K02152	MMETSP0024-20110809	36.00	1e-11	66.2
MMETSP0025-20110809 6029_1	hypothetical protein								
MMETSP0023-20110809 5224_1	unknown transcript								

Result Panel (Zoomed)

Annotation Alignment Tabular zoom in zoom out

```

>MMETSP0025-20110809|6029_1 len=180
Length = 180
Score = 66.2 bits (160), Expect = 1e-11
Identities = 36/100 (36%), Positives = 57/100 (57%)
Query: 10 ESOTGIGELMAETRASQIVAEARIGRGRNQARVEAQQIDSTRKQDEEENWALGGGEGENAAALQAEINWQWGLGRNYSQAVDVAIRCCVSEVETA
Sbjct: 77 DSDGIGKLLAVQEQAGQIVAAARAEKTLRQAKAEADAEIAAYRAQREKYGQLVLSQ 136
Query: 70 GGSEGNAAALQAEINWQWGLGRNYSQAVDVAIRCCVSEVETA
G AA L+A* ++ + Q N + ++L K
Sbjct: 137 TGDSTRARLEADCTAQIATVYVQVQANKKIVTRGLAQK 176

>MMETSP0024-20110809|5198_1 len=180
Length = 180
Score = 66.2 bits (160), Expect = 1e-11
Identities = 36/100 (36%), Positives = 57/100 (57%)
Query: 10 ESOTGIGELMAETRASQIVAEARIGRGRNQARVEAQQIDSTRKQDEEENWALGGGEGENAAALQAEINWQWGLGRNYSQAVDVAIRCCVSEVETA
+E GIG+*A E A TVA AR + R+QAK EA I +YRA+++E++ + LSG
  
```

METAREP, Overview, Fig. 4 Screenshot of the BLAST Sequence Page. A protein sequence of interest can be searched against a selection of datasets. BLAST results

can be displayed and exported in various formats including annotation (shown), alignment (shown in the zoom panel), and tabular

The INTERNAL_EMAIL_EXTENSION variable can be specified to identify internal users that register with the instance and set permissions accordingly. By default, users that register with the specified e-mail extension are granted full data access. The GOOGLE_ANALYTICS_TRACKER_ID and GOOGLE_ANALYTICS_DOMAIN_NAME variables configure the instance to synchronize Web usage with Google Analytics to track usage statistics.

The NUM_METASTATS_BOOTSTRAP_PERMUTATIONS variable sets the number of replicates to determine the null distribution for the METASTATS test and can be increased to

increase the precision of the p-values (see (White et al. 2009) for details).

To activate the METAREP blast functionality, searching and exporting of sequences, the SEQUENCE_STORE_PATH variable needs to be defined. This path points to the location on the Web server where the formatdb-formatted protein sequence files are kept (organized by project ID and dataset, Table 4). The perl/scripts/metarep_format_sequence.pl utility can be used to format sequence data according to this format. If an FTP server is available, data sharing of a collection of custom files per dataset the via dataset download option

METAREP, Overview, Table 4 METAREP sequence and FTP data organization

Feature	Root	Project directory	Dataset directory	Files
Sequence export and BLAST functionality	Sequence store root directory	12	GS695_GDQ27C301_0p1	GS695_GDQ27C301_0p1. phr GS695_GDQ27C301_0p1. pin GS695_GDQ27C301_0p1. psd GS695_GDQ27C301_0p1. psi GS695_GDQ27C301_0p1. psq formatdb.log
		12	GS695_GLDFQNX02	GS695_GLDFQNX02_viral/ GS695_GLDFQNX02_viral. phr GS695_GLDFQNX02_viral. pin GS695_GLDFQNX02_viral. psd GS695_GLDFQNX02_viral. psi GS695_GLDFQNX02_viral. psq formatdb.log
FTP export functionality	FTP root directory	12	GS695_GDQ27C301_0p1. tgz	
		12	GS695_GLDFQNX02.tgz	

can be activated by specifying the `FTP_HOST`, `FTP_USERNAME`, and `FTP_PASSWORD` variables. The software identifies FTP data by looking for the project ID folder followed by a tar-gzipped file that has a matching dataset name, i.e., `<dataset-names>.tgz`.

Example Hardware Configurations

The main requirements are driven by the amount of annotations that are to be stored in index files and served by a Solr/Lucene server. The main impact on performance for a single machine is the amount of memory available for result retrieval, caching, and operating systems for file caching. If annotations are weighted, i.e., the weight field is set to values other than 1, the CPU requirements increase (see (Goll et al. 2012), Fig. 6). We are currently running a two-server system that is served by two load

balanced Dell Power Edge R710 servers each having eight cores (2.66 GHz), 72 G RAM, and 2×600 GB HD. So far we have successfully indexed 190 M. Our HMP METAREP instance that serves over 400 million weighted annotations entries runs on a single server with two multi-threaded Xeon X7560 2.26 GHz processors with a total of 16 cores (32 threads), 256 G RAM, and 4 terabyte of disk space. For performance benchmarks, please see Goll et al. (2010), Supplementary Fig. 1, and Goll et al. (2012), Fig. 6.

Additional Resources

As part of the NIH HMP project, the software was tested with short-read annotations derived from over 14 trillion Illumina reads (Goll et al. 2012). The study includes several scenarios on how to

investigate the NIH human microbiome data including how to analyze specific metabolic markers, cluster datasets based on their metabolic profile, and identify pathways that are differentially abundant between human body habitats. The data can be accessed at www.jcvi.org/hmp-metarep. The following short video tutorial summarizes key functionality (YouTube ID:7FPJaPyLjMk). The METAREP home page at www.jcvi.org/metarep provides an anonymous login via the “Try It” button to evaluate the latest functionality for a collection of ocean samples taken from the North Pacific Subtropical Gyre (DeLong et al. 2006). The open-source code of the software and developer information including how to contribute to the open-source project is available at the project’s source code repository at <https://github.com/jcvi/METAREP>. For questions and comments, please join the mailing list at www.jcvi.org/metarep or directly send an e-mail to metarep@googlegroups.com.

References

- Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*. 2012;8:e1002358. doi:10.1371/journal.pcbi.1002358.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.
- Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*. 2011;12:356. doi:10.1186/1471-2105-12-356.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25–9. doi:10.1038/75556.
- DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. Community genomics among stratified microbial assemblages in the ocean’s interior. *Science*. 2006;311:496–503. doi:10.1126/science.1120250.
- Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11:R86. doi:10.1186/gb-2010-11-8-r86.
- Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, et al. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics*. 2010;26:2631–2. doi:10.1093/bioinformatics/btq455.
- Goll J, Thiagarajan M, Abubucker S, Huttenhower C, Yooseph S, et al. A case study for large-scale human microbiome analysis using JCVI’s Metagenomics Reports (METAREP). *PLoS ONE*. 2012;7:e29044. doi:10.1371/journal.pone.0029044.
- Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Res*. 2003;31:371–3.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68:669–85. doi:10.1128/MMBR.68.4.669-685.2004.
- Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012;486:215–21. doi:10.1038/nature11209.
- Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*. 2004;32:D235–9. doi:10.1093/nar/gkh117.
- Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, et al. IMG/M-HMP: a metagenome comparative analysis system for the human microbiome project. *PLoS ONE*. 2012;7:e40151. doi:10.1371/journal.pone.0040151.
- Meyer F, Paarmann D, D’Souza M, Olson R, Glass EM, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 2008;9:386. doi:10.1186/1471-2105-9-386.
- Milligan, Glenn W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*1980;45(3):325–342.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2012;40:D290–301. doi:10.1093/nar/gkr1065.
- Stein LD. The case for cloud computing in genome informatics. *Genome Biol*. 2010;11:207. doi:10.1186/gb-2010-11-5-207.
- Sun S, Chen J, Li W, Altintas I, Lin A, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res*. 2011;39:D546–51. doi:10.1093/nar/gkq1102.
- Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, et al. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand Genomic Sci*. 2010;2:229–37. doi:10.4056/signs.651139.
- White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*. 2009;5:e1000352. doi:10.1371/journal.pcbi.1000352.
- Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*. 2008;9:R151. doi:10.1186/gb-2008-9-10-r151.

MetaTISA: Metagenomic Gene Start Prediction with

Huaiqiu Zhu¹ and Gangqing Hu²

¹Department of Biomedical Engineering, and Center for Theoretical Biology, Peking University, Beijing, China

²Systems Biology Center, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA

Synonyms

Gene start annotation; Translation initiation site (TIS) prediction

Definition

Gene start: the start position from which a genomic sequence can be translated into protein.

Introduction

Knowledge of exact information of gene start plays an important role in identification of native purified proteins from the high-throughput proteomics (Poole et al. 2005). In addition, a correct prediction of gene start facilitates the identification of cis-regulatory signals related to translation initiation (Hu et al. 2008c) and thus facilitates the understanding of the diversity and evolution scenario of translation initiation mechanisms (Zheng et al. 2011). However, gene start annotation in widely used public databases such as GenBank and RefSeq is not of high quality in general (Nielsen and Krogh 2005). In particular, the longest open reading frame is frequently used to annotate a protein-coding gene (Besemer et al. 2001), which results in a systematical low quality in gene start annotations for GC-rich species (Nielsen and Krogh 2005; Hu et al. 2008b). Therefore, accurate gene start prediction has been an intensive research subject for many labs in the last decade.

In recent years, gene start prediction for microbial genomes has achieved high accuracies for a number of methods (Besemer et al. 2001; Zhu et al. 2004; Tech et al. 2005; Delcher et al. 2007; Makita et al. 2007; Hu et al. 2008a, 2009; Hyatt et al. 2010). Most of the methods are unsupervised and can be roughly sorted into two groups based on whether specific assumptions are made on gene start-related features. The first group involves statistic models specifically designed for the cis-regulatory signals in the vicinity of gene start such as the Shine-Dalgarno (SD) signal (Shine and Dalgarno 1974). Assumptions are then made regarding the length of the signal, the start codon usage, and the distances between the signal and start codon (Besemer et al. 2001; Zhu et al. 2004; Delcher et al. 2007; Makita et al. 2007; Hu et al. 2008a; Hyatt et al. 2010). These methods show consistently high prediction accuracies on a number of genomes such as *E. coli* and *B. subtilis*. However, the assumptions that apply to these genomes may not apply to others. The other methods build statistic model to characterize the whole sequences around gene starts and do not take specific assumptions on gene start-related genomic features. Tech et al. (2005) introduced a second-order Markov model with positional smoothing to characterize sequence properties around gene start and achieved comparable accuracies to other methods. This method however is criticized for potential dependency of the quality of initial annotation (Makita et al. 2007). Later on, Hu et al. (2009a) introduced a classification of putative start codons into three categories based on evolutionary pressures acting on the sequences: true start codons (purifying selection), false start codons in intergenic regions (minimal sequence feature preserved under neutral selection), and false start codons in coding regions (period-three oscillations in sequence content under purifying selection) (Hu et al. 2008b). The sequence feature of each group is then characterized by a non-homogeneous Markov model, and an iterative nonsupervised procedure is utilized for parameter estimations (Hu et al. 2009b). The method achieves a better accuracy than other

methods, and the prediction is independent from the quality of initial annotation (Hu et al. 2009b).

Since many of the metagenomics projects involve high-throughput proteomics to identify novel proteins followed by experimental validation, the development of gene start prediction algorithm for metagenomic fragments receives increasing attentions (Hoff et al. 2008). It is important to realize that although the current methods are successful in gene start prediction for microbial genomes, they are not directly applicable to the metagenomic projects (Hu et al. 2009a). This is largely caused by the fragmentary nature of the metagenomics sequences and their uncertainties in phylogenetic origins. A tool called MetaTISA (Hu et al. 2009a) was implemented to address this question.

“Binning Followed by Self-Training”

MetaTISA is essentially a sequential application of metagenomics binning – a process that identifies from what species a particular sequence has originated – and an unsupervised procedure for gene start prediction within each bin:

1. **Binning:** giving a set of metagenomics fragments, each fragment was assigned to a genus based on its k -mer nucleotide frequencies as described in (Sandberg et al. 2001). Briefly, a metagenomic fragment F of size l consists of $l-(k-1)$ overlapping motifs M of size k . The probability of finding fragment F in genus G_i , denoted by $p(F|G_i)$, can be estimated by a product of the probabilities of finding each motif M in genus G_i , which can be estimated from the normalized k -mer nucleotide frequencies within genus G_i . Based on Bayesian statistics, giving the occurrence of fragment F , the probability that F belongs to G_i may be expressed as $p(G_i|F) = [p(F|G_i)p(G_i)]/P(F)$, where $P(F)$ is the probability of finding fragment F , which is independent of genus, and $p(G_i)$ is a prior probability that reflects the relative abundance of genus G_i in the metagenomic sample of concern. MetaTISA assumes that the prior probability is equal

among genera and then assigns the phylogenetic origin of fragment F to the genus that reports the maximal value of $p(F|G_i)$. Since the k -mer frequencies are pre-calculated for each genus G_i , it is crucial to keep the parameters updated to maintain the classification accuracy especially when novel genera are discovered.

2. **Unsupervised gene start predictions:** fragments assigned to the same genus are supposed to have close phylogenetic origin and share a similar mechanism of translation initiation. In this regard, gene start prediction methods developed for microbial genomes may be applied. MetaTISA utilizes the methods described in Hu et al. (2009b) to accomplish with several considerations. Firstly, it trains the parameters for each genus in an unsupervised manner (also known as self-training). This offers the advantage to exclude the needs of a set of known training sets. However, for genus that receives only a few number of fragments (<200 by default), the prediction does require pre-computed parameters. But note that the parameters training for this genus is also a nonsupervised process (Hu et al. 2009b). Secondly, the predication is independent from the quality of input. For metagenomic samples, the quality of gene start prediction from a gene annotation pipeline may vary considerably across fragments bins with different GC content. Thirdly, the method estimates the probability that a putative start codon is within coding regions. This helps tell the completeness of a coding sequence within a fragment. Lastly but not the least, it outputs genus-specific parameters that may facilitate the comparison of TIS-related signals among different metagenomic samples (Noguchi et al. 2008).

Prediction Accuracies

MetaTISA is designed as a post-processor for gene prediction pipelines currently available for

metagenomes, such as MGA (Noguchi et al. 2008), GeneMark.hmm (Zhu et al. 2010), and Glimmer-MG (Kelley et al. 2012). The improvements brought by MetaTISA are demonstrated by post-processing gene predictions from MGA on metagenomic fragments simulated using 100 genomes. Two kinds of simulations with different fragment sizes are conducted: 400 bp for 454 or 700 bp for Sanger. When assessed on experimentally verified datasets, the sensitivities are improved by 6–8 % without a loss of specificities regardless of the choice of fragment length (Hu et al. 2009a). An indirect way of accuracy assessment on real metagenomic samples is to investigate the TIS-feature-associated parameters self-trained for each genus. As an example, the method is applied to post-process MGA's predictions for Human Gut Community Subject 7, and as a result it reveals expected RBS patterns such as SD signals for genus within *Firmicutes* (Hu et al. 2009a).

Availability

The tool is written in C++ and the source code is freely available under GNU GPL license. A web server (<http://mech.ctb.pku.edu.cn/MetaTISA/>) is dedicated for the user to run the program online and to receive the results by email. The web server also provides downloading service for source codes, files for pre-computed parameters, and executable version for Windows and Linux platforms.

Summary

By a sequential combination of metagenomic fragments binning and a self-training of parameters within each bin, MetaTISA significantly improves the identification of gene starts for metagenomes. Noteworthy, this “binning-followed-by-self-retraining” scheme has been successfully applied to the prediction of protein-coding sequences for metagenomes (Kelley et al. 2012).

Cross-References

- ▶ [FragGeneScan: Predicting Genes in Short and Error-Prone Reads](#)
- ▶ [MetaBin](#)

References

- Besemer J, Lomsadze A, et al. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 2001;29(12):2607–18.
- Delcher AL, Bratke KA, et al. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics.* 2007;23(6):673–9.
- Hoff KJ, Tech M, et al. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics.* 2008;9:217.
- Hu G, Liu Y, et al. New solutions of translation initiation site prediction for prokaryotic genomes. *Prog Biochem Biophys.* 2008a;35(11):1254–62.
- Hu G, Zheng X, et al. Computational evaluation of TIS annotation for prokaryotic genomes. *BMC Bioinformatics.* 2008b;9:160.
- Hu G, Zheng X, et al. ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. *Nucleic Acids Res.* 2008c;36(Database issue):D114–9.
- Hu G, Guo J, et al. MetaTISA: Metagenomic Translation Initiation Site Annotator for improving gene start prediction. *Bioinformatics.* 2009a;25(14):1843–5.
- Hu G, Zheng X, et al. Prediction of translation initiation site for microbial genomes with TriTISA. *Bioinformatics.* 2009b;25(1):123–5.
- Hyatt D, Chen GL, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11:119.
- Kelley DR, Liu B, et al. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 2012;40(1):e9.
- Makita Y, de Hoon MJ, et al. Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. *BMC Bioinformatics.* 2007;8:47.
- Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics.* 2005;21(24):4322–9.
- Noguchi H, Taniguchi T, et al. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 2008;15(6):387–96.
- Poole 2nd FL, Gerwe BA, et al. Defining genes in the genome of the hyperthermophilic archaeon

- Pyrococcus furiosus*: implications for all microbial genomes. *J Bacteriol.* 2005;187(21):7325–32.
- Sandberg R, Winberg G, et al. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.* 2001;11(8):1404–9.
- Shine J, Dalgarno L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A.* 1974;71(4):1342–6.
- Tech M, Pfeifer N, et al. TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics.* 2005;21(17):3568–9.
- Zheng X, Hu G, et al. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics.* 2011;12:361.
- Zhu H, Hu G, et al. Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics.* 2004;20(18):3308–17.
- Zhu W, Lomsadze A, et al. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38(12):e132.

Metaxa, Overview

Johan Bengtsson-Palme¹, Martin Hartmann²,

K. Martin Eriksson³ and R. Henrik Nilsson³

¹Institute of Neuroscience and Physiology,

The Sahlgrenska Academy, University of Gothenburg, Göteborg, Sweden

²Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Zurich, Switzerland

³Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden

Synonyms

16S extraction; rRNA extraction; SSU extraction; Taxonomic assignment

Definition

Metaxa is a software tool for extracting full-length and partial ribosomal small subunit (SSU; 16S/18S/12S) sequences from metagenomic datasets and for classifying the

extracted sequences to taxonomic domains and organelle of origin. Metaxa is freely available from <http://microbiology.se/software/metaxa/>.

Introduction

A common question in metagenomic studies concerns the species composition of the community sampled (Desai et al. 2012). This is frequently addressed using a specific genetic marker, typically the ribosomal RNA (rRNA) small subunit (SSU) gene sequence (also referred to as the 16S, 18S, or 12S subunit depending on the lineage under scrutiny). In some studies, the SSU gene is amplified by PCR and sequenced separately in order to study microbial diversity. However, even if the SSU sequences are not targeted for separate sequencing, it is still possible to identify and extract the SSU component of a metagenome. This task has traditionally been carried out through similarity searches against sequence databases such as GenBank (Benson et al. 2009), SILVA (Pruesse et al. 2007), GreenGenes (DeSantis et al. 2006), or RDP (Cole et al. 2007).

The complexity of the data requires frequent manual intervention to accurately sort out the origin of the sequences in such BLAST-based approaches, and the process is further complicated by the fact that the SSU gene is found not only in the core genome of bacteria, archaea, and eukaryotes but also in the chloroplasts and mitochondria of eukaryote organisms. These different gene copies, although often very similar to one another, are non-orthologous and should in most cases not be analyzed jointly. Metagenomic efforts are generally interested in the bacterial and/or eukaryote diversity in the sample, and thus any mitochondrial or chloroplast SSU sequences, bearing high similarity to bacterial SSU genes, may confound the analysis if left in the dataset. To avoid noise and bias associated with analyzing non-orthologous sequences as if they were orthologous, the sequences must be subjected to manual inspection, which is a time-consuming process further complicated by the

large number of incorrectly identified or poorly annotated reference sequences in the public sequence databases (Bidartondo 2008; Hartmann et al. 2011).

Metaxa (Bengtsson et al. 2011) is a software package that resolves the problem of extracting and sorting SSU sequences to origin in an accurate and rapid way. The end result is a set of FASTA files, each representing all SSU sequences from a particular organelle or taxonomic domain, for further analysis of species composition or other endeavors.

Methods

Extraction

The rRNA SSU gene is composed of eight to nine hypervariable (“V”) regions flanked by more conserved domains (Hartmann et al. 2010). Metaxa carries out the extraction of SSU sequence fragments from the metagenome using the HMMER package (Eddy 2010) and Hidden Markov Models (HMMs) representing the most conserved parts of the SSU gene, chiefly at the 5′ and 3′ end of each V region. These HMMs are modeled according to the same principles as those of V-Xtractor (Hartmann et al. 2010). Since the Metaxa models represent a set of highly conserved domains, false-positive matches can be all but avoided as only high-scoring profile matches are considered. Metaxa features HMM profiles representing the archaeal and bacterial 16S genes, the eukaryote 18S gene, the mitochondrial 12S and 16S genes, and the chloroplast 16S gene. These sets of HMM profiles enable Metaxa to identify and distinguish among all these classes of SSU sequences.

Classification

After extracting all SSU sequences from the query dataset, Metaxa proceeds to classify the extracted SSU sequence fragments. This is performed by comparing each fragment to a carefully selected set of reference SSU sequences from GreenGenes, SILVA, CRW (Cannone et al. 2002), and MitoZoa (Lupi et al. 2010) using BLAST (Altschul et al. 1997).

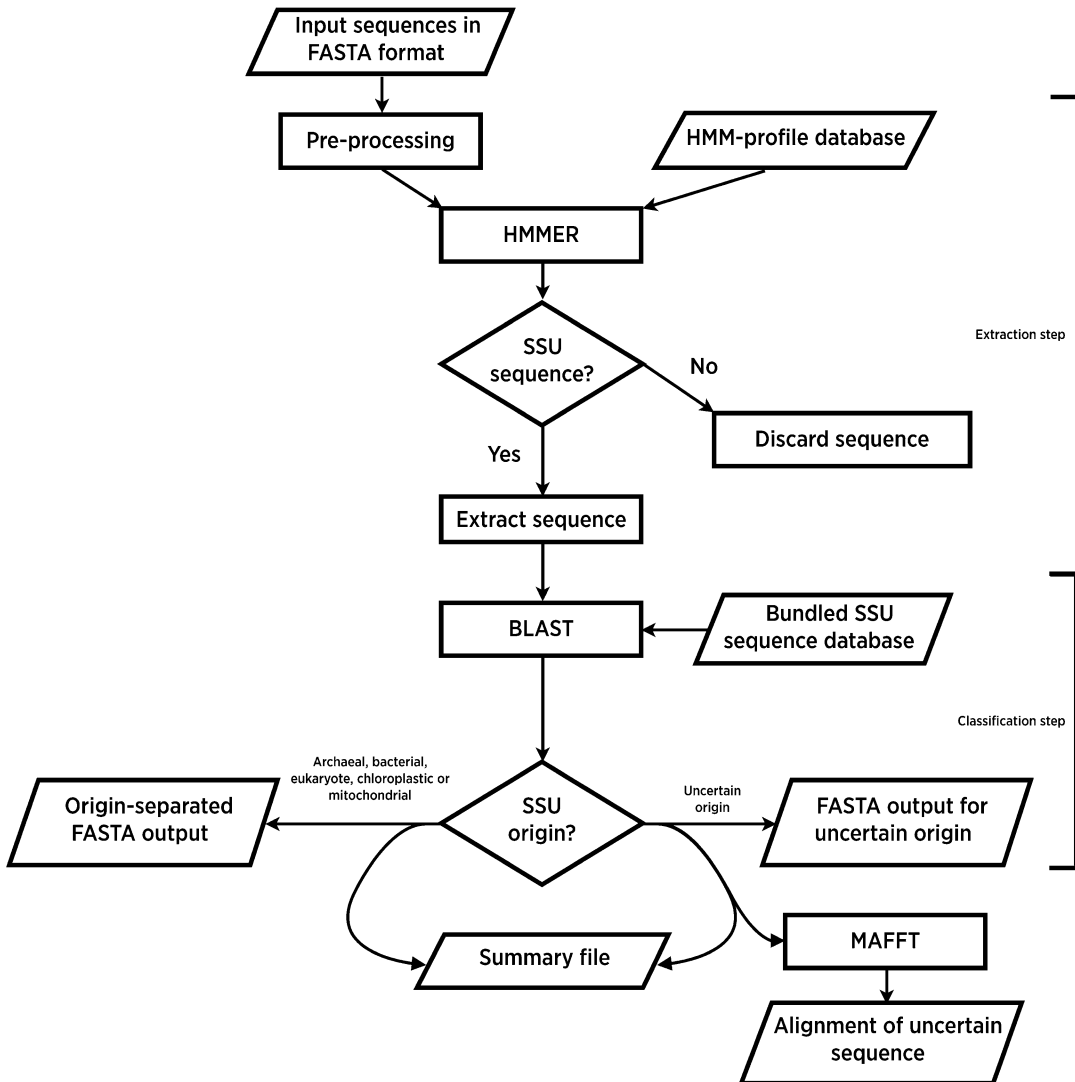
By default, the five highest-scoring BLAST matches are examined for origin in terms of organelle or taxonomic domain, and each origin is given a score based on the number of sequences among the top five BLAST hits that belong to the respective origin. The matches to the HMM profiles in the previous step are weighted together with these BLAST-based origin scores to make a final call on the most likely origin of the sequence fragment. In cases where the origin cannot be determined with certainty, but where there is a strong candidate, Metaxa assigns the sequence to the most likely origin, but flags it as potentially in need of manual inspection. If scores for origin are tied altogether, the sequence is assigned into a special “uncertain” bin. In the two latter cases, sequence alignments of the extracted fragment and the five best BLAST matches are computed automatically using MAFFT (Katoh and Toh 2008), to assist the user in the interpretation process.

Input and Output

Metaxa takes input in the FASTA format and outputs one FASTA file for each origin found. Optionally, Metaxa can also produce output in table format. The entire running process is outlined in Fig. 1.

Performance

Metaxa has been shown to classify more than 99.95 % of the core-release sequences in the SILVA database according to their annotated origin, and it has a false-positive rate of 0.00012 % (Bengtsson et al. 2011). When evaluated on simulated metagenomic data comprising three sets of 100,000 sequences with fragment lengths of 1,000, 300, and 100 bp, Metaxa processed the datasets in 112, 47, and 35 min, respectively, with very high accuracy down to typical 454 read lengths (300 bp), retaining fidelity for bacterial sequences even at read lengths as short as 100 bp (Fig. 2). This suggests that Metaxa is highly reliable for Sanger, as well as 454-derived, metagenomes, and that it is useful even on metagenomes generated using short-read



Metaxa, Overview, Fig. 1 Overview of the Metaxa running process

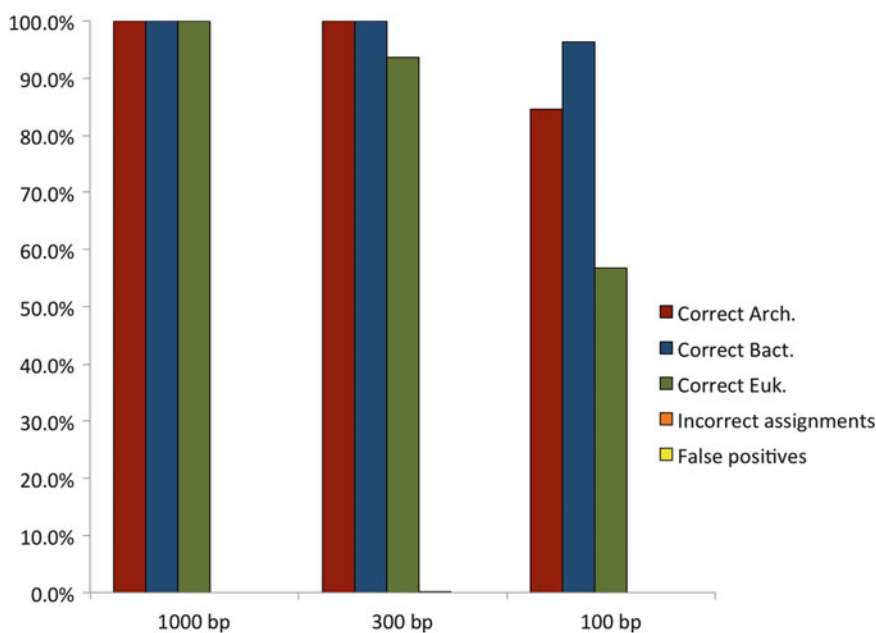
sequencing technologies, such as Illumina. Metaxa takes advantage of multiple processor cores, if available, and it has no software or hardware restriction on the number of input sequences.

Applications

Metaxa has obvious uses in deriving taxonomic inferences from metagenomic sequence sets.

For example, a set of sequences extracted using Metaxa could be used for sequence diversity analysis. However, because of the classification capabilities of Metaxa, it is also useful in sorting out PCR-amplified SSU libraries before continuing with species richness investigations such as rarefaction or species accumulation analysis. Here, the ability of Metaxa to separate chloroplast and mitochondrial SSU sequences from other SSU entries is crucial for the accuracy of the downstream analysis. Metaxa could also be





Metaxa, Overview, Fig. 2 Performance of Metaxa at different read lengths

used as a tool to verify the authenticity of annotations in SSU sequence databases and reference libraries.

Availability

Metaxa is written in Perl and released as an open-source package under the GNU GPL v. 3 license. It runs on Unix and Linux platforms, including Mac OS X. The software package can be freely downloaded from <http://microbiology.se/software/metaxa/>.

Summary

Metaxa is a high-performance software tool for extracting and classifying SSU sequences from metagenomic datasets. The accuracy of the software is very high, providing high sensitivity toward SSU fragments even at short-read lengths while maintaining a false-positive rate of about 0.00012 %. Metaxa is fast compared to, e.g., BLAST, and it takes advantage of multiple

processor cores where available. It can be used as a tool for taxonomic analysis of metagenomes as well as a classification tool for SSU amplicons. Metaxa is freely available from <http://microbiology.se/software/metaxa/>.

Cross-References

- ▶ [Microbial Diversity, Bar-Coding Approaches](#)
- ▶ [Phylogenetics, Overview](#)
- ▶ [Silva Databases](#)

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17): 3389–402.
- Bengtsson J, Eriksson KM, Hartmann M, Wang Z, Shenoy BD, Grelet G-A, Abarenkov K, Petri A, Alm Rosenblad M, Nilsson RH. Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and

- chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*. 2011;100(3):471–5.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2009;37(Database issue):D26–31.
- Bidartondo MI. Preserving accuracy in GenBank. *Science (New York)*. 2008;319(5870):1616.
- Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*. 2002;3:2.
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*. 2007;35(Database issue):D169–72.
- Desai N, Antonopoulos DA, Gilbert JA, Glass EM, Meyer F. From genomics to metagenomics. *Curr Opin Biotechnol*. 2012;23(1):72–6.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72(7):5069–72.
- Eddy S. HMMER. <http://hmmer.janelia.org> (2010). Accessed 2012-05-15.
- Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. V-Xtractor: an open-source, high-throughput software tool to identify and extract hyper-variable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *J Microbiol Method*. 2010; 83(2):250–3.
- Hartmann M, Howes CG, Veldre V, Schneider S, Vaishampayan PA, Yannarell AC, Quince C, Johansson P, Björkroth KJ, Abarenkov K, Hallam SJ, Mohn WW, Nilsson RH. V-REVCOMP: automated high-throughput detection of reverse complementary 16S rRNA gene sequences in large environmental and taxonomic datasets. *FEMS Microbiol Lett*. 2011;319(2):140–5.
- Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 2008;9(4):286–98.
- Lupi R, de Meo PD, Picardi E, D'Antonio M, Paoletti D, Castrignanò T, Pesole G, Gissi C. MitoZoa: a curated mitochondrial genome database of metazoans for comparative genomics studies. *Mitochondrion*. 2010; 10(2):192–9.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. 2007;35(21): 7188–96.

Microbial Diversity, Bar-Coding Approaches

James A. Foster

Department of Biological Sciences, Institute for Bioinformatics & Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA

Introduction

Amplicon fingerprints are useful for ecological studies of microbial communities. Most studies to date have used these techniques for determining how many species are present (richness, or alpha diversity) in what ratios (beta diversity), which populations or species are present, and what metabolic or ecological functions the community and its constituents may provide. These data inform downstream analyses to determine the response of microbial ecosystems to environmental change, the relationship between human microbiota and health, the ecological succession, the co-evolutionary constraints within and between communities and their environments, and more (Foster et al. 2012a).

This encyclopedia entry focuses on bacterial fingerprinting, since it has a longer history and is more mature than fingerprinting techniques for other kingdoms of life. But these techniques are in principle applicable to all microbial organisms, including archaea and eukarya such as fungi, diatoms and tiny arthropods, and viruses (assuming they are organisms). Amplicons for bacteria have been in use since the beginning of the molecular revolution and their gene products have been well characterized. However, potential amplicons exist for all organisms. As dominant bacterial life is on Earth, it is by no means the only microbial realm of interest. Nonetheless, it is the focus of this entry.

The terminology herein is taken from the bacterial ecology literature. A *population* is a collection of individuals of the same type. In sexual organisms, a population is typically a collection of individuals from the same species.

In asexual organisms, however, the species concept is problematic. In any case, one may be interested in discriminating to a subspecific or strain level, or indeed to higher levels. Thus, the definition of a population is relative to the specific question under investigation. A *community* is a collection of co-occurring populations. Therefore, the number of distinct populations in a community is the richness of that community. The diversity of a community includes the relative abundance of populations and their potential interactions.

Amplicon fingerprinting techniques have developed in tandem with new sequencing technologies. Current fingerprinting approaches are particularly well adapted to modern high-throughput sequencing and have largely replaced older techniques based on electrophoresis or capillary sequencers. The older approaches are still useful for crude estimates using older, and therefore inexpensive and less used, equipment. However, as the cost of new sequencing technologies drops, more modern amplicon fingerprinting approaches are likely to continue to replace their predecessors.

Amplicon fingerprinting techniques are *culture independent*, meaning that it is unnecessary to grow cultures of individual populations or communities before extracting DNA. This is particularly significant in the microbial world, since most bacteria and archaea cannot currently be grown in the lab. Estimates show that as much as 97 % of existing microbial biodiversity is currently uncultivable (Whitman et al. 1998). These techniques enable ecological and functional analysis of communities that largely consist of otherwise inaccessible “biotic dark matter.”

Choosing Amplicons

With bacteria, the amplicon of choice has long been the gene for the small RNA subunit of the ribosome, known as 16S rDNA for its size (16 Svedberg units). Nearly universal primers exist for several regions of this gene. The secondary structure of 16S rDNA is well characterized

and highly conserved, providing a reliable guide for fast and accurate alignment of large sets of sequences (Nawrocki et al. 2009). This gene is strongly conserved, since it is a critical part of the replicative machinery in bacteria (and some archaea). So it is in principle useful for recognizing deep phylogenetic divergences. And finally the 16S rDNA gene shows little evidence of horizontal transfer, which makes it more useful as a phylogenetic marker. Woese and Fox first demonstrated the utility of 16S rDNA analysis with their discovery that archaea are a distinct kingdom of life (Woese 2004; Woese and Fox 1977).

Several hypervariable regions in the 16S rDNA gene provide enough sequence variation to distinguish bacterial populations, sometimes to the strain level. Hypervariable regions typically contain loops in the rRNA secondary structure, which change more as species evolve, since they are not as structurally constrained as stems. Reliable primers exist for nine regions, known as V1 through V9, that were short enough to be completely sequenced by Sanger sequencing when the primers were developed (Kim et al. 2011). Hypervariable regions differ in the specificity and precision with which they can distinguish different types of organisms, so the choice of amplicon primers is study specific (Schloss 2010; Bazinet and Cummings 2012). As newer sequencing technologies have increased the length of genetic fragments that can be sequenced, it has become standard practice to amplify from one end of one region to an end of another region. For example, V35 and V69, which span regions 3–5 and 6–9, respectively, are common in the literature.

Since it has become possible to sequence much larger fragments, it has become common to attach “bar code adapters” to primers. This makes it easier to multiplex samples from several different experimental treatments into single sequencing runs and then separate the data algorithmically. In theory, one could improve resolution of fingerprinting techniques by multiplexing several primers for multiple hypervariable regions, as if fingerprinting multiple fingers at the same time. However, most projects currently work with only single sets of primers. However,

very soon it will be feasible to sequence the entire 16S rDNA gene, which of course will comprise all hypervariable regions, making the choice of primers irrelevant for microbial community fingerprinting. An intriguing possibility will be to multiplex fingerprinting from multiple genes that expand analysis beyond the bacterial kingdom, for example, multiplexing 16S rDNA and 18S rDNA amplicons.

Databases of full 16S rRNA sequences exist for hundreds of thousands of microbes (Cole et al. 2007; DeSantis et al. 2006). A typical workflow searches these databases for putative homologues to amplicons. The annotations for these hits then inform likely taxonomic and functional associations (Kuczynski et al. 2010).

But modern databases have serious limitations. It is rarely possible to classify bacteria below the family level, since there are vastly more different populations than have been observed. As cultivation-independent sequencing methods grow more popular, new sequences in the databases tend to be from unclassified, and therefore unannotated, populations. Annotations in existing databases are highly biased toward pathogenic or other human-associated organisms. Very closely related genera, species, and strains can differ dramatically in their metabolic potential and preferred ecological habitats. Finally, different species vary widely in their 16S rDNA copy numbers, making it easy to confuse dosage effects and within-individual sequence variation with species abundances.

Other genetic targets may serve the same function as 16S does for microbial ecology, provided they exhibit sufficient variation, stability, and vertical inheritance. For example, the RNA polymerase β -subunit gene, *rpoB*, is a single-copy gene and has been recommended as an alternative to 16S rDNA. Other highly conserved house-keeping genes such as cytochrome B (*cytB*), those responsible for electron transport in aerobic organisms, may be more appropriate for plant studies or deep resolution of *Cyanobacteria*. And of course, eukaryotes and some Archaea do not have 16S ribosomal subunits, so a more appropriate gene is their small subunit analogue, the 18S rDNA gene. Currently, these alternatives

do not have databases comparable to those available for 16S rDNA and have fewer useful primers.

No fingerprinting technique based on a single gene, however, carefully chosen, can hope to distinguish all microbes or fully elucidate all microbial metabolic and ecological functions. Even when it becomes feasible to routinely sequence entire 16S rDNA genes from individual cells, the gene-based amplicon analysis will only produce gene genealogies rather than organismal phylogenies or full metabolic profiles. Multiplexing amplicon processing for several genes may improve phylogenetic resolution. But as it becomes feasible to sequence entire genomes for whole communities with shotgun metagenomics or single-cell genomics, it will become unnecessary to choose target amplicons at all.

Fingerprinting Techniques

Fragment-based techniques use the length of amplicon fragments as fingerprints. The spectra of these lengths indicate which microbial populations were in the original sample, assuming that there is sufficient variation in the amplicon fragments. We present the three most common fragment-based techniques here.

Temperature gradient and denaturing gradient gel electrophoresis (TGGE and DGGE) separate the DNA fragments by size using standard gel electrophoresis (Fischer and Lerman 1979). The resulting band patterns are then the community fingerprints. Presumably, more complex patterns represent more complex communities and patterns from distinct populations contributing additively to the overall pattern so that one can decompose the community fingerprint into constituent populations.

Automated ribosomal intergenic spacer analysis (ARISA) determines the spectra of the intergenic spacer region (ITR) between small and large ribosomal subunit genes in bacteria (Fisher and Triplett 1999). The flanking genes are highly conserved, making ITS a reasonable amplicon. Moreover, the length ITS is highly

variable between bacterial species, so a spectrum of ITS lengths is a reasonable fingerprint.

Terminal restriction fragment length polymorphism (TRFLP) analysis binds fluorescent markers to the amplicon PCR primers before restriction, marking the restriction fragments adjacent to the primer (Schütte et al. 2008). One can then separate the labeled fragments by size, for example, in a capillary sequencer. The spectra of the lengths of these fragments are then the fingerprint for the study sample.

All three length-based fingerprinting techniques have inherent biases and limitations, and all three are still commonly used. A PubMed search on 12 July 2012 for the terms “DGGE,” “ARISA,” and “TRFLP” returned 5658, 119, and 107 hits, respectively, with several recent citations indicating current use of all three techniques.

Bioinformatics has been critical for interpreting fragment-based amplicon fingerprint data. A common approach has been to perform *in silico* analyses of existing databases, to determine length spectra for known sequences. This provides a kind of “reverse telephone book” with which one can translate empirical fingerprints into possible population compositions. Two typical tools for this sort of analysis, focused on TRFLP and still in heavy use, are the Microbial Community Analysis (MiCA) suite and the TFLP Analysis Program (TAP-TRFLP) (Shyu et al. 2007; Cole et al. 2009).

Sequence-based fingerprinting techniques use the amplicon sequences themselves as fingerprints, rather than their length spectra. Current sequencing technologies, also known as next-generation sequencing, have made it feasible to sequence millions of amplicons in a single run. Different sequencing technologies vary in their sequencing accuracy, typical type of sequencing errors, and length of amplicon (Foster et al. 2012b). Consequently, the vast majority of current amplicon fingerprinting projects use amplicon sequences rather than derived data such as lengths.

Bioinformatics to analyze sequence-based fingerprints is a very active area of research. New and improved algorithms are constantly

emerging for cleaning and quality control of raw data, detecting erroneous sequences (such as chimeras), aligning sequences, clustering fingerprints by similarity, searching for similar annotated sequences in existing databases, and more. Two software packages aggregate state-of-the-art algorithms and pipelines to bring the state of the art to the typical user, namely, Quantitative Insights Into Microbial Ecology (QIIME) and MOTHUR (Caporaso et al. 2010; Schloss et al. 2009). Both packages are compatible with most computing platform and are updated regularly with the newest algorithms from the research community. Both have extensive tutorials and reference documentation. MOTHUR is open source. Both packages perform most standard diversity analyses and produce datasets that can be imported into the R statistical environment for further analysis (Beck et al. 2011).

To summarize, amplicon choice remains important to fingerprinting analyses, though fragments of the 16S rDNA gene remain the amplicon of choice for bacterial community diversity studies. Amplicon sequences are becoming the fingerprints of choice, though derived data such as length spectra for restriction fragments or interspacer regions are still widely used. Future sequencing technologies are sure to change the fingerprinting landscape significantly. Finally, amplicon fingerprinting analysis requires extensive bioinformatic support, and appropriate tools are available.

Cross-References

- ▶ [Culture Collections in the Study of Microbial Diversity, Importance](#)
- ▶ [Metagenomics, Metadata, and Meta-analysis](#)
- ▶ [Microbial Ecology in the Age of Metagenomics: An Introduction](#)
- ▶ [New Computational Methodologies to Understand Microbial Diversity](#)
- ▶ [Next-Generation Sequencing for Metagenomic Data: Assembling and Binning](#)
- ▶ [Protein-coding Genes as Alternative Markers in Microbial Diversity Studies](#)

References

- Bazin AL, Cummings MP. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*. 2012;13(1):92. doi:10.1186/1471-2105-13-92.
- Beck D, Settles M, Foster JA. OTUbase: an R infrastructure package for operational taxonomic unit data. *Bioinformatics (Oxford, England)*. 2011;27(12):1700–1. doi:10.1093/bioinformatics/btr196.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Noah F, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6. doi:10.1038/nmeth.f.303.
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM, et al. The Ribosomal Database Project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*. 2007;35:D169–72. doi:10.1093/nar/gkl889.
- Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*. 2009;37:D141–5. doi:10.1093/nar/gkn879.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72(7):5069–72. doi:10.1128/AEM.03006-05.
- Fischer SG, Lerman LS. Length-independent separation of DNA Restriction Fragments in two-dimensional gel electrophoresis. *Cell*. 1979;16(1):191–200.
- Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol*. 1999;65(10):4630–6.
- Foster JA, JH Moore, Gilbert JA, Bunge J. Microbiome studies: analytical tools and techniques. In: Russ B Altman, A Keith Dunker, Lawrence Hunter, Teri E Klein (eds), *Pac Symp Biocomput*. 2012a;200–2. World Scientific, Singapore .
- Foster JA, Bunge J, Gilbert JA, Moore JH. Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life. *Brief Bioinform*. 2012b. doi:10.1093/bib/bbr080.
- Kim M, Morrison M, Yu Zhongtang. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. 2011;84(1):81–7. doi:10.1016/j.mimet.2010.10.020
- Kuczynski J, Liu Z, Lozupone C, McDonald D. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods*. 2010;7(10):813–9. doi:10.1038/nmeth.1499.
- Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*. 2009. doi:10.1093/bioinformatics/btp157.
- Schloss PD. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol*. 2010. doi:10.1371/journal.pcbi.1000844.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537–41. doi:10.1128/AEM.01541-09.
- Schütte UME, Abdo Z, Bent SJ, Shyu C, Williams CJ, Pierson JD, Forney LJ. Advances in the use of Terminal Restriction Fragment Length Polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Appl Microbiol Biotechnol*. 2008;80(3):365–80. doi:10.1007/s00253-008-1565-4.
- Shyu C, Soule T, Bent SJ, Foster JA, Forney LJ. MiCA: a web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphisms of 16S and 18S rRNA genes. *J Microb Ecol*. 2007;53(4):562–70.
- Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*. 1998;95(12):6578–83.
- Woese CR. A new biology for a new century. *Microbiol Mol Biol Rev MMBR*. 2004;68(2):173–86. doi:10.1128/MMBR.68.2.173-186.2004.
- Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*. 1977;74(11):5088–90.

Microbial Ecology in the Age of Metagenomics: An Introduction

Jianping Xu

Department of Biology, McMaster University, Hamilton, ON, Canada

Introduction

Microbial ecology is an interdisciplinary science related to microbiology and ecology. Its investigations range from analyzing the diversity of microorganisms within and among the different ecological niches on Earth to understanding the interrelationships among microorganisms, between microorganisms and macroorganisms, and between microorganisms and their abiotic environmental factors. Microbial diversity and

the interactions between microbes and other organisms can be analyzed at morphological, structural, physiological, and/or genetic levels. The recent advances in high-throughput technologies, especially in genome sequencing, are reshaping our understandings of microbial ecology. This entry introduces the fundamental concepts and issues in microbial ecology, with a brief focus on how metagenomics tools are impacting microbial diversity studies.

Microorganisms and Microbiology

A microorganism refers to any life form that can't be easily seen by the human naked eye. Microorganisms encompass morphologically, structurally, and phylogenetically very diverse forms of life and traditionally include both acellular life forms such as viruses and cellular life forms in all three domains, the Bacteria, Archaea, and Eukarya (Woese 1987). Organisms in Bacteria and Archaea are completely microbial. Even in Eukarya, macroorganisms such as animals and plants represent only parts of two of at least eight superkingdoms within this domain, while the remaining six or more superkingdoms are exclusively microbial (Baldauf 2003). While most microorganisms can't be seen at all by the naked eye, for many microorganisms, certain stages of their life cycles can be easily visualized. For example, mushrooms, the sexual reproductive structure of certain groups of fungi, are a common occurrence on forest floors at certain times of the year.

Microorganisms were first seen and described by Antonie van Leeuwenhoek in 1676 when he used a microscope to examine a variety of natural and human-made objects. Subsequent developments in methodologies for growing, purifying, and studying microorganisms ushered in a golden era of microbiology, which is still going strong today. Microorganisms have now been found in virtually every habitable niche on Earth, from hot springs to salt lakes, from frozen environments in the Antarctica and glaciers at the top of mountains to

hydrothermal vents at the bottom of deepest oceans. Current estimates put the number of microbial cells on Earth at around 5.0×10^{30} , about eight orders of magnitude greater than the number of stars in the observable universe. Indeed, despite their small sizes, the large number of microbial cells on Earth makes microorganisms the single largest carbon sink, more than those from plants and animals. Their large number, broad ecological distribution, and vast diversity of metabolic pathways unparalleled by macroorganisms make microbes indispensable and central to our considerations of global geochemical cycles and environmental issues.

Most of the early methodologies for studying microorganisms are still widely used today, and many discoveries about the fundamental features of life were made using microorganisms as model systems. Among the many practical contributions of microbiology, microbiological discoveries have significantly impacted (and are continuing to impact) the control and prevention of diseases in plants, animals, and humans. However, techniques and methodologies alone were insufficient for establishing microbiology as a fledging field of scientific investigation. Reductionist approaches and guidelines for hypothesis testing such as the Koch's postulates for identifying the causative agents of infectious diseases were pivotal for the development of microbiology. Interestingly, with the rapid developments both in high-throughput experimental tools (e.g., Xu 2014) and in bioinformatics software capable of analyzing large and diverse datasets, holistic views about microorganisms are beginning to attract significant scientific attention. Indeed, aside from the traditional subdisciplines such as microbial cell biology, biochemistry, physiology, genetics, ecology, and evolutionary biology, microbiology now also includes microbial genomics, systems microbiology, microbial community ecology, and ecosystem microbiology. In addition, the diverse subdisciplines of microbiology have become integral components of agriculture, forestry, animal husbandry, fishery, mining, environmental sciences, and medicine.

Microbial Ecology

Broadly speaking, microbial ecology is the scientific discipline that examines the relationships between microorganisms and their environments. Ecologically oriented studies of microbes were performed as soon as their existence was realized. However, the term microbial ecology came into frequent use only in the early 1960s, and its emergence as an independent field of investigation was propelled by both the awakening public interest in environmental issues and the increasing recognition of the essential roles of microbes in Earth's geochemical cycling and in human welfare.

At present, microbial ecological investigations can be grouped into three broad types: (i) identifying the taxonomic, structural, and functional diversities of microorganisms in natural ecological niches; (ii) analyzing the relationships among microorganisms, between microorganisms and macroorganisms (plants and animals including humans), and between microorganisms and environmental factors (such as nutrients, temperature, pH, pressure, oxygen); and (iii) investigating the mechanisms that generate and maintain the diversity of microorganisms and their relationships with each other and with their biotic and abiotic factors in natural environments. Among the three types of research activities, most metagenomics studies of microbial ecology have focused on microbial diversity in natural environments.

Below is a brief introduction to metagenomics and how metagenomics approaches have shaped our understanding of microbial diversity. For the impact of metagenomics tools on the other two aspects of microbial ecology, please refer to other entries in this encyclopedia.

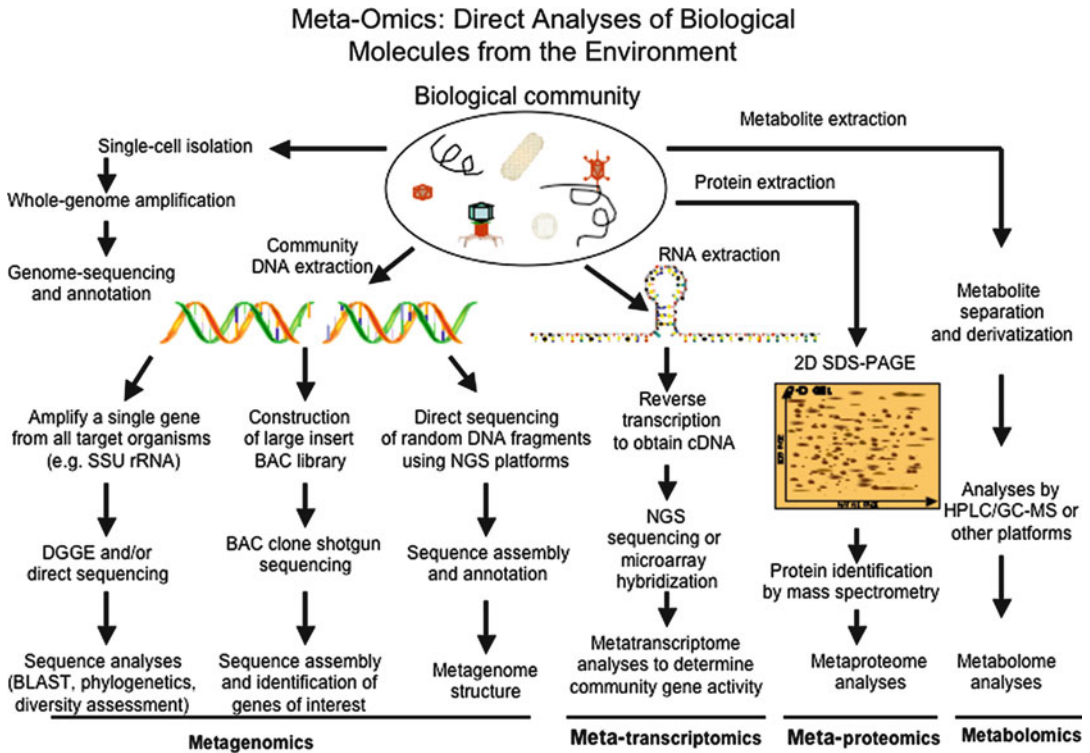
Metagenomics

Metagenomics refers to the field of study that analyzes genetic materials obtained directly from environmental samples. Several other terms, such as environmental genomics, ecological genomics, and community genomics, have emerged over the

years to describe direct analyses of environmental DNA (Marco 2009). However, metagenomics has emerged as the favorite term and the prefix "meta-" is now used to describe the direct analyses of environmental RNA, proteins, and metabolites, corresponding respectively to meta-transcriptomics, meta-proteomics, and metabolomics (Fig. 1). Together, the direct analyses of biological molecules from natural environments constitute the field of "meta-omics" (Fig. 1).

The different subfields of meta-omics analyze complementary sets of biological molecules directly from the environments that together help provide holistic views of the natural biological communities. For example, analyses of environmental DNA samples can provide estimates of the taxonomic and genome diversities of organisms in ecological niches in nature, the extracted RNA, protein, and metabolites provide information about the functions of the environmental genomes, including the degrees to which genes are transcribed and translated, and the types and amount of metabolites are generated in natural ecological niches. In addition, to properly analyze and integrate the diverse biological datasets, effective "meta-programs" are also needed and several such programs are currently available (de Bruijn 2011).

Because biological materials (e.g., different types of microbial cells) can be very different from each other in terms of their size, morphology, and structure, obtaining DNA (and/or RNA, protein, and metabolites) directly from environmental samples that can realistically reflect their native biological states may require extensive sample treatments. Such treatments may include sorting biological samples (including different types of cells and viral particles) based on sizes, removing materials that inhibit downstream reactions, and applying different extraction methods that permit the lysis of cells with specific types of cell walls. Once the pools of targeted biological materials are obtained, additional treatments of these materials may be needed before they are channeled into high-throughput analytical platforms. Below is a brief overview of the applications of metagenomic tools on estimates of microbial diversity.



Microbial Ecology in the Age of Metagenomics: An Introduction, Fig. 1 Legend: an overview of meta-omics: the direct analyses of biological molecules such as DNA, RNA, protein, and metabolites using high-

throughput technologies. To effectively utilize such data, suits of “meta-programs” are required to analyze and integrate the diverse meta-datasets (Modified from Xu 2010)

Estimates of Microbial Genetic Diversity Using Metagenomic Data

Depending on the objectives of research, microbial diversity in the environment can be expressed as a quantitative measure using several common indices such as phylogenetic diversity, species diversity, genotype diversity, gene diversity, and nucleotide diversity. Above the species level, microbial diversity can be quantified based on evolutionary distances among the observed taxonomic groups from a specific environment. Below the species level, microbial diversity can be described using population genetic parameters such as nucleotide diversity, gene diversity, and genotype diversity. Nucleotide diversity, gene diversity, and genotype diversity refer respectively to the probability that two randomly drawn bases at a specific site of the genome, alleles of a specific gene locus, and genotypes in

a population will be different (Xu 2010). At the species level, microbial diversity is measured as species diversity. There are various measures of species diversity. One commonly used refers to the frequency that two randomly drawn individuals in an environment will be different species. This measure takes into account both the number of species (species richness) and the frequency of each species (species abundance) in the environment. Conceptually, this measure of species diversity is similar to those used for nucleotide diversity, gene diversity, and genotype diversity.

Microbial species diversity is among the most commonly analyzed and compared in microbial ecological studies. The earliest and still one of the most common metagenomics methods for estimating species diversity of prokaryotes (including both Bacteria and Archaea) in natural environments is the direct analyses of sequence variation at the 16S ribosomal RNA gene

(Pace et al. 1985). These analyses may involve the polymerase chain reaction (PCR), denaturing gradient gel electrophoresis (DGGE), cloning, and sequencing. A broadly accepted criterion to delineate prokaryote species is that two strains belong to the same species if their 16S rRNA genes show $\geq 97\%$ sequence similarity (de Bruijn 2011). In eukaryotic microbes such as fungi, a similar criterion ($\geq 97\%$ sequence similarity) is often used, albeit for a different DNA fragment, the internal transcribed spacer (ITS) regions of the ribosomal RNA gene cluster (Schoch et al. 2012). However, in more recent analyses, direct sequencing of extracted environmental DNA using NGS technologies is increasingly used. These analyses suggest that the cultured microbes from most ecological niches represent $< 1\%$ of the true microbial species richness in their respective niches and that many of these uncultured microbes belong to distinct and previously unknown phylogenetic groups (de Bruijn 2011). Metagenomic analyses, especially those based on NGS technologies (Xu 2014), have generated very large datasets from environments including the human body (e.g., the human microbiome initiative; <http://nihroadmap.nih.gov/hmp/>) and the oceans (the Global Ocean Sampling surveys; <http://www.jcvi.org/cms/research/projects/gos/overview/>). Scientists from many countries participate in these large-scale projects.

The species diversity studies based on DNA sequences at the 16S rRNA gene are increasingly complemented by other types of data that augment our understanding of microbial diversity in natural environments. One type of such data is genetic variation among strains within a species. With high-throughput DNA sequencing, genetic variants of a gene fragment from different strains of the same species in the same ecological niche can be reliably identified (de Bruijn 2011). With sufficient genome coverage, it's also possible to uncover genome variants. Such information allows direct comparisons of gene frequencies and genotype frequencies among microbial populations from diverse ecological niches, including the inferences of the modes of reproduction in nature (Xu 2010). The second type of

complementary data is the messenger RNA sequences obtained from environmental samples. In combination with DNA sequence data, the mRNA data allow inferences of the potential physiological activities of the different groups of microorganisms in natural environments (de Bruijn 2011).

Summary

This entry serves as an introduction to microorganisms, microbiology, microbial ecology, and metagenomics. The impact of metagenomics on estimates of microbial diversity was briefly discussed. With the increasing application of high-throughput technologies in analyzing biological materials (DNA, RNA, proteins, and metabolites) directly from environments, the future of microbial ecology is looking brighter than ever.

Cross-References

- ▶ [Microbial Diversity, Bar-Coding Approaches](#)

References

- Baldauf SL. The deep roots of eukaryotes. *Science*. 2003;300:1703–6.
- Bruijn F. Handbook of molecular microbial ecology I: metagenomics and complementary approaches. New Jersey: Wiley/Blackwell; 2011. p. 113–22.
- Marco D. Metagenomics: theory, methods and applications. Norfolk: Caister Academic Press; 2009.
- Pace NR, Stahl DA, Olsen GJ, Lane DJ. Analyzing natural microbial populations by rRNA sequences. *ASME News*. 1985;51:4–12.
- Schoch CL*. The Fungal Barcode Consortium (one of 100 collaborators). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A*. 2012;109:6241–6.
- Woese CR. Bacterial evolution. *Microbiol Rev*. 1987;51:221–71.
- Xu J. Microbial population genetics. Norfolk: Caister Academic Press; 2010.
- Xu J. Next-generation sequencing: technologies and applications. Norfolk: Caister Academic Press; 2014.

Microbial Ecosystems, Protection of

Paul L. E. Bodelier
Netherlands Institute of Ecology
(NIOO-KNAW), Wageningen, Netherlands

Synonyms

Conservation of microbial diversity and ecosystem functions provided by microbes; Preservation of microbial diversity and ecosystem functions provided by microbes

Definition

The use, management, and conservation of ecosystems in order to preserve microbial diversity and functioning.

Introduction

Ecosystems collectively determine biogeochemical processes that regulate the Earth system. Loss of biodiversity is generally regarded as detrimental to ecosystems and ecosystem functioning and therefore has been a central issue for environmental scientists during the last decades (Hooper et al. 2012). Microorganisms (i.e., bacteria, archaea, protozoa, and fungi) comprise a major part of the total biomass of organisms inhabiting on Earth and represent the largest source of biodiversity. They play critical roles in biogeochemical processes and ecosystem functioning and are fundamental to many ecosystem services (e.g., soil health, wastewater treatment, nutrient recycling, human health, carbon sequestration, etc.) (see Table 1) (Ducklow 2008). Considering the challenges we are facing with overexploitation of the planet, climate change, pandemics, increasing demands in food production, and need for renewable energy and resources, it is remarkable that microbes and their diversity are absent in the ongoing debates about global biodiversity loss and conservations

policy, despite various pleas to do so (Cockell and Jones 2009). The biodiversity-ecosystem function (BEF) research inherently requires the investigation of the relationship between species assemblies and ecosystem processes, a link which is difficult to make with microbes. High diversity, rapid generation times, high adaptability due to genome rearrangements, and ubiquitous distribution have led to the notion that microbial communities are highly redundant and omnipresent and therefore inextinguishable. However, the latter is a misconception driven by a number of gaps in our understanding of the functioning of microbial communities and the relevance of microbial diversity in ecosystem functioning.

Knowledge Gaps in Understanding Microbial BEF

Definition of Species

Considering the IUCN Red List of species and the associated criteria to get on this list (<http://www.iucn.org/>), it is quite obvious that microbes have not made it in there yet. A species is a fundamental unit of biological organization, but its relevance for microbes is debated. The inability to define taxonomic units equivalent to animal and plant species is also one of the most fundamental problems hampering the study of the BEF matter in microbial communities. The first problem is that the isolation and cultivation of microbes in order to assess their geno- and phenotype has led to the description of only 7,000 species whereas DNA based-methods have identified more than 100 prokaryotic phyla to be present in ecosystems (Pace 2009). Hence, approx. only 1 % of the actual microbial biodiversity is represented as cultured organisms while the characteristics and functions of the remaining 99 % are unknown. Next to this, bacterial taxonomy employs universal thresholds of DNA-sequence difference to help demarcate species. However, the sequence-identity cutoff value used to demarcate species has led to “species” that are enormously diverse in their genome content physiology and ecology. Hence, what is

Microbial Ecosystems, Protection of, Table 1 Major groups of microbes and ecosystem services they provide. The last column depicts the ecosystem service category as was defined in the Millennium Ecosystem Assessment 2005

Microbial group	Process	Ecosystem service	Ecosystem service category
Heterotrophic bacteria/Archaea	Organic matter breakdown, mineralization	Decomposition, nutrient recycling, climate regulation, water purification	Supporting and regulating
Photoautotrophic bacteria	Photosynthesis	Primary production, carbon sequestration	Supporting and regulating
Chemo(litho) autotrophic	Specific elemental transformations (e.g., NH_4^+ , S_2^- , Fe_2^+ , CH_4 oxidation)	Nutrient recycling, climate regulation, water purification	Supporting and regulating
Unicellular phytoplankton	Photosynthesis	Primary production, carbon sequestration	Supporting and regulating
Archaea	Specific elemental transformation (e.g., metals, CH_4 formation, NH_4^+ oxidation), often in extreme habitats.	Nutrient recycling, climate regulation, carbon sequestration	Supporting and regulating
Protozoa	Mineralization of other microbes	Decomposition, nutrient recycling, soil formation	Supporting
Fungi	Organic matter breakdown and mineralization	Decomposition, nutrient recycling, soil formation, primary production (i.e., mycorrhizal fungi)	Supporting
Viruses	Lysis of hosts	Nutrient recycling	Supporting
All	Production of metabolites (e.g., antibiotics, polymers), degradation of xenobiotics, genetic transformation and rearrangement	Production of precursors to industrial and pharmaceutical products	Provisional
All	Huge diversity, versatility, environmental and biotechnological applications	Educational purposes, getting students interested in science	Cultural

From Bodelier 2011

regarded as a “species” in microbiology would definitely not be comparable to species in macroecology (animals and plants), and commonly the term “operational taxonomic units” is used in microbiology. The situation will improve due to better cultivation methods and insights, resulting in increased coverage of phylogenetic lineages with cultured representatives. Next to this, metagenomic, metaproteomic, and even single-cell genomic techniques enable the characterization of functions of not-yet-cultivated organisms in their environment (Raes and Bork 2008). These novel techniques will facilitate the development and application of novel concepts in environmental microbiology which may bridge the gap with macroecology, bypassing the species hang-up in order to develop generic concepts and theories in microbial ecology. Recently, the

concept of ecological coherence of taxa higher than the species level was put forward, which suggests that deeper clades of various ranks may be used as alternative ecologically meaningful units in microbial ecology (Philippot et al. 2010). With the vast amount of metagenomic data available from an increasing variety of environments and the advent of comparative genomics, the field of microbial ecology is undergoing a paradigm shift away from tax-oriented concepts of community analysis that have been inherited from macroorganism ecology toward trait-centered and/or systems biology-oriented approach in which functional units (protein-coding genes, enzymes, metabolites) are the key components of the overall ecosystem (Green et al. 2008). Using functional traits and environmental gradients can bring general

patterns into community ecology, and a trait-centered perspective would be a tractable way for microbial ecology to address the significance of microbial diversity for ecosystem functioning. Considering the fact that in plant sciences BEF studies are also incorporating traits rather than species richness only, the trait-centered approach may offer options for convergence of macro- and microbial ecology which will be essential for including microbes in conservation policy.

Lack of Microbial Biogeography?

The conventional view of microbial distribution of species through space and time has been dominated for decades by the “Baas-Becking” hypothesis “everything is everywhere, but the environment selects.” The lack of dispersal limitations of microorganisms would ensure a global distribution, but that local deterministic factors would determine the relative abundance of “latent” and “flourishing” species. This view is in sharp contrast with plants and animals which show clear taxa-area relationships and biogeography. The Baas-Becking legacy is likely one of the main reasons why microbial diversity is not on the biodiversity-conservation agenda. However, in the last decade there are a number of studies demonstrating species-area relationships, biogeography, and spatial patterns at various scales for microbes (see Zhou et al. 2008). Next to this, microbial endemism has been reported as well, while studies using high-throughput sequencing technology clearly demonstrated the presence of habitat-specific communities shaped by edaphic factors and historical contingencies. A meta-analysis of all currently available 16S rRNA gene sequences revealed clear environmental distributions on the genus or species level with soil and freshwater as least selective habitats, while marine, animal, and thermal habitats were the most selective (Tamames et al. 2010). The emerging pattern in microbial biogeography studies is definitely that not all microbial communities occur everywhere and that local conditions can lead to unique associations of microbes. However, whether microbes obey the same distribution and community assembly rules as macroorganisms can only be

answered when it is possible to study complete microbial populations at ecologically relevant scales.

Inability to Link Species Diversity to Function

Connecting individual microbial species to the biogeochemical processes they catalyze is a prerequisite for assessing BEF relationships in microbial communities. However, considering the lack of a species concept, the metabolic versatility, the large number of unknown species, and the scale issue involved, this is the central problem area in the field of environmental microbiology. The majority of studies in the literature have relied on correlating changes in activity to changes in community composition or diversity, and only a few articles can actually show a causal relationship. A myriad of techniques have been developed for linking diversity and function (see Wagner 2009). However, many of these techniques were based on the analyses of ribosomal RNA or mRNA transcripts of functional genes, indicating only the potential to be involved in specific processes. The use of stable isotope probing (SIP) has evoked a major breakthrough in environmental microbiology (see Murrell and Whiteley 2011). The general approach is that stable isotopically ($^{13}\text{C}/^{15}\text{N}$) labeled substrates are incorporated into taxonomically relevant molecules (RNA/DNA, lipids, proteins). Only the microbes which have actively been incorporating the stable isotopes are detected when analyzing RNA/DNA or PLFA using GC-IRMS (gas chromatography-isotope ratio mass spectrometry) or proteins using GC-MS or LC-MS (liquid chromatography-mass spectrometry). The major disadvantages of SIP are the use of unnaturally high substrate concentrations in case of DNA- and RNA-based SIP, the different label uptake rates per species, and cross feeding. More recent work brought improvements in the shortcomings of traditional SIP studies by using magnetic bead capturing of mRNA, Raman spectroscopy, and NanoSIMS (secondary ion beam mass spectrometry) (see Murrell and Whiteley 2011) also in combination with metagenomic techniques uncovering active species of which no cultured representatives are available or discovering

unknown pathways or genes involved in biogeochemical processes (see Chen and Murrell 2010). The most recent addition to the SIP repertoire combined microarray detection and NanoSIMS, attaining low label incorporation levels and high phylogenetic resolution without PCR amplification of the target community (Mayali et al. 2012). The challenge in applying SIP-based techniques will be in BEF experiments, where experimental designs allowing for causal and mechanistic conclusions require high sample throughput.

Resistance, Resilience, and Redundancy of Microbial Communities

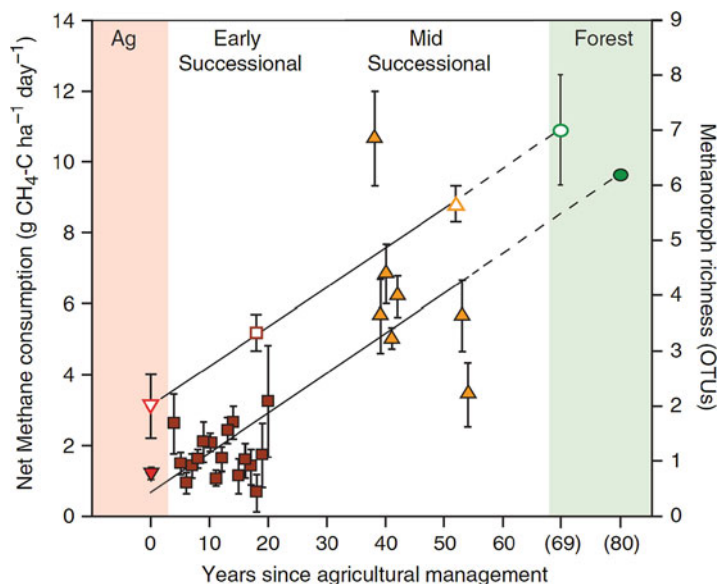
The absence of microbial diversity in BEF debate, conservation issues, and global biogeochemical process models is also caused by the paradigm of microbial omnipresence, high adaptability, and functional redundancy. Indeed, resilience after reduced diversity and redundancy of species carrying out similar functions has been demonstrated (see Bodelier 2011). But is this the rule? A number of studies have demonstrated a direct relationship between diversity and ecosystem process rate (see Bodelier 2011). Recently, a comprehensive meta-analysis demonstrated that out of 110 studies, more than 70 % demonstrate that microbial community composition was not resistant (i.e., the degree to which community composition remains unchanged when disturbed) against disturbances (fertilization, CO₂ increase, temperature, carbon amendment) (Allison and Martiny 2008). This held true for broad taxonomic groups (fungi, bacteria, archaea) as well as narrow groups with specific functions (methane oxidation, nitrification). The same study demonstrated that the resilience (i.e., the rate at which microbial community composition returns to its original composition after being disturbed) is in the order of years. Fertilization even led to differences in communities of N-cycling microbes (nitrifiers, denitrifiers) for more than 50 years (Hallin et al. 2009). Similar long-lasting effects have also been observed for methane-consuming microbes. Microbes consuming atmospheric methane are responsible for 6–10 % of global methane consumption. The process is sensitive to agricultural practices, and recovery after land abandonment can take decades

which coincides with an increase in diversity of these microbes (Fig. 1; Levine et al. 2011).

Aspects of community composition other than richness per se have been demonstrated to regulate the stability of biogeochemical processes. The initial evenness of redundant community members was demonstrated to be important in resistance to salt stress in denitrifying communities (Wittebolle et al. 2009), indicating that relative abundance of the populations in a community is an important determinative factor for process stability, even in redundant communities. Functional redundancy *sensu stricto* is difficult to assess in microbial communities, since it requires the contribution of individual community members to processes and separation between diversity and environmental factors. The stability of a particular function (e.g., methane conversion) in time is very likely affected by more properties or traits of species than the expression of that one particular functional gene only, e.g., response to inhibitors or general adaptation of species to a particular environment. Moreover, populations of interacting microbes on microbial relevant scales may not consist of many different species also due to spatial arrangement or isolation, e.g., along roots, soil pores, plant leaves, biofilms, or microbial flocs in sewage treatment. The growing body of experimental evidence suggests that microbial communities can be sensitive to disturbances and that resilience is linked to diversity. However, the majority of studies are descriptive, correlative, or strongly reductionist in nature, not allowing for causal or mechanistic conclusions.

Closing the Gaps

It is obvious that the omission of microbial communities from the BEF debate and in the managing and conservation of ecosystems is due to a lack of understanding of the functioning and composition of environmental microbial communities. The controversy between huge diversity and redundancy on the one hand and the lack of knowledge on 99 % of that diversity on the other hand leads to the fact that we do not know what we have to protect and what might have been lost



Microbial Ecosystems, Protection of, Fig. 1 The recovery of methanotroph diversity and atmospheric CH₄ consumption following row-crop agriculture. Increase in methanotroph diversity (open symbols) and CH₄ consumption (closed symbols) as a function of time since cessation of agriculture. The data clearly show that agricultural use diminishes methanotrophic diversity as well as function and that it can take decades before recovery takes place. All measurements (diversity and

consumption) are annual averages with error bars representing standard errors. Land-use treatments are as follows: agricultural management of historically tilled lands (AG), early successional fields abandoned from agriculture in 1989 (ES), successional forests abandoned from agriculture in the 1950s (SF), managed grasslands on never-tilled soil (MG), and deciduous forests (DF) (From Levine et al. 2011, with permission)

already. This controversy hampers the examination of the importance of microbial diversity for ecosystem functioning. Consequently, BEF studies in environmental microbiology are largely of descriptive nature and disconnected to ecological concepts. Approaches have been “top-down” or “bottom-up,” treating species/genotypes, community traits, and interactions as a “black box” (see Bodelier 2011). However, the rapid methodological developments of the last decades are narrowing down the limitations which kept environmental microbiology at the descriptive level. The “omic” techniques enable studying community ecology and physiology of known as well as unknown microbial species, and a systems biology approach for microbial communities is not out of reach (Raes and Bork 2008). In situ adaptation of community members as well as in situ profiling of whole genome transcripts and proteins of individual species is feasible. Next to this, methodology and concepts are emerging,

enabling individual-based physiology and ecology and even interactions on microbial relevant scale. Theoretical and conceptual approaches from macroecology are being applied to understand microbial community structure and to link it to ecosystem processes (Bodelier 2011). Ultrahigh-throughput community assessment methods will facilitate processing of large number of samples and replicates in order to obtain sufficient information allowing for experimental designs which yield mechanistic understanding of environmental microbial communities, eventually leading to the opening of the “black box.”

Microbial Community Conservation

The fact that there are no microbial species on the Red List nor are microbial communities in nature conservation policy does not mean that there are no initiatives toward conservation of microbial

communities. From the medical as well as biotechnological perspective, there is a need for the preservation of microbial genetic diversity which is mainly done by storing isolated and described microbial species in public culture collections (e.g., ATCC (<http://www.lgcstandards-atcc.org/>), DSMZ (<http://www.dsmz.de/>), and NCIMB (<http://www.ncimb.com/>)). However, since most of the diversity is represented in uncultured and not characterized microbes as part of environmental communities, we run the risk of losing genetic diversity of which we do not know its value yet, on itself a good reason for conservation. Well-known examples of biotechnological spin-off of environmental microbial communities have led to conservation efforts. The discovery of the heat-resistant Taq polymerase enzyme, used in PCR reactions, in the bacterium *Thermus aquaticus* (http://en.wikipedia.org/wiki/Thermus_aquaticus) in hot springs in Yellow Stone National Park, has led to declaring these hot springs as conservation targets in order to preserve the microbial genetic potential mainly for biotechnological applications (<http://serc.carleton.edu/microbelife/topics/biopropecting.html>), thereby making these hot springs the first environmental microbial conservation areas. Another development that contributes substantially to the “protection” of environmental microbial communities is the TEEB (The Economics of Ecosystems and Biodiversity) initiative which expresses the value of ecosystems, ecosystem services, and biodiversity in monetary values (<http://www.teebweb.org/>). Although this valuing of ecosystems is controversial and anthropogenically centered, it definitely created awareness for biodiversity among policy makers, politicians, and industry. The assessment of Earth ecosystems, biodiversity, and ecosystem services by 1,300 experts (Millennium 2005) identified key areas of ecosystem protection and conservation in order to keep our planet habitable. In all of these ecosystems, microbes play pivotal roles, a fact which is generally being recognized. Especially soils are a main focus when it comes to microbial processes because of the many ecosystem services soils and soil microbes provide and

because of the fact that soils harbor the largest source of microbial biodiversity. It is within the soil conservation that many initiatives are taken toward conservation of soil biodiversity like the EU soil framework directive in development (<http://ec.europa.eu/environment/soil/biodiversity.htm>) where also microbes are explicitly taken into account (Gardi et al. 2009). Combined with the already existing EU habitat conservation legislation (http://ec.europa.eu/environment/nature/natura2000/index_en.htm), important habitats containing a large part of microbial diversity on Earth are conserved. However, we still need to know what it is that needs to be preserved and what we can potentially lose or affect by climate change, habitat destruction, land-use change, urbanization, etc. This requires inventories of microbial diversity and functioning. Despite the serious limitations in methods to assess the sheer endless microbial diversity, there are a number of initiatives going on that come as close as possible. The Earth Microbiome Project (<http://www.earthmicrobiome.org/>) is an initiative to assess functional microbial diversity in more than 200,000 environmental samples which will be collected and analyzed in a coordinated way and will be complemented with essential metadata which can be used to infer ecological or biogeographical aspects of the communities in the database. A similar initiative has already been in place for a number of years focusing on marine microbial communities (<http://www.coml.org/international-census-marine-microbes-icomm>), while the TerraGenome project specifically focuses on soils (<http://www.terragenome.org/>). Hence, many steps on the “roadmap toward microbial conservation,” as put forward by Cockell (Cockell and Jones 2009) a number of years ago, have been taken. Projects attempting to make microbial diversity inventories are initiated and scientific approaches to link microbial species to ecosystem functions are being developed. Nevertheless, “the Red List” species approach will definitely not be applicable to microbes as already pointed out above. Hence, we need different approaches and concepts regarding “conservation units” for microbial

communities which are useful and understandable for policy makers and politicians. Habitat conservation is a good starting point, but probably we can also put forward “vulnerable” nonredundant environmental microbes which are carrying out important ecosystem functions like methane oxidizers which may be affected by anthropogenic disturbance, diminishing their functioning in the environment for decades (see Fig. 1). Next to this, educating the public, policy makers, and politicians on the importance and sheer uniqueness of microbes and microbial communities will be of utmost importance in the process of getting microbes on the conservation agenda. If not protecting them for their valuable functions, we should do it for the sake of ethics (Cockell 2011).

Summary

Despite the eminent role microbes and microbial communities play in all ecosystems on Earth, they are not considered in conservation policy or legislation. This is due to utter lack of fundamental knowledge on crucial issues concerning environmental microbial communities. The species-oriented approach in conservation biology is not emendable to microbes where there are difficulties in defining species and where more than 99 % of all species present in the environment are not known. Next to this, we have no idea what the importance is of microbial diversity for ecosystem functioning because of the lack of methodology to do so. The most important problem is probably the notion that microbes are so abundant, diverse, and resilient that they are not threatened by extinction. However, rapid developments in the field of environmental microbiology, mainly in the application of genomic and isotopic techniques, have revolutionized our knowledge and demonstrate that microbes display biogeography and are sensitive to environmental disturbance and that for a number of environmentally relevant processes, community composition is linked to ecosystem functioning. Hence, microbes are not “untouchable” and omnipresent, but in order to get them onto the conservation agenda, we have to be able to assess which

microbes are present in environments in order to be able to monitor changes with possible consequences for ecosystem functions. There are many initiatives underway seeking to make inventories of functional diversity of microbial communities in marine, terrestrial, and freshwater habitats. This knowledge will facilitate assessing impacts and consequences of anthropogenic disturbances on microbial communities and their functioning in the future and pave the way for the protection of environmental microbial communities. For the time being, we have to rely on habitat conservation guidelines and legislation to ensure maintenance of microbial communities.

References

- Allison SD, Martiny JBH. Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci U S A*. 2008;105:11512–9.
- Bodelier PLE. Toward understanding, managing, and protecting microbial ecosystems. *Front Microbiol*. 2011;2(80).
- Chen Y, Murrell JC. When metagenomics meets stable-isotope probing: progress and perspectives. *Trends Microbiol*. 2010;18(4):157–63.
- Cockell CS. Microbial rights? *EMBO Rep*. 2011;12(3):181. 181.
- Cockell CS, Jones HL. Advancing the case for microbial conservation. *Oryx*. 2009;43(4):520–6.
- Ducklow H. Microbial services: challenges for microbial ecologists in a changing world. *Aquat Microb Ecol*. 2008;53(1):13–9.
- Gardi C, et al. Soil biodiversity monitoring in Europe: ongoing activities and challenges. *Eur J Soil Sci*. 2009;60(5):807–19.
- Green JL, Bohannan BJM, Whitaker RJ. Microbial biogeography: from taxonomy to traits. *Science*. 2008;320(5879):1039–43.
- Hallin S, et al. Relationship between N-cycling communities and ecosystem functioning in a 50-year-old fertilization experiment. *ISME J*. 2009;3(5):597–605.
- Hooper DU, Adair EC, Cardinale BJ, Byrnes JEK, Hungate BA, Matulich KL, Gonzalez A, Duffy JE, Gamfeldt L, O’Connor MI. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*. 2012. doi:10.1038/nature11118.
- Levine UY, et al. Agriculture’s impact on microbial diversity and associated fluxes of carbon dioxide and methane. *ISME J*. 2011;5(10):1683–91.
- Mayali X, Weber PK, Brodie EL, Mabery S, Hoeprich PD, Pett-Ridge J. High-throughput isotopic analysis of RNA microarrays to quantify microbial resource use. *ISME J*. 2012;6:1210–21.

- Millennium, Ecosystem, Assessment 2005. Ecosystems and human well-being: general synthesis. United Nations. www.millenniumassessment.org/en/synthesis.aspx
- Murrell JC, Whiteley AS, editors. Stable isotope probing and related technologies. American Society for Microbiology (ASM); Washington DC, 2011.
- Pace NR. Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev.* 2009;73(4):565–76.
- Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S. The ecological coherence of higher bacterial taxonomic ranks. *Nat Rev Microbiol.* 2010;8:523–9.
- Raes J, Bork P. Systems microbiology – timeline – molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol.* 2008;6(9):693–9.
- Tamames J, et al. Environmental distribution of prokaryotic taxa. *BMC Microbiol.* 2010;10.
- Wagner M. Single-cell ecophysiology of microbes as revealed by Raman microspectroscopy or secondary ion mass spectrometry imaging. *Annu Rev Microbiol.* 2009;63:411–29.
- Wittebolle L, et al. Initial community evenness favours functionality under selective stress. *Nature.* 2009;458(7238):623–6.
- Zhou JZ, et al. Spatial scaling of functional gene diversity across various microbial taxa. *Proc Natl Acad Sci U S A.* 2008;105(22):7768–73.

Mining Metagenomic Datasets for Antibiotic Resistance Genes

Lisa Durso

Agroecosystem Management Research Unit,
US Department of Agriculture, University Of
Nebraska, Lincoln- East Campus, Lincoln,
NE, USA

Synonyms

Anthropogenic and human associated; Horizontal gene transfer and lateral gene transfer; Whole-genome sequencing and metagenomic sequencing

Definition

Metagenomics refers to samples in which the entire bacterial or microbial community DNA is

used and to high-throughput DNA sequencing of microbial community DNA. These sample types and methods can be used to gather information on genes that code for antibiotic resistance.

Introduction

Antibiotics are medicines that are used to kill, slow down, or prevent the growth of susceptible bacteria. They became widely used in the mid-twentieth century for controlling disease in humans, animals, and plants and for a variety of industrial purposes. Antibiotic resistance is a broad term. Depending on the classification scheme used, there are between eight and twenty different classes of antibiotics, with multiple compounds in each class. These different categories represent different basic chemical structures and modes of action – some antibiotics will inhibit cell wall synthesis, for example, while others target portions of the ribosome and a cell's protein processing machinery. Just as there are many types of antibiotics, there are also many types of antibiotic resistance. Some types of resistance are specific for an individual antibiotic, while others, such as multidrug resistance efflux pumps, can confer resistance to multiple different kinds of antibiotics. It is also likely that there are naturally occurring antibiotics that have yet to be described.

Antibiotic resistance is a normal and natural phenomenon that can be documented even in ancient (permafrost from 30,000 years ago) and pristine habitats such as Antarctica and the Sargasso Sea (Allen et al 2009; D'Costa et al. 2011; Durso et al. 2012). In addition to naturally occurring antibiotic resistance, there is no doubt that anthropogenic or human-associated use of antibiotics for health, food production, veterinary, and industrial purposes has dramatically impacted resistance. The continued emergence of antibiotic-resistant, opportunistic, and pathogenic infections in health-care settings has become a major public health concern, especially the emergence of bacteria that are resistant to multiple antibiotics or multiple classes of antibiotics. Yet few details are known about how

antibiotic resistance genes move through environmental, agricultural, and clinical settings. Metagenomics provides one tool to start characterizing antibiotic resistance genes across habitats.

The term “metagenomic” has multiple meanings. Historically it was used to describe the kind of sample that was collected and referred to collecting DNA or genomic information not just from a single organism or isolate but from a whole community, a metagenome, consisting of both cultured and uncultured organisms (metagenomic samples). More recently, the term metagenomic has come to describe a specific type of analysis that relies on high-throughput nucleic-acid sequencing of either 16S rDNA or whole-community DNA (metagenomic sequencing). In addition to providing metagenomic sequencing information, the new high-throughput sequencing methods can be used to profile whole-community RNA profiles (metatranscriptome) and whole-community protein profiles (metaproteome). This entry will examine studies using both metagenomic samples and the use of metagenomic sequencing to gather information on functional genes that code for antibiotic resistance. Although the focus here will be on mining metagenomic data for information on antibiotic resistance genes, it is acknowledged that functional and gene-based metagenomic studies complement experiments involving gene expression, protein production, and phenotypic characterization of individual and community resistance.

The Antibiotic Resistome

The concept of an antibiotic “resistome” was first proposed in 2006 by D’Costa et al. to describe the sum total of all antibiotic resistance genes across the globe and all genetic elements that could give rise to antibiotic resistance genes (D’Costa et al. 2006). It includes pathogenic bacteria that cause illness, as well as opportunistic and non-pathogenic bacteria. This concept provides a framework that unites antibiotic resistance in human, animal, and plant clinical applications, with the broader pool of antibiotic-resistant

bacteria present in the environment, including food, water, and soil. D’Costa et al. (2006) cultured spore-forming bacteria from soil and screened them against 21 antibiotics, including both old and new antibiotics and naturally occurring and synthetic antibiotics. Based on their results, they identified the soil as a reservoir of antibiotic resistance genes and proposed the idea of a pan-microbial resistome. Contrary to the general public perception that use of antibiotics in human medicine and agriculture is the root cause of antibiotic resistance, the antibiotic resistome hypothesis supports the idea of a naturally occurring global pool of antibiotic resistance and suggests that the environment (especially soil) serves as a reservoir of antibiotic resistance elements. In this model antibiotic resistance elements can be enriched and selected for by anthropogenic antibiotic use. However, unlike previous models, the concept of the antibiotic resistome expands the focus from the selection of pathogens via the direct use of antibiotics in clinical settings to a global pool of antibiotic resistance that can potentially be transferred from harmless bacteria into human, animal, and plant pathogens. Later work by the same group (Wright 2007; D’Costa et al. 2011) as well as others (Riesenfeld et al. 2004; Henriques et al. 2006; Aminov and Mackie 2007; Mori et al. 2008) provides supporting evidence for the natural occurrence of antibiotic resistance, especially in soil, and the global distribution of antibiotic resistance genes.

Conceptually, the relationship between increased anthropogenic use of antibiotics and increases in the number and types of antibiotic-resistant bacteria and antibiotic resistance genes is clear. On a practical level, many of the details regarding the ecology of antibiotic resistance and antibiotic resistance genes in the environment remain unknown. These include fate and transport of naturally occurring and anthropogenically induced antibiotic-resistant genes within and between environmental, agricultural, and clinical settings as well as rates of gene transfer, rates of gene expression, and impact of naturally occurring and anthropogenically introduced antibiotic concentrations on short- and long-term microbial community structure.

Antibiotic Resistance Genes

The genes that code for antibiotic resistance are carried either as part of the regular bacterial chromosome, which is passed vertically to individual daughter cells, or as part of mobile genetic elements such as plasmids and transposons which can be transferred both vertically to daughter cells and horizontally to other strains or species of bacteria. These antibiotic resistance genes, sometimes called antibiotic resistance determinants or antibiotic resistance elements, code for a variety of different kinds of proteins involved in inactivating the antibiotic, removing the antibiotic from the cell, or modifying the target of the antibiotic so that it is not recognized by the drug. For any specific antibiotic, there may be multiple types of resistance mechanisms. Many of these mechanisms are complex and require the coordination of a suite of genes, so that for any individual antibiotic, there are multiple different antibiotic resistance genes.

There are two basic approaches to mining metagenomic datasets for antibiotic resistance genes: those that are database dependent and those that are discovery driven. The database-dependent systems are good for comparative studies that screen large numbers of samples or large number of genes and examine similarities or differences in antibiotic resistance gene patterns across samples. These methods rely on previously sequenced antibiotic resistance genes to provide the information used to design primers or to provide a list against which new sequences are compared. The limitation of database-dependent methods is that a particular gene must already have been sequenced in order to be in the database, and researchers can only screen against genes that have already been discovered, characterized, and deposited in the database. Discovery-driven methods, while time-consuming and low-throughput, can be used to describe novel antibiotic resistance genes. In this approach, DNA fragments from metagenomic samples are cloned into hosts such as *E. coli*, or constructs such as bacterial artificial chromosomes (BACs), and then screened for a particular phenotype. The collection of DNA fragments in the new host is

called a library. In the case of antibiotic resistance, the clone or BAC libraries are plated onto media containing a specific amount of antibiotic. If they grow in the presence of the antibiotic, they are considered resistant. If they do not grow, they are considered sensitive. In human medicine and clinical settings, there are well-defined standard methods that specify, by organism and antibiotic, the concentration needed to be considered resistant. In environmental and experimental settings, these standards do not exist, and there is no consistent definition across studies.

Studying Antibiotic Resistance Genes from Metagenomic Samples

Metagenomic samples can be mined for known as well as uncharacterized antibiotic resistance genes using functional screening of metagenomic clone libraries. After creating the libraries, clones are plated onto media containing the antibiotic of interest. Colonies that grow in the presence of the antibiotic are assumed to be carrying an antibiotic-resistant gene from the original sample. The inserts from the resistant clones can be sequenced, and the sequences compared against database of known antibiotic resistance genes. As early as 1997, these methods were used to characterize the diversity of quinolone resistance genes in soil (Waters and Davies 1997). This functional metagenomic approach has been used to target specific classes of antibiotic resistance genes, as well as more general surveys of antibiotic resistance where libraries are screened against multiple antibiotics. For example, tetracycline resistance has been assayed from human mouth, and organic pig samples (Diaz-Torres et al 2003; Kazimierczak et al. 2009) and β -lactamase genes have been extracted from samples such as tropical surface waters and Alaskan soil (Henriques et al. 2006; Allen et al. 2009).

The mining of functional genes focuses on two main types of samples. When trying to determine baseline levels of antibiotic resistance and evolutionary relationships of individual genes, pristine samples and those dating from before the use of

antibiotics are used (D'Costa et al. 2011). When searching for novel antibiotic resistance genes, complex samples are used, especially those with increased levels of antibiotic compounds such as feces or activated sludge (Sommer et al. 2009; Mori et al. 2008). It is also possible to use publicly available information to screen for potentially novel antibiotic resistance genes. Both the National Center for Biotechnology Information (NCBI) and the MG-RAST server (Meyer et al. 2008) have extensive DNA sequence datasets that are available to the public. Once identified via the public databases, antibiotic resistance genes of interest can then be characterized using other methods (Toth et al. 2010).

Studying Antibiotic Resistance Genes Using Metagenomic Sequencing Methods

One tool that is useful for exploring antibiotic resistance in metagenomic samples is MG-RAST (Meyer et al. 2008). MG-RAST, developed at Argonne National Laboratory and the University of Chicago, provides metagenomic data analysis tools for both public and private metagenomic sequencing sets. There are hundreds of publicly available metagenomes on the MG-RAST website (<http://metagenomics.anl.gov>). These can be accessed directly using the sample ID number or via the “browse metagenome” function. Researchers may submit their own metagenomic datasets to the site for analysis, with processing priority given to datasets that will be made immediately available to the public. After normalization, both taxonomic and functional data are extracted from the submitted sequences and made available for visualization via the website. There are many different classification schemes that are available for organizing data on MG-RAST. One system, called SEED (Overbeek et al. 2005), is designed to classify functional genes across genomes using a standardized system for categorizing genes or gene fragments. The SEED system of organization is hierarchical in nature, and each of the primary functional groups or systems is

composed of subsystems. Examples of primary SEED functional groups are “cell wall synthesis,” “nitrogen metabolism,” and “virulence.” Within the virulence functional category is a subset of genes that are associated with “resistance to antibiotic and toxic compounds” (RATC). Drilling even further down into this particular functional group, gene fragments are binned by categories such as “aminoglycoside adenylyltransferases,” “beta-lactamase resistance,” and “resistance to fluoroquinolones.”

After identifying a metagenome, a list of antibiotic resistance genes can be accessed using the “analysis” icon. Under “Data Type” choose “Functional Abundance” and “Hierarchical Classification.” The Data Selection annotation source should be “subsystems” and the Data Visualization option should be “table.” Then, hit the “generate” button. After processing the data, a table will be displayed with three functional classification levels displayed, along with abundance and quality data. The abundance results are clickable, and open a window that lists the taxonomic assignments of each of the hits, as well as a link to the actual sequence and M5nr nonredundant protein data. The M5nr database allows classification of the fragment across multiple classification schemes.

Because metagenomic sequencing is performed on whole-community DNA without a PCR step, the data generated can be considered quantitative. So in addition to describing which antibiotic resistance genes are present, metagenomic analysis can quantify the relative amounts or proportions of individual genes and/or gene classes – both within any individual sample and across samples from different habitats. As with all methods associated with tracking antibiotic resistance in the environment, there are no standard methods (Allen et al. 2010). However, control metrics available through MG-RAST, in particular a new metric called duplicate read inferred sequencing error estimation (DRISEE; Keegan et al. 2012), can serve as screening tools to decide on minimum quality standards for inclusion or exclusion of specific metagenomic samples for analysis.

These metagenomic sequencing tools can be used to start addressing questions related to the

ecology of antibiotic resistance in specific habitats and across ecosystems. Metagenomic analysis of 45 microbiomes across the globe revealed functional gene profiles that correlated with environment (Dinsdale et al. 2008). This idea was expanded to antibiotic resistance genes, providing an antibiotic resistance “fingerprint” for some samples (Durso et al. 2011). A metagenomic analysis of public datasets was performed specifically comparing RATC genes from agricultural and nonagricultural metagenomes (Durso et al. 2012). Among the 26 metagenomes studied, the total percent of RATC gene fragments (based on all classified fragments) ranged from 0.7 % for the Sargasso Sea sample to 4.4 % for the dog. The fecal samples (dog, fish, three human, and cattle) had the highest overall percent of RATC genes, while the marine samples (Chesapeake, Galapagos, Zanzibar, Gulf of Mexico, Key West, Madagascar, Gulf of Maine, and Sargasso Sea) had the lowest overall percent of RATC genes. In addition to having the highest proportion of RATC genes, the dog metagenome also displayed the highest diversity of RATC classes (31 classes) and the Sargasso Sea displayed the lowest diversity (7 classes). Using MG-RAST, individual classes of antibiotic resistance genes could be examined. The fish metagenome, for instance, had over ten times as many genes coding for mercuric reductase and mercury resistance (3.9 %), compared to the average for the other metagenomes (0.31 %), while the day 29 kimchi metagenome, a Korean fermented vegetable, had high levels of the two-protein Gram-positive multidrug resistance compared to the other metagenomes examined. In both of these examples, the metagenomic data reflect what we already know about the biology of these systems and suggest that metagenomic RATC data can be used to distinguish fundamental differences in microbial community ecology from diverse microbiomes.

In addition to information on specific antibiotic resistance genes, analysis of metagenomic sequencing data can also provide taxonomic information about a sample. The use of the 16S ribosomal RNA gene to classify bacteria is well known. Some of the fragments in a metagenomic sample that code for 16S rRNA genes can be

pulled out and used for taxonomic purposes. In addition, MG-RAST has the ability to link protein-coding fragments with taxonomic assignments using SEED and other systems. Currently, the only way to access this linked information for individual reads from MG-RAST is through the “assignment” column on the functional gene table, so it is time-consuming to assemble this linked data, even for a single metagenome. Grouped data are more easily accessible in MG-RAST using the “workbench” function. In the functional table, the last column contains a box titled “to workbench.” Reads belonging to specific functional groups can be selected, and then a second taxonomic-specific analysis can be run exclusively on the reads in the workbench.

Using these methods, information can be gathered on which bacteria are likely carrying specific antibiotic resistance genes and how the bacterial communities may change over time or space. Some types of antibiotic resistance, such as beta-lactamase, MDR efflux pumps, and fluoroquinolone resistance, are broadly distributed across many (>10) taxa, while other types of resistance genes such as tetracycline and vancomycin resistance are more taxonomically restricted (4 or 5 taxa each). Within individual antibiotic resistance classes, the taxonomic distribution of specific genes or gene classes varies by metagenome. For example, MDR efflux pump genes are associated mainly with Clostridia in animal agriculture metagenomes but are more frequently assigned to Gammaproteobacteria in coastal marine samples. Metagenomic sequencing enables researchers to track the change in microbial communities over time. One set of publicly available metagenomes follows the fermentation of kimchi over the course of a month. The antibiotic resistance gene profiles associated with the kimchi change dramatically as the fermentation progresses, and these specific changes can be tracked using metagenomic sequencing.

The strengths of these metagenomic sequencing methods are that they allow researchers to identify and gain a quantitative understanding of functional gene relationships across samples and geographies. It should be kept in mind that there are many places where artifacts of processing of

either the sample itself or the resulting sequence data can influence the results. Although these sequence-based metagenomic data are excellent for getting oriented in a system and providing an overview of what is potentially there, the output is of fairly low resolution and requires follow-up using other methods before detailed conclusions can be drawn. Nonetheless, there is great value in the information that these kinds of techniques can provide. Like the Lewis and Clark expedition, which mapped the entire US western frontier based on sampling a single route covering much less than 1 % of today's public roads in the area, data generated by metagenomic sequencing methods provide a first step in exploring previously unknown territory. For antibiotic resistance, they offer the capacity to examine the prevalence of antibiotic gene distribution on a global scale and the opportunity to begin to compare distribution of specific antibiotic resistance genes across samples and time.

Summary

The ecology of antibiotic resistance genes in the environment remains largely unexplored. Metagenomic tools provide the opportunity to identify novel antibiotic resistance genes, explore the epidemiology of antibiotic-resistant genes across multiple habitats, and begin to define relationships between antibiotic resistance genes and the bacteria that likely carry them. The availability of public metagenomic datasets affords all researchers an opportunity to ask and answer questions about antibiotic resistance.

References

- Allen H, Cloud-Hansen K, Wolinski J, et al. Resident microbiota of the gypsy moth midgut harbors antibiotic resistance determinants. *DNA Cell Biol.* 2009;28(3):109–17.
- Allen H, Donato J, Wang H, et al. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev.* 2010;8:215–59.
- Aminov R, Mackie R. Minireview: evolution and ecology of antibiotic resistance genes. *FEMS Microbiol Lett.* 2007;271:147–61.
- D'Costa V, McGrath K, Hughes D, et al. Sampling the antibiotic resistome. *Science.* 2006;311:374–7.
- D'Costa V, King C, Kalak L, et al. Antibiotic resistance is ancient. *Nature.* 2011;477(7365):457–61.
- Diaz-Torres M, McNab R, Spratt D, et al. Novel tetracycline resistance determinant from the oral metagenome. *Antimicrob Agents Chemother.* 2003;47(4):1430–2.
- Dinsdale E, Edwards R, Hall D, et al. Functional metagenomic profiling of nine biomes. *Nature.* 2008;452:629–33.
- Durso L, Harhay G, Bono J, et al. Virulence-associated and antibiotic resistance genes of microbial populations in cattle feces analyzed using a metagenomic approach. *J Microbiol Methods.* 2011;84(2):278–82.
- Durso LM, Miller DN, Wienhold BJ. Distribution and quantification of antibiotic resistant genes and bacteria across agricultural and non-agricultural metagenomes. *PLoS One.* 2012;7:e48325.
- Henriques I, Moura A, Alves A, et al. Analysing diversity among β -lactamase encoding genes in aquatic environments. *FEMS Microbiol Ecol.* 2006;56:418–29.
- Kazimierczak K, Scott K, Kelly D, et al. Tetracycline resistance of the organic pig gut. *Appl Environ Microbiol.* 2009;75(6):1717–22.
- Keegan K, Trimble W, Wilkening J, et al. A platform-independent method for detecting errors in metagenomic sequencing data: DRISSEE. *PLoS Comput Biol.* 2012;8(6):e1002541. doi:10.1371/journal.pcbi.1002541.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9:386.
- Mori T, Mizuta S, Suenaga H, et al. Metagenomic screening for bleomycin resistance genes. *Appl Environ Microbiol.* 2008;74(21):6803–5.
- Overbeek R, Begley T, Butler R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33:5691–702.
- Riesenfeld C, Goodman R, Handelsman J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol.* 2004;6(9):981–9.
- Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science.* 2009;325:1128–1131.
- Toth M, Smith C, Frase H, et al. An antibiotic-resistance enzyme from a deep-sea bacterium. *J Am Chem Soc.* 2010;132:816–23.
- Waters B, Davies J. Amino acid variation in the GYRA subunit of bacteria potentially associated with natural resistance to fluoroquinolone antibiotics. *Antimicrob Agents Chemother.* 1997;41(12):2766–9.
- Wright G. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev.* 2007;5:175–86.

Mining Metagenomic Datasets for Cellulases

David J. Rooks and Alan J. McCarthy
Microbiology Research Group, Institute of
Integrative Biology, Biosciences Building,
University of Liverpool, Liverpool, UK

Synonyms

Environmental DNA; Glycosyl hydrolases;
Metagenomes; Metatranscriptomes

Definition

Metagenomic (DNA) or metatranscriptomic (cDNA) sequence datasets generated using DNA or RNA extracts are obtained directly from environmental samples. These include soil, water, gut contents, and degrading organic matter/plant biomass and biofilms; laboratory-incubated microcosms or mesocosms in which cellulose-degrading microorganisms are enriched also serve as sources of nucleic acids for the preparation of sequence datasets. Genes encoding glycosyl hydrolases and specifically those likely to be active against cellulose (cellulases) can be sought, most efficiently in the large sequence datasets generated by the application of pyrosequencing technologies.

Cellulose and Its Biodegradation

Cellulose is the most abundant form of photosynthetically fixed carbon in the biosphere. It is a fibrous linear homopolymer of glucose in the form of cellobiose (dimer) units linked by β -1,4-glycosidic bonds, and it occurs naturally in plants, some fungi, protozoa, and one group of animals – the urochordates (Lynd et al. 2002). Native cellulose is a highly crystalline polymer due to the formation of rigid microfibrillar structures stabilized by inter- and intramolecular hydrogen bonds and van der Waals interactions

between the polymer chains, and this is largely responsible for its recalcitrance. This network of bonds leads to a mostly uniform arrangement of fibers, and the resultant crystalline cellulose lacks enzyme-accessible surface morphologies, further enhancing resistance to hydrolysis (Zhou et al. 2009). Cellulose usually occurs naturally in close physical association with hemicelluloses, which are heteropolysaccharides that, in terrestrial plants, form the lignocarbhydrate matrix enveloping cellulose fibers and essentially constituting the plant cell wall structure. Cotton is the only naturally occurring pure form of highly crystalline cellulose. For microorganisms to hydrolyze and metabolize insoluble polymeric cellulose, extracellular cellulases must be produced and in multiple forms that act synergistically. The two primary models are those in which the enzymes are truly secreted, versus the cellulosome, a surface-bound multimeric complex of polypeptides comprising catalytic and non-catalytic components; the cellulosome has been likened to a polysaccharide processing nanomachine (Fontes and Gilbert 2010). There is a possible third model in which cellulose is bound to the bacterial cell surface and further processed in the periplasmic space (see Ransom-Jones et al. 2012). Three major types of enzymatic activities are found: (i) endoglucanases, (ii) exoglucanases (cellobiohydrolases), and (iii) β -glucosidases (cellobiases). The evidence for oxidative attack on cellulose has often been equivocal, but there are now data that establish the involvement of an enzyme (GH61) in cellulose depolymerization (Quinlan et al. 2011).

Cellulase Structure and Function

Cellulases are generally glycosidic hydrolase (GH) enzymes that utilize the same mechanism of acid-base catalysis with inversion or retention of glucose anomeric configuration (Davies and Henrissat 1995). Cellulases are modular enzymes composed of independently folding, structurally and functionally discrete units, referred to as either domains or modules (Henrissat et al. 1998), and

are the most diverse enzymes that catalyze a single reaction. Automated data mining suggests that there are 15 glycoside hydrolase families that contain cellulases; families are defined by amino acid sequence similarity (CAZy – see below). Structural studies show that cellulases have eight different protein folds and contain a carbohydrate-binding module, which is usually linked to a catalytic-binding domain (Shoseyov et al. 2006). Glycosyl hydrolases with open active sites typically exhibit endocellulolytic activity (endoglucanases) and cleave β 1–4 links at amorphous sites in the polysaccharide chain to generate chain ends and ultimately oligosaccharides of various lengths (Horn et al. 2006). Those with tunnelloid active sites exhibit exocellulolytic activity and are cellobiohydrolases that act in a processive manner on the reducing or nonreducing ends to liberate either glucose or cellobiose as major products. β -Glucosidases convert cellobiose to glucose, completing the highly synergistic and complete enzymatic depolymerization of cellulose.

The Carbohydrate-Active Enzyme Database (CAZy)

Identification of cellulase genes per se can be achieved by interrogating DNA sequence databases to identify homologies or, more ambitiously, to look for new types or classes of enzymes among the genes of unknown function that invariably dominate metagenome sequence datasets. The former is facilitated by the Carbohydrate-Active Enzyme (CAZy) database (<https://www.cazy.org>) (Cantarel et al. 2009), a comprehensive repository of CAZymes that is an almost unique resource for enzyme discovery. At present, CAZy covers approximately 300 protein families, including glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and carbohydrate-binding modules (CBMs). All known cellulases are found within the CAZy database and are denoted by two enzyme commission numbers: EC 3.2.1.4 (endoglucanase) and EC 3.2.1.91 (cellobiohydrolase). Families

GH5 and GH9 have the largest number of biochemically characterized cellulases; however, this could be largely due to the abundance of these cellulases in the limited number of model cellulolytic organisms that have been studied (Sukharnikov et al. 2011). The database is frequently updated to provide rich sets of manually curated information on all groups of CAZymes, i.e., names, GenBank accession numbers, EC designations, 3D structure, and taxonomy, and the information can serve as an invaluable resource to identify CAZyme genes or gene fragments in both genomes and metagenomes. Although the collection of enzyme data in CAZy is invaluable for enzymologists, annotations could be significantly improved; the term “characterized” in CAZy is applied equally to proteins that have been analyzed biochemically and to those for which function has been computationally predicted (Stam et al. 2006).

Metagenomics

The vast majority of microorganisms in the biosphere have yet to be cultivated and remain an untapped source of enzymes for biotechnological applications. The current impetus to find novel cellulases for applications, particularly in biomass refining, stems from the importance of utilizing cellulose as a substrate for second-generation biofuel production. The requirement for synergy and the low specific activity of cellulases in native cellulose saccharification processes remains a major challenge. Environmental microbiology research was changed radically by molecular biology, with the greatest effort directed toward describing true phylogenetic/functional diversity in natural microbial communities by PCR amplification of marker genes. However, cellulase genes, although well defined at the protein sequence level, can rarely be simply amplified in this way because the extent of nucleotide sequence variation does not enable the design of appropriate oligonucleotide primers for PCR. Subsequently, the development of quantitative PCR and the use of environmental RNA as the template moved this field forward,

but we are now firmly in the era of environmental metagenomics, made possible by pyrosequencing technology (next-generation sequencing). Thus, metagenomics, the direct sequencing of DNA fragments from environmental samples, enables mining of the vast genetic resource held in the genomes of uncultured microorganisms that dominate natural microbial communities. Currently, a single pyrosequenced metagenome can comprise up to 15 gigabases in reads of up to 600 bp (Illumina). Alternatively, the metagenome can be cloned into a suitable vector that can accommodate large inserts (20–40 kb) and subsequently screened for cellulases (Rooks et al. 2012). Functional screening in an expression host (usually *E. coli*) using Congo red staining of carboxymethyl cellulose (CMC) (McDonald et al. 2012) has successfully recovered cellulases from a diversity of metagenomes, including those from soil (Voget et al. 2006), the buffalo rumen (Duan et al. 2009), the termite hindgut (Warnecke et al. 2007), and the human intestine (Qin et al. 2010).

Metatranscriptomes

Metagenomics provides information on the potential metabolic and functional capacity of a microbial community. However, these DNA-based analyses cannot differentiate between expressed and non-expressed genes. Environmental transcriptomics (metatranscriptomics) retrieves and sequences mRNAs from the microbial community to provide an unbiased perspective on gene expression in situ. Due to the difficulties inherent in the processing of environmental RNA to maintain integrity and ultimately recover the high-quality mRNA from the predominantly ribosomal RNA background, publications are relatively few in number. The first report of a pyrosequenced metatranscriptome from a complex microbial community was by Leininger et al. (2006) who demonstrated that archaeal transcripts of the key enzyme (amoA) in ammonia oxidation were several orders of magnitude more abundant in soils than the bacterial equivalent, suggesting that it is

members of the Archaea that are the primary drivers of nitrification. Gilbert et al. (2008) identified a large number of novel highly expressed sequence clusters from marine microbial communities, the majority of which were orphaned, thus demonstrating the utility of the metatranscriptomic approach in the discovery of novel genetic variants. Damon et al. (2012) addressed the global activities of soil eukaryotes by sequencing $2 \times 10,000$ cDNAs synthesized from polyadenylated mRNA directly extracted from forest soils. A total of 2,076 sequences were putative homologues to genes for different enzyme classes; specific annotation identified enzymes active on major plant biomass polymers, with glycoside hydrolases representing 0.5 % of the total. Finally, a metatranscriptomic analysis targeted specifically at fungal glycoside hydrolases induced by the addition of cellulosic substrates to soil, generated 47 putative cellulase sequences spanning 13 families identified within a cDNA sequence dataset comprising over 56,000 protein-coding sequence fragments (Takasaki et al. 2013). Therefore, despite the inherent difficulties of extracting, enriching, and processing mRNA from environmental samples, for which technological solutions are emerging, metatranscriptomics offers the advantage of targeting genes that are active in the environment and therefore functionally competent and exploitable.

Bioinformatics and Screening

In environmental metagenomics, determining the true microbial community structure that will lead to the discovery of new taxa, and hence novel enzymes, has been the most important driver. The bioinformatic tools and approaches available tend to reflect this emphasis on taxonomy. MEGAN (Huson et al. 2007) is a data management program used in the taxonomic analysis of large sequencing datasets, processing the results of comparisons between a known database and metagenome-derived sequences. In the context of this entry, information on the presence/abundance of known taxa of cellulose degraders

can be provided, and it is always an analysis worth doing. To identify novel cellulases, more sophisticated bioinformatic approaches are required to search for domains and motifs indicative of enzymes with cellulose binding and/or catalytic functions. Sequence comparisons among proteins with suggestive domain architectures or genomic contexts in metagenomic DNA have the potential to identify novel cellulases; the discovery of a new carbohydrate-binding module in metagenomic DNA by Mello et al. (2010) is a particularly good example of what can be achieved by the continuing development of bioinformatic tools.

With complete sequences and their genomic context if located within larger sequenced DNA fragments, homology-based approaches can be extended. Firstly, structural modeling of members of likely cellulase families can identify those with unusual binding and catalytic sites that may therefore exhibit functional novelty. Secondly, domains of unknown function, which are likely to be putative cellulase or cellulase-related sequences because they are consistently linked by genome context, can be characterized through distant homology, non-homology, and structure-based approaches. This is exemplified by the identification of a novel cellulase from a sequenced marine bacterial genome through signature domains that assemble enzymes into plant cell wall degradative complexes (Bras et al. 2011).

Sequences with matches indicative of cellulases can of course be identified by BLAST searches against the CAZy database (see above) and through functional annotation pipelines such as SEED (Overbeek et al. 2005) and MG-RAST (Meyer et al. 2008) to provide taxonomic affiliations for functional and hypothetical protein-encoding genes. However, identification of even distant relationships for the short sequence read output (<500 bp) that is characteristic of pyrosequencing is a bioinformatic challenge. The danger of simply searching against databases of known cellulase gene sequences is that true novelty will be missed and the metagenomes will only be

mined for variants of these known cellulases. Much longer sequences, ideally complete genes, are the best source material for bioinformatic prediction of potential cellulase function, and metagenomic/metatranscriptomic datasets can provide the probes to identify such genes and their neighbors in contemporaneously produced fosmid or bacterial artificial chromosome (BACS) libraries (Rooks et al. 2012). Subsequent cloning, overexpression, and purified protein production then provide sufficient material for the detailed structure/function characterization, combining classical biochemistry and structural biology approaches, necessary to establish that a novel cellulase has been teased out from the metagenome.

Future Prospects

Firstly, the tandem approach of using environmental RNA and DNA as the starting material to generate complimentary metatranscriptomes and metagenomes, thus benefitting from the specific advantages of each, is becoming more feasible with developments in ribosomal RNA depletion and messenger RNA enrichment techniques. Four hundred and fifty four pyrosequencing, which had predominated due to the relatively long read lengths (ca. 800 bp) that could be obtained, is receding to be replaced by next-generation sequencing technology that can deliver ever-increasing read lengths (currently ca. 500 bp by using paired end reads) in combination with read numbers in the 10^7 range. All of this in an economically competitive environment in which sequencing run costs continue to decrease. The bioinformatic bottleneck remains in terms of computer processing capacity, and thus time, but specifically in relation to mining metagenomes for genes encoding enzymes, the future is the ability to reliably predict and model protein structure and function *in silico* and thus identify truly novel cellulases among those numerous translated metagenomic sequences that lack homology with any known protein-encoding sequences.

References

- Bras JL, Cartmell A, Carvalho AL, et al. Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis. *Proc Natl Acad Sci U S A*. 2011;108:5237–42.
- Cantarel BL, Coutinho PM, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 2009;37:233–8.
- Damon C, Lehembre F, Oger-Desfeux C, et al. Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PLoS ONE*. 2012;7:e28967.
- Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. *Structure*. 1995;3:853–9.
- Duan CJ, Xian L, Zhao GC, et al. Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J Appl Microbiol*. 2009;107:245–56.
- Fontes CM, Gilbert HJ. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Ann Rev Biochem*. 2010;79:655–81.
- Gilbert JA, Field D, Huang Y, et al. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE*. 2008;3:e3042.
- Henrissat B, Teeri TT, Warren RA. A scheme for designating enzymes that hydrolyse the polysaccharides in the cell walls of plants. *FEBS Lett*. 1998;425:352–4.
- Horn SJ, Sikorski P, Cederkvist JB, et al. Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *PNAS*. 2006;103:18089–18094.
- Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17:377–86.
- Leininger S, Urich T, Schlöter M, et al. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*. 2006;442:806–9.
- Lynd LR, Weimer PJ, van Zyl WH, et al. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev*. 2002;66:506–77.
- McDonald JE, Rooks DJ, McCarthy AJ. Methods for the isolation of cellulose-degrading microorganisms. *Methods Enzymol*. 2012;510:349–74.
- Mello LV, Chen X, Rigden DJ. Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Lett*. 2010;584:2421–6.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma*. 2008;9:386–92.
- Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 2005;33:5691–702.
- Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;54:59–65.
- Quinlan RJ, Sweeney MD, Lo Leggio L, et al. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. *Proc Natl Acad Sci U S A*. 2011;108:15079–84.
- Ransom-Jones E, Jones DL, McCarthy AJ, et al. The fibrobacteres: an important phylum of cellulose-degrading bacteria. *Microb Ecol*. 2012;63:267–81.
- Rooks DJ, McDonald JE, McCarthy AJ. Metagenomic approaches to the discovery of cellulases. *Methods Enzymol*. 2012;510:375–94.
- Shoseyov O, Shani Z, Levy I. Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol Mol Biol Rev*. 2006;70:283–95.
- Stam MR, Danchin EG, Rancurel C, et al. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng Des Sel*. 2006;19:555–62.
- Sukharnikov LO, Cantwell BJ, Podar M, et al. Cellulases: ambiguous nonhomologous enzymes in a genomic perspective. *Trends Biotechnol*. 2011;29:473–9.
- Takasaki K, Miura T, Kanno M, et al. Discovery of glycoside hydrolase enzymes in an avicel-adapted forest soil fungal community by a metatranscriptomic approach. *PLoS ONE*. 2013;8:e55485.
- Voget S, Steele HL, Streit WR. Characterization of a metagenome-derived halotolerant cellulase. *J Biotechnol*. 2006;126:26–36.
- Warnecke F, Luginbuhl P, Ivanova N, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. 2007;450:560–5.
- Zhou W, Schuttler HB, Hao Z, et al. Cellulose hydrolysis in evolving substrate morphologies I: a general modeling formalism. *Biotech Bioeng*. 2009;104:261–74.

Mock Community Analysis

Sarah Highlander
Genomic Medicine, J. Craig Venter Institute,
La Jolla, CA, USA

Definitions

Mock community: A defined mixture of microbial cells and/or viruses or nucleic acid molecules created *in vitro* to simulate the composition of a microbiome sample or the nucleic acid isolated therefrom.

Microbiome: The microbes (bacteria, archaea, fungi, protists, and viruses) that inhabit

a specific environment or host, such as all the microbes that live in and on the human body.

Metagenome: The complete DNA (genomic) content of a microbiome sample. The term “metagenome” was first used by Handelsman et al. to describe the “collective genomes of soil microflora” (Handelsman et al. 1998).

Introduction

Although a few studies have reported creation of mock communities for environmental microbial systems, this review will be restricted to mock communities that have been developed for studies of the human microbiome.

Microbial Mock Communities

For the human sampling aspect of the Human Microbiome Project (HMP), the clinical centers at Baylor College of Medicine and the Washington University School of Medicine were tasked with obtaining microbiome samples from 18 different body sites. These samples were in the form of saliva, tooth scrapings, buccal swabs, vaginal swabs, nasal swabs, skin scrapings, feces, etc. Each sample had a different physical and microbial composition, yet it was necessary to have a standard and uniform method of DNA extraction for each. The method selected included chemical lysis with sodium dodecyl sulfate (SDS) and mechanical disruption by bead beating followed by column purification of the DNA from the cell lysate (http://www.hmpdacc.org/doc/HMP_MOP_Version12_0_072910.pdf) using the MO BIO PowerSoil DNA Isolation Kit (Carlsbad, CA). As a means to evaluate the DNA purification protocol, we created a mock cell community that consists of 22 bacterial strains and one archaeal strain, mostly representing strains found at different sites within the human body (Table 1). The strains were selected as having different features such as different cell wall compositions (gram positive, gram negative, spore formers, encapsulated, thick cell wall),

Mock Community Analysis, Table 1 Strains in the HMP mock cell community (BEI HM-280)

Genus species	Strain number
<i>Acinetobacter baumannii</i>	ATCC 17978
<i>Actinomyces odontolyticus</i>	ATCC 17982
<i>Bacillus cereus</i>	ATCC 10987
<i>Bacteroides vulgatus</i>	ATCC 8482
<i>Bifidobacterium adolescentis</i>	DSM 20083
<i>Clostridium beijerinckii</i>	ATCC 51743
<i>Deinococcus radiodurans</i>	ATCC 13939
<i>Enterococcus faecalis</i>	ATCC 47077
<i>Escherichia coli</i>	ATCC 700296
<i>Helicobacter pylori</i>	ATCC 700392
<i>Lactobacillus gasseri</i>	ATCC 33323
<i>Listeria monocytogenes</i>	ATCC BAA-679
<i>Methanobrevibacter smithii</i>	ATCC 35061
<i>Neisseria meningitidis</i>	ATCC BAA-335
<i>Porphyromonas gingivalis</i>	ATCC 33277
<i>Propionibacterium acnes</i>	DSM 16379
<i>Pseudomonas aeruginosa</i>	ATCC 47085
<i>Rhodobacter sphaeroides</i>	ATCC 17023
<i>Staphylococcus aureus</i>	ATCC BAA-1718
<i>Staphylococcus epidermidis</i>	ATCC 12228
<i>Streptococcus agalactiae</i>	ATCC BAA-611
<i>Streptococcus mutans</i>	ATCC 700610
<i>Streptococcus pneumoniae</i>	ATCC BAA-334

aerobe or anaerobe, high and low percent G+C, and having completely sequenced genomes. The strains were grown under appropriate growth conditions to late logarithmic or stationary phase and then mixed at an equal ratio at a concentration of 10^8 cells/ml. The cell mix is available, free of charge, from BEI Resources (www.beiresources.org). A similar mixture, formulated in 40 % glycerol (BEI HM-281), was also created to be used as a viable mock community for single cell studies.

We extracted DNA from the mock cells community using the HMP standard DNA isolation protocol, then performed 454 amplicon sequencing of the 16S ribosomal RNA variable regions, V1-V3 regions. We failed to detect any *M. smithii* or bifidobacterial reads and recovered less than 1 % of the total reads corresponding to the following input organisms: *Acinetobacter*

baumannii, *Actinomyces odontolyticus*, *Clostridium beijerinckii*, *Deinococcus radiodurans*, *Helicobacter pylori*, *Lactobacillus gasseri*, *Rhodobacter sphaeroides*, or *Streptococcus* spp. In contrast, the relative abundance of *Neisseria* reads was approximately 35 % and the relative abundance of *Bacillus* and *Enterococcus* reads isolated from the mock community was approximately 15 % for each genera. These observations are likely due to a combination of the relative ability of an organism to be lysed and the percent match of the 16S primers to rRNA gene targets. For example, it is known that the 534R primer has numerous mismatches to actinobacterial 16S rRNA genes, particularly to the bifidobacteria, and an evaluation of primer mismatches to other members of the mock community revealed F27 mismatches to *Acinetobacter*, *Pseudomonas*, and *Escherichia* and numerous 534R mismatches to *Methanobrevibacter*, as described below.

Although we did not use our mock cell community for rigorous testing of lysis and DNA extraction methods for metagenomics, Yuan et al. have performed a systematic evaluation of common DNA extraction methods (Yuan et al. 2012), using a mock community composed of equal cell counts of 11 bacterial species chosen to represent different human body sites: *E. coli*, *S. aureus*, *P. aeruginosa*, *S. agalactiae*, *Corynebacterium tuberculostrictum*, *Lactobacillus iners*, *Lactobacillus crispatus*, *Atopobium vaginae*, *Gardnerella vaginalis*, and *P. acnes*. They compared six different DNA methods that combined different lysis (enzymatic, chemical, and bead beating) and DNA purification (silica column or phenol/chloroform plus isopropanol precipitation) methods. DNA yield and DNA integrity (shearing) were evaluated. Microbial abundance was measured by 454 sequencing of the 16S rDNA V1-V2 regions using a mixture of forward primers that were chosen to prevent bias against *Lactobacillus* spp. and *Gardnerella* spp. (Yuan et al. 2012), followed by statistical analyses that included accommodation for differences in 16S rRNA gene copy number per organism. Extraction methods that included

bead beating (as included in the HMP protocol) delivered the best representation of the community structure. Addition of mutanolysin, but not lysozyme, or lysostaphin, also enhanced recovery of the expected proportions of 16S rRNA gene sequences. In sum, *L. iners* was overrepresented using all techniques and the two gram-negative organisms, *E. coli* and *P. aeruginosa*, were underrepresented in all. Thus, the authors caution that none of the methods tested returned the actual representation of the input mock community.

In another comparison of extraction methods, Willner et al. created a mock community of 12 strains that included organisms relevant to respiratory infections and cystic fibrosis (CF) (Willner et al. 2012). The goal was to use this mock community to compare and evaluate methods for DNA extraction prior to their application to clinical bronchoalveolar lavage (BAL) samples obtained from CF patients. They also developed an *in silico* simulation of the mock BAL community using the software package Grinder (<http://sourceforge.net/projects/biogrinder/>). The mock community was composed of the following bacterial species from actively growing stocks (relative proportions in parentheses): *P. aeruginosa* (1), *Burkholderia cepacia* (0.1), *S. aureus* (0.1), *Haemophilus influenzae* (0.1), *Moraxella catarrhalis* (0.01), *S. epidermidis* (0.01), *Klebsiella pneumoniae* (0.01), *N. meningitidis* (0.001), *Burkholderia multivorans* (0.001), *Legionella pneumophila* (0.0001), *S. pneumoniae* (0.0001), and *Neisseria gonorrhoeae* (0.00001). Aliquots of the mock community were extracted using a “CTAB method,” a “saline protocol,” using the NucleoSpin Tissue Kit (pellet and liquid protocols) and the MO BIO PowerSoil Kit. Community abundance was evaluated by 454 16S rDNA sequencing of the V8-V9 regions using a degenerate 1,114 F3 primer (Willner et al. 2012). Data were normalized to 900 reads per sample. At this level, few (<1 %) to no streptococcal reads were detected in most of the preparations, and no *Legionella* reads were detected in any of the preparations. In contrast,

the abundance of *Neisseria* reads was greater than that predicted by the *in silico* model. Unfortunately, each sample included *Escherichia* and *Dechloromonas* as contaminants and the CTAB samples had a high percentage of *Stentrophomonas*. This made it difficult to draw conclusions about the efficiency and reproducibility of the methods employed.

Diaz et al. created two different oral bacterial mock cell mixes, one in even cell distribution and one in unequal distribution, using seven species that are representative of the tooth surface (Diaz et al. 2012): *Streptococcus oralis*, *S. mutans*, *Lactobacillus casei*, *Actinomyces oris*, *Fusobacterium nucleatum*, *P. gingivalis*, and *Veillonella* sp. Late logarithmic phase cultures were mixed in an even distribution based on cell counts and in an uneven mixture that replicated the proportions found in the oral cavity. These cell communities were lysed using a single protocol that included lysozyme treatment, overnight proteinase K digestion, and column purification of the DNA. As a control, genomic DNAs from the seven bacteria were mixed, in equal proportion based on 16S rRNA gene copy number. All DNAs were used for 454 sequencing of the V1-V2 region of the 16S rRNA gene. Very few *S. mutans* or *P. gingivalis* reads were recovered, despite efficient sequencing of the control DNA, suggesting that the lysis method was inefficient for these two members of the mock community.

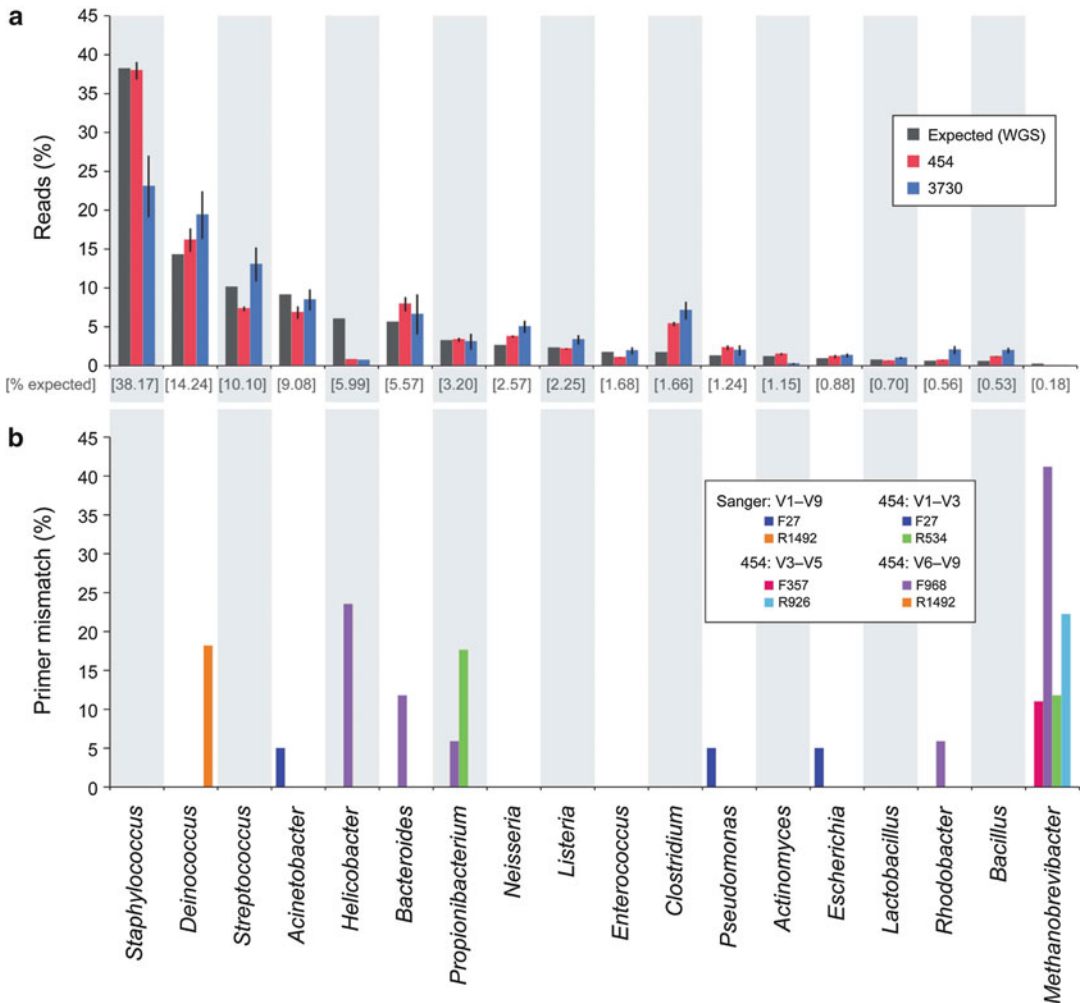
DNA Mock Communities

While mock communities composed of mixtures of cells were intended to be used to evaluate different DNA extraction methods, they also revealed biases in 16S rRNA gene amplification, sequencing, and classification. DNA mock communities have been created in attempts to address these issues, to examine sensitivity and presence of chimeric sequences, to serve as controls for protocol development, etc. DNA mock communities can be composed of mixtures of genomic DNA, of plasmid clones of genes (usually the 16S rRNA gene), or of PCR amplicons. Generally,

these mock communities were created as controls and calibrators for 16S rRNA gene variable region sequencing on next-generation sequencing platforms, but they are useful in the context of metagenomic sequencing as well.

Turnbaugh et al. used genomic DNA from 67 gut bacterial strains (e.g., containing the genera *Bifidobacterium*, *Collinsella*, *Bacteroides*, *Prevotella*, *Clostridium*, *Dorea*, *Roseburia*, *Ruminococcus*, *Streptococcus*, *Citrobacter*, *Enterobacter*, *Proteus*, and *Providencia*) to create even and uneven mixtures as calibrators for 454 16S rDNA sequencing of the V2 region for a twin study of gut microbiomes (Turnbaugh et al. 2010). Following quality filtering, pyrosequencing, denoising and chimera removal, the estimated diversity (at 97 % species cutoff) of the three uneven mock communities was 75, 58, and 63, respectively, which was remarkably close to the 62 phylotypes expected in the community. Diversity was not estimated for the even communities, although the ratio of observed-to-expected sequences by phylotype was tabulated and reported. This revealed an absence of bifidobacterial reads, due to multiple primer mismatches, and overabundances of sequences mapping to other genera, including the *Bacteroides* and the clostridia. The authors acknowledged that these observations could be the result of a number of factors including variations in 16S rRNA gene copy number per strain and DNA quality.

During the development of standardized 454 16S rDNA sequencing protocols for the HMP, we created mock DNA communities using genomic DNAs from 21 of the strains listed in Table 1 (*B. adolescentis* and *P. gingivalis* were not included) plus *Candida albicans* MYA-2876. DNAs were prepared from individual cultures and each DNA preparation was validated for purity by Sanger paired-end sequencing of 384 full-length 16S rDNA clones obtained from each. Genomic DNAs were combined, based on 16S rRNA gene copy number, to form even or staggered mock communities. The even communities theoretically contained 10^5 16S rDNA copies from each species per amplification reaction, and the staggered communities had 16S rDNA copies that ranged from

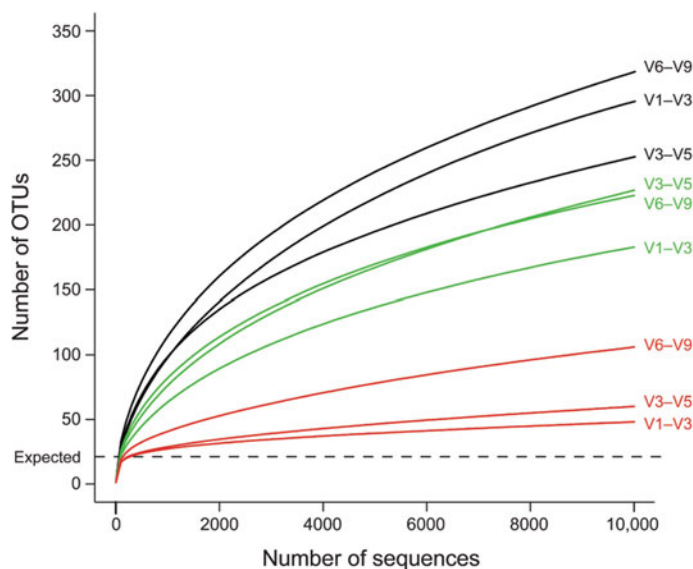


Mock Community Analysis, Fig. 1 Deviation from the expected for the 16S rDNA sequencing of the 20 bacterial + one archaeal mock community. **(a)** Distribution of reads over the 18 genera; expected frequencies (gray) were determined by whole-genome shotgun sequencing of the mock community, and observed frequencies were

determined by 454 reads (red) or Sanger 3,730 reads (blue). Error bars represent standard error. **(b)** Lowest percent mismatch between primer and 16S rDNA gene copy by organism, sequencing technology, and variable gene region (Jumpstart Consortium Human Microbiome Project Data Generation Working Group 2012)

10^3 to 10^6 copies from a particular species per reaction. All reactions contained approximately 1,000 copies of the *C. albicans* 18S rRNA gene (Haas et al. 2011) (Jumpstart Consortium Human Microbiome Project Data Generation Working Group 2012). These mock communities were used to develop an improved chimera detection tool, called ChimeraSlayer, and revealed a high level of chimerism in short variable region products (Haas et al. 2011). The

mock community was essential to validate and benchmark methods for 16S rDNA sequencing by all four genome sequencing centers involved in the HMP (the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute, and the Washington University Genome Sequencing Center) and revealed clear cases of primer mismatches that caused some genera to be underrepresented (Fig. 1).



Mock Community Analysis, Fig. 2 Quality filtering and chimera checking and removal improve estimates of community diversity as evaluated by rarefaction analysis. Operational taxonomic units (*OTUs*) are plotted versus number of 454 sequence reads for three 16S rDNA variable region windows, V1-V3, V3-V5, and V6-V9, before

quality filtering (*black*), after quality filtering (*green*), and after quality filtering and chimera removal (*red*). The expected number of *OTUs* in the mock community was 18 (*dotted line*) (Jumpstart Consortium Human Microbiome Project Data Generation Working Group 2012)

Sequencing the mock community clearly illustrated the need for quality filtering and chimera checking of 454 data of variable regions as illustrated in Fig. 2. Without filtering, the diversity of the 21 species (18 operational taxonomic units) in the community is estimated number in the 100s. Following quality filtering and chimera removal, the community richness is only a few-fold higher than expected, especially for the V1-V3 and V3-V5 regions of the 16S rDNA gene.

The HMP mock community also revealed examples of misclassification, poor classifiability (lowest in V6-V9), and unexplained overrepresentation of some genera (Jumpstart Consortium Human Microbiome Project Data Generation Working Group 2012). This mock community has been used to evaluate how quality filtering impacts taxonomic classification of reads generated on the Illumina platform (Bokulich et al. 2013), and a modified version has been used to develop

a dual-index method for 16S amplicon sequencing on the Illumina MiSeq platform (Kozich et al. 2013).

Another type of mock DNA community is a set of plasmid clones of nearly full-length 16S rRNA gene fragments (Wu et al. 2010). Here, PCR amplicons from *Clostridium difficile*, *Bacteroides fragilis*, *S. pneumoniae*, *Desulfovibrio vulgaris*, *Campylobacter jejuni*, *Rhizobium vitis*, *Lactobacillus delbrueckii*, *E. coli*, *Treponema* sp., and *Nitrosomonas* sp. were cloned into pTOPO vectors and then used to create even and staggered DNA mixtures, which were then used as templates for 454 sequencing of the V1-V2 regions of the 16S rRNA gene. The authors report that correct proportions of input 16S rDNA sequence type were recovered following 454 sequencing and analysis, although different polymerases used for replicates of the staggered community gave slightly different results. Use of cloned 16S rRNA genes as controls is convenient, although genes on

supercoiled high copy number plasmids are not likely to be good surrogates for chromosomal ribosomal genes.

Summary

DNA mock communities have identified problems with use of “universal” 16S rRNA gene primers for amplification of variable regions for microbiome sequencing and have revealed flaws in taxonomic classification systems, where known sequences were classified incorrectly (Jumpstart Consortium Human Microbiome Project Data Generation Working Group 2012). They have also shown the critical requirement for stringent read quality filtering and chimera removal of 16S rDNA sequencing reads, which has helped to reduce estimates of the size the “rare biosphere” of human microbiome.

Mock communities of cells have proved valuable as controls for development of uniform DNA extraction methods for microbiome samples and DNA mixtures continue to be important as calibrators for 16S rDNA and metagenomic sequencing on changing high-throughput platforms. Neither type of mock community is perfect. Cell mixtures are easily contaminated; they may have incorrect cell counts due to clumping, dead cells, or the presence of bacteriophage; and are limited to species that can be grown without difficulty. DNA mixtures may also be contaminated, so it is important to validate the purity of the preparations prior to mixing, and mixtures based on 16S rDNA copy number may be skewed if calculations or assumptions are incorrect, particularly if the genomes of the input DNAs are not finished. Cell communities are plagued with the same issues of amplification bias and misclassification discovered with DNA using DNA communities.

Despite the flaws inherent in mock communities, they are useful as a uniform benchmark for microbiome and metagenome technology development and evaluation. The concept could be expanded to include mock communities of viruses, and fungi. Further, one could imagine developing mock communities composed of different types of

molecules such as RNAs or peptides or known components of the metabolome as being useful controls for microbiome work.

Cross-References

- ▶ [Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes](#)
- ▶ [Extraction Methods, Variability Encountered in](#)

References

- Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*. 2013;10:57–9.
- Diaz PI, Dupuy AK, Abusleme L, et al. Using high throughput sequencing to explore the biodiversity in oral bacterial communities. *Mol Oral Microbiol*. 2012;27:182–201.
- Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 2011;21:494–504.
- Handelsman J, Rondon MR, Brady SF, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. 1998;5:R245–9.
- Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE*. 2012;7:e39315.
- Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*. 2013;79:5112–20.
- Turnbaugh PJ, Quince C, Faith JJ, et al. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA*. 2010;107:7503–8.
- Willner D, Daly J, Whiley D, et al. Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples. *PLoS ONE*. 2012;7:e34605.
- Wu GD, Lewis JD, Hoffmann C, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol*. 2010;10:206.
- Yuan S, Cohen DB, Ravel J, et al. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS ONE*. 2012;7:e33865.

Molecular Ecological Network of Microbial Communities

Ye Deng¹ and Jizhong (Joe) Zhou^{2,3,4}

¹Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

²Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

³Department of Environmental Science and Engineering, Tsinghua University, Beijing, China

⁴Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Synonyms

Microbial community network; Microbial ecological network; Microbial interaction network; Microbial network; Molecular ecological network

Definition

The network of microbial communities constructed using molecule-based experimental data, especially metagenomic data (e.g., microarray hybridization, sequencing), is referred to as molecular ecological network (MEN). It aims to understand the interaction of members in a given community. If the molecules are based on phylogenetic gene markers (e.g., 16S small subunit ribosomal DNA), the network is defined as phylogenetic molecular ecological networks (pMENs).

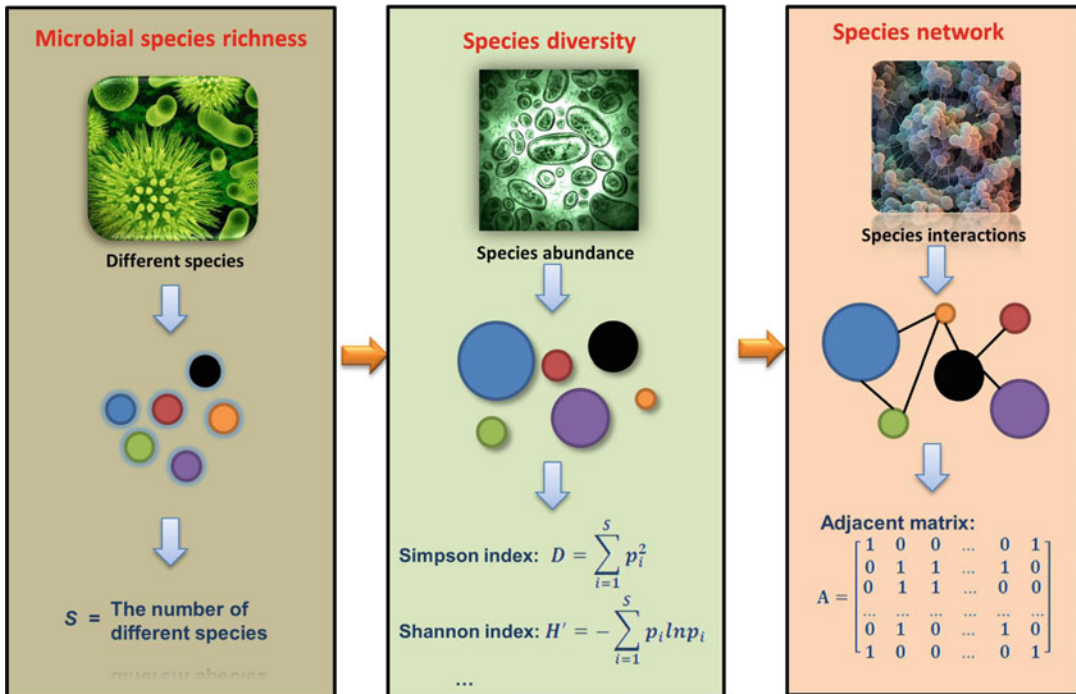
Introduction

In microbial ecology, the majority of data analytical efforts are focused on revealing the composition and diversity of a microbial community and also the changes across space, time series, and/or with experimental treatments. The conventional analytical approaches usually use species richness and α -diversity to depict a community structure, and several diversity indexes, such as

Simpson index and Shannon index, are used to measure the level of α -diversity (Fig. 1). Besides, the difference between two communities is often estimated by β -diversity, and more multivariate statistical techniques are used to describe the community patterns and associations with environmental factors, such as ordination and regression methods (Deng 2013). Compared to the intensive studies in community compositions and diversities, there is much less attention on the interaction and network relationships among microbial species (Zhou et al. 2010).

In natural environment, the microbial species rarely live independently; instead, a large amount of organisms tend to exist sympatrically and synchronously through various types of symbiotic relationships. Their relationship could be positive (mutualism and commensalism) or negative (competition, predation, and amensalism) to the partner species (Faust and Raes 2012), and all relationships among the species form a complicated interaction web. These relationships can be simply exhibited as a network structure (Fig. 1), in which each node represents a species and the edge linking two nodes represents the interaction between these two species. More complex relationships in the community could be integrated into the network model as well. For instance, the strength of relationship could be represented to the edge weights, and regulatory or beneficial relationships could be represented to the edge directions. Additionally, the abundance of species could be visualized as the sizes of nodes. Therefore, a comprehensive network structure could adequately depict the inherent relationships within a microbial community.

Since the end of the last century, the ecological network studies have been started and well developed in the macro-ecology (Montoya et al. 2006). Food web structures have been intensively studied due to their crucial contribution to the stability of creaturely communities (Pimm 2002). Meanwhile, the mutualistic networks have also evoked a lot of attention (Bascompte and Jordano 2007). Through those interactions, an ecosystem is capable of accomplishing its systems-level functions which could not be achieved by individual populations. Therefore,



Molecular Ecological Network of Microbial Communities, Fig. 1 The study of microbial ecology from species richness and diversity to interaction network

explaining the ecological network structures, dynamics, and mechanisms has become an essential part in ecology. However, the studies on interactions among microbial species are much more difficult than those studies in macro-ecology, majorly due to their incredibly high species diversity. Besides, most natural microbial species are uncultivable and also invisible to the naked eyes, which makes it more challenging to define network structure in a microbial community. Here, the definition of phylogenetic molecular ecological network (pMEN) for microbial community, the network inference, and the common network properties are first introduced, and then several key ecological questions are able to be addressed through network analysis.

Phylogenetic Molecular Ecological Network for Microbial Communities

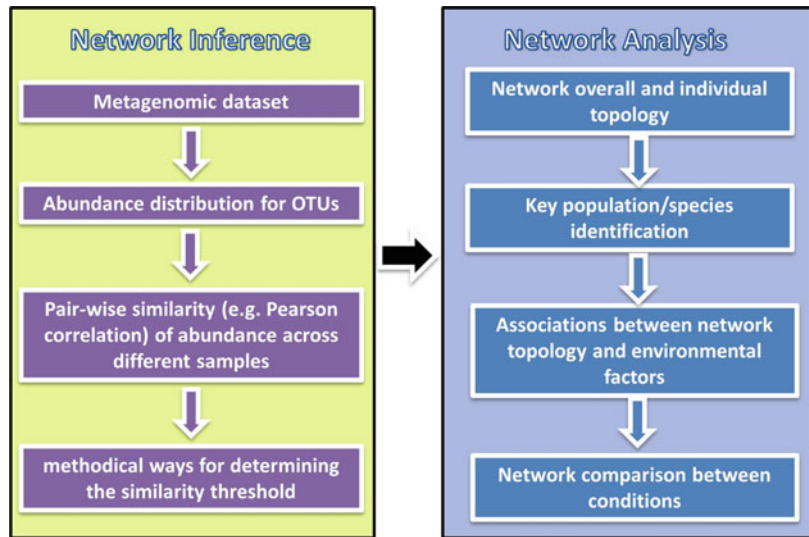
Owing to the technique innovation of molecular biology, the modern microbial taxonomy often

relies on phylogenetic molecular markers, such as ribosomal RNA (rRNA) genes or some highly conserved coding genes (e.g., *nifH*, *amoA*, *gyrB*). In microbial diversity surveys, consequently, the definition of operational taxonomic unit (OTU) is used to delimit the microbial taxa by the similarity of those sequences (Achtman and Wagner 2008). Each OTU then represents a certain taxon, such as a species or a genus. The composition and diversity of microbial communities actually are based on molecular OTUs rather than individual species. Recently, due to the rapid development of high-throughput sequencing technology, large amounts of microbial diversity surveys have been carried out in various environmental habitats through small subunit (SSU) rRNA sequencing projects. These massive, community-wide, replicated metagenomic data provide unprecedented opportunities to infer the interaction networks in microbial communities (Raes and Bork 2008).

As a result, an ecological network generated from metagenomic data really reflects the

Molecular Ecological Network of Microbial Communities, Fig. 2

The common steps of molecular ecological network (MEN) analysis. Two major parts are included: network inferences and network analyses. In each of them, several key steps are listed



relationships among molecular OTUs. Therefore, such molecule-based ecological networks in microbial communities are referred to as molecular ecological networks (MEN) (Zhou et al. 2010). The networks derived from functional gene markers are referred to as functional molecular ecological networks (fMEN) (Zhou et al. 2010), and those based on phylogenetic gene markers as phylogenetic molecular ecological networks (pMENs) (Zhou et al. 2011).

Network Inference Approach

For metagenomic data analysis, the abundance of each gene marker in a sample is measured by the number of sequences for sequencing data or hybridization signal intensity for microarray data. Thereafter, the determined gene richness and abundance are used to describe the composition and structure of this microbial community. Based on such experimental data, a network graph can be constructed to illustrate the interactions of different gene markers (species) (Fig. 2). The way of constructing the connection diagram from the behavior of its components is known as network inference or reverse engineering (Faust and Raes 2012).

Various approaches for network inference have been developed and widely used in both

genomic biology and ecology (Barabasi and Oltvai 2004; Faust and Raes 2012). Based on the mathematical algorithms, they can be classified into Bayesian network, relevance network, and ordinary and partial differential equation methods [reviewed by De Jong (2002)]. Besides, some graphical theory-based methods were recently developed (Kramer et al. 2009). Among them, the relevance network method is the most commonly used approach due to its simple calculation procedure and high noise tolerance (Deng et al. 2012). For the relevance network method, a similarity is first measured between each two OTUs. This similarity measurement can be Pearson, Spearman, biweight, and jackknife correlations or mutual information (Hardin et al. 2007).

For network inference, another critical step is to identify a true link (Fig. 2). The key question is how similar a true link should be. The most commonly used way to choose the similarity threshold is based on biological knowledge which could confirm some true interactions by previous experimental discovery and then use similar values between those interactions to determine the threshold for other links. The constructed network through this arbitrary threshold is subjective rather than objective (Barabasi and Oltvai 2004). There are also a couple of methodical ways for determining the similarity threshold,

such as the significance level of correlation (p value), false discovery rate (FDR), permutation test, and random matrix theory (RMT)-based methods. Among them, p value-based and permutation test-based methods give the least strict threshold and lead to large amounts of links in a messy network that could be alike to random network. The FDR-based method has the strictest threshold, which could generate a loose network, and a lot of true interactions might be ignored. RMT-based algorithm has advantages in this step (Luo et al. 2007). This method is able to automatically identify a threshold based on the inherent property of the similarity matrix. The results indicated it is robust to reveal the meaningful relationships through high-throughput data in both genomics and ecology (Luo et al. 2007; Zhou et al. 2010, 2011).

Network Properties

After the species interactions have been inferred, many pMENs are formed for the communities in different habitats, such as soil, ocean, groundwater, and human guts (Deng et al. 2012; Faust and Raes 2012). Several common topological properties, such as small world, scale-free, and modularity (Table 1) were also observed in all kinds of pMENs, like other biological networks from food webs in macro-ecology to complex regulation networks in molecular biology. These common network properties are important for the robustness and stability of complex systems (Barabasi and Oltvai 2004; Kitano 2004; Zhou et al. 2010, 2011).

The scale-free is a most notable characteristic in complex systems. It is used to describe the finding that most nodes in a system have few directly linked nodes (neighbors), while few nodes have a large amount of neighbors (Table 1). It implies the roles of species in the microbial community might be quite different. A few microbial species could be generalists with higher connectivity which are inclined to have closer relationships with environmental traits than other species (Zhou et al. 2011; Deng et al. 2012).

The “small world” is used to depict that any two nodes in a network can be connected just by passing a few of linked neighbors (Table 1). It is originally referred in sociology that 6° of separation between us and everyone else on this planet. This property usually reflects the efficiency of system and may be valuable for microbial communities. In the small-world community, the energy, materials, and information can be easily transported within the entire system. In the microbial community, this characteristic drives efficient communications among different members so that relevant responses can be taken rapidly to environmental changes (Zhou et al. 2011; Deng et al. 2012).

The modularity property is used to demonstrate that a network could be degraded to sub-networks, also called modules, according to its structure (Table 1). Each module in gene regulatory networks is considered as a functional unit, which consists of several elementary genes and performs an identifiable task (Luo et al. 2007). Modularity in an ecological community may reflect habitat heterogeneity, physical contact, functional association, divergent selection, and/or phylogenetic clustering of closely related species (Olesen et al. 2007; Zhou et al. 2010). Also microorganisms in the same module could have similar ecological niches (Zhou et al. 2011; Faust and Raes 2012).

Except these three common properties, there are many other topological indexes that could be used to measure the organization and structure of microbial networks, such as clustering coefficient, hierarchy, density, transitivity, and connectedness [definitions and descriptions seen in Deng et al. (2012)]. All these could become valuable indexes to measure the microbial structure for the studies of microbial ecology.

Network Interpretation Aspects

Once the network graph is drawn, we should disclose the ecological meanings behind this structure. Several key ecological questions can be revealed through network analysis procedures (Fig. 2).

Molecular Ecological Network of Microbial Communities, Table 1 The most commonly used topological indexes and properties for complex networks

Network property	Mathematic measurement	Ecological implication
Connectivity	$k_i = \sum_{j=1}^m a_{ij}$ where m is the number of all neighbors (linked nodes) of node i and a_{ij} is the strength between nodes i and j . For the unweighted network, k_i equals the number of neighbors	It was used to describe the number of interactions of each node, also named as node degree. In most complex systems, the nodes with the highest connectivity always played crucial roles and were usually considered as network centers. In pMEN, the study found that nodes with higher connectivity were inclined to have closer relationships with environmental traits (Zhou et al. 2011; Deng et al. 2012)
Scale-free	$P(k) \sim k^{-\gamma}$, where $P(k)$ is the number of nodes with k degrees, k is connectivity, and γ is a constant	In most cases, the connectivity distributions of pMEN and other complex systems follow this power law, indicating most nodes in a network have few neighbors, while few nodes have large amount of neighbors. This phenomenon suggests the most species in the communities are peripherals, but a few of the species could be generalists and play more important roles than others
Small world	$GD = \frac{\sum d_{ij}}{n(n-1)}$ GD is the abbreviation of the average geodesic distance, where d_{ij} is the shortest path between nodes i and j , and n is the total number of all nodes	A smaller GD means all the nodes in the network are closer, indicating each two nodes in the network could be connected by a small number of acquaintances, and so-called small-world network. Most pMENS are small-world network, which imply that the energy, materials, and information can be easily transported through entire systems (Deng et al. 2012)
Modularity	$M = \sum_{b=1}^{N_M} \left[\frac{l_b}{L} - \left(\frac{K_b}{2L} \right)^2 \right]$ where N_M is the number of modules in the network, l_b is the number of links among all nodes within the b^{th} module, L is the number of all links in the network, and K_b is the sum of degrees (connectivity) of nodes which are in the b^{th} module	Modularity property was used to demonstrate a network which could be naturally divided into subcommunities, so-called modules. A modularity value can be calculated by Newman's method (Newman 2006) whose value is between 0 and 1. Modularity in an ecological community may reflect habitat heterogeneity, physical contact, functional association, divergent selection, and/or phylogenetic clustering of closely related species (Olesen et al. 2007; Zhou et al. 2010)

Identify Key Populations/Species in the Community Based on Network Topology

In a scale-free network, the roles of nodes for the community could be quite different. Most nodes are just peripheral, and they have less contribution to the network structure and stability. But a few of the nodes may be located in the core of the network, and if it is removed from the network, it will largely change the network structure. These key nodes could be identified by multiple network indexes, such as connectivity, stress

centrality, betweenness, eigenvector centrality, clustering coefficient, and vulnerability [definitions and descriptions seen in Deng et al. (2012)]. The nodes with higher indexes may carry out different functions for the network structure. For example, the nodes with highest connectivity are commonly regarded as centers in the network, while the nodes with highest betweenness usually serve as bridges to connect other nodes. Therefore, in pMEN these key nodes representing species also could play different roles for the

microbial community. Previous results already showed the nodes with higher connectivity were inclined to have closer relationships with environmental traits in pMEN (Zhou et al. 2011; Deng et al. 2012), indicating they could be more important to response the environmental change than other species.

The key nodes also can be determined based on the nodes' roles in their own modules. In a module-separated network, the node topological roles can be defined by two parameters, within-module connectivity (z_i) and among-module connectivity (P_i) (Guimera and Amaral 2005), and the roles of nodes could be illustrated in a ZP plot [seen in Deng et al. (2012) Fig. 4c]. According to values of z_i and P_i , the roles of nodes were classified into four categories: peripherals, connectors, module hubs, and network hubs. In a pMEN, peripherals could represent specialists, while module hubs and connectors are close to generalists and network hubs as supergeneralists (Deng et al. 2012).

Associations Between Network Structures and Environmental Factors

The correlations between pMEN topology and environmental factors can be examined in both direct and indirect ways. Indirectly, the OTU significance (GS) is firstly calculated, which is the square of the correlation coefficient (r^2) between OTU abundance profile and environmental factor. A higher GS value indicates this species better fits the variance of environmental factor than the other species with lower GS values. Thereafter, the correlation between GS and nodes' topological indexes (e.g., connectivity, betweenness) is able to measure the relationship of network topology with environmental factors (Deng et al. 2012). The correlation can be calculated either by using Pearson correlation for single GS or Mantel and partial Mantel tests for multiple GS of environmental factors (Zhou et al. 2011; Deng et al. 2012).

The correlations between module-based eigengenes and environmental factors are able to detect the modules' response to

environmental variance. Module eigengene is the most representative variable for all the OTUs within a module through singular value decomposition (SVD) (Langfelder and Horvath 2007). Eigengene network analysis is feasible to reveal the network organization in module levels and directly test the correlation between modules and environmental factors (Deng et al. 2012). Because the taxa in a module could be functionally associated with overlapping ecological niches (Faust and Raes 2012), the module eigengenes are able to distinguish the module functions by associations with environmental factors.

Network Comparisons for Microbial Communities Under Different Conditions, Locations, or Across Time Series

To analyze how the environment affects network structures and species interactions, the network is constructed and compared under different experimental conditions, geographic locations, or across time succession. Various network indexes can be evaluated among different communities in terms of the network sensitivity and robustness, but since only a single value is available for each network, it is unable to perform standard statistical analyses to assess statistical significance of differences. Thus, the randomized networks are introduced to generate a null model for each identified network. Different methods can randomize the network differently; however, the commonly used Maslov-Sneppen method keeps the numbers of nodes and links unchanged but rewires the positions of all links in the pMEN so that the sizes of networks are the same and the randomly rewired networks are comparable with the original ones (Maslov and Sneppen 2002). This method has been typically used for ecological network analyses. For each identified network, usually a total of 100 randomized networks are implemented, and therefore, all network indexes could be generated 100 times. Then the average and standard deviation for each index of all random networks are obtained. The statistical

Z-test is able to test the differences of the indices between the MEN and random networks. Meanwhile, for the comparison between the network indices under different conditions, the Student t-test can be employed by the standard deviations derived from corresponding randomized networks (Deng et al. 2012).

Except for the overall network topology, network comparisons also can be performed in different levels and aspects, such as node overlaps, module preservations, topological roles of individual nodes, and network hubs among different networks (Zhou et al. 2011). The changes among different levels suggested the switches of species interactions under different conditions that could be ecologically important for microbial community to deal with the environmental changes.

Summary

The current studies of microbial networks are still limited but rapidly growing. From the network inferences to network interpretations, there are a lot of fundamental ecological concerns that have not been well addressed, such as how well the modeled networks reflect the real interactions among microbial species, whether these interactions are casual or fixed under different environmental conditions, and how to classify the types of interactions among microbial species (i.e., mutualistic or trophic). Cautions must be taken for the interpretation of the underlying mechanisms that shape microbial communities through the present network analysis.

Nevertheless, by taking the advantage of rapid technical revolution, microbial ecology studies can be performed at a new level, network inferences. Consequently, through the analysis of network structures, previously ignored interactions among microbial species could be revealed and their responses to environmental changes could be disclosed. With the development and complement on methodology, the studies of microbial interaction networks will evoke more and more attention in the near future.

References

- Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol.* 2008;6(6):431–40.
- Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5(2):101–15.
- Bascompte J, Jordano P. Plant-animal mutualistic networks: the architecture of biodiversity. *Annu Rev Ecol Evol Syst.* 2007;38:567–93.
- De Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol.* 2002;9(1):67–103.
- Deng Y. Microarray data analysis. In: He Z, editor. *Microarrays: current technology, innovations and applications.* Norwich: Horizon Scientific Press; 2013.
- Deng Y, Jiang YH, et al. Molecular ecological network analyses. *BMC Bioinforma.* 2012;13(1):113.
- Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol.* 2012;10(8):538–50.
- Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature.* 2005;433(7028):895–900.
- Hardin J, Mitani A, et al. A robust measure of correlation between two genes on a microarray. *BMC Bioinforma.* 2007;25:8.
- Kitano H. Biological robustness. *Nat Rev Genet.* 2004;5(11):826–37.
- Kramer N, Schafer J, et al. Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinforma.* 2009;24:10.
- Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol.* 2007;1:54.
- Luo F, Yang Y, et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinforma.* 2007;8:299.
- Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science.* 2002;296(5569):910–3.
- Montoya JM, Pimm SL, et al. Ecological networks and their fragility. *Nature.* 2006;442(7100):259–64.
- Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci U S A.* 2006;103(23):8577–82.
- Olesen JM, Bascompte J, et al. The modularity of pollination networks. *Proc Natl Acad Sci U S A.* 2007;104(50):19891–6.
- Pimm SL. *Food webs:* University of Chicago Press; 2002.
- Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol.* 2008;6(9):693–9.
- Zhou J, Deng Y, et al. Functional molecular ecological networks. *mBio.* 2010;1(4):e00110–69.
- Zhou J, Deng Y, et al. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO₂. *mBio.* 2011;2(4):e00122–11.

Monitoring Lactic Acid Bacterial Diversity During Shochu Fermentation

Akihito Endo

Department of Food and Cosmetic Science,
Faculty of Bioindustry, Tokyo University of
Agriculture, Abashiri, Hokkaido, Japan

Synonyms

Beneficial microbe; Lactic acid bacteria

Definition

The composition of lactic acid bacteria during fermentation of Japanese traditional distilled spirit is reviewed.

Introduction

Shochu is a Japanese distilled spirit made from several starchy materials. Fermentation of alcoholic beverages is usually carried out by combination of several microorganisms. Lactic acid bacteria (LAB) are well known to play beneficial roles in several food fermentations, including dairy products, vegetables, and meat. This bacterial group is also beneficial for fermentation of beverages, e.g., aroma production and reduction of acid level in wines and growth prevention of spoilage microorganisms in sake. Recent culture-dependent and culture-independent study revealed that LAB can be seen in fermentation mashes of *shochu*. In the present chapter, lactic acid bacterial diversity during *shochu* fermentation is briefly reviewed.

Shochu Fermentation

Shochu is a popular Japanese distilled spirit and is mainly produced in the south Kyushu area of

Japan. Rice, sweet potato, or barley is usually used as the main ingredient. The fermentation process consists of three stages, i.e., *koji* production, yeast-seed production, and alcoholic fermentation. *Koji* mold (*Aspergillus niger* or *A. kawachii*) saccharifies starches to glucose by amylases, and yeast, *Saccharomyces cerevisiae*, is responsible for alcoholic fermentation. Such microorganisms are inoculated into the fermentation as starters. Rice or barley is usually used as a *koji* ingredient, and the main ingredient is added to the mash at the beginning of alcoholic fermentation stage. During the saccharification, the mold produces large amounts of citric acid, resulting in significant decrease of pH. The pH in the yeast-seed production stage is therefore between 3.0 and 3.5 and that in the alcoholic fermentation stage between 4.0 and 4.5. Alcoholic concentration at the end of fermentation is 14–17 % (v/v). Because of such harsh environment for bacterial survival, bacterial diversity had been considered to be poor, and very few studies have done for this microbiota so far.

LAB Diversity in Fermentation Mashes of *Shochu*

Culture-dependent and culture-independent studies have been carried out to study for LAB diversity in fermentation mashes of *shochu*. LAB population in yeast-seed is generally low, i.e., below 10^5 CFU/ml of mash as determined by culturing and real-time quantitative PCR. Poor LAB diversity (0–2 species in each mash) is seen in the yeast-seed. *Lactobacillus plantarum*, *L. paracasei*, *Weissella confusacibaria*, *Leuconostoc citreum*, and *Enterococcus faecium* have been found in the mashes by denaturing gradient gel electrophoresis (DGGE) combined with LAB-specific primers (Endo 2005; Endo and Okada 2005b). LAB diversity in alcoholic fermentation stage is dependent on the variety of main ingredients. Sweet potato generally produces higher population and richer diversity of LAB than rice or barley. This might be due to differences of nutrition between the main

ingredients. Sweet potato mashes contain 10^4 – 10^8 CFU/ml of LAB and rice or barley mashes contain 10^4 – 10^5 CFU/ml or less. *Lactobacillus brevis*, *L. fermentum*, *L. helveticus*, *L. hilgardii*, *L. kefir*, *L. nagelii*, *L. paracasei*, *L. pentosus*, *L. plantarum*, *Leuconostoc mesenteroides*, *Leuc. citreum*, *Leuc. lactis*, *Lactococcus lactis*, *Enterococcus faecium*, *Pediococcus pentosaceus*, and *W. confusalis/cibaria* have been found in alcoholic fermentation mashes made from sweet potato (Endo 2005; Endo and Okada 2005b). *Lactobacillus satsumensis* is a novel species found in the mashes (Endo and Okada 2005a). Of such species, *W. confusalis/cibaria* is the most seen species. Several species found in this fermentation can be also seen in wine fermentation. This might be due to similar harsh environments (high alcohol content and low pH) in the fermentation of the two alcoholic beverages. Mashes made from rice or barley contain poorer LAB diversity than those made from sweet potato.

An interesting DNA sequence, which was characterized as uncultured *Leuconostoc* sp., was found in yeast-seed by DGGE profile. BLAST analysis of the sequence revealed low similarities (below 95 %) against known *Leuconostoc* spp. but high similarities (99.3 %) against uncultured *Leuconostoc* spp. (accession nos. EU469745 and AJ405013) (Endo 2005, 2011), suggesting the presence of unknown LAB in *shochu* mashes. Population of the organism is approximately 10^8 CFU/ml as determined by qPCR. Because of its predominance in the yeast-seed, it might be an acid-tolerant *Leuconostoc* sp., although *Leuconostoc* spp. are known to be acid sensitive.

Most of LAB strains found in *shochu* mashes were resistant to 10–15 % (v/v) of alcohol, and, moreover, they were able to grow at pH 3.5 (Endo 2011). Very few strains were able to grow at pH 3.0. Most of the strains metabolize citrate when in the presence of glucose. These characteristics suggest that LAB seen in *shochu* mashes have adapted to their habitat. Citrate metabolism by LAB produces several aroma compounds, including diacetyl, acetoin, and acetic acid. These compounds have both positive and

negative impacts on the quality of the final product. Proper management of LAB might therefore introduce *shochu* having better quality.

Summary

Shochu is a Japanese traditional distilled spirit made from starchy materials. During the fermentation, *Aspergillus* spp. works for saccharification of ingredients and *Saccharomyces cerevisiae* plays alcoholic fermentation. *Aspergillus* spp. produces large amounts of citric acid during the fermentation and preserves the fermentation from spoilage. LAB have generally low population and poor diversity at the beginning of fermentation (yeast-seed stage), but their population and diversity increase at the latter fermentation (alcoholic fermentation stage). *W. confusalis/cibaria*, *Lactobacillus* spp., and *Leuconostoc* spp. are usually seen in the fermentation. Such LAB have characteristics to survive in alcoholic and acidic environment, suggesting that LAB have adapted to their habitat.

Cross-References

- ▶ [Culturing](#)
- ▶ [Evaluating Putative Chimeric Sequences from PCR-amplified Products](#)
- ▶ [Phylogenetics, Overview](#)

References

- Endo A. Lactic acid bacterial diversity during *shochu* fermentation. PhD thesis, Tokyo University of Agriculture; 2005.
- Endo A. Diversity of lactic acid bacteria in fermented products. *Jpn J Lactic Acid Bact.* 2011;22:87–92.
- Endo A, Okada S. *Lactobacillus satsumensis* sp. nov., isolated from mashes of *shochu*, a traditional Japanese distilled spirit made from fermented rice and other starchy materials. *Int J Syst Evol Microbiol.* 2005a;55:83–5.
- Endo A, Okada S. Monitoring the lactic acid bacterial diversity during *shochu* fermentation by PCR-denaturing gradient gel electrophoresis. *J Biosci Bioeng.* 2005b;99:216–21.

MRL and SuperFine+MRL

Tandy Warnow

Institute for Genomic Biology, University of Illinois, IL, USA

Synonyms

Phylogeny = phylogenetic tree = tree;

MRL = matrix representation with likelihood;

MRP = matrix representation with parsimony

Introduction

The estimation of evolutionary trees is one of the basic challenges in biology (Felsenstein 2003), but current methods have great difficulties with large datasets – often due to computational issues. For example, methods like maximum likelihood (ML) and maximum parsimony (MP) are highly accurate techniques when they can be properly run, but both are NP-hard (a technical term that has the consequence that exact algorithms are not likely to be found except through exhaustive search techniques). As a result, ML and MP analyses on large datasets either cannot be run at all, or take a very long time to run, or return poor results. Since most accounts of the number of species suggest that the Tree of Life itself will involve many millions of species, truly large-scale phylogenetic estimation is beyond the reach of current methods. Instead, an alternative approach has been proposed, in which different research groups would calculate phylogenetic trees on subsets of the species set and then these trees would be combined into a tree on the full dataset. The techniques that combine trees into a tree on the full taxon set are called “supertree methods,” and the resultant large tree is called a “supertree.”

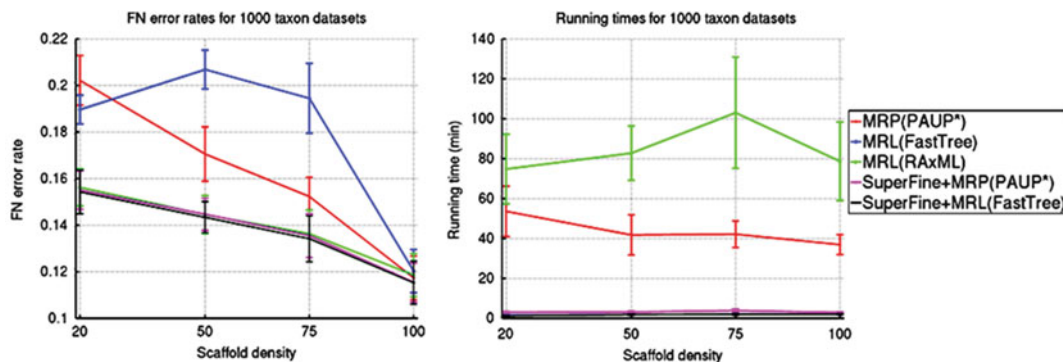
There are many supertree methods (surveyed in Bininda-Emonds 2004), but matrix representation with parsimony (MRP), developed in Baum (1992) and Ragan (1992), is the most well known and most frequently used. MRP operates in two

steps: first the input source trees are each represented by a matrix over $\{0,1,?\}$, where each row represents a species and each column represents a branch in the source tree. These matrices are then concatenated together to form the “MRP matrix.” Finally, this matrix is analyzed using maximum parsimony heuristics, where maximum parsimony is the NP-hard optimization problem that seeks to find a tree on the species set with the smallest number of total changes.

In Swenson et al. (2011), MRP was compared to a collection of other supertree methods and found to be the most reliable with respect to accuracy and ability to analyze large datasets. However, that study also showed that the Quartets MaxCut (QMC) method developed by Snir and Rao (2012) was more accurate than MRP for those datasets on which QMC was able to run. An interesting variant on MRP was developed by Nguyen et al. (2012), in which the MRP matrix was analyzed under maximum likelihood, using a symmetric two-state model. This method, called MRL for “matrix representation with likelihood,” was shown to be more accurate than MRP on simulated datasets.

Thus, while MRP remains the most frequently used supertree method, MRL and QMC are new supertree methods that offer some advantages over MRP; furthermore, new supertree methods continue to be developed.

In Swenson et al. (2012), a new technique was developed called “SuperFine.” This is a meta-method that can be used with any supertree method (e.g., MRP, MRL, QMC, etc.), to produce a modified supertree method. For example, when SuperFine is used with MRP, it is referred to as SuperFine+MRP, and when it is used with MRL, it is referred to as SuperFine+MRL. SuperFine has two steps. The first step computes a “strict consensus merger” (SCM) tree from the set of input trees, where the SCM tree contains many high degree nodes (“polytomies”). The second step refines this tree by using the base method to refine each polytomy. The refinement around each polytomy is performed by encoding each of the source trees on a new leafset $\{1..d\}$, where d is the degree of the polytomy; these



MRL and SuperFine+MRL, Fig. 1 We present tree error and running times (in minutes) for supertree methods on ten replicates of 1,000-taxon datasets. The method given parenthetically indicates the heuristic used to solve MRP or MRL (e.g., PAUP* for MP and FastTree-2 (Price et al. 2010) or RAxML for ML). The scaffold density refers to the percentage of the taxa that are in the “scaffold” dataset. We show standard error for the missing

branch rates, and the standard deviation for the running times. Averages are computed for those replicates with sufficient taxonomic overlap to perform an accurate supertree analysis: $n = 10$ for all scaffold densities except $n = 7$ for the 20% scaffold density and $n = 9$ for 50% scaffold density (reproduced (with permission from the publisher) from Nguyen et al. (2012, 7:3))

smaller source trees are then passed to the base supertree method, which computes a supertree on $\{1..d\}$, and this supertree replaces the polytomy. The refinements around the polytomies can be performed in parallel since they are independent. Hence, the second step is not only very fast, but very easily parallelized. In Swenson et al. (2012), they showed that SuperFine+MRP gave much more accurate trees than MRP and was also much faster. They also compared SuperFine+QMC and QMC and showed similar improvements. Finally, Nguyen et al. (2012) compared SuperFine+MRL and MRL and showed similar improvements. Thus, SuperFine is a method that can improve supertree methods.

A comparison between these different methods (SuperFine+MRP, SuperFine+MRL, MRP, and MRL) is shown in Fig. 1. The experiment involves gene trees that evolve within a species tree under a birth-death process, and so may not contain all the taxa; however, some genes are universal and so contain all the taxa. These genes are then used to evolve sequences under different sequence evolution models. There are two types of gene trees – “clade-based” gene trees that are restricted to clades in the species tree and “scaffold” trees that are used

to link the clade-based trees together. The scaffold trees are produced by random sampling of the taxa and then using the universal genes to construct a source tree. Thus, the clade-based source trees contain a subset of the taxa in a clade, while the scaffold trees contain a random subset of the taxa, but may – in some cases – contain all the taxa. The scaffold density refers to the percentage of the taxa in the scaffold tree. Scaffold source trees for the supertree problem are produced by selecting a scaffold density and then concatenating the alignments from the universal genes on the randomly selected scaffold taxa and computing a maximum likelihood tree on the concatenated alignment. Similarly, the clade-based source trees are computed by selecting a clade and then finding the genes that provide the best coverage for that clade (from the clade-based genes), concatenating the alignments, and computing a maximum likelihood tree on the concatenated alignments. Finally, supertrees are computed on the source trees using MRP, MRL, SuperFine+MRP, and SuperFine+MRL. The resultant species trees are then compared to each other with respect to the missing branch rate and running time. Figure 1 shows results obtained on 1,000-taxon simulated datasets and demonstrates that SuperFine+MRP

and SuperFine+MRL provide the best accuracy of all methods and are much faster than the other methods. It also shows that MRL outperforms MRP with respect to accuracy under low scaffold density conditions. Finally, MRP is “solved” using MP heuristics in PAUP* (Swofford 2003), while MRL is “solved” using either FastTree-2 (Price et al. 2010) or RAxML (Stamatakis 2006). Note that the choice of ML heuristic has an impact on the running time and accuracy of the MRL method. Note also that SuperFine+MRL (FastTree) matches the accuracy of SuperFine+MRP(PAUP*) but is much faster.

Summary

The construction of a large phylogeny, potentially spanning the Tree of Life, is considered to be one of the hardest computational problems in biology. A central approach to this problem involves using supertree methods, which combine source trees, each on a subset of the species, into a tree on the full set of species. While many supertree methods have been developed, MRP (matrix representation with parsimony) is the most well known and most frequently used supertree method. However, newer supertree methods – including MRL (matrix representation with likelihood) and QMC (Quartets MaxCut) – have been introduced that provide comparable or better accuracy to MRP. Finally, a new technique for “boosting” supertree methods has been developed. This method, called “SuperFine,” operates in two steps, where the first step constructs a consensus tree from the source trees and the second step uses the base supertree method to refine the consensus tree. Simulations show that SuperFine improves the accuracy of MRP, MRL,

and QMC while also reducing the running time used by these methods. Thus, SuperFine is a general purpose meta-method for improving supertree estimation.

Funding

This work was supported by NSF grant DEB 0733029 to T.W.

References

- Baum BR. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*. 1992;41:3–10.
- Bininda-Emonds O, editor. *Phylogenetic supertrees: combining information to reveal the tree of life*. Dordrecht: Kluwer Academic Publishers; 2004.
- Felsenstein J. *Inferring phylogenies*. Sunderland: Sinauer Associates; 2003.
- Nguyen N, Mirarab S, Warnow T. MRL and SuperFine+MRL: new supertree methods. *Algorithm Mol Biol*. 2012;7:3.
- Price M, Dehal P, Arkin A. FastTree 2 – approximately maximum likelihood trees for large alignments. *PLoS ONE*. 2010;5:e9490.
- Ragan MA. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol*. 1992;1:53–8.
- Snir S, Rao S. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol Phylogenet Evol*. 2012;62:1–8.
- Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
- Swenson MS, Suri R, Linder CR, et al. An experimental study of Quartets MaxCut and other supertree methods. *Algorithm Mol Biol*. 2011;6:7.
- Swenson M, Suri R, Linder CR, et al. SuperFine: fast and accurate supertree estimation. *Syst Biol*. 2012;61:214–27.
- Swofford DL. *PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4*. Sinauer Associates. 2003.

N

New Computational Methodologies to Understand Microbial Diversity

Haiwei Luo
Department of Marine Sciences, University of Georgia, Athens, GA, USA

Synonyms

Bioinformatic methods for exploring genetic diversity; Methods for metagenomic sequence analysis

Definition

Microbial diversity is broadly defined as genetic variation in natural microbial populations.

Introduction

Metagenomics studies the genetic materials of a natural microbial community recovered from an environmental sample. A typical metagenomic study involves two major steps, including an initial experimental stage for genetic material extraction and sequencing and a following stage using standard bioinformatic tools for molecular sequence analysis. The present review, however, focuses on several recently developed computational methods that are designed to explore ecological diversity of

microbial populations through analyzing published metagenomic databases. Although these methods have only been used to mine metagenomic data sets from the oceans, they can be easily adapted to those from any other environments.

An Ensemble Machine Learning Method to Predict Protein Subcellular Localization of Metagenomic Sequences

Bacteria consume dissolved organic matter (DOM) through hydrolysis, transport, and intracellular metabolism, and these activities occur in distinct subcellular localizations. Therefore, investigation of protein and proteome subcellular localization is likely to improve our understandings about how bacteria interact with DOM.

Many computational algorithms have been developed to predict the subcellular localization of proteins. These algorithms employ a variety of supervised machine learning techniques and different information sources to make predictions. They can be generally classified into three types. One type of methods explores the presence/absence of signal peptides or specific protein domains, such as SignalP (Dyrlov Bendtsen et al. 2004) and Phobius (Käll et al. 2007). These methods require protein sequences to be complete. Metagenomic peptides, however, are often fragmentary, making these methods not applicable. The second type, such as Proteome Analyst, uses localization information from

well-annotated homologous sequences identified by BLAST. It is not suitable to make discoveries of a protein family with different subcellular localizations. The third type of methods builds machine learning models (e.g., support vector machine) and predicts protein localization using features, such as amino acid/dipeptide compositional bias, physicochemical properties of amino acids, and others. Since these sequence features are derived from whole protein sequences, most algorithms in this category are minimally affected by the incompleteness of peptide sequences. Examples are CELLO (Lu et al. 2004), SUBLOC (Hua and Sun 2001), and PSLDOC (Chang et al. 2008). Only the third approach is useful in the case of metagenomic peptides which are often fragmentary.

Because all algorithms have their own bias, the predictions from individual algorithms in the third category are frequently inconsistent. This is related to the fact that sorting signals targeting different subcellular locations usually share some similarities. For example, sorting signals targeting the periplasm and outer membrane both have N-terminal positively charged regions. In this case, prediction algorithms usually have some ambiguity for distinguishing these neighboring compartments. When an algorithm predicts a protein as a periplasmic protein with the highest confidence, it also implies that the protein has a probability of being located in its neighboring compartments, including the cytoplasm, inner membrane, outer membrane, and extracellular space, with higher probability assigned to the locations closest to the periplasm. Indeed, neighboring compartments are usually reported as suboptimal predictions by the component algorithms (CELLO, SUBLOC, and PSLDOC). The MetaP algorithm proposed recently considers neighborhood relations among subcellular localizations and also suboptimal predictions. It thus has the benefit of resolving conflicting predictions by the base algorithms and achieves higher precision and accuracy of prediction (Luo et al. 2009).

The predicted location of MetaP for a sequence s is the one that has the maximum sum of weighted voting for that subcellular localization. The prediction can be denoted formally

as $P_s = \operatorname{argmax}_i \sum_{j=1}^N P(i, j)$ where N is the total

number of base predictors and i is the index of a predicted subcellular compartment: cytoplasmic ($i = 1$), cytoplasmic membrane ($i = 2$), periplasmic ($i = 3$), outer membrane ($i = 4$), and extracellular ($i = 5$). $P(i, j)$ denotes the voting weight of the prediction of the j th element predictor for compartment i . It is defined as

$$P(i, j) = \sum_{k=0}^{M_j} 2^{-|C_k - i|} \cdot W_K,$$

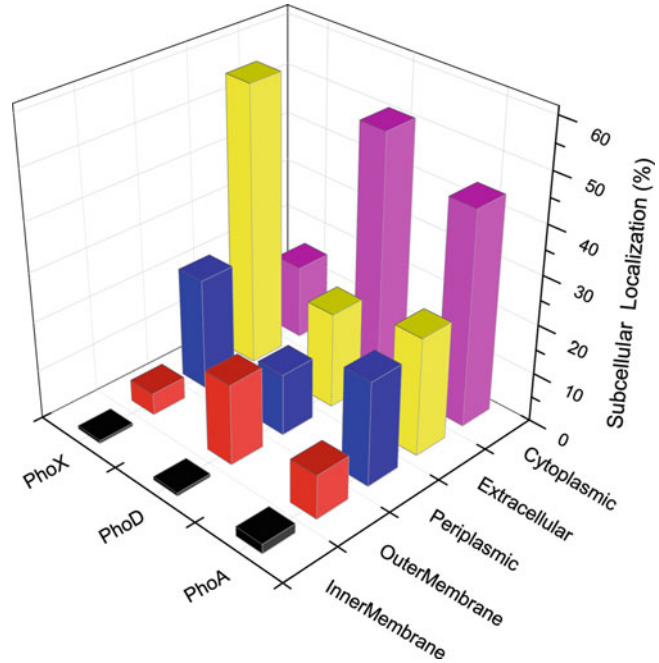
where M_j is the number of predictions of the j th predictor. It means that the voting weight of a prediction by the j th predictor for compartment i depends on the offset of the index C_K of its predicted class with regard to the index i as well as its normalized score W_K . The voting weight W_K for K th prediction is defined on the basis of its relative score by comparison with all other predictions made by this algorithm. Because raw scores of predictions from different component base algorithms are not directly comparable, the raw score S_K is converted into a normalized probability $p(K) = p(S \leq S_K)$ by calculating the percentage of predictions with lower raw scores among all predictions for a given algorithm. W_K is then defined as $W_K = p(K)$.

The performance of MetaP and the component algorithms was evaluated using sets of testing sequences whose localizations were verified by experiments (Menne et al. 2000). For the purpose of testing the accuracy of fragmentary protein prediction, the N-terminal of the testing sequences is removed. This benchmark test showed that MetaP makes more accurate predictions of fragmentary peptide sequences than any component method.

MetaP was applied to several protein families of alkaline phosphatases using the Global Ocean Sampling (GOS) metagenomic data sets (Luo et al. 2009). Alkaline phosphatases are major hydrolytic enzymes of organic phosphoesters which are the dominant forms of dissolved organic phosphorus in the ocean and providing an important source to meet bacterial phosphorus requirements. It was thought that marine bacterial alkaline phosphatases are exclusively

New Computational Methodologies to Understand Microbial Diversity,

Fig. 1 Subcellular localization distributions of APases recovered from the GOS metagenomic database (figure adapted from Luo et al. 2009)



ectoenzymes. However, MetaP predicted that about 40 % of the alkaline phosphatases are located in the cytoplasm (Fig. 1). Further bioinformatic analysis suggested that the cytoplasmic alkaline phosphatases might play a role in hydrolyzing the imported small organic phosphorus compounds. In addition, application of MetaP to a metatranscriptomics data set showed diel variations in the fraction of transcripts encoding inner membrane and periplasmic proteins compared to cytoplasmic proteins (Fig. 2), suggesting a close coupling of photosynthetic extracellular release and bacterial consumption (Luo 2012).

An Evolutionary Genetic Method to Classify Metagenomic Reads Taxonomically

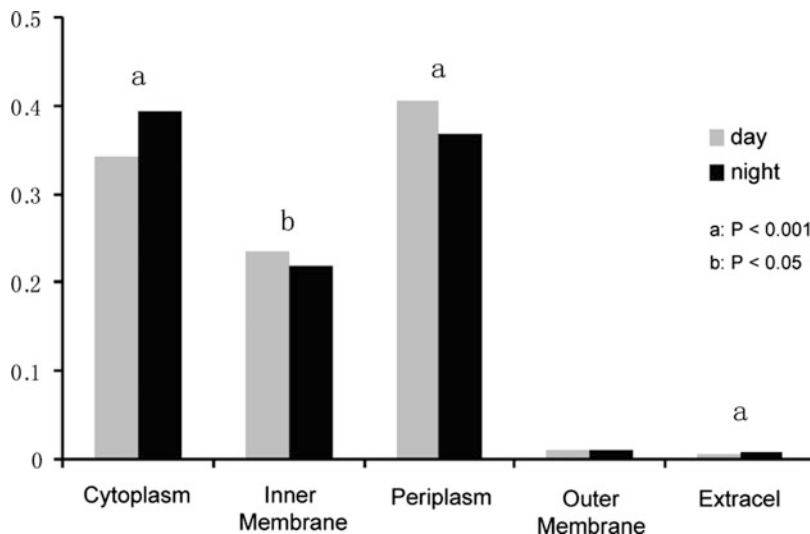
Metagenomic DNA represents genetic potential of the microbial community in an environment. Due to its unbiased nature, a majority of a metagenomic sample consists of DNA from those abundant microbial lineages. Therefore, it provides raw material for studying genome content of the abundant taxa in the nature. On the

other hand, metagenomic DNA is a mixture from all microbes in the sample, making it difficult to study genome content of a specific microbial lineage in a systematic way. It is therefore important to develop high-throughput computational approaches to systematically classify metagenomic genes taxonomically. This would lead to an improved understanding of the ecological functions of the abundant taxa in the nature.

Definitively assigning sequences from diverse metagenomic data sets to taxonomic groups is problematic, however. Most applications rely on BLAST-based (Altschul et al. 1997) identification of best hits to an annotated sequence database. While the BLAST best hit approach is easy to use, its accuracy is decidedly influenced by the composition of the annotated database. Thus, a substantial fraction of best BLAST hits may not be the closest relatives phylogenetically, an issue that is exacerbated when taxonomic groups are not evenly represented in the database (Koski and Golding 2001). A second type of methods employs machine learning principles to classify metagenomic reads based on the nucleotide sequence characteristics (McHardy et al. 2007). These methods are also subject to the high

New Computational Methodologies to Understand Microbial Diversity,

Fig. 2 Differential gene expression in protein subcellular localizations between day and night in surface waters of North Pacific Subtropical Gyre. The letter above the bar indicates the significance level: (a), $P < 0.001$; (b), $P < 0.05$ (figure adapted from Luo 2012)

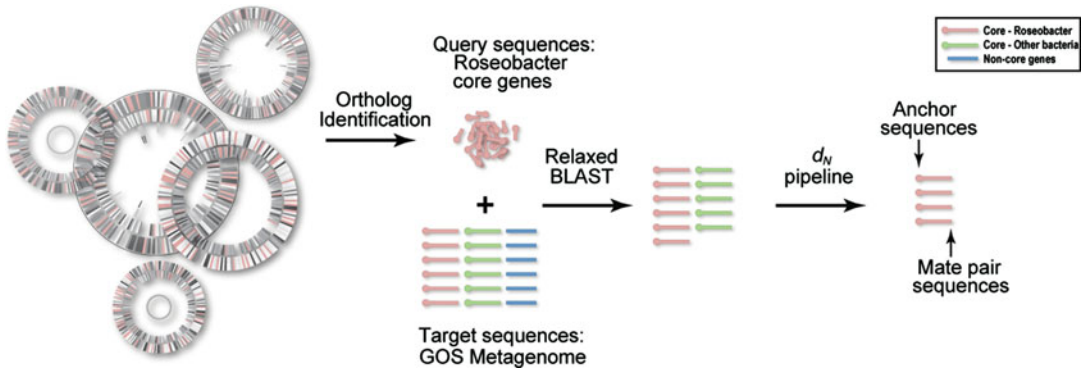


false-positive issue, which cannot meet the needs of many ecological studies.

A bioinformatic approach is recently developed to assign metagenomic gene fragments to taxonomic groups by computing evolutionary distances of protein-coding DNA sequences (Luo et al. 2012). In a protein-coding DNA sequence, point mutation occurs both in synonymous sites which do not change the corresponding amino acid sequence and in non-synonymous sites which change the encoded amino acids. Thus, the evolutionary distances of protein-coding DNA sequences can be represented using synonymous (d_s) and non-synonymous (d_n) substitution rate. More specifically, d_s is the number of synonymous substitutions per synonymous site, and d_n for the number of non-synonymous substitutions per non-synonymous site. Since synonymous mutations are largely invisible to natural selection, synonymous sites are easily saturated with substitutions. In contrast, most non-synonymous mutations are deleterious, and many of them have been eliminated by purifying selection. Thus, d_n is much smaller than d_s in a vast majority of genes (Luo and Hughes 2012). Often, marine microbial ecologists are interested in highly diverged lineages (e.g., *Roseobacter*, SAR11, *Vibrio*, *Prochlorococcus*). At this level of divergence, the synonymous sites are saturated with

substitutions among some members of the lineage (Luo and Hughes 2012). Therefore, d_n is used to measure the evolutionary distances of protein-coding genes. The d_n pipeline assigns a metagenomic gene to a microbial clade (e.g., the marine *Roseobacter* clade) based on the requirement that the mean evolutionary distance between a metagenomic gene and each of the reference orthologous genes from the clade members is smaller than the mean of all pairwise comparisons among the reference orthologous genes in that clade. Mathematically, the requirement can be expressed using $\frac{\sum_1^n d_{N,ref-meta}}{n} < \frac{\sum_1^{\binom{n}{2}} d_{N,ref-ref}}{\binom{n}{2}}$, in which n is the number of reference orthologous genes, $d_{N,ref-meta}$ is d_n between a reference gene and the metagenomic gene fragment, and $d_{N,ref-ref}$ is d_n between two reference genes.

The d_n pipeline takes in alignments, each consisting of reference orthologous genes belonging to the core genome of a monophyletic microbial clade and one metagenomic gene fragment with unknown taxonomic affiliation. Identification of putative gene fragments from metagenomic reads requires in silico translation of the reads in six reading frames and then selection of all fragments with a certain minimal length (e.g., 60 amino acids) between stop



New Computational Methodologies to Understand Microbial Diversity, Fig. 3 A flowchart of preprocessing steps for d_N pipeline for high-confidence phylogenetic classification of metagenomic DNA fragments. The circles on the leftmost are *Roseobacter* genomes, in which pink-colored parts represent core genomes. The gene fragments in the GOS metagenome are categorized into three parts, in which pink colored are

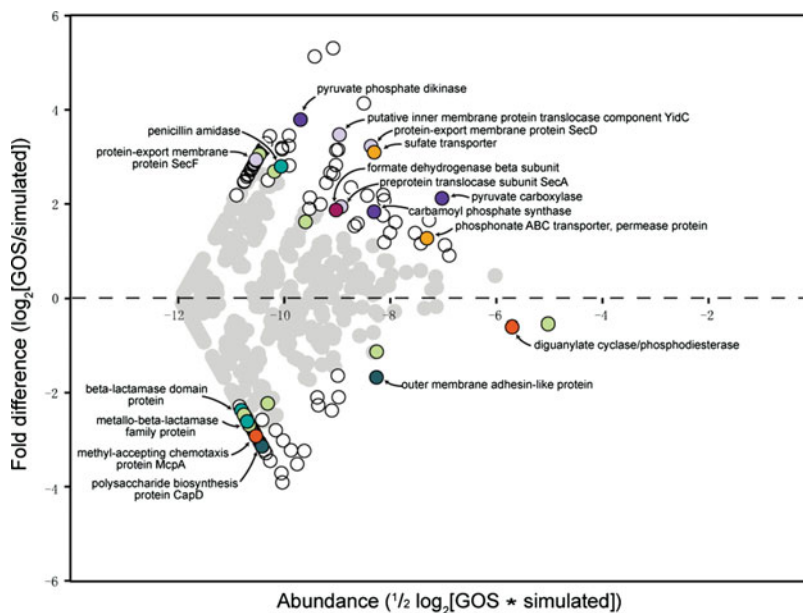
Roseobacter core genes, green colored are other bacterial genes homologous to the *Roseobacter* core genes, and blue colored are not homologous to the core genes which are not recovered by a BLAST similarity search. The d_N pipeline is designed to filter out other bacterial genes in green, but a few true *Roseobacter* sequences are missing because of the conservative nature of the d_N pipeline (figure adapted from Luo et al. 2012)

codons. Then, BLAST identifies a set of putative metagenomic gene fragments that are homologous to the reference genes (Fig. 3). Each of the homologous metagenomic gene fragments will be aligned to the reference genes at the amino acid level, and the DNA sequences are imposed on the alignment. Next, the PAML software (Yang 1997) computes d_N for each pairwise comparison in the DNA alignment.

The output of the d_N pipeline is a set of metagenomic gene fragments that are assigned to the microbial clade. Validation of the d_N pipeline using phylogenetic analyses showed that the false-positive rate is smaller than 1%. Since these classified metagenomic gene fragments are homologous to the core genomes of a microbial clade which encode for biological functions that are essential to basic cellular functionality, they are unlikely to provide valuable information about ecologically relevant processes. However, depending on the library design for sequencing, a read may be partnered with a pair end read, both of which are from the same DNA molecular, and the pair end read may carry an ecologically relevant gene. Thus, an important extension of the d_N pipeline is to examine the pair end of the assigned reads. Here, the metagenomic gene fragment that is directly identified by the d_N pipeline is named

“anchor sequence,” and its pair end read is named “mate pair sequence.” These assigned metagenomic genes are by no means a comprehensive list of genes affiliated with this microbial clade, since they can be only identified if they are core genes or physically linked to a core gene of that clade.

This whole procedure, including preprocessing, the d_N pipeline, and the mate read analysis, was applied to assign metagenomic genes in the Global Ocean Sampling (GOS) data sets (Rusch et al. 2007) to the marine *Roseobacter* clade. The major finding is that the uncultivated *Roseobacter* populations differ systematically in several genomic attributes from their cultured representatives, including fewer genes for signal transduction and cell surface modifications but more genes for Sec-like protein secretion systems, anaplerotic CO₂ incorporation, and phosphorus and sulfate uptake (Fig. 4). Several of these trends match well with characteristics previously identified as distinguishing r- versus K-selected ecological strategies in bacteria, suggesting that the r-strategist model assigned to cultured roseobacters may be less applicable to their free-living oceanic counterparts. Thus, genomic analyses of cultured roseobacters appear to be biasing our view of the lineage’s



New Computational Methodologies to Understand Microbial Diversity, Fig. 4 Differential representation of gene families in oceanic compared to cultured roseobacters (M versus A plot). Families plotting above the line are enriched, and those plotting below the line are depleted in the oceanic roseobacters. Non-gray symbols represent gene families with significant differential representation between the two metagenomes. Colors indicate gene families with similar functions: *dark purple*,

anaplerotic CO₂ incorporation; *light purple*, Sec secretion system; *dark orange*, signaling; *light orange*, nutrient transport; teal, antibiotic synthesis or resistance; *maroon*, C1 metabolism; *dark green*, cell surface properties; and *light green*, hypothetical. This plot shows differential representation for just one of three simulated metagenomic data sets that were constructed, all of which had congruent results (figure adapted from Luo et al. 2012)

ecology toward a stronger r-selected ecological model than is merited (Luo et al. 2012).

A Statistical Modeling Approach for Comparative Metagenomics

An important goal of metagenomics is to explain the genetic potential of the microbial community in the context of the environmental gradients. One way of approaching this goal is to reveal correlations between environmental gradient and gene abundance. Although standard statistical tests such as regression analysis have been successful in correlating gene abundance and geochemical parameters in large-scale sampling efforts, exploring smaller data sets requires designing sophisticated statistical modeling approaches. The following example illustrates it (Luo et al. 2011).

Phosphonate contains a stable carbon-phosphorus (C-P) bond, comprising 25 % of the high-molecular-weight dissolved organic phosphorus in the ocean. Phosphonates are degraded by two types of enzymes, C-P lyases and hydrolases. The C-P lyase is a multienzyme complex, and the corresponding genes are only expressed when inorganic phosphate becomes limited, suggesting that the activity of C-P lyase genes is regulated by phosphate concentrations. In the ocean, there is a vertical gradient of phosphate level, in which phosphate is depleted in the upper euphotic zone (<100 m), reaches its maximum at the base of mesopelagic zone (1,000 m), and has a minor decrease in the bathypelagic zone (>1,000 m). Only the phosphate level in the upper euphotic zone can be a limiting factor to biological productivity. Thus, the depth profile of ocean water column provides a natural platform to test microbial adaptation to phosphate gradient

by correlating the vertical gradient of phosphate and C-P lyase gene abundance in different depths.

Examination of a recently available metagenomic data set containing thousands of sequences at each of seven depths (10 m, 70 m, 130 m, 200 m, 500 m, 770 m, 4,000 m) in the North Pacific Subtropical Gyre showed that the lytic executor genes (phnG, phnH, phnI, phnJ, phnM) of the C-P lyase complex are exclusively found in the upper euphotic zone. To validate the pattern of C-P lyase executor genes being present in the surface ocean metagenomes but absent in deeper samples, a statistical approach was designed. Testing the significance of the existence or absence of executor genes in the two depth regions is equivalent to testing the following two hypotheses: (1) executor genes exist in surface waters (≤ 70 m), and (2) executor genes are absent at depths ≥ 130 m. The basic method applied was the one-sample test on proportions. Specifically, a 95 % confidence interval was set up to indicate the range of possible true proportions (or population proportions) of executor genes, which was defined as $p \pm (1.96 \times \sqrt{p(1-p)/N})$. Here, N is the number of observed genes in total. The symbol p denotes the sample proportion. In this context, sample proportions are the proportions of the executor genes among all the genes collected and mathematically defined as the ratio between the number of observed executor genes and the number of observed genes in total (i.e., sample size N) at each depth category (either ≤ 70 m or ≥ 130 m) in the water column. The confidence interval was set up based on the normality assumption of proportions when the sample size (N) is large, which is certainly a valid assumption for this data set. If zero is not included in the interval, then the existence of the executor genes can be confirmed with 95 % of confidence. In the case of the upper euphotic zone, the 95 % of confidence does not include zero, confirming that executor genes exist in surface waters (≤ 70 m).

The above method would not be applicable if sample proportions of the executor genes are

zeros because any 95 % confidence intervals would include zero. This indicates that it is unlikely to see any executor genes at the designated water depths where these genes are not found. However, the absence of the executor genes could indicate that these genes are rare and the sample size is not large enough or the number of observed executor genes might have been miscounted, i.e., misclassification might have occurred.

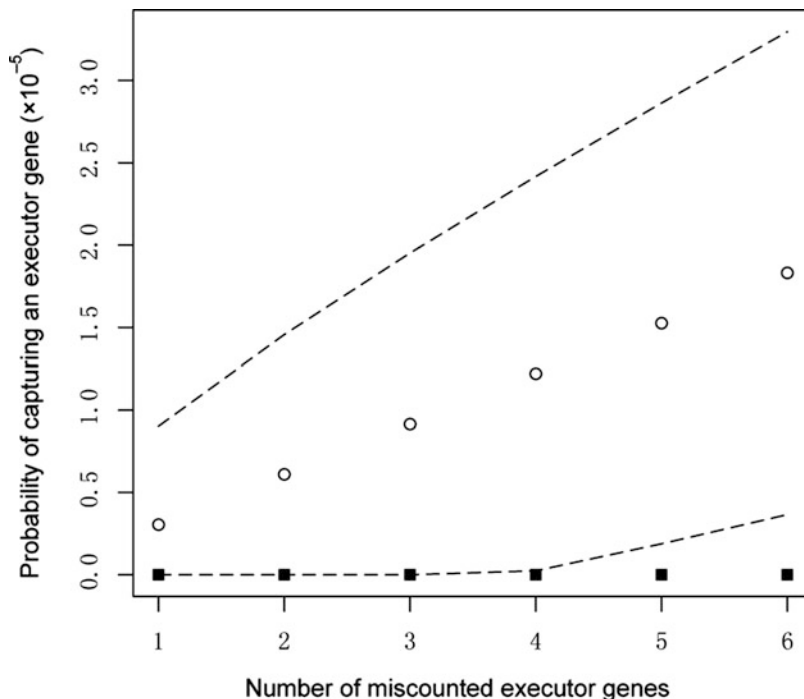
To take these possibilities into account, the following two directions are considered for further investigation: (1) increasing the sample size so that it is large enough to collect one executor gene or (2) increasing the number of observed executor genes to correct possible misclassifications. Biologically, these two directions are independent, but mathematically they are linked. The following mathematical principle proves that if direction 2 cannot lead to the confirmation of the true existence of executor genes, then the existence cannot be confirmed through direction 1 either.

N_{130} is used to denote the current sample size at depth 130 m and deeper. A larger sample size is denoted as $N_{130} + X$ with $X > 0$. If the executor genes are miscounted by 1, a 95 % confidence interval can be set up with the lower bound given as $1/N_{130} - 1.96 \times \sqrt{1/N_{130} \times (1 - 1/N_{130})/N_{130}}$. If this lower bound of confidence interval includes zero, then it must be negative, i.e., $1/N_{130} < 1.96 \times \sqrt{1/N_{130} \times (1 - 1/N_{130})/N_{130}}$. This gives $(N_{130} - 1)/N_{130} > 1/1.96^2$. Now the sample size increases to $N_{130} + X$. Since $(N_{130} + X - 1)/(N_{130} + X) > (N_{130} - 1)/N_{130}$ and $(N_{130} - 1)/N_{130} > 1/1.96^2$, we then have $(N_{130} + X - 1)/(N_{130} + X) > 1/1.96^2$.

To test direction 2, the numbers of miscounted genes were specified as $N_{mis} = 1, 2, 3, 4, 5, 6$. Again, the sample proportions and the 95 % confidence intervals are calculated for each case. It can be seen that zero was always included in the 95 % confidence intervals unless four or more executor genes are misclassified (Fig. 5), which is unlikely to happen in practice due to the characteristic of the rarity of these executor genes.

New Computational Methodologies to Understand Microbial Diversity,

Fig. 5 Assumed sample proportions (denoted by circles) and 95 % confidence intervals (dashed lines). Y-axis: to get the probability, multiply the y-values by 10^{-5} ; X-axis: the number of assumed miscounted executor genes among $N_{130} = 327, 741$ genes. The filled squares are the locations of zeros (figure adapted from Luo et al. 2011)



Therefore, executor genes are indeed absent at depths ≥ 130 m.

At this point, the two earlier proposed hypotheses have been tested. The proposed statistical modeling approach can be widely applied to resolve biological questions regarding correlations between environmental gradient and gene abundance when the data sets are not large enough to do linear regression analysis.

Summary

In the past decade, large-scale metagenomic data sets have been released to the public community, and this trend is likely to be continued. Many biological questions may be answered by analyzing these data sets with appropriate computational approaches. Some of the promising methods are illustrated above, which are based on machine learning techniques, molecular evolutionary principles, and statistical modeling approaches. These studies are examples of future research directions of computational metagenomics.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402.
- Chang J-M, Su EC-Y, Lo A, Chiu H-S, Sung T-Y, Hsu W-L. PSLDoc: protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins.* 2008;72:693–710.
- Dyrlov Bendtsen J, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: signalP 3.0. *J Mol Biol.* 2004;340:783–95.
- Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics.* 2001;17:721–8.
- Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic Acids Res.* 2007;35:W429–32.
- Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol.* 2001;52:540–2.
- Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, et al. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics.* 2004;20:547–56.
- Luo H. Predicted protein subcellular localization in dominant surface ocean bacterioplankton. *Appl Environ Microbiol.* 2012;78:6550–7.

- Luo H, Hughes AL. dN/dS does not show positive selection drives separation of polar-tropical SAR11 populations. *Mol Syst Biol.* 2012;8.
- Luo H, Benner R, Long RA, Hu J. Subcellular localization of marine bacterial alkaline phosphatases. *Proc Natl Acad Sci USA.* 2009;106:21219–23.
- Luo H, Zhang H, Long RA, Benner R. Depth distributions of alkaline phosphatase and phosphonate utilization genes in the North Pacific Subtropical Gyre. *Aquat Microb Ecol.* 2011;62:61–9.
- Luo H, Löytynoja A, Moran MA. Genome content of uncultivated marine Roseobacters in the surface ocean. *Environ Microbiol.* 2012;14:41–51.
- McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4:63–72.
- Menne KML, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics.* 2000;16:741–2.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern tropical pacific. *PLoS Biol.* 2007;5:e77.
- Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics.* 1997;13:555–6.

New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE

Hidetoshi Takami

Microbial Genome Research Group, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

Synonyms

Functional potential evaluator

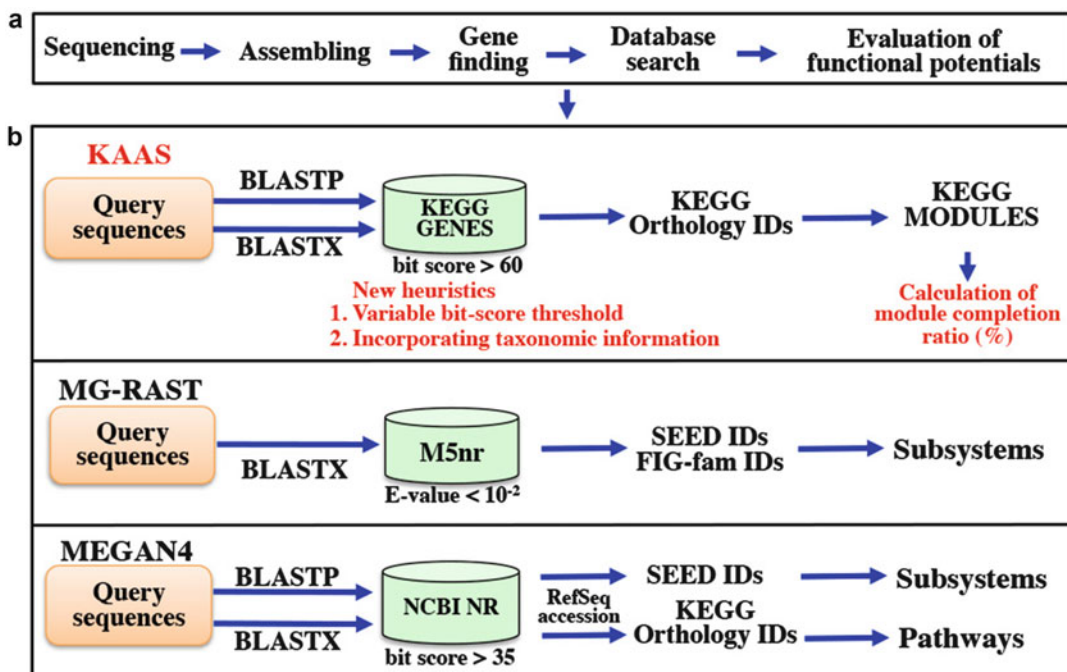
Definition

Although one of the main goals of genomic analysis is to elucidate the comprehensive functions (functionome) in individual organisms or a whole community in various environments, a standard evaluation method for discerning the functional

potentials harbored within the genome or metagenome has not yet been established. Thus, a new evaluation method for the potential functionome, based on the completion ratio of Kyoto Encyclopedia of Genes and Genomes (KEGG) functional modules, was developed. Basic methodology and application of this method for comparative functional genomics and metagenomics are expounded in this entry.

Introduction

One of the main goals of genomic and metagenomic analyses is to extract the comprehensive functions (functionome) harbored in an individual organism or a whole community in various environments. However, evaluating the potential functionome is still difficult when compared with the functional annotation of individual genes or proteins, i.e., based on a similarity search against a reference database such as the NCBI-NR database of non-redundant protein sequences, usually employing a variant of the BLAST program, or on the protein domain search against a protein family database such as PFAM. This is mainly because a standard methodology for extracting functional category information, such as individual metabolism, energy generation, and transportation systems, has not yet been fully established. Traditionally, clusters of orthologous groups (COGs) have been used for functional classification of proteins, particularly in microbial genome sequencing projects. The COG database provides 17 functional categories for orthologous groups in order to facilitate functional studies and serves as a platform for functional annotation of newly sequenced genomes and studies on genome evolution. Although the COG functional categories are often used within Standards in Genomic Sciences (<http://standardsingenomics.org/index.php/sigen>) as a standard analysis, through combination with the Integrated Microbial Genomes (IMG) system (Markowitz et al. 2012), no large functional differences are usually observed in such broad categories, even between phenotypically different organisms and also whole microbial



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 1 Outline of the methodology. (a) Workflow from sequencing to evaluation of the potential functions. (b) Detailed workflow of the three annotation servers, KAAS, MG-RAST, and MEGAN4, using query sequences after gene finding process of sequenced data; KAAS and MEGAN4 use BLASTP and BLASTX for amino acid and nucleotide query sequences, respectively, and MG-RAST uses only BLASTX. All use different

databases, i.e., KEGG GENES for KAAS, M5nr (Willke et al. 2012) for MG-RAST (M5nr includes the SEED as a subset), and NCBI-NR for MEGAN4, and different default threshold values for the BLAST hits. Each server converts the hit entries to the corresponding orthology IDs for functional annotation and pathway/module/subsystem mapping. *Red-colored texts of KAAS indicate its improvements in the current study.* This figure has been modified from the previous one (Takami et al. 2012)

communities in different environments. Thus, it is difficult to differentiate the functional potentials between different genomes and metagenomes by analysis based on COG classification.

Recently, more detailed and comprehensive functional categories facilitated in KEGG (Kanehisa and Goto 2000) and SEED (Overbeek et al. 2005) have been used for comparative genomics and as metagenomics tools to highlight functional features represented by KAAS (KEGG Automatic Annotation Server) (Moriya et al. 2007), MG-RAST (Meyer et al. 2008), and MEGAN (Huson et al. 2011) (Fig. 1). They all employ a similarity-based method for functional annotations, but utilize different databases for protein sequences, default threshold values, and

orthology IDs for mapping annotated sequences to functional categories depending on their desired outputs, namely, pathways in KEGG or subsystems in SEED. Notably, KAAS has been applied to protein-coding sequences from several metagenomic samples, and their annotated KEGG pathways and other classifications are already available. The outputs of these systems include functional distributions of each sample by hierarchical classification using KEGG and/or SEED and comparisons between several samples when necessary. However, it is still difficult to evaluate the functional potentials via the current classification systems (such as pathway map-based analysis) because the functional information from different organisms such as microbes, plants, and animals has been mixed up.

On the other hand, KEGG MODULE, a newly defined database that collects pathway modules and other functional units, presents a promising tool for functional classification (Kanehisa et al. 2008). Because the KEGG modules cover major metabolisms and physiological processes necessary for functional characterization of each categorized organisms such as plants, animals, and microbes, a new evaluation method using the KEGG MODULE database was developed to resolve the difficulties for evaluation of potential functionome and it was employed for comparative functional genomics and metagenomics (Takami et al. 2012). Based on this result, we also developed metabolic and physiological potential evaluator (MAPLE) system. The MAPLE provides a user-friendly Web interface not only for characterization of potential functionome harbored in the genomic and metagenomic sequences but also for comparative analyses for the module completion ratio (MCR) and mapping patterns to the KEGG modules (<http://www.genome.jp/tools/maple/>).

Development of New Evaluation Method for Potential Functionome

Kegg Module

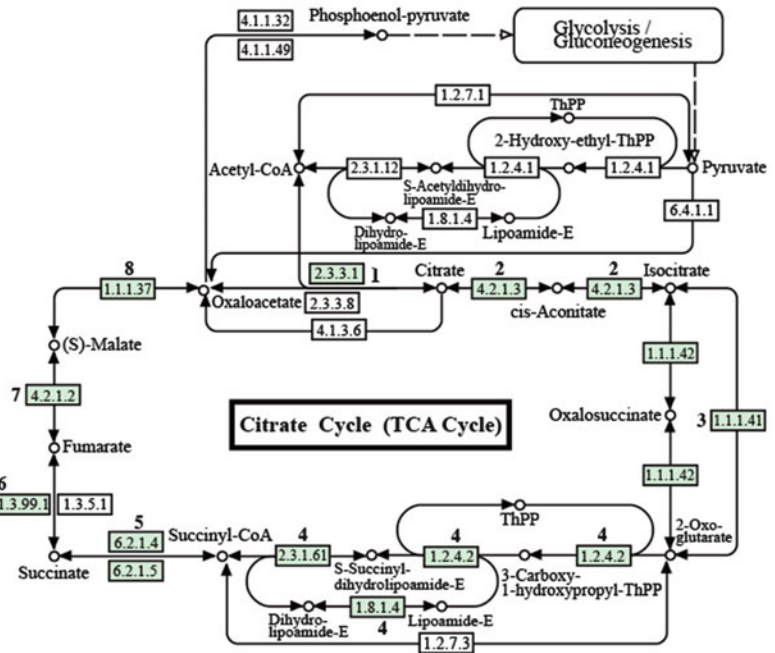
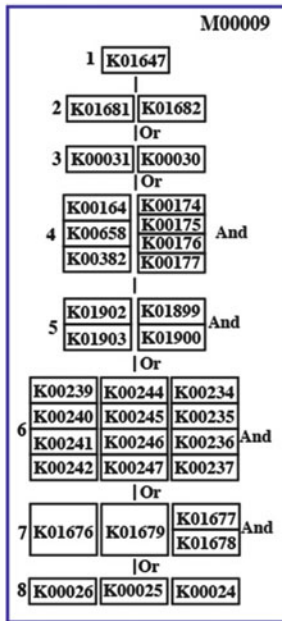
KEGG MODULE (Kanehisa et al. 2008) is a collection of pathway modules and other functional units designed for automatic functional annotation or pathway enrichment analysis. Pathway modules such as the TCA cycle core module (Fig. 2a) are tighter functional units than KEGG pathway maps and are defined as consecutive reaction steps, operon or other regulatory units, and phylogenetic units obtained by genome comparisons. Other functional units include (1) structural complexes representing sets of protein subunits for molecular machineries such as photosystems (Fig. 2b), (2) functional sets representing other types of essential sets such as aminoacyl-tRNA synthetases, and (3) signature modules representing markers of phenotypes such as enterohemorrhagic *E. coli* pathogenicity signature for Shiga toxin. The KEGG MODULE falls into 56 small functional categories (Table 1),

and the latest version is available from the KEGG FTP site (<http://www.kegg.jp/kegg/download>). Each module is defined by the combination of KO identifiers so that it can be used for annotation and interpretation purposes in individual genomes or metagenomes. Notations of the Boolean algebra-like equation for this definition include space-delimited items for pathway elements, comma-separated items in parentheses for alternatives, a plus sign to define a complex, and a minus sign for an optional item. Some modules have branching points in their reaction cascades, leading to different products or alternative reaction pathways. These modules are divided into several parts depending on the branching patterns and are redefined as submodules for accurate calculation of the completion ratio. The module completion ratio was calculated for each submodule to examine fine-grained functional categories (Takami et al. 2012).

Calculation of the Module Completion Ratio Based on a Boolean Algebra-Like Equation

The completion ratio of all KEGG functional modules in each organism was calculated based on a Boolean algebra-like equation. For this analysis, one genome was selected from each of the 1,041 available prokaryotic species as of March 2013. As one of the examples, M00009_1 is a core pathway module for the TCA cycle comprising eight components (Fig. 2a). In each KO number set, vertically connected KO identifiers indicate a complex and therefore represent “And” or “+” in the Boolean algebra-like equation, whereas horizontally located K numbers indicate alternatives and represent “Or” or “,” in the equation. When genes are assigned to all KO identifiers in each reaction according to the Boolean algebra-like equation, the module completion ratio (MCR) becomes 100%. If genes are not assigned to KO identifiers in two components, the MCR is calculated as 75% ($6/8 \times 100 = 75$). On the other hand, M00163_1 comprising six components in cyanobacteria represents a complex module for photosystem I. If genes assigned to KO identifiers in two of those components are missing, the MCR is calculated as 66.7% (Fig. 2b).

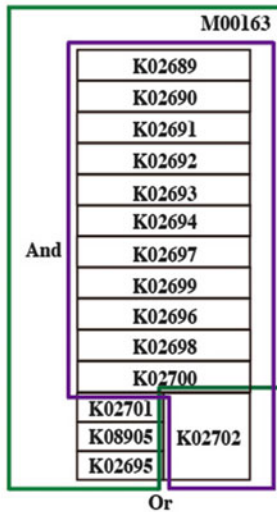
a



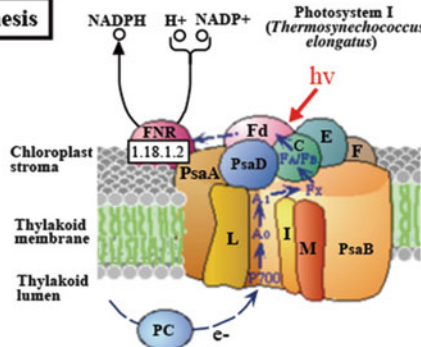
Boolean algebra-like equation:

$$K01647 (K01681, K01682) (K00031, K00030) (K00164+K00658+K00382) (K01902+K01903, K01899+K01900) (K00239+K00240+K00241+K00242, K00244+K00245+K00246+K00247) (K01676, K01679, K01677+K01678) (K00026, K00025, K00024)$$

b



Photosynthesis



Cyanobacteria

PsaA	PsaB	PsaC	PsaD	PsaE	PsaF	PsaG	PsaH
PsaI	PsaJ	PsaK	PsaL	PsaM	PsaN	PsaO	PsaX

Photosystem I

PsaA	PsaB	PsaC	PsaD	PsaE	PsaF	PsaG	PsaH
PsaI	PsaJ	PsaK	PsaL	PsaM	PsaN	PsaO	PsaX

Boolean algebra-like equation:

$$(K02689+K02690+K02691+K02692+K02693+K02694)+(K02697+K02699+K02696+K02698+K02700)+(K02701+K08905+K02695), K02702)$$

New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 2 KEGG functional modules. (a) A pathway module. The module M00009 comprising eight components is defined for the citrate cycle (TCA cycle) core module and represented as a Boolean algebra-like equation of KO

identifiers or K numbers for computational applications. The relationship between this module and the corresponding KEGG pathway map is also shown by indicating corresponding K number sets in the module and EC numbers in the pathway map using the same index. In each K number set, vertically connected

Assignment of the Query Sequences to KO Identifiers

Because KAAS is an efficient tool for assigning KO identifiers to genes from complete genomes based on a BLAST search of the KEGG GENES database combined with a bidirectional best-hit method (Moriya et al. 2007), the KAAS system is used to assign KO identifiers to protein sequences from metagenome projects and to users' own data from other genome and metagenome projects. Recently the KAAS system has just been slightly modified to improve the accuracy of KO assignments by (i) using a variable bit-score threshold instead of a fixed one (60 in the original KAAS system) to avoid missed annotations when there are sufficient high-scoring hits for KO assignment and (ii) considering taxonomic information of each KO when more than one candidate KO is obtained (Fig. 1) (Takami et al. 2012). This modification resulted in improved positive predictive value (#true positives/#all positives) by 2–5 % in the KO reassignment tests for 30 selected species. The latest stand-alone KAAS system for Linux and Mac OS X is available from the Web site of KAAS HELP (<http://www.genome.jp/tools/kaas/help.html>). This new KAAS was used for estimation of database dependency on the accuracy of the KO assignment (Fig. 3). *Escherichia coli* was selected as a representative of prokaryotic species and constructed four different types of data sets: without *E. coli* and closely related species (1,239 species), without all species within family *Enterobacteriales* (1,200 species), without all species within class *Gammaproteobacteria* (1,040 species), and without all species within phylum *Proteobacteria* (755 species). The draft genome of *E. coli* from infants in Trondheim, Norway, (accession, ERX127960) was used for this analysis because the assembled genome from the short-read sequences produced by a 454 GS

FLX Titanium sequencer contains several sequencing errors. The amino acid sequences of complete CDSs identified from the draft genome were randomly fragmented to 50, 60, 80, 100, 120, 150, and 200 residues in length, and each fragment was subjected to verification of database dependency based on the accuracy of KO identifier assignment (Fig. 3). In general, because most microbes thriving in natural environments are uncultivable, many genes in environmental metagenomes do not show significant similarity to those from known species in the public genome database. Especially when microbial genomes belonging to the same phylum as the query microbe are missing in the genome database, the accuracy rate of KO assignment to proteins phylogenetically distant from known phyla is expected to be low. In fact, when all species within phylum *Proteobacteria* were not included in the data set, the accuracy rate of KO assignment to full proteins of *E. coli* decreased to 80 %, but the accuracy rate of approximately 70 % was maintained even in the proteins fragmented to about 100 residues (Fig. 3). Considering these results, even if the genes from unidentified phyla of the so-called candidate division are included in the metagenomes, the KAAS system can presumably assign KO identifiers to genes longer than 300 bp (100 amino acids) with an accuracy rate of approximately 70 %.

Distribution Patterns of the Module Completion Ratio in 1,256 Prokaryotic Species

KEGG modules are modular functional units derived from the KEGG pathways and are categorized into pathway modules, structural complexes, functional sets, and genotypic

New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 2

(continued) K numbers indicate a complex and therefore represent “And” or “+” in the Boolean algebra-like equation, whereas horizontally located K numbers indicate alternatives and represent “Or” or “,” in the equation. (b) A structural complex module. The structural complex

module M00163 comprising six components is defined for the type I photosystem. The Boolean algebra-like equation and the corresponding KEGG pathway map are also shown. This figure has been redrawn with the updated KEGG module database from the previous one (Takami et al. 2012)

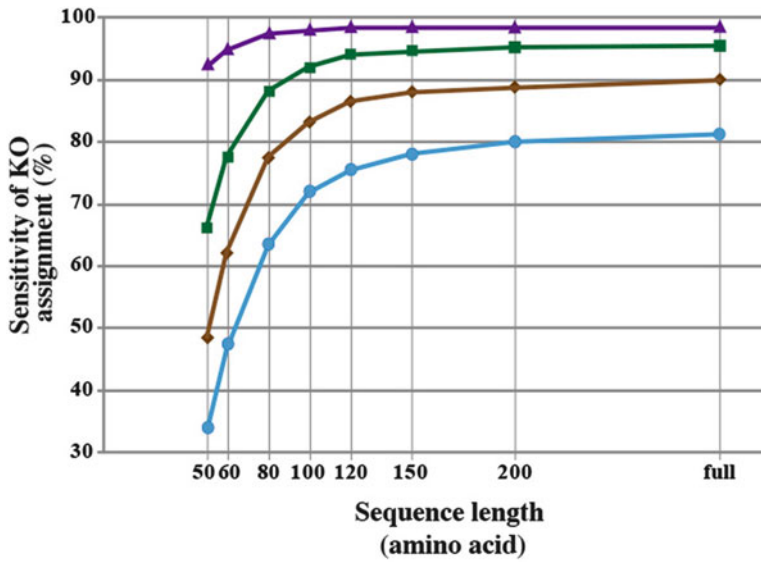
New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE

Table 1 Breakdown of small functional categories of the KEGG modules

Pathway modules	Structural complex modules
Cofactor and vitamin biosynthesis	Saccharide and polyol transport system
Central carbohydrate metabolism	Phosphotransferase system (PTS)
Aromatics degradation	ATP synthesis
Lipid metabolism	Phosphate and amino acid transport system
Aromatic amino acid metabolism	Mineral and organic ion transport system
Carbon fixation	ABC-2 type and other transport systems
Methane metabolism	Bacterial secretion system
Glycan metabolism	Metallic cation, iron-siderophore, and vitamin B12 transport system
Sterol biosynthesis	RNA processing
Fatty acid metabolism	Ubiquitin system
Lysine metabolism	Spliceosome
Other carbohydrate metabolism	Protein processing
Glycosaminoglycan metabolism	Repair system
Terpenoid backbone biosynthesis	DNA polymerase
Cysteine and methionine metabolism	Peptide and nickel transport system
Nitrogen metabolism	Replication system
Branched-chain amino acid metabolism	RNA polymerase
Lipopolysaccharide metabolism	Proteasome
Purine metabolism	Photosynthesis
Pyrimidine metabolism	Carbohydrate metabolism
Polyamine biosynthesis	Ribosome
Alkaloid and other secondary metabolite biosynthesis	Glycan metabolism
Sugar metabolism	
Other terpenoid biosynthesis	Functional set modules
Serine and threonine metabolism	Two-component regulatory system
Arginine and proline metabolism	Aminoacyl-tRNA
Phenylpropanoid and flavonoid biosynthesis	Nucleotide sugar
Sulfur metabolism	
Histidine metabolism	Signature modules
Other amino acid metabolism	Pathogenicity

signatures. Each KEGG module is designed for automatic functional annotation by a Boolean algebra-like equation of KEGG Orthology IDs. However, it remains uncataloged as to which species possess common modules or if certain modules demonstrate universality or rareness between specific species, phyla, etc. Specific information regarding the phylogenetic profiles of each module holder would be especially useful for annotating metagenomes. Thus, the distribution patterns of the completion ratios of the KEGG modules were examined in the 1,256 prokaryotic species whose genomic sequences have been completed. Although distribution of the module completion ratios in the 1,256 species varied greatly depending on the kind of module, it could be categorized into four patterns (universal, restricted, diversified, and nonprokaryotic) regardless of the module type (pathway, structural complex, signature, or functional set), when considering 70 % of all species to represent a majority measurement for the patterns (Table 2 and Fig. 4).

Pattern A defined as “universal” comprised modules completed by more than 70 % of the 1,256 species (Fig. 4a). Of 226 pathway modules containing submodules, modules grouped into pattern A account for only 7.5 % (Table 2) and mainly belong to the categories of central carbohydrate metabolism and cofactor and vitamin biosynthesis. Pattern B defined as “restricted” comprised modules completed by less than 30 % of the species (Fig. 4b) and accounted for 17.3 % of all the pathway modules, and 37 modules were rare modules completed by less than 10 % of the 1,256 species (Table 2). Pattern C defined as “diversified” accounted for 40.3 % of all the pathway modules and comprised modules ranging widely in completion ratios. M00012_1 (the glyoxylate cycle comprising five components) is one of the representatives of pattern C (Fig. 4c). One or several KO identifiers were assigned to each reaction in this module; however, KO identifiers, except for K01637 and K01638 assigned to the third and fourth components, were also assigned to other pathway modules such as the TCA (Krebs) cycle (M00009_1), first carbon oxidation (M00010_1), reductive



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 3 Effect of database dependency on accuracy of the KO assignment. Purple triangles show the results using the data set without proteins from the genera *Escherichia*, *Salmonella*, *Shigella*, and *Yersinia* (1,239 species). Similarly, green squares, brown diamonds, and blue dots show the results without proteins from the order *Enterobacteriales* (1,200 species), class *Gammaproteobacteria* (1,040 species), and phylum *Proteobacteria* (755 species), respectively. KO identifiers specific to the

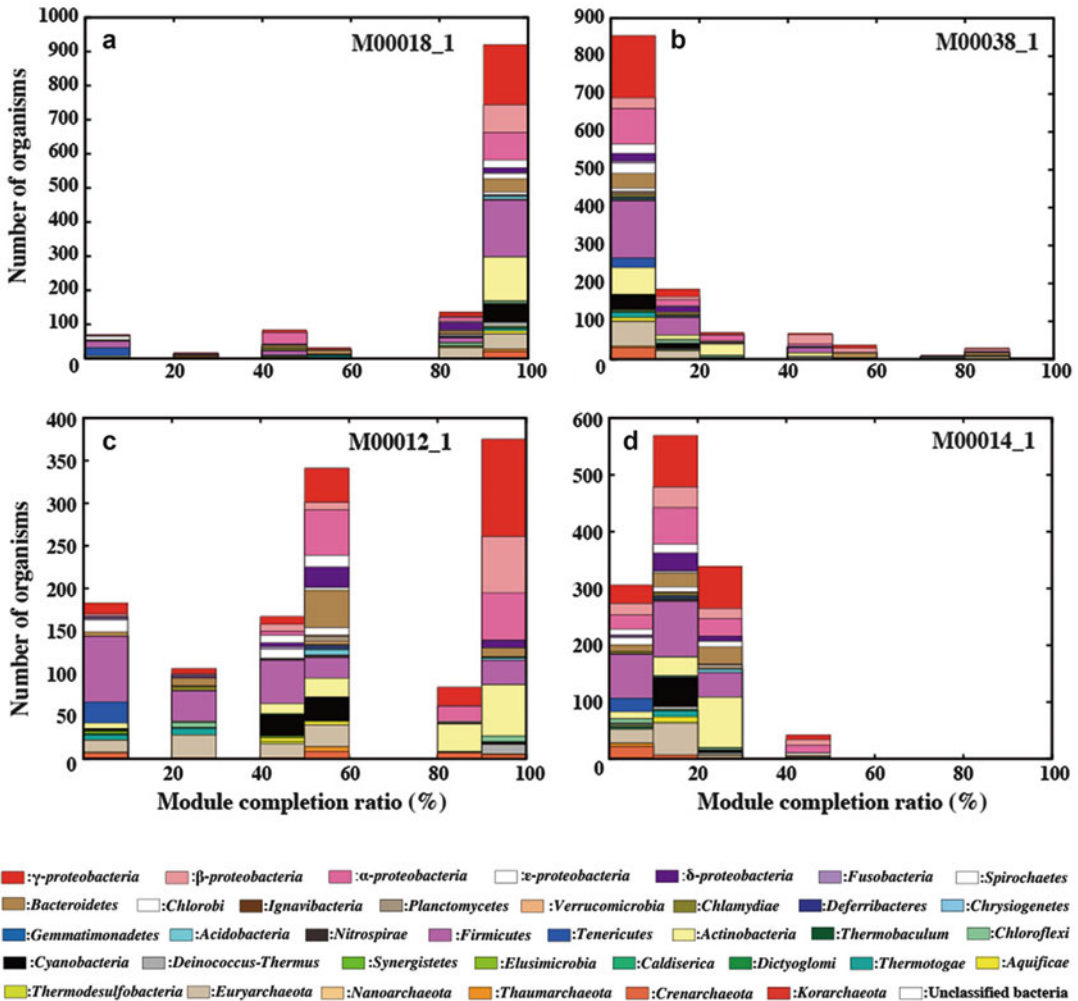
genera *Escherichia*, *Salmonella*, *Shigella*, and *Yersinia* (16 KO identifiers), order *Enterobacteriales* (90), class *Gammaproteobacteria* (203), or phylum *Proteobacteria* (370) were removed in advance from the protein data set. Here, the accuracy is defined by the sensitivity $TP/(TP + FN)$, where TP and FN are the numbers of true positives and false negatives, respectively. The truncated proteins were also used to confirm the effect of amino acid (a.a.) sequence lengths on the accuracy of KO assignments as described in the text. This figure has been slightly modified from the previous one (Takami et al. 2012)



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Table 2 Classification of the KEGG modules based on the module completion ratio of 1,256 prokaryotes

Completion pattern	Definition of module type	Pathways [226]		Structural complexes [331]		Functional sets [86]		Signatures [9]	
		No. of modules (%)		No. of modules (%)		No. of modules (%)		No. of modules (%)	
		Total	rare	Total	rare	Total	rare	Total	rare
A	Universal	17 (7.5)	0 (0)	9 (2.7)	0 (0)	1 (1.2)	0 (0)	0 (0)	0 (0)
B	Restricted	39 (17.3)	37 (47.4)	133 (40.2)	99 (81.1)	77 (89.5)	67 (97.1)	8 (88.9)	8 (88.9)
C	Diversified	91 (40.3)	41 (52.6)	70 (21.1)	23 (18.9)	5 (5.8)	2 (2.9)	1 (11.1)	1 (11.1)
D	Nonprokaryotic	79 (35.0)	0 (0)	119 (36.0)	0 (0)	3 (3.5)	0 (0)	0 (0)	0 (0)

[] shows total number of the KEGG modules containing branched modules. “Rare” indicates the modules completed by less than 10 % of 1,256 prokaryotic species. Universal, the modules completed by more than 70 % of 1,256 prokaryotic species. Restricted, the modules completed by less than 30 % of 1,256 prokaryotic species. Diversified, the modules that varies in the module completion ratio among 1,256 prokaryotic species. Nonprokaryotic, the modules not to be completed by any prokaryotic species



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 4 Typical completion patterns to the KEGG modules by 1,256 prokaryotic species. (a) Universal modules. The modules completed by more than 70 % of 768 prokaryotic species. M00018_1, which is threonine biosynthesis (aspartate-homoserine-threonine), is one of the examples of the pattern A-1. (b) Restricted modules completed by less than 30 % of 768 prokaryotic species. M00038_1, which is tryptophan metabolism, is one of the examples of the pattern B. C: Diversified modules. These

are the modules that vary in the module completion ratio among 1,256 prokaryotic species. M00012_1, which is glyoxylate cycle, is one of the examples of the pattern C. D: Nonprokaryotic modules completed by no prokaryotic species. M00014_1, which is glucuronate pathway, is one of the examples of the pattern D. Breakdown of taxonomic variations that complete each KEGG module is summarized in Table 3. This figure has been redrawn with the updated KEGG module and genome databases from the previous one (Takami et al. 2012)

TCA cycle (M00173_1), and C4-dicarboxylate cycle (nicotinamide adenine dinucleotide (NAD)⁺-malic enzyme type) (M00171_1). Some KO IDs assigned to many of the modules, categorized into pattern C, were also assigned to several other independent modules. Thus, when

the module completion ratio is low, the relationship between the module completion ratio of the targeted module and others to which the same KO identifiers are assigned should be considered. Pattern D, which accounted for 35.0 % of all pathway modules, comprised nonprokaryotic

modules that are not completed by prokaryotic species (Fig. 4d).

Of the 331 structural complex modules containing submodules redefined from modules with various complex patterns, 133 modules were categorized into pattern B (47.4 %) and 99 were rare modules (Table 1). Pattern C accounted for only 21.1 % in the structural modules compared with 40.3 % in the pathway modules. Thus, it was hypothesized that most of the structural complex modules, except for pattern D, are shared only in limited prokaryotic species.

Nonprokaryotic modules account for 35 % of pathway and 36 % of structural complex modules, respectively, and other modules were classified into various taxonomic patterns such as prokaryotic, Bacteria specific, and Archaea specific based on the MCR profiles (Table 3). These four patterns indicate the universal and unique nature of each module and also the versatility of the KO identifiers mapped to each module. Thus, the four criteria and taxonomic classification for each module should be helpful for the interpretation of results based on module completion profile.

Application of the Evaluation Method for Potential Functionome to Genomic and Metagenomic Analyses

Comparative Functionome Analysis of Bacilli Based on the KEGG Modules

Bacillus and its related species in genera such as *Oceanobacillus* and *Geobacillus* reclassified from genus *Bacillus* (*Bacillus*-related species) are known to thrive in a wide range of environmental conditions: pH 2–12, temperatures between 5 and 78 °C, salinity from 0 % to 30 % NaCl, and pressures from 0.1 Mpa (atmospheric pressure) to at least 30 MPa (pressure at a depth of 3,000 m) (Takami 2006). The genome structure of these species within family *Bacillaceae* is comparatively similar, and the core structure comprising more than 1,400 orthologous groups is well conserved among *Bacillaceae* (Uchiyama 2008). Therefore, moderately related bacillar genomes from eight species with different

phenotypic properties were selected to test our evaluation method for potential functionome using KEGG modules, in order to differentiate the functional potentials harbored in their genomes.

The gene products from eight bacillar genomes were assigned to KO identifiers constructing each module in 139 pathway, 112 structural complex, and 25 functional set modules. There was a significant difference in the module completion ratio by eight bacilli in terms of at least 25 pathway, 40 structural complex, and 15 functional set modules (Fig. 5a, b). In particular, the completion ratio in *Oceanobacillus iheyensis*, a mesophilic, extremely halotolerant alkaliphile, was very low in three modules for NAD biosynthesis, phosphatidylethanolamine biosynthesis, and biotin biosynthesis. These three modules were completed by all bacilli except for *O. iheyensis* although they are categorized into one of the diversified modules (pattern C). Conversely, the module for tryptophan biosynthesis belonging to pattern C was completed by only *O. iheyensis*, although other species partially completed them. Through these results it was evident that *O. iheyensis* differs from other bacilli in its metabolic potentials.

Some of the completed structural complex modules were found to be shared in bacilli with the same phenotypic properties or to be independently species specific (Fig. 5b). For example, the *Firmicutes*-specific modules for the teichoic acid transport system were shared only among three mesophilic neutrophiles (*B. subtilis*, *B. amyloliquefaciens*, and *B. licheniformis*), although this module is widely shared in other genera such as *Staphylococcus*, *Clostridium*, and *Listeria* within phylum *Firmicutes*. On the other hand, two other modules, the iron (III) transport system and phosphonate transport system which are shared in many prokaryotic species within various phyla and belonged to pattern C, were shared only among three mesophilic alkaliphiles (*B. halodurans*, *B. pseudofirmus*, and *O. iheyensis*). Although it has been previously reported that the orthologous genes for the phosphonate transport system were shared between *O. iheyensis* and *B. halodurans*

New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Table 3 Breakdown of taxonomic patterns of the KEGG modules

Pathway [226]		Structural complex [331]	
Major taxonomic pattern	Number (%)	Major taxonomic pattern	Number (%)
Nonprokaryote	79 (35.0)	Nonprokaryote	119 (36.0)
Prokaryote	50 (22.1)	Bacteria	55 (16.6)
Bacteria	30 (13.3)	Prokaryote	51 (15.4)
<i>Proteobacteria</i>	27 (11.9)	<i>Proteobacteria</i>	36 (10.9)
<i>Euryarchaeota</i>	10 (4.4)	<i>Firmicutes</i>	17 (5.1)
<i>Proteobacteria/Actinobacteria</i>	5 (2.2)	<i>Actinobacteria</i>	5 (1.5)
<i>Firmicutes</i>	4 (1.8)	<i>Cyanobacteria</i>	5 (1.5)
<i>Proteobacteria/Firmicutes/Actinobacteria</i>	3 (1.3)	Archaea	4 (1.2)
<i>Chloroflexi</i>	2 (0.9)	<i>Proteobacteria/Firmicutes</i>	4 (1.2)
<i>Crenarchaeota</i>	2 (0.9)	<i>Euryarchaeota/Crenarchaeota</i>	3 (0.9)
<i>Cyanobacteria</i>	2 (0.9)	<i>Euryarchaeota/Crenarchaeota/Nanoarchaeota</i>	3 (0.9)
<i>Actinobacteria/Crenarchaeota</i>	1 (0.4)	<i>Proteobacteria/Firmicutes/Fusobacteria</i>	3 (0.9)
<i>Chlamydiae/Cyanobacteria</i>	1 (0.4)	<i>Euryarchaeota</i>	2 (0.6)
<i>Chloroflexi/Deinococcus-Thermus/Euryarchaeota</i>	1 (0.4)	<i>Firmicutes/Tenericutes/Actinobacteria</i>	2 (0.6)
<i>Euryarchaeota/Crenarchaeota</i>	1 (0.4)	<i>Proteobacteria/Actinobacteria</i>	2 (0.6)
<i>Firmicutes/Euryarchaeota</i>	1 (0.4)	<i>Proteobacteria/Aquificae</i>	2 (0.6)
<i>Proteobacteria/Acidobacteria</i>	1 (0.4)	<i>Proteobacteria/Firmicutes/Actinobacteria</i>	2 (0.6)
<i>Proteobacteria/Actinobacteria/Acidobacteria</i>	1 (0.4)	<i>Actinobacteria/Cyanobacteria</i>	1 (0.3)
<i>Proteobacteria/Actinobacteria/Bacteroidetes</i>	1 (0.4)	<i>Actinobacteria/Verrucomicrobia/Nitrospirae</i>	1 (0.3)
<i>Proteobacteria/Actinobacteria/Cyanobacteria</i>	1 (0.4)	<i>Firmicutes/Fusobacteria</i>	1 (0.3)
<i>Proteobacteria/Cyanobacteria</i>	1 (0.4)	<i>Firmicutes/Spirochaetes</i>	1 (0.3)
<i>Proteobacteria/Firmicutes</i>	1 (0.4)	<i>Proteobacteria/Actinobacteria/Deinococcus-Thermus</i>	1 (0.3)
<i>Proteobacteria/Verrucomicrobia</i>	1 (0.4)	<i>Proteobacteria/Actinobacteria/Verrucomicrobia</i>	1 (0.3)
Functional set [86]		<i>Proteobacteria/Bacteroidetes/Aquificae</i>	1 (0.3)
Major taxonomic pattern	Number (%)	<i>Proteobacteria/Chlamydiae</i>	1 (0.3)
<i>Proteobacteria</i>	26 (30.2)	<i>Proteobacteria/Chlorobi</i>	1 (0.3)
<i>Firmicutes</i>	19 (22.1)	<i>Proteobacteria/Chlorobi/Deferribacteres</i>	1 (0.3)
Bacteria	11 (12.8)	<i>Proteobacteria/Cyanobacteria</i>	1 (0.3)
<i>Actinobacteria</i>	6 (7.0)	<i>Proteobacteria/Cyanobacteria/Chlorobi</i>	1 (0.3)
<i>Cyanobacteria</i>	6 (7.0)	<i>Proteobacteria/Firmicutes/Deferribacteres</i>	1 (0.3)
Nonprokaryote	3 (3.5)	<i>Proteobacteria/Firmicutes/Spirochaetes</i>	1 (0.3)
Prokaryote	3 (3.5)	<i>Proteobacteria/Tenericutes</i>	1 (0.3)
<i>Firmicutes/Fusobacteria</i>	2 (2.3)	<i>Proteobacteria/Thermodesulfobacteria</i>	1 (0.3)
<i>Proteobacteria/Nitrospirae</i>	2 (2.3)	Signature [9]	
<i>Firmicutes/Tenericutes/Thermotogae</i>	1 (1.2)	Major taxonomic pattern	Number (%)
<i>Proteobacteria/Acidobacteria/Deferribacteres</i>	1 (1.2)	<i>Proteobacteria</i>	5 (55.6)
<i>Proteobacteria/Acidobacteria/Planctomycetes</i>	1 (1.2)	<i>Euryarchaeota</i>	1 (11.1)

(continued)

New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Table 3 (continued)

Pathway [226]		Structural complex [331]	
<i>Proteobacteria/Chrysiogenetes/Firmicutes</i>	1 (1.2)	<i>Proteobacteria/Actinobacteria</i>	1 (11.1)
<i>Proteobacteria/Cyanobacteria</i>	1 (1.2)	<i>Proteobacteria/Thaumarchaeota</i>	1 (11.1)
<i>Proteobacteria/Firmicutes/Chlamydiae</i>	1 (1.2)	<i>Proteobacteria/Verrucomicrobia/Nitrospirae</i>	1 (11.1)
<i>Proteobacteria/Nitrospirae/Deferribacteres</i>	1 (1.2)		
<i>Proteobacteria/Spirochaetes/Verrucomicrobia</i>	1 (1.2)		

[] shows total number of the KEGG modules containing branched modules

(Takami et al. 2012), it could be easily visualized using our new evaluation method that this system was also shared in other mesophilic and alkaliphilic *B. pseudofirmus*, whose genome sequence has been completed recently. Although how the differentiated functional modules confer phenotypic properties directly or indirectly is still unclear, a series of the above results should be helpful in better understanding of the physiological properties.

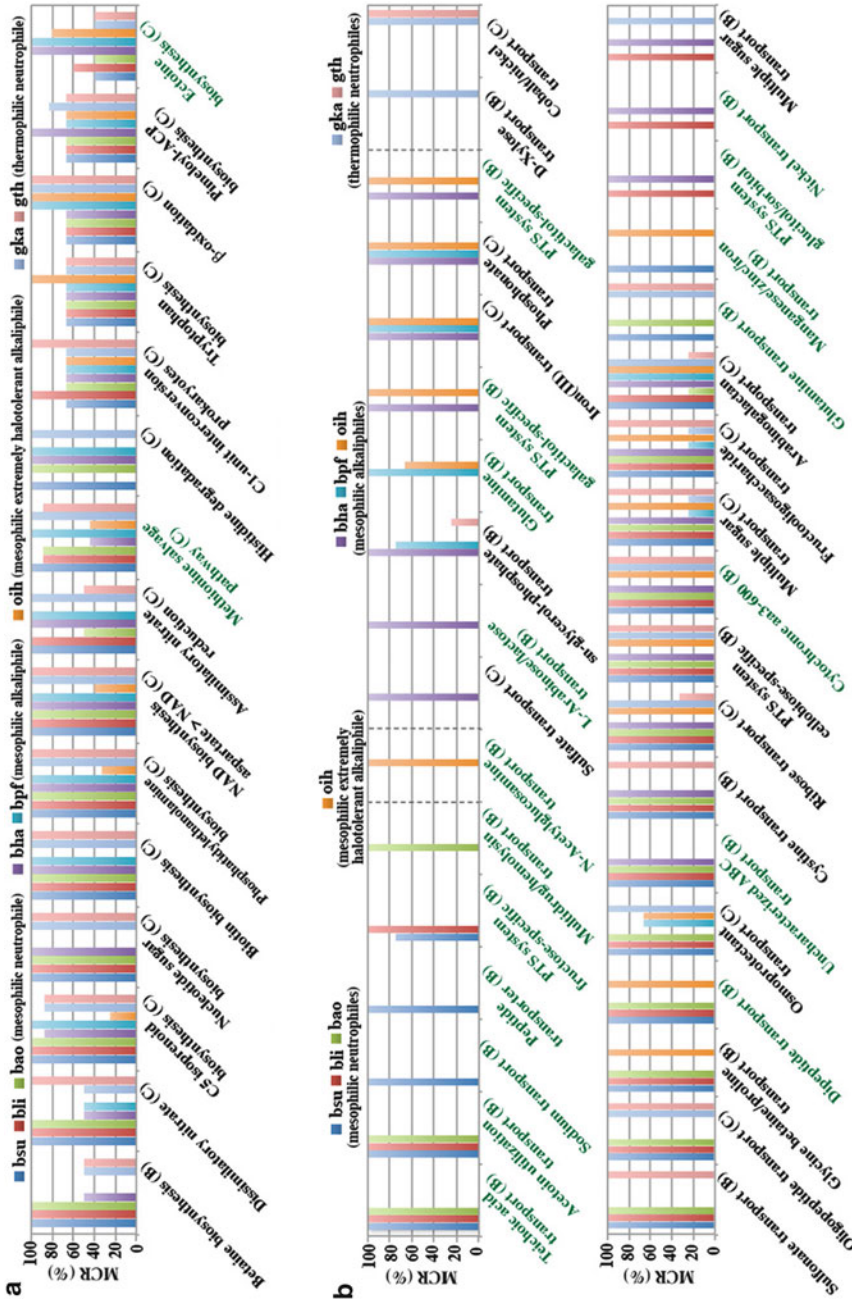
Comparative Functionome Analysis of Humans and Human Gut Microbiomes

The completion ratio of each KEGG module was compared between humans and human gut microbiomes to illustrate their metabolic linkage. The metagenomic data of gut microbiomes from 13 healthy Japanese individuals, previously reported on, was used (Kurokawa et al. 2007). There was a significant difference in the module completion ratios of 13 individuals in terms of at least 33 pathway modules (Fig. 6a).

The most complete 16S rRNA gene sequence-based enumerations available in human gut microbiomes indicate that more than 90 % of phylotypes belong to just two of the 70 known divisions of Bacteria, the *Bacteroidetes* and the *Firmicutes*, with the remaining phylotypes distributed among eight other phyla (Eckburg et al. 2005). Pairwise comparison of the completion ratio of the KEGG module clearly demonstrated the well-recognized functional complementation of the gut microbiome to the human host, which includes essential amino acid and vitamin biosynthesis. The contributors completing the modules for vitamin production are *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and

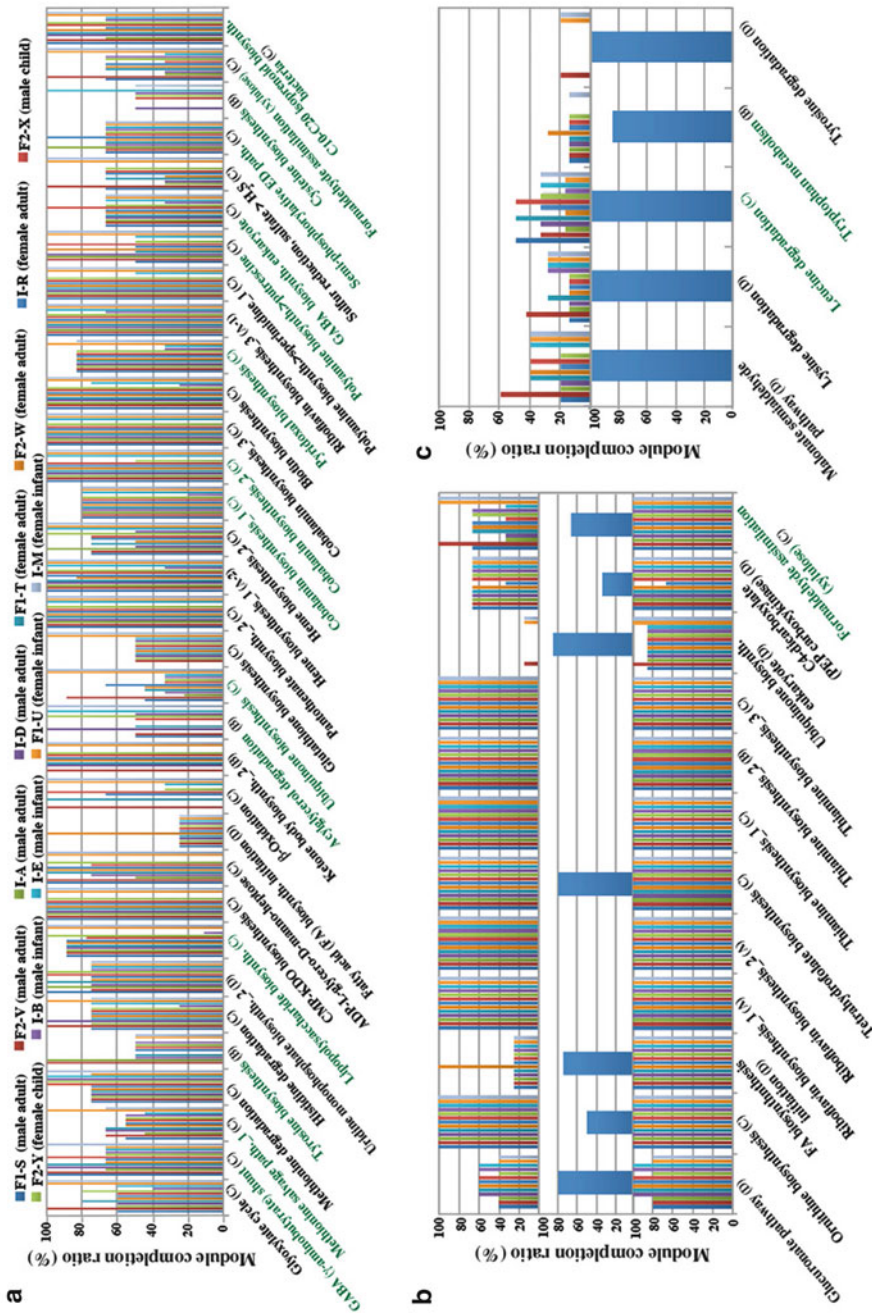
Gammaproteobacteria. Completion patterns of the KEGG module for these amino acids and vitamins mainly fall into patterns C and D except for riboflavin biosynthesis belonging to one of the universal modules A, indicating that these modules are involved in the nutritional supply for the gut microbiome as well as for the host (Fig. 6b). Interindividual variation was also evident in the completion ratio of the module for vitamins. For example, the module belonging to pattern C for pyridoxal (vitamin B6) biosynthesis was mainly attributable to *Bacteroidetes* in adults and *Gammaproteobacteria* in infants; however, its completion ratio in two male infants (In-B and In-E) was extremely low (33.33 %) (Fig. 6a). Interindividual variations in completion ratios were also observed in modules for polyamine biosynthesis, for example, putrescine, spermidine, and spermine (Takami et al. 2012). Similarly, the completion ratio of the KEGG modules for γ -aminobutyric acid (GABA) varied among individuals, and *Gammaproteobacteria* mainly contributed to GABA production (Fig. 6a). Because these polyamines and GABA are essential biological substances that act as cell growth promoters and inhibitory neurotransmitters, respectively, in humans, these variations may be linked to susceptibilities to certain diseases. Indeed, a recent report on metabolic changes in gut microbiomes after bariatric surgery for obese patients demonstrated their potential for polyamine production in the gut; elevated protein putrefaction because of the bypassed food passage promoted putrescine and GABA production from gut microbiota (Li et al. 2011).

Interestingly, gut microbiomes showed preference for amino acid catabolism. The gut



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 5 Comparison of module completion patterns in eight phenotypically different *Bacillus*-related species. (a) Pathway modules showing remarkable differences appeared among the eight species. (b) Structural complex modules showing remarkable differences appeared among the eight species. *Upper plot* indicates common or specific modules in the species possessing each phenotype. Green letters show rare modules completed by less than 10 % of 1,256 prokaryotic

species. Alphabet in parentheses shows the patterns of completion profile based on the module completion ratio as shown in Table 2 and Fig. 4. *bsu*, *B. subtilis*; *bao*, *B. amyloliquefaciens*; *bli*, *B. licheniformis*; *bha*, *B. halodurans*; *B. pseudofirmus*; *oih*, *O. iheyensis*; *gka*, *G. kaustophilus*; and *gth*, *G. thermoglucosidarius*. This figure has been redrawn with the updated KEGG module database from the previous one (Takami et al. 2012)



New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE, Fig. 6 Comparison of module completion patterns in humans and human gut microbiomes from 13 healthy individuals. (a) Typical pathway modules showing remarkable differences in the module completion ratio appeared among human gut microbiomes from 13 healthy individuals. (b) Typical pathway modules possessing complementary relationships between humans and human gut microbiomes in the module completion ratio. (c) Typical pathway modules for which the completion ratio in the human gut microbiome is very low in contrast to that in humans. Detailed information of the 13 individuals has been previously described (Kurokawa et al. 2007). This figure has been redrawn with the updated KEGG module database from the previous one (Takami et al. 2012)

microbiome did not seem to utilize exogenous lysine, leucine, and aromatic amino acids such as tryptophan and tyrosine (Fig. 6c). To our knowledge, this is a novel finding on the nutritional preference of gut microbes. This may be one of the mutualistic representations of gut microbiomes to avoid nutritional competition with the host because these aromatic amino acids are precursors of various biological substances such as catecholamines, melatonin, serotonin, thyroid hormones, and NAD. Thus, the new evaluation method based on the KEGG modules is expected not only to highlight the metabolic linkage between host and commensal microbes but also to identify microbiome-based biomarkers for particular diseases.

Summary

A new evaluation method for potential functionomes based on the KEGG modules was developed. Using this new method, significant difference in module completion ratio by eight bacilli in terms of at least 25 pathway, 40 structural complex, and 15 functional set modules was highlighted, although how the differentiated functional modules confer phenotypic properties directly or indirectly is unclear thus far. Because the coverage of KEGG modules over whole metabolic and signaling networks is continuously increasing, differences in module completion ratio will provide some important clues to the understanding of phenotypic properties. Furthermore, variations in the functional potential of human gut microbiomes from 13 healthy individuals could be characterized by the pathway and structural complex module units, and the complementarity between biochemical functions in human hosts and nutritional preferences in human gut microbiomes identified.

Functional annotations to metagenomic sequences remain difficult because metagenomic data targeting various environments still contains incomplete genes from various unidentified species, absent in a reference database. In this entry, the KAAS system was used for functional annotation to the human metagenomes and also

applied to estimate database dependency on the accuracy of the KO assignment using the *E. coli* draft genome. As a result, the KAAS system could correctly assign to KO groups with an accuracy rate of approximately 80 %, even if the gene hosts were not classified into known phyla within the reference database. Thus, this method will work well for comparative functional analysis in metagenomics, able to target unknown environments containing various uncultivable microbes within unidentified phyla, although further verification studies on database dependency for metagenomics should be performed. Based on this method, we developed the metabolic and physiological potential evaluator (MAPLE) and provided a user-friendly Web interface not only for the characterization of potential functionome harbored in the genomic and metagenomic sequences but also for comparative analyses for the MCR and mapping patterns to the KEGG modules (<http://www.genome.jp/tools/maple/>).

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Human Gut Microbial Genes by Metagenomic Sequencing](#)
- ▶ [KEGG and GenomeNet, New Developments, Metagenomic Analysis](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)

References

- Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. *Science*. 2005;308:1635–8.
- Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. *Genome Res*. 2011;21:1552–60.
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
- Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and environment. *Nucleic Acids Res*. 2008;36:D480–4.

- Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiome. *DNA Res.* 2007;14:169–81.
- Li JV, Ashrafian H, Bueter M, et al. Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk. *Gut.* 2011;60:1214–23.
- Markowitz VM, Chen I-MA, Palaniappan K, et al. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 2012;40:D115–22.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics.* 2008;9:386.
- Moriya Y, Itoh M, Okuda S, et al. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 2007;35:W182–5.
- Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33:5691–702.
- Takami H. Genomic diversity of extremophilic Gram-positive endospore-forming *Bacillus*-related species. In: Williams CR, editor. *Trends in genome research.* New York: NOVA Publisher; 2006. p. 25–85.
- Takami H, Taniguchi T, Moriya Y, et al. Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC Genomics.* 2012;13:699.
- Uchiyama I. Multiple genome alignment for identifying the core structure among moderately related microbial genomes. *BMC Genomics.* 2008;9:515.
- Willke A, Harrison T, Wilkening J, et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics.* 2012;13:141.

Next-Generation Sequencing for Metagenomic Data: Assembling and Binning

Henry C. M. Leung, Yi Wang, S. M. Yiu and Francis Y. L. Chin

Department of Computer Science, The University of Hong Kong, Hong Kong, China

Introduction

Microorganisms contribute the largest number of living cells in the world. The activity of different microbes forms a microbial ecosystem which

affects all species in the world. Traditional method of studying microorganisms requires culturing a single kind of microbe and studying each microbe based on next-generation sequencing (NGS) technology by its genome one at a time (Perna et al. 2001). However, as a single kind of microbe usually cannot live alone and over 99 % of microbes cannot be cultivated in the laboratory (Rappe and Giovannoni 2003; Eisen 2007), traditional culture-based method cannot analyze the interactivity of a microbial community well. Metagenomic, which studies all microbes in a community as a whole, is introduced for solving the problem. Based on the NGS technology (Shendure and Ji 2008), instead of sequencing each single cultivated microbe one by one, metagenomic sequences all microbes in an environment sample as a community directly without cultivation (Weinstock 2012; Gilbert and Dupont 2011; Hunter et al. 2012; Tremaroli and Backhed 2012; Wooley et al. 2010). Thus, genomes of microbes that cannot be studied before can now be obtained and be analyzed.

However, the complexity of a microbial community is high. There can be tens of thousands kinds of microbes in a single sample. As genomes of these microbes coexist in the sample, reads (DNA short fragments) obtained from genomes of different microbes are mixed and required to be separated after NGS step. More seriously, as the abundance of different microbes in a sample can vary with several orders of magnitudes (Qin et al. 2010), few reads are sequenced from the low-abundance species which may be treated as erroneous reads. Thus, several approaches have been developed for analyzing metagenomic data depending on the property of samples and research objectives.

Sequencing Biomarker

Traditional sequencing techniques, e.g., Sanger (Sanger and Coulson 1975), have a relatively low throughput. Thus, it is impossible to sequence the whole genome sequences of all microbes in a sample, especially for the low-abundance species. Instead of sequencing the whole genome,

biologists usually design primers for capturing short regions in the genomes of various microbes, e.g., fingerprinting polymerase chain reaction (PCR) on 16S rRNA genes. Each 16S rRNA gene is a 1.5-kilobase-long gene for encoding part of the prokaryotic ribosome. Although each genomic sequence varies among different bacteria, there are some conserved regions (for the ribosome function) in the 16S rRNA gene such that primer can be designed for capturing the 16S rRNA gene for different bacteria. Moreover, species with 97 % identical in the 16S rRNA gene usually are in the same operational taxonomic unit (OUT) (Weinstock 2012). Thus, sequencing the 16S rRNA genes can determine which kinds of bacteria in a sample and their relative abundances (16S rRNA genes of high-abundance bacteria will be sequenced more than those of low-abundance bacteria resulting more reads covering these genes). Instead of 16S rRNA, 18S rRNA gene encodes eukaryotic ribosome and can also be sequenced for identifying eukaryotes in a sample.

However, as the read lengths of most popular sequencing techniques are shorter than 1.5 kb (typical length of a 16S rRNA gene), biologists can only sequence a portion of 16S rRNA genes, and the accuracy of identification depends on the read length. Traditional Sanger sequencing techniques can produce 1-kb-long read which can cover a larger portion of 16S rRNA genes. However, its throughput is low such that 16S rRNA genes of many species may not be sequenced and the relative abundances of species may not be estimated well. One of the next-generation sequencing techniques, 454 pyrosequencing, can produce several orders more reads than the Sanger sequencing technique, but the read length is about 400 bases, which can cover only a short portion of 16S rRNA gene, and thus the sensitivity of identifying different microbes in a sample will decrease. The Illumina platform, another next-generation sequencing technique, can produce several orders more reads than 454 pyrosequencing; however, the read length is at most 250 bases, thus resulting to lower sensitivity than 454 pyrosequencing.

Besides the problem of sequencing the whole 16S rRNA gene with high throughput, there is another problem of analyzing metagenomic data using 16S rRNA genes (or 18S rRNA genes). Microbe can transfer gene from one to another without reproduction process, horizontal gene transfer, and thus the 16S rRNA gene of one kind of microbe may be transferred to another microbe and introduces problems in analyzing metagenomic data. In real situation, microbes can have multiple copies of 16S rRNA genes, varying from 1 to 15 (Case et al. 2007; Klappenbach et al. 2001), and horizontal gene transfer makes the abundances difficult to be estimated. Recently, other housekeeping genes, e.g., *rpoB*, *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM*, are used (in addition of 16S rRNA gene) for identifying different species in a metagenomic sample.

Sequencing Whole Genome

Since using a single or only several biomarkers to represent a species may have a problem, another way to analyze metagenomic data is sequencing the whole genomes of different microbes in the sample. With the help on the high-throughput next-generation sequencing techniques, biologists can sequence the whole genomes of all microbes in a sample with reasonably high sequencing depth.

Assembling Reads

As the read lengths of next-generation sequencing are much shorter than the genomes of microbes, analyzing sequenced reads directly is difficult especially for Illumina platform. One possible way is assembling overlapped short reads to longer contigs before analysis (Mende et al. 2012). Although there are many existing assembling algorithms (Vyahhi et al. 2012; Peng et al. 2010) designed for genomic data, they cannot be applied on metagenomic data directly because of the following results:

1. Abundances of different microbes vary in metagenomic data. Since erroneous reads

introduce arbitrary for assembling, existing genomic assemblers try to determine erroneous reads and remove them before assembling. Based on the assumption that erroneous reads are sampled fewer times than correct reads, these genomic assemblers usually consider those reads or length k substring of reads, called k -mers, with low sampling rate (multiplicity) as erroneous reads and k -mers. These erroneous reads are removed before assembling. However, since the abundance of microbes vary a lot in metagenomic data, correct reads and k -mers from low-abundance microbes could be sampled much fewer than the erroneous reads and k -mers from high-abundance microbes. These genomic assemblers fail to remove erroneous reads and k -mers and produce either very short contigs or incorrect long contigs.

2. Common regions across different microbes. Due to horizontal gene transfer and the existence of common housekeeping genes, some common patterns could appear in multiple genomes. As the read length can be shorter than these common patterns, genomic assemblers cannot determine the genomic sequences of microbes near their common patterns. Although similar problem also appears in assembling genomic data, the number of common patterns in metagenomic genomic is much more than those in genomic data (Peng et al. 2011). As a result, shorter or erroneous contigs will be produced by existing genomic assemblers.
3. Huge data size. As the number of microbes in a metagenomic data is huge, a high sequencing depth is required to obtain enough reads (say $10\times$ coverage) from each microbe (especially for the low-abundance microbes). Thus, the total amount of input reads (e.g., 200G nucleotides in the metagenomic data of cow stomach (Qin et al. 2010), over 100G of nucleotides required for studying soil metagenome (Frisli et al. 2013)) for assembling metagenomic data can be much more than the genomic data. How to store and process this huge amount of reads becomes a big problem.

Due to the above problem, several assemblers have been developed for assembling metagenomic data, including Genovo (Laserson et al. 2011) for 454 pyrosequencing and MetaVelvet (Namiki et al. 2012), Ray Meta (Boisvert et al. 2012), Meta-IDBA (Peng et al. 2011), and IDBA-UD (Peng et al. 2012) for the Illumina platform. Since the length of 454 pyrosequencing read is longer than those constructed by Illumina platform and the number of input reads is much smaller than those by Illumina platform, Genovo stores all the input reads and calculates their pairwise overlapped relationship. It then calculates the probability of a set of reads sampled from the same contigs based on Bayesian approach and applies a series of hill climbing to obtain a set of contigs with the highest likelihood. However, this approach fails when the number of input reads increases (Boisvert et al. 2012). Because of the huge amount of input reads, MetaVelvet, Ray Meta, Meta-IDBA, and IDBA-UD all assemble contigs using de Bruijn graph approach. A de Bruijn graph represents the connection of a set of reads using k -mers, length k strings of the read. Each k -mer in the reads is represented by a vertex, and there is an edge from vertex u to vertex v if and only if k -mers u and v appear in at least one read consecutively, i.e., the length- $(k-1)$ suffix of u is the same as the length- $(k-1)$ prefix of v . Thus, a contig is represented by a path in the de Bruijn graph. Because of the existence of sequencing error and common regions among different genomes, paths representing different genomes may overlap and the de Bruijn becomes complicated. Existing metagenomic assemblers apply different approach to decompose the de Bruijn graphs or determine contigs directly from the de Bruijn graph. Meta-IDBA decomposes the de Bruijn graph based on the observation that there are more interconnections between k -mers sampled from the same genome than k -mers from sampled different genomes. After decomposition, paths representing different genomes will be separated and can be reconstructed easier. MetaVelvet decomposes the de Bruijn graphs based on the multiplicities of k -mers.

By determining some local peaks in the distribution of multiplicities of k -mers, MetaVelvet decomposes the de Bruijn graph according to the multiplicities. As k -mers sampled from different genomes may have similar multiplicities and k -mers sampled from the same genome could have different multiplicities (due to sequencing bias), IDBA-UD calculates the average multiplicity of k -mers in the same contig and uses it to determine erroneous k -mers and k -mers sampled from different genomes. As the threshold is determined locally, it can decompose the de Bruijn more accurate than Meta-IDBA and MetaVelvet using global thresholds. Ray Meta uses another approach to construct the contigs. Instead of decomposing the de Bruijn graph, it applies a heuristics-guided graph traversal to reconstruct the contig. Although all the above assemblers try to reconstruct contigs from metagenomic data, short contigs (several thousand nucleotides) and chimera contigs (misassembles contigs from different genome together) could be resulted because of the high diversity of metagenomic data.

Since the number of k -mer is large, researches have been performed for investigating storage of de Bruijn graph using less memory. Several efficient data structures have been developed based on bloom filter (Chikhi and Rizk 2012; Pell et al. 2012). A bloom filter uses a hash table and several hash functions to store the existence of k -mers. When storing a k -mer, each hash function will calculate an address based on the pattern of k -mer, and all these addresses will be set to 1 in the hash table. Thus, the existence of a k -mer in the reads can be determined by checking several bits in the hash table. Although there may be some false-positive k -mers, the number of false positives is small when the hash table is large enough and there are multiple hash functions.

Binning

After reconstructing contigs, each long contig can be aligned to known reference genomes in the database for identifying the microbes in the samples (Huson et al. 2011). Even when there is no similar reference genome in the database, gene sequence may be predicted (Rho et al. 2010)

using different classifiers which help analyzing the metabolism of the unknown microbes. However, for the contigs sampled from microbes without genome reference and low-abundance microbes without enough reads for assembling long contigs, binning approach is required. Note that since the most microbes cannot be cultivated and their genomes are still unknown, many reads and contigs cannot be aligned to reference genome in the database.

Binning reads and contigs is to cluster reads and contigs sampled from the same microbes using the common property on the reads. Composition-based methods use generic features, e.g., GC content, codon usage, dinucleotides distribution, and 4-mer distribution to classify reads sampled from different genomes. Existing supervised or semi-supervised binning algorithm (Brady and Salzberg 2009; McHardy et al. 2006) can construct a classifier to determine the source of reads based on reference genome in the database. Compared with alignment-based methods, these algorithms do not require the exact reference genome. Instead, classifier can be constructed from a similar genome in the database such that more reads can be binned. However, as there are limited number of reference genomes in the database, many reads still cannot be classified correctly. Some binning algorithms are designed to cluster reads sampled from the same genome using properties on reads directly without any reference genomes.

MetaCluster 3.0 (Yang et al. 2010) clusters reads based on 4-mer distribution. Given two long reads from the same genome, the occurrence frequencies of different 4-mers on the two reads should be similar (Zhou et al. 2008). MetaCluster 3.0 calculates the pairwise spearman distance of reads based on 4-mer distributions and clustering reads using k -mean clustering methods. However, MetaCluster 3.0 can only handle metagenomic data with similar abundances and long read length (500 bp or more). In order to bin short reads of length about 100 bp, AbundanceBin (Wu and Ye 2011) and TOSS (Tanaseichuk et al. 2012) consider the occurrence frequency of k -mers ($k = 25$) in all the reads. k -mers that occur frequently should be sampled

from high-abundance microbes, while k -mers that occur rarely should be sampled from low-abundance microbes. Based on this assumption, AbundanceBin and TOSS can bin reads according to the k -mer frequencies. However, when the abundances of two microbes are similar (abundance ratio within 1:3), these algorithms fail to separate the reads sampled from the two microbes. MetaCluster 4.0 further improves MetaCluster 3.0 by combining overlapped short reads to long virtual contigs and estimates the 4-mer or 5-mer distribution of the virtual contigs. As the lengths of virtual contigs are much longer than the short reads, 4-mer distribution of the virtual contigs can be estimated accurately. By constructing a huge number of small clusters and merging cluster with similar 4-mer distribution, MetaCluster 4.0 (Wang et al. 2012a) can handle metagenomic data with microbes of different abundances. However, these unsupervised binning algorithms cannot handle low-abundance microbes well because they cannot distinguish reads sampled from these low-abundance microbes from the error reads sampled from high-abundance microbes. MetaCluster 5.0 (Wang et al. 2012b) is designed for binning reads from both high- and low-abundance microbes. It performs binning with two rounds. In the first rounds, its target is to bin reads sampled from high-abundance microbes using restricted parameters for constructing virtual contigs and clustering reads. Reads sampled from low-abundance microbes can be handled in the second round using less restricted parameters. By applying multiple rounds of binning, MetaCluster 5.0 can bin reads from microbes with sequencing depth as low as $6\times$ in a metagenomic dataset containing 100 microbes. However, it still cannot bin reads sampled from microbes with sequencing depth lower than $6\times$.

Conclusion

Assembling and binning reads are two important procedures for analyzing metagenomic data. The high biodiversity and large variations in abundances of genomes in metagenomic data make

the problems challenging. A common practice for analyzing metagenomic data is to assemble short reads to longer contigs. Then try to identify microbes in the sample by aligning the contigs and unassembled reads to reference genomes. As most of the microbes have no reference in the database, the unaligned reads and contigs should be binned together using generic features, e.g., GC content, codon usage, dinucleotide distribution, and 4-mer distribution. Previous study shows that binning contigs instead of reads can improve the accuracy of binning. It is because the long contigs carry more generic information than the short reads. However, few researches have been performed on studying how to improve the result of assembling using binning. Moreover, researchers usually use the information of reference genomes by alignment and supervising binning. In fact, similar genomes in the database may be used to improve the performance of de novo assembling. As the performance of existing de novo assemblers and binning algorithms on real biological data is not satisfied, further researches on combining assembling, binning, and the use of reference genomes may be a possible way to improve the performance of analyzing metagenomic data.

References

- Boisvert S, Raymond F, Godzaridis E, et al. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 2012;13(12):R122.
- Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods.* 2009;6:673–6.
- Case RJ, Boucher Y, Dahllöf I, et al. Use of 16s rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol.* 2007;73:278–88.
- Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a bloom filter. *Algorithm Bioinforma.* 2012;7534:236–48.
- Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol.* 2007;5(3):e82.
- Frisli T, Haverkamp TH, Jakobsen KS, et al. Estimation of metagenome size and structure in an experimental soil microbiota from low coverage next-generation sequence data. *J Appl Microbiol.* 2013;114(1):141–51.
- Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci.* 2011;3:347–71.

- Hunter CI, Mitchell A, Jones P, et al. Metagenomic analysis: the challenge of the data bonanza. *Brief Bioinform.* 2012;13(6):743–6.
- Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 2011;21:1552–60.
- Klappenbach JA, Saxman PR, Cole JR, et al. rrndb: the ribosomal RNA operon copy number database. *Nucleic Acid Res.* 2001;29:181–4.
- Laserson J, Jojic V, Koller D. Genovo: de novo assembly for metagenomes. *J Comput Biol.* 2011;18(3):429–43.
- McHardy AC, Martin HG, Tsirigos A, et al. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2006;4:63–72.
- Mende DR, Waller AS, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE.* 2012;7(2):e31386.
- Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012;40(20):e155.
- Pell J, Hintze A, Canino-Koning R, et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc Natl Acad Sci.* 2012;109(33):13272–7.
- Peng Y, Leung HC, Yiu SM, et al. IDBA- a practical iterative de Bruijn graph de novo assembler. *Res Comput Mol Biol.* 2010;6044:426–40.
- Peng Y, Leung HC, Yiu SM, et al. Meta-IDBA: a de novo assembler for metagenomic data. *Bioinformatics.* 2011;27:i94–101.
- Peng Y, Leung HC, Yiu SM, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with high uneven depth. *Bioinformatics.* 2012;28:1420–8.
- Perna N, Plunkett III G, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature.* 2001;409:529–33.
- Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464(7285):59–65.
- Rappe MS, Giovannoni SJ. The uncultured microbial majority. *Annu Rev Microbiol.* 2003;57:369–94.
- Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010;38(20):e191.
- Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol.* 1975;94(3):441–8.
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008;26:1135–45.
- Tanaseichuk O, Borneman J, Jiang T. A probabilistic approach to accurate abundance-based binning of metagenomic reads. *Algorithm Bioinforma.* 2012;7534:404–16.
- Tremaroli V, Backhed F. Functional interactions between the gut microbiota and host metabolism. *Nature.* 2012;489:242–9.
- Vyahhi N, Pyshkin A, Pham S, et al. From de Bruijn graphs to rectangle graphs for genome assembly. *Algorithm Bioinforma.* LNCS. 2012;7534:249–61.
- Wang Y, Leung HC, Yiu SM, et al. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *J Comput Biol.* 2012a;19:241–9.
- Wang Y, Leung HC, Yiu SM, et al. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics.* 2012b;28:i356–62.
- Weinstock GM. Genomic approaches to studying the human microbiota. *Nature.* 2012;489:250–6.
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol.* 2010;6(2):e1000667.
- Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using *l*-tuples. *J Comput Biol.* 2011;18(3):523–34.
- Yang B, Peng Y, Henry CM, et al. Unsupervised binning of environmental genomic fragments based on an error robust selection of *l*-mers. *BMC Bioinforma.* 2010;11 Suppl 2:S5.
- Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. *BMC Bioinforma.* 2008;9(1):546.

NGS QC Toolkit: A Platform for Quality Control of Next-Generation Sequencing Data

Ravi K. Patel and Mukesh Jain
Functional and Applied Genomics Laboratory,
National Institute of Plant Genome Research
(NIPGR), New Delhi, India

Synonyms

Format converters; Illumina; NGS data quality control; NGS data trimming; Roche 454

Definition

NGS QC Toolkit is a Perl-based stand-alone program package for the quality control (QC) of next-generation sequencing (NGS) data. In addition to QC tools, it consists of many subsidiary tools for handling and processing of data obtained from Illumina and Roche 454 sequencing platforms. The open-source toolkit is freely available at <http://www.nipgr.res.in/ngsqctoolkit.html>.

Introduction

The need for fast and high-throughput sequencing has resulted into discovery of NGS technologies. The advent of these technologies has transformed the genomics research by providing an opportunity to study genetic information at a single-base resolution in cost-effective manner (Metzker 2010). However, usually several artifacts are reflected in NGS data due to technical errors and limitations associated with different NGS platforms. These sequence artifacts, including read errors, poor-quality reads, and primer/adaptor contamination, might affect downstream sequence analysis, such as *de novo* genome and transcriptome assembly, gene expression studies, and single nucleotide polymorphism detection. To avoid misleading conclusions, it is necessary to filter the NGS data for these sequence artifacts (Benaglio and Rivolta 2010).

NGS platform vendors have developed commercial QC pipelines dedicated to mitigate the effect of limitations associated with their platforms. However, even after processing through these pipelines, many sequence artifacts remain in the data. Several efforts have been made to resolve one or the other sequence artifacts, but many of them are specific to a particular sequencing platform. NGS QC Toolkit (Patel and Jain 2012) can handle many of the known sequence artifacts in Illumina and Roche 454 sequencing data. It is a stand-alone and user-friendly toolkit written in Perl programming language by employing modularized structure supported by several subroutines for various tasks, which allows better maintainability. The toolkit comprises many easy-to-use tools for quality check and filtering, trimming, generating statistics, and different file format/variant conversion for Illumina and Roche 454 sequencing data (Fig. 1).

QC Workflow

The toolkit provides dedicated tools for the QC of single-end (SE) and paired-end (PE) data from Illumina (IlluQC tools) and Roche 454 (454QC tools) sequencing platforms. Although various

parameters in these tools are set to the sensible default values, they can be adjusted by the users to optimize QC analysis, which makes these tools versatile for different NGS assays. IlluQC has the ability to identify different FASTQ file variants (Cock et al. 2009) and set the quality scoring system accordingly for further analysis. Reads are analyzed based on their quality, and the poor-quality reads not fulfilling the user-specified criteria are discarded. The filtered reads are checked for the primer/adaptor sequence contamination and the matching reads are discarded. The high-quality filtered data is exported as output along with various quality statistics. 454QC tools read FASTA files and filter reads based on the specified length cutoff at several stages in the analysis. The tool can also perform trimming of reads containing homopolymer(s) longer than specified length. Further, the quality check and primer/adaptor sequence match are performed similar to that of IlluQC tools. However, unlike IlluQC tools, 454QC tools trim respective ends of the read showing primer/adaptor match. Eventually, the high-quality reads are exported in FASTA format. Processing of Roche 454 PE data (using 454QC_PE.pl) requires an additional step of finding the linker sequence to separate and process both end reads simultaneously.

Key Characteristics

While NGS QC Toolkit shares its features with many other QC tools (Schmieder and Edwards 2011; Cox et al. 2010; Lassmann et al. 2009; Pandey et al. 2010), it also provides few unique attributes for the QC analysis of NGS data. In addition to high-quality filtered data output, it is also equipped with the modules for generating several different kinds of statistics in graphical format along with text files to help users make better understanding of the data quality (Patel and Jain 2012).

Reduced Computational Time and Storage Space Requirement

Continued improvement in NGS technologies has achieved larger read length and manifold

increase in throughput. To reduce time requirement for the QC of several gigabases of sequence data, parallel computing has been implemented in the QC tools. Significant decrease in the analysis time was evident using parallelized QC tools on multi-core computer systems (Patel and Jain 2012). Nevertheless, tools can also be run on single-core computers without any additional requirement. Another challenge with the huge NGS dataset is the increased storage space requirement, which is considerably reduced by the use of compressed (gzip) files. The high-quality filtered output data in compressed gzip files can be used directly for downstream analysis, which saves large amount of storage space.

Conservation of PE Data Integrity

PE sequencing data helps increase sequence coverage and confidence in the alignment which is very crucial for downstream analysis. However, surprisingly, not many QC pipelines maintain the pairing information of the PE data in the filtered data but the NGS QC Toolkit. QC tools analyze both reads of each pair concurrently and export the high-quality filtered PE data along with the unpaired reads (when only one read of the pair passes QC filters). In this way, QC tools maintain PE data integrity and try to retain all important high-quality sequencing data.

Homopolymer Trimming

A major artifact is introduced in Roche 454 pyrosequencing data by the use of pyrophosphate for the detection of incorporated bases. It was found that linearity of signal intensity is disturbed when longer homopolymer is encountered (Margulies et al. 2005). This artifact may affect the downstream analysis due to frameshift. 454QC tools provide an optional parameter to trim the homopolymer of the given minimum threshold length.

FASTQ Variant Detection

Use of inconsistent variants of FASTQ format by different sequencing platforms makes it tough for the users to apply appropriate tools for the analysis, because the quality scoring system varies with the variants (Cock et al. 2009). To make

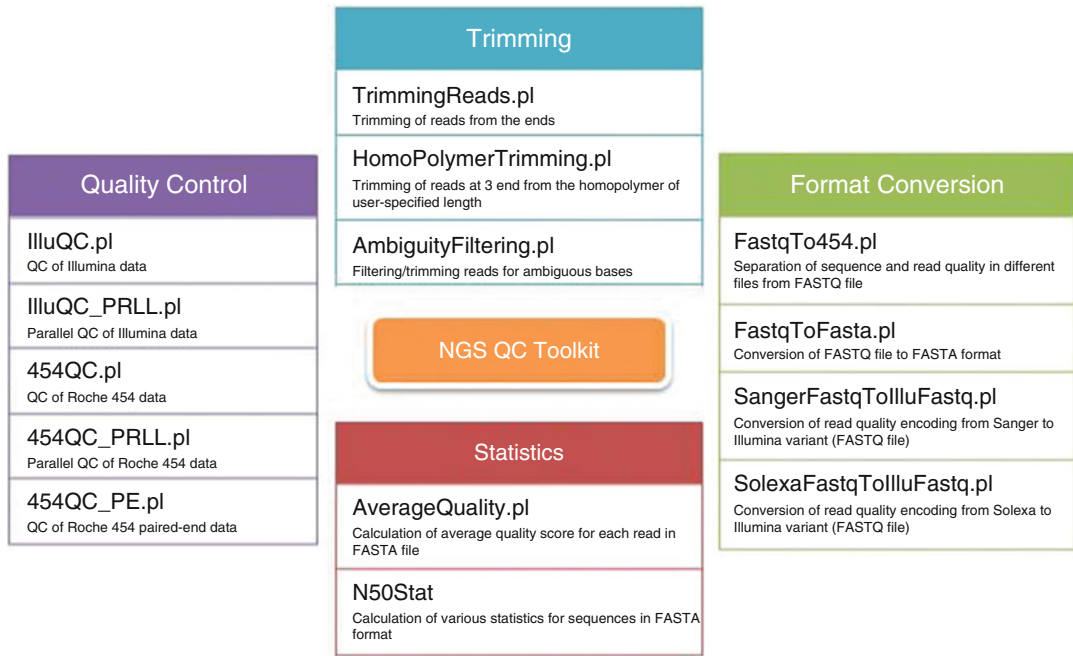
the analysis easier, IlluQC tools are programmed to first identify the input FASTQ variant automatically and set appropriate scoring system for further QC analysis.

Additional Tools

Apart from QC tools, a number of additional tools are provided in the toolkit to manage and generate statistics for the NGS data (Fig. 1). A set of sequence format converter tools offer facility to convert between different variants of the FASTQ format based on the equations described previously (Cock et al. 2009). It also provides tools for conversion between FASTQ and FASTA formats. `TrimmingReads.pl` tool is capable of trimming reads based on two criteria. It can trim given number of bases from the 5' and/or 3' end of the reads. Another mode of trimming is to trim low-quality bases from the 3' end of the reads using user-defined threshold value of quality scores. `HomopolymerTrimming.pl`, as the name suggests, clips the 3' read end from first nucleotide of the homopolymer of user-defined cutoff length. A newly introduced tool upon request from users, i.e., `AmbiguityFiltering.pl`, helps to filter reads containing ambiguous bases (N/X content) or to trim flanking ambiguous bases. A couple of tools, `AvgQuality.pl` and `N50Stat.pl`, generate statistics to help nonexpert users to access various sequence statistics.

Installation

The toolkit requires Perl interpreter and few additional Perl modules like GD (optional; required to generate QC graphs) and `String::Approx`. Users need to download NGSQCToolkit zip folder from the website. The toolkit is ready to use just after unzipping the folder. The distribution includes all the tools along with a user manual, which provides important links for the module installation and describes the tools and their usage in detail. Tools can report the missing dependencies, if required modules are not found or improperly installed.



NGS QC Toolkit: A Platform for Quality Control of Next-Generation Sequencing Data, Fig. 1 Various QC and data processing tools included in the NGS QC Toolkit

Toolkit Updates

Continuous support and updates played a crucial role in the popularity of the NGS QC Toolkit among the researchers working on NGS data analysis. It has been under active development since after it had been developed more than 3 years ago. Several updates have been made to make the toolkit compatible with the ever-evolving sequencing technologies and fulfill the requirements of users (<http://www.nipgr.res.in/ngsqctoolkit.html>).

Summary

NGS QC Toolkit is an open-source stand-alone toolkit for the QC of NGS data, which can be used on any operating system with installed prerequisites. It offers user-friendly parallel computing QC tools for the quality check and filtering of Illumina and Roche 454 sequencing data. These tools provide various parameters to optimize the QC analysis of different kinds of NGS assays.

In addition, the toolkit is comprised of numerous supplementary tools for handling/processing of NGS data. This toolkit is being regularly modified and improved to accommodate users' requirements and make it compatible with changing sequencing data file formats. It is anticipated that this toolkit will provide an easy platform to even non-bioinformaticians for QC analysis of NGS data.

Cross-References

- ▶ [A De Novo Metagenomic Assembly Program for Shotgun DNA Reads](#)
- ▶ [DNA Methylation Analysis by Pyrosequencing](#)

References

Benaglio P, Rivolta C. Ultra high throughput sequencing in human DNA variation detection: a comparative study on the *NDUFA3-PRPF31* region. *PLoS One*. 2010;5(9):e13071.

- Cock PJA, Fields CJ, Goto N, et al. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2009;38:1767–71.
- Cox MP, Peterson DA, Biggs PJ. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics.* 2010;11:485.
- Lassmann T, Hayashizaki Y, Daub CO. TagDust-a program to eliminate artifacts from next generation sequencing data. *Bioinformatics.* 2009;25:2839–40.
- Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437:376–80.
- Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet.* 2010;11:31–46.
- Pandey RV, Nolte V, Schlotterer C. CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Res Notes.* 2010;3:3.
- Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One.* 2012;7(2):e30619.
- Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics.* 2011;27:863–4.

Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome

Digvijay Verma and Tulasi Satyanarayana
Department of Microbiology, University of Delhi, New Delhi, India

Synonyms

Community genomics; Culture-independent approach; Environmental genomics; Endoxylanase; Endo- β -1,4 xylanase; Thermo-alkali-stable xylanase; Xylanase

Definition

For retrieving genes encoding thermo-alkali-stable xylanases by culture-independent (metagenomic) approach, the DNA extracted from hot and alkaline environmental samples is

restricted and the fragments are cloned. The clones are screened to select colonies with the desired xylanase gene, and the insert is sequenced and the gene is subcloned and expressed. The recombinant xylanase is purified and characterized and tested for its applicability in generating xylo-oligosaccharides from agro-residues and pulp bleaching.

Introduction

Hemicellulosic components are integral part of lignocellulosic residues and the second most abundant renewable polymer of plant cell walls after cellulose. Xylan is the main constituent in hemicelluloses of lignocellulosic agro-residues. β -1,4-linked xylosyl residues form the backbone of xylan that makes it a homopolysaccharide. Since xylan contains several groups such as arabinosyl, acetyl, and glucuronosyl residues that are present in the side chains, xylans are heteropolysaccharides (Hori and Elbein 1985; Coughlan and Hazlewood 1993). Heteropolymeric xylan requires synergistic action of multiple xylanolytic enzymes for complete degradation. The complex xylanolytic system includes endoxylanase (1,4- β -D-xylan xylanohydrolase; EC 3.2.1.8), β -xylosidase (1,4 β -D-xylan xylohydrolase; EC 3.2.1.37), α -glucuronidase, α -L-arabinofuranosidase, and acetyl xylan esterase. The CAZY database (http://www.cazy.org/fam/acc_GH.html) classified xylanases into six glycosyl hydrolase families GH5, GH8, GH10, GH11, GH30, and GH43 (Collins et al. 2005). Family 10 and 11 xylanases are however widely distributed in nature. Owing to low molecular weight and substrate stringency, family 11 xylanases are considered as true xylanases, while GH10 xylanases share broad substrate specificity with higher molecular weight.

Xylanases have successfully been used in various industries like ramie fiber degumming, food processing, and textile, biofuels, feed, and paper/pulp industries. However, xylanases must be alkalistable and thermostable to withstand the

extreme conditions prevailing in the paper industries in the pre-bleaching of kraft pulp. Although several xylanases have been reported from a large number of microorganisms, most of them do not have adequate thermostability and alkalistability for their utility in paper and pulp industries. Majority of xylanases have been obtained from the culturable 0.1–1 % of the total microbial diversity existing in natural environments. The culture-independent metagenomic approaches permit retrieval of genes encoding useful enzymes from environmental samples without involving laborious and elaborate methods of cultivation of microbes. The immense demand for alkalistable and thermostable xylanases encouraged us to adapt this innovative strategy for retrieving genes that encode thermo-alkali-stable xylanases from environmental metagenomes.

In this investigation, a metagenomic library was constructed and screened for clones with xylanase activity. Xylanase-encoding gene (*Mxyl*) (accession no. AFP81696) was subcloned and expressed, and the recombinant xylanase was purified and characterized. To the best of our knowledge, this is the first report on retrieving thermo-alkali-stable GH 11 family xylanase by a metagenomic approach.

Methodology

Collection of Samples and Construction of Metagenomic Library

The samples of compost soil were collected in sterile polyethylene bags from the vicinity of a hot water spring near Fukuoka Japan and stored at 4 °C. The pH of the samples is in the acidic range (3.0–4.5). Soil DNA was extracted according to Verma and Satynarayana (2011). Metagenomic DNA was processed for constructing the metagenomic library. Five µg of metagenomic DNA was partially digested with 0.5 U of restriction enzyme *Sau3AI*. The fragments of 3–12 kb were eluted from agarose gel (1.2 %, w/v) by gel extraction kit according to manufacturer's protocol (Macherey-Nagel,

Germany). Hundred nanogram of insert DNA and 300 ng of *Bam* HI digested and dephosphorylated p18GFP vector were ligated by using T4 DNA ligase overnight at 16 °C. The ligation mixture was transformed into competent *E. coli* DH10B cells by heat shock method. The metagenomic library was spread and screened for xylanase activity on 0.3 % (w/v) RBB-xylan (4-O-methyl-D-glucurono-D-xylan-remazol brilliant blue R) (Sigma, St. Louis, MO, USA) LB-ampicillin agar plates. The transformants were grown at 37 °C overnight and observed for the zone of xylan hydrolysis.

Screening for Xylanase and Sequence Analysis

The pure clone (TSDV-MX1) showing clear zone of xylanase hydrolysis was sequenced using M13 forward and reverse primers followed by different internal primers using Applied Biosystem 373 stretch automated sequencer (Applied Biosystems, Foster City, CA, USA) at Nucleic Acid Sequencing Facility of the University of Delhi South Campus, New Delhi (India), for obtaining full sequence of the insert. The ORFs were identified by using the NCBI's open reading frame (ORF) finder tool (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). BLASTN and BLASTP of NCBI were used to align the nucleotide and amino acid sequences, respectively. Multiple alignments of the amino acids were carried out using the CLUSTALW program (<http://www.ebi.ac.uk/clustalW>). The phylogenetic analysis was done using MEGA 2.1 with neighbor-joining strategy.

Construction and Expression of Plasmids *pET28a-Mxyl* and *pET22b-Mxyl*

The xylanase gene was amplified and ligated into the digested vectors followed by transformation into competent *E. coli* XL1 blue cells to obtain *pET28-Mxyl* and *pET22-Mxyl*. The recombinant constructs were confirmed by colony PCR followed by double digestion of the construct with restriction enzymes. The clones having xylanase gene were transformed into *E. coli* BL21(DE3) and processed for sequencing.

The recombinant plasmid having the accurate sequence was then transformed into *E. coli* BL21 (DE3) competent cells for the expression of recombinant proteins from *pET28a-Mxyl* and *pET22b-Mxyl*. The expression was induced by adding isopropyl- β -D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and the culture was further cultivated at 30 °C. The samples were collected at 1 h intervals for determining the enzyme titers. Localization of the recombinant protein was determined by collecting the intracellular, extracellular, and periplasmic fractions from the cells followed by assay for xylanase (Verma and Satyanarayana 2012).

Site-Directed Mutagenesis

Multiple sequence alignment of recombinant xylanase with the known xylanases revealed Glu₁₁₇ and Glu₂₀₉ to be catalytically important residues. Experimentally it has been proved by site-directed mutagenesis using GeneArt site-directed mutagenesis kit (Invitrogen, Carsband, USA). Two point mutations (Glu₁₁₇Asp and Glu₂₀₉Asp) were created in the metagenomic xylanase gene and expressed in *E. coli* BL21(DE3) cells. The induced mutations were confirmed by sequencing.

Xylanase Assay

Xylanase was assayed according to Archana and Satyanarayana (1997) at 80 °C and pH 9.0. One unit of xylanase is defined as the amount of enzyme required to liberate 1 μ mole of reducing sugar as xylose ml⁻¹ min⁻¹ under the assay conditions.

Purification and Biochemical Characterization of rMxyl

The rMxyl was purified by affinity chromatography using Ni²⁺-NTA agarose (Novagen, Germany) (Verma and Satyanarayana 2012). The characteristics of the recombinant xylanase like the effect of pH, temperature, metal ions, inhibitors and detergents on enzyme activity, thermostability, and substrate specificity have been studied. Kinetic properties of the recombinant

enzyme (K_m and V_{max}) on different xylans from birchwood, beech wood, and oat spelt were calculated from Lineweaver-Burk double reciprocal plots.

Saccharification of Agro-residues/Hydrolysis of Xylan

One percent (w/v) standard xylo-oligosaccharides (X2–X6) and agro-residues (wheat bran, corncobs, and sugarcane bagasse) were treated with recombinant xylanase (10 U–20 U/g) to find out the hydrolysis of XOs and lignocellulosic substrates. All the substrates (wheat bran, corncobs, and sugarcane bagasse) were suspended in glycine-NaOH buffer (pH 9.0) and incubated at 80 °C. Aliquots at the desired intervals were collected and analyzed on silica-based TLC plates (Merck, Germany) to determine the hydrolysis products. The saccharification of agro-residues was determined using DNSA reagent (Miller 1959).

Results

Construction of metagenomic library, DNA sequencing, and bioinformatics analysis.

When 5.0 μ g of high molecular weight (20–30 kb) metagenomic DNA was digested with Sau3AI and the fragments were ligated into p18GFP vector with an efficiency of 3.6×10^4 clones per μ g of DNA in constructing the library, the insert sizes were in the range of 3.0–8.0 kb with an average size of 5.5 kb. On screening, a clone having xylanase gene was spotted on RBB xylan containing LB-amp plate. The full sequence of the insert showed the size of 6.231 kbp that revealed its prokaryotic origin on blast analysis. The complete insert contained nine transcriptional units with a complete ORF of 1,077 bp long xylanase gene. The sequence showed putative sequences of –35 (CACGCCA), –10 (TAAAAA), and ribosomal binding sites (AGGGG) at the upstream of xylanase gene followed by complete ORF having ATG and TAA as start and stop codons, respectively (Fig. 1). The xylanase displayed five conserved regions (I–V) of GH11 xylanase having two

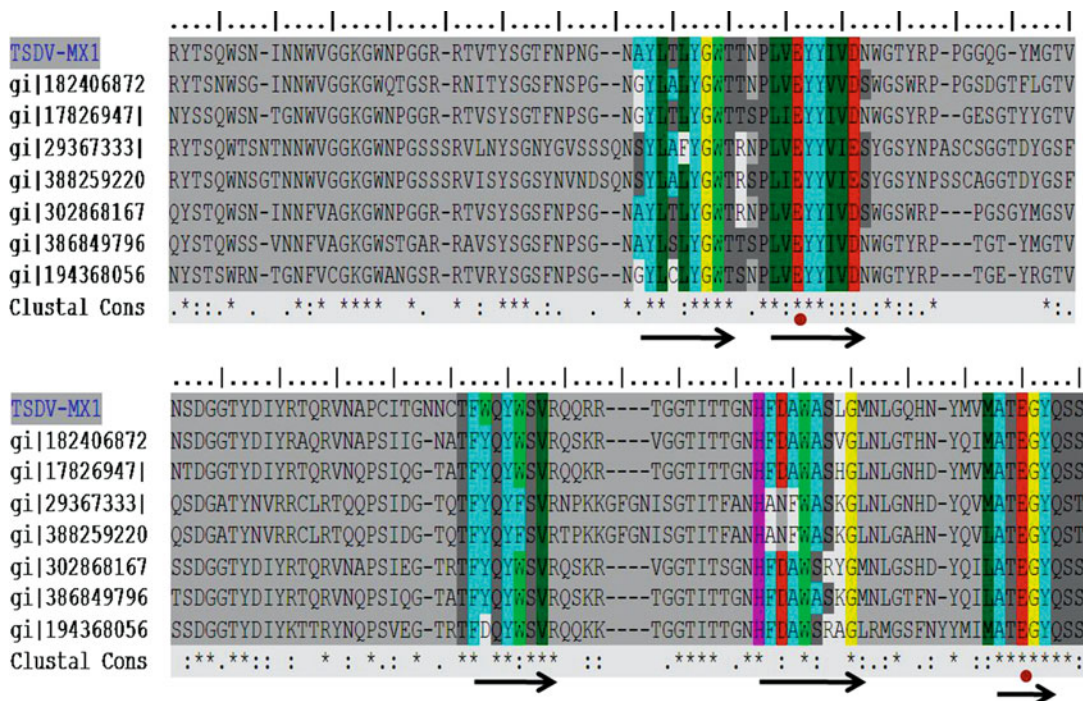

```

tctcatttctgctccatcactggctgacacgccagcaccaagcggc
                                     -35
gatgcgcatcttggtgaccc gtcacaaccaaaaaggggaagcttc
                                     -10   RBS
1  atgacagcgagtttgaggaagatcgcggttcacgagtaagagtgtc
  M T A S L R K I A F T S K S V
46 gctgcggaatcgtcggcatggccgctgtacgtgtcccctgcc
  A A A I V G M A A L Y V S P A
91 gatgcacagtgcctcaccaacaatcaaacggcactcagcagcggc
  D A Q C L T N N Q T G T H D G
136 tactactactcgttctggaaggacagcggcaacgtcagttctgc
  Y Y Y S F W K D S G N V T F C
181 ttgcagagcggcggccgatacacgtctcagtgagcaatatcaac
  L Q S G G R Y T S Q W S N I N
226 aactgggtcggcggcaagggttggatcccggcggacgacgcacg
  N W V G G K G W N P G G R R T
271 gtcacctactcggcagctcaatccgaatggcaacgcatactg
  V T Y S G T F N P N G N A Y L
316 acgctgtacggatggacgacgaatccgctcgtcagtgactacatc
  T L Y G W T T N P L V E Y Y I
361 gtcgataactggggcacctatcgcccggcggcgaaggctac
  V D N W G T Y R P P G G Q G Y
406 atgggcacgggtcaacagcagcggcggtacgtacgacatctatcgt
  M G T V N S D G G T Y D I Y R
451 acgcagcgcgtgaatgcgcggtgcatcaccggcaacaactgcacg
  T Q R V N A P C I T G N N C T
496 ttctggcagtagtgagcgttcgccagcagagaaggaccggcggc
  F W Q Y W S V R Q Q R R T G G
541 acgatcacgaccggcaaccacttcgacgcggtgggccagcctcggc
  T I T T G N H F D A W A S L G
586 atgaacctcggccaacacaactacatggtgatggcagcagggtt
  M N L G Q H N Y M V M A T E G
631 tatcagagcagcggcagctccgacatcacctggggcggcaccagc
  Y Q S S G S S D I T V G G T S
676 agcggcggcagcagcagcggaggcagcagcagcagcagcagcagc
  S G S S S S G G S S S S S S S S
721 agcagcagcggcgggtggcggcagcaagacgatcgtggtcggcggc
  S S S G G G S K T I V V R A
766 cgcggctccaccggcggcagcagatcagcctgcgcgtgaacaac
  R G S T G G E Q I S L R V N N
811 cagaccgtacagaactggacgctgggcacggcatgcagaactac
  Q T V Q N W T L G T G M Q N Y
856 acggccacgaccaacctgagcggcggcatcacctgctcacttcacg
  T A T T N L S G G I T V H F T
901 aacgataacggcggcggcgtgacgttcaggtggattacatccaggtg
  N D N G A R D V Q V D Y I Q V
946 aacggccagattcgtcaatccgagcagcagagctacaacacgggc
  N G Q I R Q S E Q Q S Y N T G
991 ttgtatgccaacggcggcgttggcggcggcggcgtggtatagcgagtg
  L Y A N G R C G G G G Y S E W
1036 atgcactgcaacggcggcctcggctacggaaacacgcgctaa 1077
  M H C N G A I G Y G N T P *

```

Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Fig. 1 Deduced amino acid sequence of recombinant xylanase (rMyl) and its nucleotide sequence. The red underlined region is leader sequence;

cyan-highlighted regions represent GH11 catalytic domain. Gray-highlighted regions are compositionally biased regions that were not used in database search and proposed as linker regions. Bluish-green-highlighted region depicts substrate binding domain



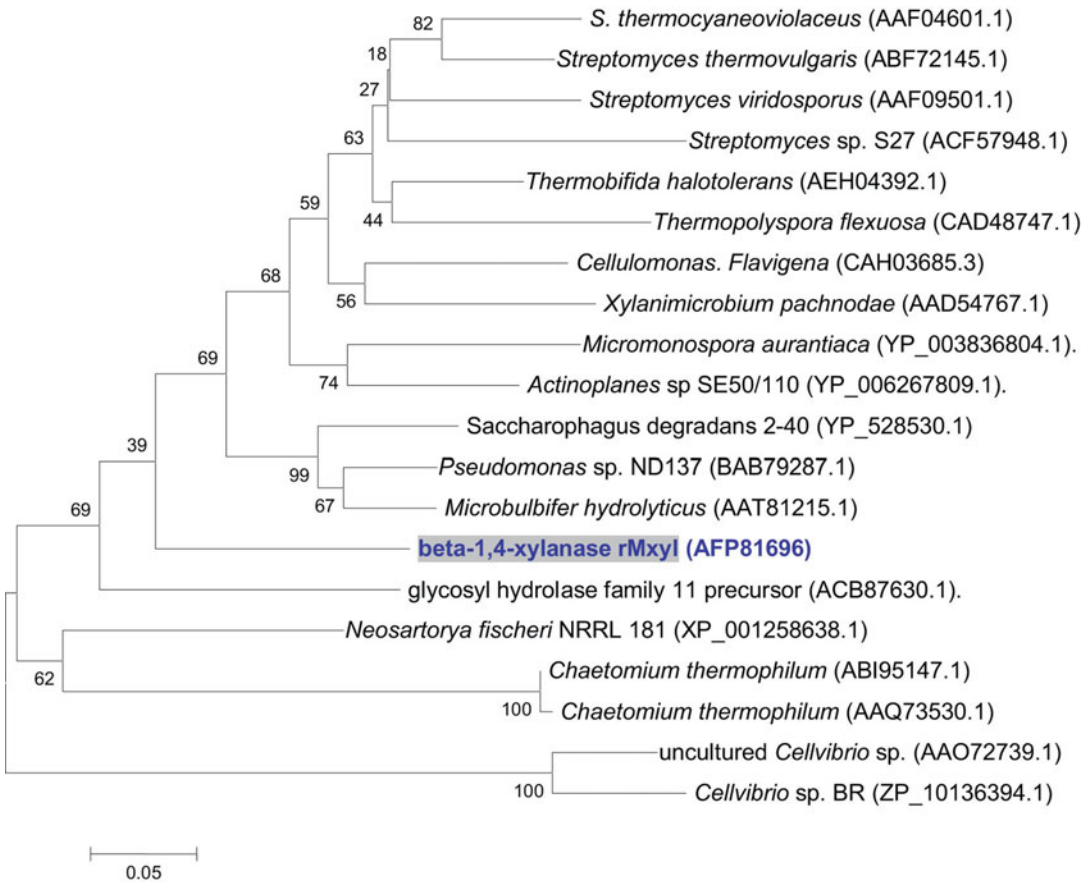
Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Fig. 2 Multiple sequence alignment of xylanase with other xylanases available in database. GenBank accession number and source of microorganisms were given as follows: 182406872 (glycosyl hydrolase family 11 precursor [uncultured bacterium]), 17826947 (*Pseudomonas* sp. ND137), 29367333 (uncultured *Cellvibrio* sp.), 388259220 (*Cellvibrio*

sp. BR), 302868167 (*Micromonospora aurantiaca* ATCC 27029), 386849796 (*Actinoplanes* sp. SE50/110), 194368056 (*Streptomyces* sp. S27). Five signature sequences: **I** (AYLTLYGW), **II** (VEYYIVDN), **III** (FWQYWSV), **IV** (HFDASWASLG), and **V** (MATEGYQSS) of GH11 family are colored. Two catalytically important residues (**Glu 117** and **Glu 209**) are marked with black circle

catalytically important residues (Glu₁₀₉ and Glu₂₁₇) present in signature sequence II and V (Fig. 2). Amino acid homology showed maximum identity (79 %) with the xylanase gene of an uncultured bacterium and *Actinoplanes* sp. SE50/110 followed by a metagenomic GH11 xylanase (71 %). It shared 63–75 % homology with xylanases produced by *Streptomyces* spp. The xylanase retrieved in this investigation exhibits 75, 67, and 64 % similarity with the endo-1,4 β -xylanases of *Cellulomonas fimi*, *Micromonospora aurantiaca* 27029, and *Amycolatopsis mediterranei* U32, respectively. It, however, has lower homology with the xylanases of *Microbulbifer hydrolyticus* (63 %), *Pseudomonas* sp. ND137 (62 %), uncultured *Cellvibrio* sp. (58 %), *Cellvibrio mixtus* (57 %), and *Aspergillus fumigatus* AF293 (52 %) (Fig. 3).

Expression of the Xylanase Gene in *E. coli* and Localization of the Encoding Recombinant Xylanase (rMxyl)

Xylanase gene was successfully cloned into pET28a and pET22b vectors. The recombinant plasmids were expressed in *E. coli* BL21(DE3) on induction with 1.0 mM IPTG at A₆₀₀ of 0.6–0.7 and 30 °C. At higher level of expression, it led to the formation of inclusion bodies, which could be solubilized using 6.0 M urea. The highest titer of the recombinant enzyme was achieved in 4–6 h. The construct (*pET28a-Mxyl*) expressed a high proportion of xylanase in cytoplasmic fraction (83 %), followed by periplasmic (9 %) and extracellular (8 %) fractions after 4–5 h of induction. When xylanase gene was cloned and expressed in pET22b(+) vector, a high proportion



Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Fig. 3 Phylogenetic tree of recombinant xylanase. rMxyl showed highest homology with xylanase of *Cellulomonas fimi* ATCC 484 followed by uncultured

microbial GH 11 xylanase. Neighbor-joining (NJ) tree is constructed by using MEGA 4.0 software. Bootstrap values ($n = 1,000$ replicates) are represented as percentage. The scale bar depicts the allowed changes per amino acid position

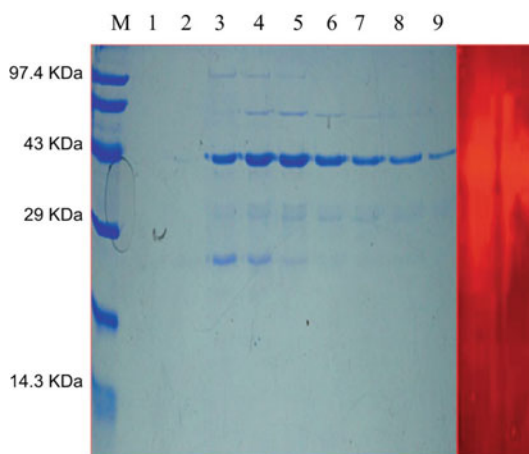
of intracellular enzyme (>60 %) was produced in the initial 3 h of induction, and thereafter, it declined. The periplasmic xylanase was optimum at 12 h, while the extracellular fraction gradually increased and it reached a peak (29 %) in 24 h.

Site-Directed Mutagenesis

Muteins having Glu₁₁₇Asp and Glu₂₀₉Asp completely lost the activity. These two glutamates are highly conserved residues in the signature sequences LVEYYIVDN and MATEGY, and these are responsible for catalytic activity of GH 11 xylanase.

Purification, Biochemical Characterization, and Zymogram Analysis of rMxyl

The recombinant xylanase was purified by Ni²⁺-NTA resin affinity chromatography and the purified recombinant protein could be eluted using imidazole (100–400 mM). The protein appeared as a single band of 40 kDa against the protein markers on 15 % SDS-PAGE, and the recombinant xylanase revealed as a clear band of xylan hydrolysis by zymogram analysis (Fig. 4). The xylanase exhibited broad range of pH (6.0–12.0) with optimum at 9.0, and it retained ~55 % residual activity at pH 10.0 (Fig. 5a).



Novel Alkalistable and Thermostable Xylanase-Encoding Gene (*Mxyl*) Retrieved from Compost-Soil Metagenome, Fig. 4 Analysis of rMxyl using SDS-PAGE (15 % polyacrylamide gel). (a). Lane 1 protein marker, Lane 2 and 3 are washes with 20 and 30 mM imidazole. Recombinant xylanase was eluted using different concentrations of imidazole (100, 200, 250, 300, 400, 450, 500 mM). Purified xylanase showed molecular mass of ~42 kDa on staining with Coomassie Brilliant Blue R-250. (b). Zymogram analysis of purified xylanase using Congo red staining method

The rMxyl is active in the temperature range between 40 °C and 100 °C (Fig. 5b) with optimum at 80 °C and retains more than 90–95 % activity after exposure to 60 °C and 70 °C for 3 h. The enzyme has a $T_{1/2}$ of 2.0 h at 80 °C and 15 min at 90 °C (Fig. 5c). The recombinant enzyme did not lose activity after 3 h exposure to pH 8.0 and 9.0, and thereafter, it declined (50 % residual activity after 4 h). Approximately 20–45 % loss in activity was recorded on either side of the pH optimum after 1 h incubation (Fig. 5d). Mg^{2+} , Sn^{2+} , and Fe^{2+} stimulated rMxyl activity, while Hg^{2+} and Mn^{2+} strongly inhibited enzyme activity even at 1 mM. Other metal ions exerted varied inhibitory action on xylanase. More than 30 % activity was lost in the presence of Mn^{2+} (Table 1). NBS and PMSF inhibited the activity to a significant extent even at 1 mM concentration. β -ME and DTT strongly inhibited enzyme activity. A stimulatory effect EDTA was recorded on xylanase activity.

Most of the metal ions did not affect enzyme activity at 1 mM concentration. Xylanase

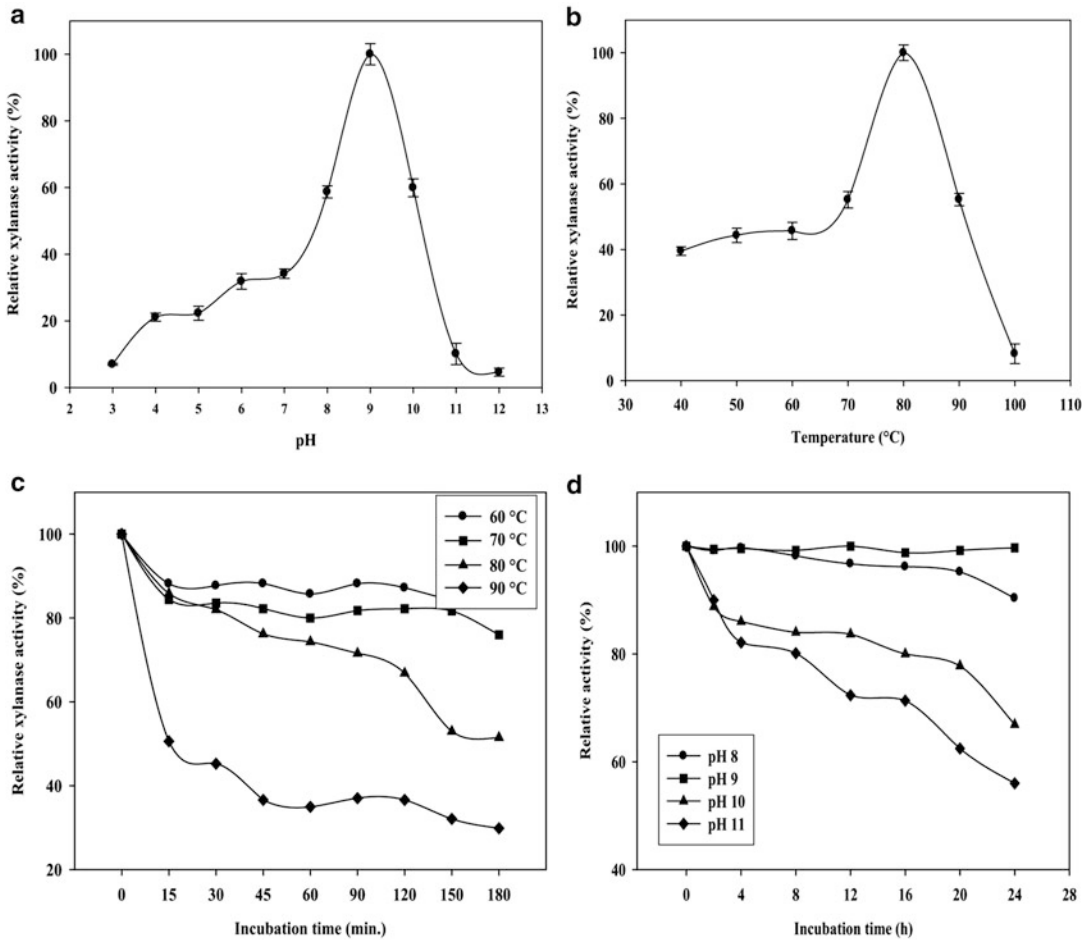
activity was, however, significantly inhibited at higher concentration by Pb^{2+} , Ag^{2+} , Ca^{2+} , Mn^{2+} , Ba^{2+} , Cd^{2+} , and Co^{2+} . In the presence of Hg^{2+} , enzyme lost activity completely. Similarly, trace amounts of β -mercaptoethanol (β -ME) and dithiothreitol (DTT) completely inhibited the xylanase activity. Inhibition in the presence of N-bromosuccinimide (NBS) signifies the role of tryptophan in catalysis, while EDTA confirms it as a non-metalloenzyme.

Saccharification of Agro-residues/ Hydrolysis of Xylan

The rMxyl hydrolyzed xylan from various sources. The enzyme activity was very high in birchwood xylan (relative activity 100 %) in comparison with that on xylan from beech wood (97 %) and arabinoxylan (80 %). There was no activity on carboxymethylcellulose (CMC) and other non-xylan polysaccharides (starch, pullulan, and chitin). The K_m and V_{max} values of the enzyme on birchwood xylan are 8.0 ± 1.21 mg/ml and 300 ± 09.12 μ mol/min/mg, respectively. The saccharification of wheat bran was high (15.2 %) as compared to that of corncobs (9.89 %) and sugarcane bagasse (4.71 %). Various xylo-oligosaccharides were detected in the hydrolysates (Fig. 6).

Discussion

Although several xylanases have been reported from diverse microbiota using traditional culture-dependent approaches, majority of them do not endure the extreme temperature and alkaline conditions prevailing in industrial processes. An alternate strategy was, therefore, adapted to retrieve a thermo-alkali-stable xylanase gene (*Mxyl*) by culture-independent metagenomic approach. The metagenomic library constructed with the DNA extracted from the compost-soil samples yielded a clone that produced xylanase. Although, the compost soils are in the acidic pH range, an alkalistable and thermostable endoglucanase had been reported from rice straw compost (Son-Ng et al. 2009).



Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Fig. 5 Effect of pH and temperature on the activity and stability of rMxyl. (a and b) The recombinant xylanase incubated in various buffers (pH 3–12) and temperatures (40–100 °C) and assayed for xylanase activity. (c) Recombinant xylanase was incubated in

glycine-NaOH buffer without substrate and kept at various temperatures. Aliquots were collected at various time interval and store at 0 °C for calculating residual activity. (d) Similarly enzyme was incubated in various buffers (pH 8–11) and aliquots of different time intervals were used xylanase assays

The culture-independent approach has started yielding the useful biocatalysts from the hidden Pandora’s Box of non-culturable microbial diversity. The protein encoded by xylanase gene comprises 358 amino acids, of which 16 are acidic and 21 basic. The predicted molecular weight, pI, and instability index of recombinant xylanase are ~40 kDa, 8.8, and 33.44 respectively. The xylanase contained a 43-amino-acid-long leader sequence at the N-terminal region followed by a catalytic domain (44th–212th) of GH11 family interrupted

by a short stretch of arginine- and threonine-rich non-catalytic region (WSVRQ₂R₂TG₂TIT₂). In addition, serine-rich Q linker region (S₂GS₂DITVG₂TS₂G₂TS₂G₂S₃G₂S₁₀G₄) has also been detected from amino acid 213 to 248 just after catalytic domain. Such repeated amino acids make linker regions that usually discriminate catalytic domain from carbohydrate-binding domain (Gilkes et al. 1991). Moreover, linkers have also been reported as integral parts of various xylanases that connect thermo-stabilizing domains, surface

Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Table 1 Effect of modulators on rMxyl activity

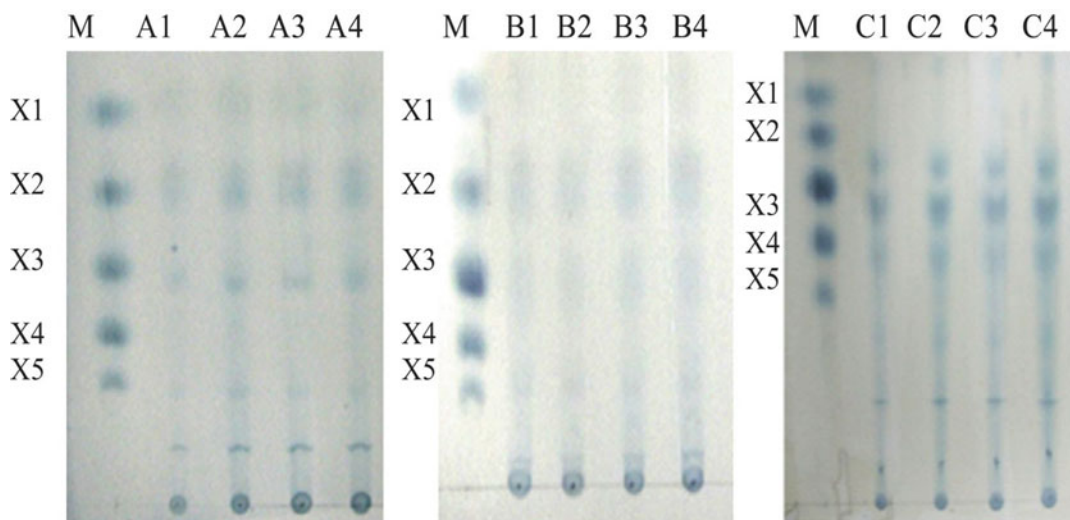
Metal ions	1 mM	5 mM	10 mM
Mg ²⁺	106.45 ± 1.05	99.65 ± 0.98	87.38 ± 0.45
Fe ²⁺	108.65 ± 0.75	116.01 ± 0.27	93.67 ± 1.32
Sn ²⁺	110.43 ± 0.67	76.12 ± 0.44	45.17 ± 0.63
Ni ²⁺	91.21 ± 0.22	79.01 ± 1.34	32.84 ± 0.43
Zn ²⁺	91.67 ± 0.32	76.64 ± 0.78	32.89 ± 0.89
Pb ²⁺	81.33 ± 0.67	20.78 ± 0.32	09.65 ± 0.67
K ⁺	81.21 ± 1.08	20.62 ± 0.12	12.67 ± 0.45
Ag ²⁺	73.48 ± 0.53	54.55 ± 0.69	27.83 ± 0.98
Ca ²⁺	72.43 ± 0.43	35.45 ± 0.21	12.09 ± 0.19
Mn ²⁺	71.76 ± 0.63	27.34 ± 1.32	09.67 ± 0.27
Ba ²⁺	66.45 ± 0.67	23.91 ± 0.34	18.65 ± 0.33
Cd ²⁺	54.67 ± 0.43	29.33 ± 0.49	12.87 ± 0.65
Co ²⁺	59.15 ± 1.23	29.63 ± 0.65	12.54 ± 1.12
Na ⁺	61.43 ± 0.78	39.75 ± 1.06	27.35 ± 0.78
Cu ²⁺	29.12 ± 0.18	15.76 ± 0.76	10.09 ± 0.87
Hg ²⁺	0	0	0
Inhibitors	1 mM	5 mM	10 mM
NBS	46.66 ± 0.12	35.67 ± 0.09	20.12 ± 0.11
IAA	103.45 ± 0.54	89.75 ± 0.32	69.85 ± 1.56
β-ME	0	0	0
DTT	0	0	0
EDTA	105.65 ± 1.23	107.19 ± 1.01	89.98 ± 0.56
Detergents	0.1 % (v/v)	0.5 % (v/v)	
Tween 20	103.45 ± 1.32	105.67 ± 0.98	
Triton X100	108.32 ± 0.96	104.05 ± 0.92	
SDS	97.34 ± 1.32	65.89 ± 0.19	
Control	100 ± 0.12	100 ± 0.23	100 ± 0.67

layer homology domains, and dockerin domains which play a role in stabilizing the protein. Amino acid homology and hydrophobic cluster analysis categorized this high molecular weight xylanase into GH11 family. Metagenomic origin, distinct characteristics, lower homology, and higher molecular weight (>30 kDa) make this a novel xylanase. The integrated N-terminal pelb signal sequence in pET22b(+) directed the enzyme to periplasm that further led to secretion into the extracellular environment.

The site-directed mutagenesis of two residues of glutamate to aspartate resulted in a complete loss of xylanase activity due to disruption in double-displacement mechanism. In order to take the advantage of thermostability of the

recombinant xylanase, it was subjected to high temperature prior to purification by Ni²⁺-NTA agarose resins. This step reduced the extra load of non-His-tagged, less thermostable, and contaminant host proteins (Mamo et al. 2006; Verma and Satyanarayana 2012).

The rMxyl exhibits optimum activity at higher temperature (80 °C) and pH (9.0) which is similar to xylanases produced by *Dictyoglomus thermolacticum*, *Thermotoga maritima*, *Bacillus stearothermophilus*, and *Geobacillus thermoleovorans* having optimal activity at or above 80 °C (Uchino and Fukuda 1983; Mathrani and Ahring 1992; Khasin et al. 1993; Verma and Satyanarayana 2012). The activity and stability of rMxyl at higher pH are the crucial properties of



Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome, Fig. 6 Profile of xylo-oligosaccharides liberated by the action of rMxyl. Lane (A1–A4)*: spots of X1, X2, and X3 were detected from wheat bran. Lane (B1–B4)*: hydrolysate from corncobs showed prominently

X2 and X3. While X3, X4, and X5 were detected from hydrolysate of sugarcane bagasse (C1–C4)*. Lane M: standards of various XOs. X1 xylose, X2 xylobiose, X3 xylotriose, X4 xyloptetraose, X5 xylopentaose. *: 1/2/3/4 time intervals of 5, 15, and 30 min and 1 h, respectively

xylanases for their applicability in paper processing industry. The shelf-life of rMxyl is more than 3 months at 4 °C, which retains greater than 90 % activity. The recombinant xylanase is optimally active at 80 °C and pH 9.0 that distinguishes it from already reported xylanases. The xylanase of *Thermotoga maritima* has T_{opt} of 90 °C, but it gets inactivated fast at pH 6.0 (Yoon et al. 2004). Similarly the alkalistability at higher pH is reported in many xylanases but are active at lower temperatures (Khasin et al. 1993). The recombinant xylanase of GH10 family from *Bacillus halodurans* showed both properties together having optima at 75 °C and pH 9.0, but it losses 50 % activity at 65 °C after 4 h and gets inactivated very fast at 80 °C (Mamo et al. 2006). The metagenomic xylanase, on the other hand, has good thermostability at higher temperatures (60 °C, 70 °C and 80 °C) with only 20–30 % loss after 3 h exposure. The most significant aspect of this investigation is obtaining a highly alkalistable (pH_{opt} . 9.0) and thermostable (T_{opt} . 80 °C) xylanase from environmental samples by a metagenomic approach.

Cations (Mg^{2+} , Sn^{2+} , and Fe^{2+}) stimulated the rMxyl activity while 1 mM, Hg^{2+} , and Mn^{2+} significantly inhibited the activity. The inhibition of xylanase by Hg^{2+} suggests the presence of tryptophan residues that oxidize indole ring, thereby inhibiting the xylanase activity. The inhibition of xylanase activity by Cu^{2+} is similar to the majority of the xylanases (Matteotti et al. 2012). In *Glaciecola mesophila* KMM 241, EDTA caused ~25 % enhancement in activity (Guo et al. 2009). NBS inhibition suggests the involvement of tryptophan in xylanase activity. Total loss of xylanase activity by β -ME and DTT suggests the distortion of disulfide linkages present between cysteine residues (Maalej et al. 2009; Matteotti et al. 2012). Detergents exerted a slight stimulatory effect on the recombinant xylanase which is a common feature of the other xylanases. However, rMxyl was inhibited by SDS.

The rMxyl hydrolyzed birch wood and beech wood xylans efficiently. The structural similarity of beech wood and birch wood xylans may be the reason for the high activity. The enzyme exhibited almost similar activities on oat spelt

xylan and arabinoxylan. Oat spelt xylan is a type of arabinoxylan very rich in arabinose (xylose/arabinose = 66:34) (Gruppen et al. 1992; Kormelink and Voragen 1993). Interestingly the rMxyl liberated xylo-oligosaccharides from xylan in just 5 min and it was sustainable on prolonged incubation. Several xylanases have been reported from various microorganisms that liberate xylo-oligosaccharides following xylan hydrolysis. Alkaline xylanases show better action on agro-residues by lowering the steric hindrance caused by cellulose and enhancing the solubility of hemicellulosic materials (Gruppen et al. 1992). The metagenomic xylanase finds application in food industry for the production of xylo-oligosaccharides as prebiotics (Vazquez et al. 2000).

Summary

Most of the xylanases retrieved by culture-dependent and culture-independent approaches exhibit optimal activity in the pH and temperature ranges of 6.0–8.0 and 40–60 °C, respectively. The xylanase (rMxyl) obtained in this investigation through metagenomic approach displays alkalistability as well as thermostability. This is the first report on the xylanase with twin stabilities obtained through a culture-independent approach. A very low similarity in amino acid sequence of the enzyme with other known xylanases makes it a novel xylanase. The possibility of obtaining thermo-alkali-stable xylanase from composts may lead to an intense search for similar enzymes in this and other related niches.

References

- Archana A, Satyanarayana T. Xylanase production by thermophilic *Bacillus licheniformis* A99 in solid-state fermentation. *Enzyme Microb Technol.* 1997; 21:12–7.
- Collins T, Gerday C, Feller G. Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiol Rev.* 2005;29:3–23.
- Coughlan MP, Hazlewood GP. β -1,4 D-xylan-degrading enzyme systems: biochemistry, molecular biology and applications. *Biotechnol Appl Biochem.* 1993;17: 259–89.
- Gilkes NR, Henrissat B, Kilburn DG, et al. Domains in microbial 4-glycanases: sequence conservation, function, and enzyme families. *Microbiol Rev.* 1991; 55:303–15.
- Gruppen H, Hamer RJ, Voragen AGJ. Water unextractable cell wall material from wheat flour. 2. Fractionation of alkali extracted polymers and comparison with water extractable arabinoxylans. *J Cereal Sci.* 1992;16:53–67.
- Guo B, Chen X, Sun C, et al. Gene cloning, expression and characterization of a new cold-active and salt tolerant endo- β -1, 4-xylanase from marine *Glaciecola* mesophila KMM 241. *Appl Microbiol Biotechnol.* 2009;84:1107–15.
- Hori H, Elbein AD. The biosynthesis of plant cell wall polysaccharides. In: Higuchi T, editor. *Biosynthesis and biodegradation of wood components.* Orlando: Academic; 1985. p. 109–35.
- Khasin A, Alchanati I, Shoham Y. Purification and characterization of a thermostable xylanase from *Bacillus stearothermophilus* T-6. *Appl Environ Microbiol.* 1993;59:1725–30.
- Kormelink FJM, Voragen AGJ. Degradation of different [(glucuron)arabino] xylans by a combination of purified xylan-degrading enzymes. *Appl Microbiol Biotechnol.* 1993;38:688–95.
- Maalej I, Belhaj I, Masmoudi NF, Belghith H. Highly thermostable xylanase of the thermophilic fungus *Talaromyces thermophilus*: purification and characterization. *Appl Biochem Biotechnol.* 2009; 158:200–12.
- Mamo G, Delgado O, Martinez A, et al. Cloning, sequencing analysis and expression of a gene encoding an endoxylanase from *Bacillus halodurans* S7. *Mol Biotechnol.* 2006;33:149–59.
- Mathrani IM, Ahring BK. Thermophilic and alkaliphilic xylanase from several *Dictyoglomus* isolates. *Appl Microbiol Biotechnol.* 1992;38:23–7.
- Matteotti C, Bauwens J, Brasseur C, et al. Identification and characterization of a new xylanase from gram-positive bacteria isolated from termite gut (*Reticulitermes santonensis*). *Protein Expr Purif.* 2012;83:117–27.
- Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugars. *Anal Chem.* 1959; 31:426–8.
- Son-Ng I, Li CW, Yeh Y, et al. A novel endoglucanase from the thermophilic bacterium *Geobacillus* sp. 70PC53 with high activity and stability over broad range temperatures. *Extremophiles.* 2009; 13:425–35.
- Uchino F, Fukuda O. Taxonomic characteristics of an acidophilic strain of *Bacillus* producing thermophilic acidophilic amylase and thermostable xylanase. *Agric Biol Chem.* 1983;47:965–7.

- Vazquez MJ, Alonso JL, Dominguez H, et al. Xylo-oligosaccharides: manufacture and applications. *Trends Food Sci Technol.* 2000;11:387–93.
- Verma D, Satyanarayana T. Cloning, expression and applicability of thermo-alkali-stable xylanase of *Geobacillus thermoleovorans* in generating xylo-oligosaccharides from agro-residues. *Bioresour Technol.* 2012;107:333–8.
- Verma D, Satyanarayana T. An improved protocol for DNA extraction from alkaline soil and sediment samples for constructing metagenomic libraries. *Appl Biochem Biotechnol.* 2011;165:454–64.
- Yoon HS, Han NS, Kim CH. Expression of *thermotoga maritima* endo-b-1, 4-xylanase gene in *E. coli* and characterization of the recombinant enzyme. *Agric. Chem. Biotechnol.* 2004;47:157–160.

Novel Approaches to Pathogen Discovery in Metagenomes

Jun Hang

Viral Diseases Branch, WRAIR, Silver Spring, MD, USA

Synonyms

Community genomics; Metagenomics and pathogen identification; Microbiome and virome; Pathogenomics

Definitions

Pathogen discovery: identification of causative microbial or viral agent(s) for an illness or asymptomatic infection. The identification may refer to etiological diagnosis for individuals, epidemiology investigation on population scale, and animal or environmental surveillance on orphan pathogens.

Metagenomics: genomic study on a population of biologically or functionally close microorganisms as a whole community, without separation of components into pure culture isolates.

Introduction

The best-known statement on pathogen discovery probably is the so-called Koch's postulates, in

which isolation of disease causative microbe and determination of its etiological features are of the essence (Falkow 2004; Lipkin 2010). There are fascinating and tragic stories in medical history of human volunteers or doctors who sacrificed health or even their lives to test pathogens on themselves to satisfy the postulates. The principles guided the development of clinical microbiology and remain the important guidelines, if not the rules, even in the era of molecular biology and genomics. Nevertheless, study has shown that the vast majority of microorganisms cannot be readily grown or are not cultivable at all (Handelsman 2004). It is also true for pathogens; in other words, there are numerous varieties of potential pathogens that exist and evolve in the environment; it is just a matter of time when and where they will emerge or reemerge to cause sporadic cases or outbreak. In addition, the manifestation of some diseases is contributable to coexistence of multiple organisms or imbalanced microbial community at host tissues.

Technique approach for pathogen diagnostics evolves along with scientific discovery and technology innovation on microbiology as well as other disciplines. A variety of techniques are used in clinical labs, including the traditional microbiology tests, rapid serological assays, and various molecular assays. They are well designed and validated with reliable sensitivity and specificity (Lipkin 2010). Many of them are automated for improved speed, convenience, and accuracy. However, in spite of the great effectiveness and robustness, threat from emerging pathogens remains real. In particular, because of the rising globalization and drastic climate changes, novel pathogens and new variant strains have more often appeared and spread. There are chances that a highly virulent pathogen may escape detection by conventional methods and can cause a widespread outbreak and public health crisis with dramatic economic loss and social consequences.

To answer the emergent challenge, novel approaches utilizing the advanced technologies have been developed to effectively identify pathogens as well as elucidate pathogenesis mechanism in comprehensive way (Lipkin 2010; Olsen et al. 2012). Metagenomics analyzes all genomic

information in a specifically defined population. The deep and comprehensive metagenomic information allows individual organisms of interest to be interrogated in the context of the whole community and with its phylogenetic relatives (Joseph and Read 2010). The significant strategy has transformed the way we perceive microbial world. The related laboratory and bioinformatics approaches were successfully used in identifying causal pathogens for outbreaks and providing vital insights into the source and/or evolutionary origins (Koser et al. 2012). Approaches based on rich knowledge from metagenomics are vigorously implemented to pathogen discovery and are believed to be clear path of future perspectives of the clinical diagnostics (Eisen and MacCallum 2009; Olsen et al. 2012).

Strategy and Schemes

Genomic approaches to detection of pathogen in clinical specimens are either based on known genomic information (sequence dependent) or designed to capture unique and disease-relevant as well as redundant and irrelevant sequences altogether (sequence independent) (Olsen et al. 2012). Metagenomics was initially developed in the era of Sanger sequencing (Fredericks and Relman 1996; Handelsman 2004) and truly thrived with the emerging of the next-generation sequencing (NGS) technologies which make DNA sequencing much less expensive and hugely productive (Petrosino et al. 2009). It is now feasible and affordable to either sequence a number of amplicons at exceedingly high depth to capture rare variants or sequence all DNA and/or RNA by design in a complex sample. NGS allows direct sequencing of microbial contents without microbiological cultivation for isolation and enrichment. Numerous molecular biology techniques for sample preparation prior to sequencing and bioinformatics tools for data mining and analyses were developed (Thomas et al. 2012). Experiment design and the choice of technical and

analytical approaches are vital for the sensitivity and accuracy for pathogen discovery.

16S Ribosomal RNA Gene Sequencing for Human Microbiota Assessment and Identification of Bacterial Pathogens

Bacterial 16S rRNA gene sequence has long been used to classify bacteria down to taxonomy levels of genus or lower. In contrast to amplification, cloning, and sequencing of full length 16S rRNA genes by Sanger method, NGS enables massive acquisition of a million or more 16S rRNA gene segment sequences in a single run to decipher bacterial composition (species richness and abundance) in a community (Kuczynski et al. 2012). Sequence across two to three variable regions has been suggested to contain taxonomic information unique enough for classification. Roche 454 pyrosequencing is currently the method of choice due to its relatively long read length and low sequence error rate. Read length average 300–500 bases for Roche GS FLX Titanium system and 500–800 bases for the recent FLX + system. FLX + application on amplicon sequencing is currently under development and yet to be validated for 16S sequencing which will achieve longer read length without comprising sequence quality. Different from genome sequencing in which reads are assembled by overlapping to obtain a consensus sequence, in 16S-based metagenomic analysis, 454 sequencing reads are classified individually, i.e., each read is one operational taxonomic unit (OTU). Therefore, high-performance sample preparation and sequencing procedures, stringent data processing, and analytical pipeline are critical for achieving and maintaining accuracy and sensitivity. Many studies to compare materials and methods for optimization have been published (Kuczynski et al. 2012). One significant open resource is the Data Analysis and Coordination Center (DACC) from the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP) and is available at website

<http://www.hmpdacc.org/>. The fundamental knowledge on healthy human microbial communities and the developed metagenomics techniques and analytic tools are being brought into the clinical arena with encouraging successes on making diagnoses of difficult diseases and complex outbreaks (Loman et al. 2012). Pathogenomics is showing its power and clinical importance by revealing genomics and metagenomics basis for complicated syndromes which cannot be explicitly understood with conventional clinical tests. In consequence, improved therapeutic practices, reduced medication costs, and more-informed disease prevention measures can result in dependable public health protection.

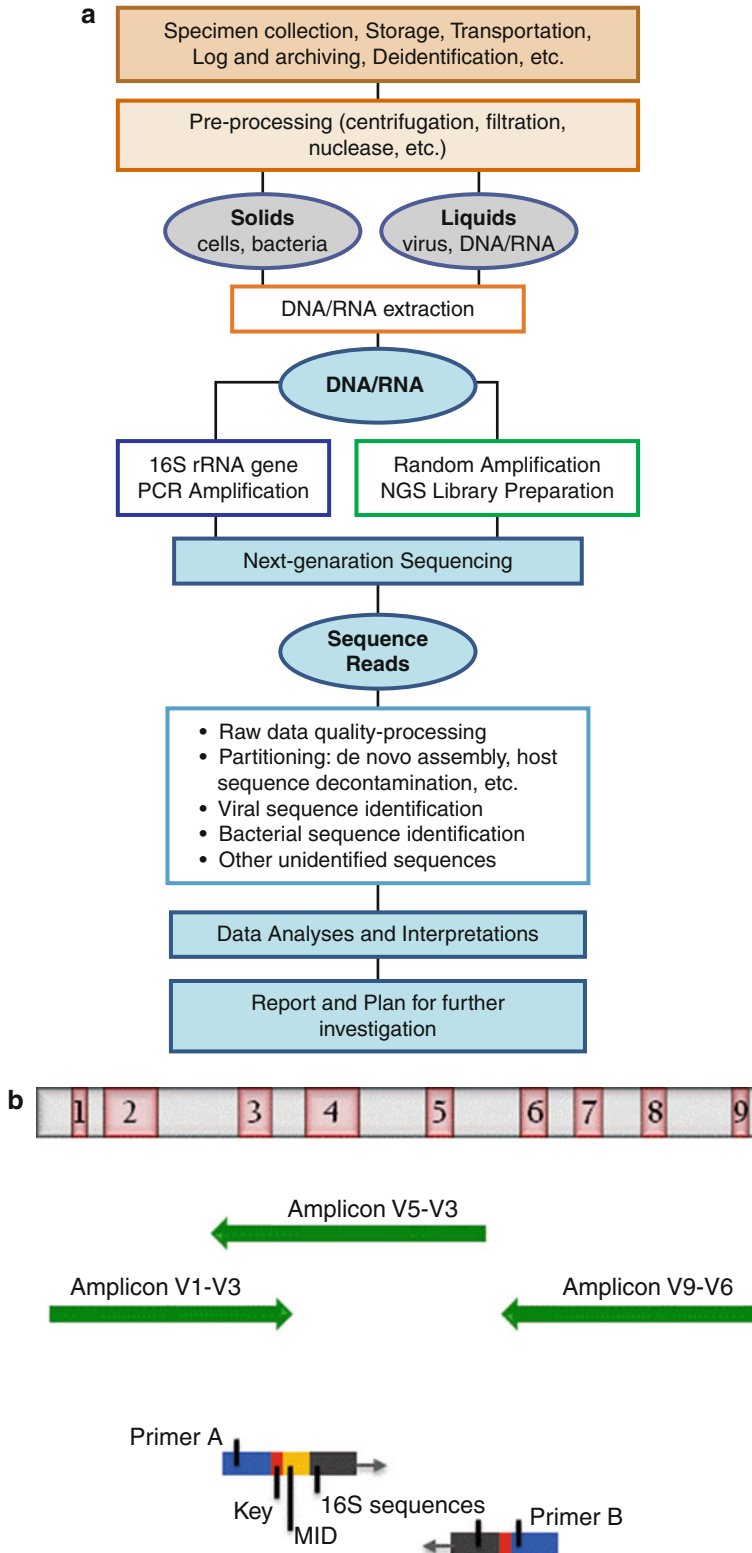
Microbial Metagenomics and Single-Cell Sequencing

There are considerable and ongoing efforts to characterize collective whole genomes in an entire community. For example, several research teams used Illumina's NGS technology and shotgun sequencing approach to generate several hundred gigabases of microbial sequences for extensive cataloging of genes in human gut microbes (Qin et al. 2010; Arumugam et al. 2011). From a number of studies, the depth and comprehensiveness of our knowledge on human microbiomes is unprecedented, and it would not be possible without having the advanced NGS technologies and the associated sophisticated bioinformatics tools (Kuczynski et al. 2012; Thomas et al. 2012). However, such a shotgun unselective metagenomic strategy requires tremendous computational power and may not be efficient or cost-effective enough for a routine pathogen diagnostics practice. There are multiple approaches that may facilitate the overcoming of the hurdles for the wide use of metagenomics in clinical settings. The HMP and other international programs aim to build a database of fully annotated complete genome sequences for bacteria of clinical and human health relevant. The high-quality reference genome database with rich and definitive genomic, genetic, functional,

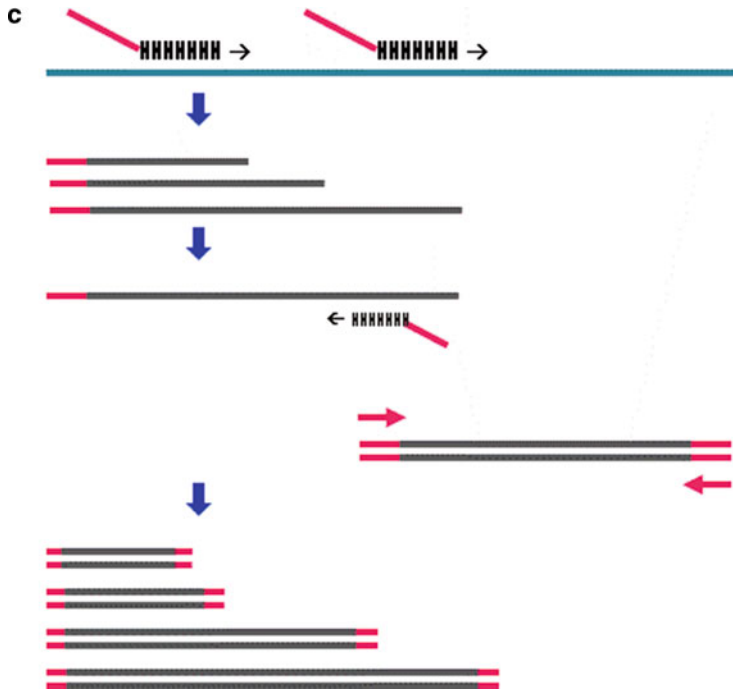
and phenotypic information will be the key to a metagenomics-based clinical test (Joseph and Read 2010). Other components essential to the feasibility include streamlined sample processing and sequencing system with automation, convenient data collection and management procedure, and efficient bioinformatics pipeline in concert with reference information for sequence analysis and amenable to integration with medical record and interactive communication to worldwide disease networks and specific study consortiums (Koser et al. 2012).

In addition to the promising clinical use of whole-genome metagenomics, the scientific and technical resources gaining from metagenomics quests have a multitude of utilities that can make existing pathogen discovery methods design and perform better (Fournier and Raoult 2011). For instance, with the comprehensive genomic information on the microbial community corresponding to the specimens, multilocus sequence typing (MLST), PCR-based molecular assays, microarray-based assays, etc. can be made more specific for the targets with reduced nonspecificity. Moreover, assay results can be interpreted with better estimation of probability of miss-calling and the false-positive, therefore concluded with increased confidence.

Another promising approach is single-cell genome sequencing for pathogen discovery. Individual microorganisms or parasites are physically isolated out of a complex community, i.e., clinical matrix, either under microscopy by morphology or using devices such as flow cytometry cell sorting. Both methods are well established and already routinely used in clinical laboratory. Harvested single cell or a homogenous pool of cells are then subjected to amplification and sequencing. Multiple displacement amplification (MDA) from a single cell has been shown robust and faithful for downstream sequencing and microarray applications. Studies showed 95 % or higher genome coverage by using single-cell genomic sequencing (Pallen et al. 2010). The culture-free approach coupled with lab-on-chip microfluidic cell harvesting and processing automation may make its way to become suitable for clinical diagnostic use.



Novel Approaches to Pathogen Discovery in Metagenomes, Fig. 1 (continued)



Novel Approaches to Pathogen Discovery in Metagenomes, Fig. 1 Pathogen discovery workflow. (a) Flow diagram of the main procedures to pathogen identification. (b) 16S-based targeted metagenomics for determination of bacterial composition. *Top panel* shows 16S rRNA gene and hypervariable regions 1–9. *Center panel* shows three amplicons commonly used in 16S-based metagenomic sequencing. The *arrows* indicate the sequencing direction. *Bottom panel* shows fusion primers for PCR amplification of 16S rRNA gene segments. Primer A/B and key sequences are compatible to the

choice of downstream NGS platform, e.g., Roche/454 GS. Amplicon(s) for each sample can be barcoded individually using sequences such as 454's 10-nt Multiplexing identifier (*MID*) sequences. (c) Unbiased random amplification. Random reverse transcription is primed by random hexamers or octamers tailed with specific sequence. Subsequent random PCR uses the random primers and the primer matching with the specific sequence. The double-stranded random amplicons can be sequenced with NGS for viral sequence identification

N

Unbiased Random Amplification and Sequencing for Viral Pathogen Identification

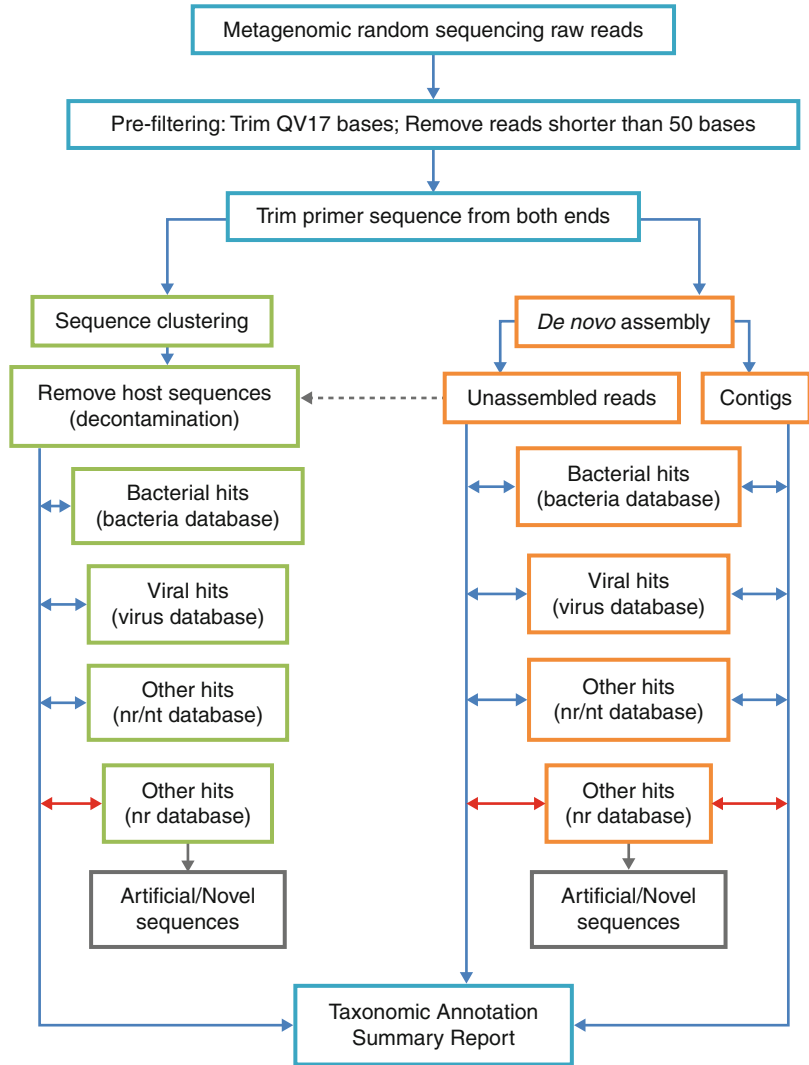
While microbial metagenomics for bacterial pathogen identification is still at its early stage, viral metagenomics has become a robust approach for hunting novel viral pathogens when viral culture and molecular assays cannot make the diagnosis (Djikeng and Spiro 2009; Mokili et al. 2012). Because of the vast number and variety of viruses in nature and the high frequency of evolution events including nucleotide mutagenesis, sequence recombination, and segment reassortment, it is not rare that a novel

virus or a new virus variant escaped initial detections or was misdiagnosed and caused an outbreak. De novo approach with no requirement for known sequence is therefore advantageous for viral pathogen discovery. One technique breakthrough is to identify novel viral sequence by unbiased random amplification and massive sequencing with NGS platforms. The process illustrated in Fig. 1 includes the following major steps: sample preparation which may require extra preprocessing to enrich viruses and reduce non-virus contents, random reverse transcription and anchored random PCR amplification, sequencing the random amplicons by NGS, and data mining for identification of viral pathogen

Novel Approaches to Pathogen Discovery in Metagenomes,

Fig. 2 Bioinformatics strategy for identification of viral sequences.

Metagenomic sequencing data are processed using streamlined multiple sequence analyzing tools to search for disease-related sequence hits. Two typical analysis paths are shown as examples. It is crucial for the efficiency to reduce redundancy (e.g., sequence assembly) and human genome sequences (decontamination) prior to database alignment while retaining relevant sequences. Reduced volume sequences are subjected to thorough alignments to the specified as well as mega databases



sequences. The significance of the culture-independent viral metagenomic approach was shown in studies in which novel viruses responsible for unresolved infections were identified. The discoveries by metagenome sequencing also led to subsequent confirmation with PCR, successful viral isolation by choosing suited cells, and complete viral genome sequences for rapid molecular tests for epidemiology and surveillance. Another noteworthy use of unbiased metagenomic sequencing is to detect coinfection of viruses or virus variants with estimation of the relative abundance for personalized medicine. For example, sensitive and accurate

monitoring of low-level drug-resistant HIV variants is clinically relevant for proactive health care of HIV-infected population (Gega and Kozal 2011).

There are a variety of protocols which were originated from the same technical approach but designed differently based on individual circumstances. The considerations include enriching viral contents in complex matrices by pretreatment with nucleases to degrade nonviral naked nucleic acids or concentration of viral particles by filtration or centrifugation, DNase treatment prior to reverse transcription to reduce genomic DNA, the removal of ribosomal RNA,

size selection of amplification products, and adjusting clonal amplification conditions to sequence random amplicons of broad range of sizes. To find virus sequence reads in metagenomic sequencing of clinical samples, capable bioinformatics workflow is needed to achieve the sensitivity, specificity, and speed. The workflow may comprise a set of data processing operations which can be chosen from tools such as de novo assembly, sequence clustering, decontamination (e.g., removal of human sequences), NCBI BLAST tools, etc. Two exemplary workflows are shown in Fig. 2. Nevertheless, further simplified and streamlined sample preparation and sequencing procedure which can be readily reproduced in clinical laboratory, good data management and sharing practices, and diagnostic-specific bioinformatics solution will be essential for viral pathogen discovery by means of metagenomics.

Summary

The capability on pathogen discovery is driven by technology innovation. Koch's postulates evolved from its original microbial form to the molecular postulates (Falkow 2004) and currently "the metagenomic version" (Mokili et al. 2012). Multidisciplinary strategy and methodology of metagenomics open a new era of pathogen discovery: analyze pathogenesis in comprehensive ecology and community views; delineate etiology with information on pathogen coinfection, virulent variants and concurrent factors, and individualized therapy with the considerations of metagenomes for optimal efficacy; and avoid misuse of antibiotics and antiviral drugs (Relman 2011). Next-generation sequencing is not only the ultimate sequence-based approach for pathogen identification but also a solution to stimulate clinical microbiology and molecular diagnostics when a novel pathogen is encountered. Despite the sound "proof-of-principle" as well as advancements on both technical and analytical means, substantial individual and concerted efforts are needed on translating

pathogen discovery on metagenomes from explorative research to standardized clinical practices.

References

- Arumugam M, Raes J, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473(7346):174–80.
- Djikeng A, Spiro D. Advancing full length genome sequencing for human RNA viral pathogens. *Futur Virol*. 2009;4(1):47–53.
- Eisen JA, MacCallum CJ. Genomics of emerging infectious disease: a PLoS collection. *PLoS Biol*. 2009;7(10):e1000224.
- Falkow S. Molecular Koch's postulates applied to bacterial pathogenicity – a personal recollection 15 years later. *Nat Rev Microbiol*. 2004;2(1):67–72.
- Fournier PE, Raoult D. Prospects for the future using genomics and proteomics in clinical microbiology. *Annu Rev Microbiol*. 2011;65(65):169–88.
- Fredericks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev*. 1996; 9(1):18–33.
- Gega A, Kozal MJ. New technology to detect low-level drug-resistant HIV variants. *Futur Virol*. 2011; 6(1):17–26.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68(4):669–85.
- Joseph SJ, Read TD. Bacterial population genomics and infectious disease diagnostics. *Trends Biotechnol*. 2010;28(12):611–8.
- Koser CU, Ellington MJ, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog*. 2012;8(8): e1002824.
- Kuczynski J, Lauber CL, et al. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet*. 2012;13(1):47–58.
- Lipkin WI. Microbe hunting. *Microbiol Mol Biol Rev*. 2010;74(3):363–77.
- Loman NJ, Misra RV, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(5):434–9.
- Mokili JL, Rohwer F, et al. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol*. 2012;2(1):63–77.
- Olsen RJ, Long SW, et al. Bacterial genomics in infectious disease and the clinical pathology laboratory. *Arch Pathol Lab Med*. 2012;136(11):1414–22.
- Pallen MJ, Loman NJ, et al. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol*. 2010;13(5):625–31.
- Petrosino JF, Highlander S, et al. Metagenomic pyrosequencing and microbial identification. *Clin Chem*. 2009;55(5):856–66.

- Qin J, Li R, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464(7285):59–65.
- Relman DA. Microbial genomics and infectious diseases. *N Engl J Med*. 2011;365(4):347–57.
- Thomas T, Gilbert J, et al. Metagenomics – a guide from sampling to data analysis. *Microb Inform Exp*. 2012;2(1):3.

Nucleotide Composition Analysis: Use in Metagenome Analysis

Isaam Saeed
Optimisation and Pattern Recognition Group,
Melbourne School of Engineering, The University
of Melbourne, Parkville, Australia

Synonyms

Binning; Genome signature; Nucleotide frequency

Definition

The composition of nucleotide bases in a microbial genome is not random and is instead biased toward different compositional structures that vary between organisms. These biases occur as identifiable patterns in oligonucleotide base composition, and it is by these patterns that otherwise anonymous metagenomic sequences are grouped into inferred populations. This allows for more in-depth analysis of the functional potential of a sampled microbial community in the context of constituent members (inferred populations).

Introduction

The composition of nucleotide bases in a microbial genome is not random and is instead biased toward different compositional structures that vary between organisms. These biases occur as identifiable patterns in oligonucleotide base

composition, and it is by these patterns that otherwise anonymous metagenomic sequences can be grouped into inferred populations enabling in-depth functional analysis.

Extensive sequencing of microbial DNA made possible the large-scale analysis of this genome base composition. Such analyses have revealed that the various patterns in base composition may be related to specific molecular machinery within microbial cells that help shape base composition. These biasing effects are thought to be mediated by the processes of DNA repair and replication, mutations, and base-step conformational tendencies that operate in concert to give rise to the characteristic base composition of different microbial genomes (Karlin et al. 1997).

Since the sequencing methodology of metagenomics does not preserve the association between sequenced reads and their genome of origin, functional analysis of a metagenome can only provide an overall snapshot of what a microbial community can potentially do. However, if the association between a sequence in a metagenome and the original genome (or population) from which it was sampled from can be inferred, then the resulting functional analysis can probe deeper into the inner workings of a microbial community. Processing sequences in this manner prior to functional analysis is referred to as binning. There are currently two major ways to address the binning problem (McHardy and Rigoutsos 2007): the first classifies sequences using a database of preexisting knowledge of microbial organisms; and the second groups related sequences based on the common patterns that arise from biases in the base composition of microbial genomes. The latter approach reflects the exploratory nature of metagenomics, given that the majority of microorganisms cannot be cultivated in a laboratory environment and therefore they may not be represented in current databases as yet.

When considering the use of patterns (or genome signatures) in nucleotide base composition for binning, there are two major factors that will influence the quality of the resulting set of identified groups (inferred populations). The first is the taxonomic resolution of patterns to be

used, which is governed by the between-genome distinctness of a pattern. The second is the accuracy at which these patterns can be grouped, which is governed by the within-genome conservation of a pattern.

A Simple Binning Strategy Using GC Content

It is well established that there are differences in GC content between various microbial genomes. The benefit of this to binning is that these biases can often be used as a representative pattern to group related sequences that share similar GC content. Although localized GC content can vary throughout a genome, if large enough sequences are available in a metagenome, then the assumption that the observed GC content is representative of the full genome composition still holds. It should be noted that GC content is not a unique property of individual genomes and if it is used it will group sequences coarsely (in terms of microbial taxonomy). In such a scenario, if GC content is significantly different between the identified groups, then it can be assumed that these groups are unrelated, but it is not conclusive to say that the sequences within each group are related unless further analysis is conducted (due to the nonuniqueness property of GC content). To increase the taxonomic resolution of binning using GC content, for example, it can often combine with other complementary features. An approach of this sort was used to group sequences of the metagenome of an acid mine drainage biofilm (Tyson et al. 2004), where GC content was combined with local assembly depth to distinguish the dominant populations that shared similar GC content.

Generally, metagenomes with a small number of dominant species tend to be easier to assemble, and the resulting contig lengths can often make GC content a viable option to group sequences in such data sets. However, there are still limitations to this approach, and for more complex metagenomes, GC content has been superseded by higher-order statistics of base composition, referred to as oligonucleotide (or n-mer) frequencies.

Nucleotide Frequency

With the advent of large-scale, high-throughput DNA sequencing, the increased sample size of sequenced DNA molecules provided a foundation for extensive statistical evaluation of nucleotide composition in different genomes. Further studies of genomic composition, in light of the increasing number of available genome sequences, pioneered the use of higher-order statistics to describe signatures in microbial genomes. The underlying principle of these signatures is based on the observation that specific oligomers are under-/overrepresented in different genomes and that the similar biases occur in related genomes. Nucleotide frequency is among the most widely used ways of representing these biases and is calculated by counting all occurrences of fixed length oligos (or n-mers) within a sequence and then normalizing by the total number of oligos in that sequence to arrive at an estimate of the oligonucleotide frequency content. The features of microbial genomes based on nucleotide frequency, which have been successfully applied to metagenomic studies, include the following: the dinucleotide odds ratio, codon signatures/trinucleotide frequencies, and tetranucleotide frequencies.

Dinucleotide Odds Ratio

Among the earliest of these nucleotide frequency signatures that was found to be biologically relevant was the dinucleotide odds ratio, which was based on early *in vitro* studies on differences in dinucleotide content between various organisms (Karlin et al. 1997). This signature considered the dinucleotide frequency content of a sequence and factored out the effect of mononucleotide frequencies using a normalization scheme based on a Markov model, as given by

$$\rho_i^2 = \frac{f_{XY}}{f_X f_{Y'}}$$

where X and Y represent the first and second mononucleotide in the dinucleotide to be normalized and f represents the frequency of mono-/dinucleotides. The derived statistic, also referred to as the dinucleotide odds ratio, could adequately describe biases specific to various

microbial organisms. For example, it was observed that there is a general TpA avoidance mechanism across various microbial genomes and a CpG underrepresentation in thermophilic microbes. Distances between sequences represented using the dinucleotide odds ratio are evaluated using the Manhattan distance, also referred to as the δ^* distance. When this odds ratio differs from 1, the resulting statistic provides a means to estimate the under/overrepresentation of specific dinucleotides, given by the limits 0.78 and 1.23, respectively (Karlin et al. 1997). Although several genome-wide biases were found when using this odds ratio statistic, the discrimination between larger sets of microbial genomes (more representative of real-world metagenomes) is still better handled by higher-order frequencies.

Codon Signatures/Trinucleotide Frequency

Gene sequences are relatively conserved within a genome, as any changes at critical locations may cause the gene product to be defective. This motivated the use of signatures based on this knowledge to capture more representative patterns in microbial genomes. Codon usage in the gene sequences is thought to be mediated by the overall genome composition and is also related to the flexibility of the choice of codons due to the degeneracy of the genetic code. Trinucleotide frequencies (i.e., frequencies of all possible 3-mers: AAA, AAC, AAT, ..., GGG) can be used to capture some of these biases, and alternatively, an extension to the dinucleotide odds ratio is also able to capture these codon signatures using dinucleotides (Karlin et al. 1998). This codon signature is constructed as follows:

$$\begin{aligned}\gamma_{XY}(1,2) &= \frac{f_{XY}(1,2)}{f_X(1)f_Y(2)} \\ \gamma_{XY}(2,3) &= \frac{f_{XY}(2,3)}{f_X(2)f_Y(3)} \\ \gamma_{XY}(3,4) &= \frac{f_{XY}(3,4)}{f_X(3)f_Y(4)}\end{aligned}$$

where the indices represent the nucleotide base at the first, second, or third nucleotide within a codon (with index 4 referring to the first base of the next codon). This signature requires at least

50 full-length genes from a given genome to make a stable estimate, and these ratios can be biased within a genome (not only between genomes), depending on the set of gene classes that comprise the genes in a genome. Due to these issues, such signatures may cause difficulty in grouping sequences for the purpose of binning.

Tetranucleotide Frequency

Tetranucleotide frequency (all possible combinations of 4-mers, of which there are 256) offers greater discrimination between species in a metagenome than lower-order nucleotide frequencies. For this reason, tetranucleotide frequency is perhaps the most widely used in clustering metagenomic sequences. Moreover, it has been found to capture a species-specific signature (a reasonably strong phylogenetic signal at lower taxonomic ranks), which makes it not only a more powerful alternative to clustering metagenomic sequences but also offers biologically meaningful groupings of sequences (Teeling et al. 2004). This was also confirmed by (Mrazek 2009) who correlated 16S rRNA distances with various signatures and found that tetranucleotide frequency was able to outperform other feature sets. It has also been found that tetranucleotide frequency can be used to find conserved signatures flanking 16S rRNA genes, which can in turn be used to assign classes to the identified groups of sequences (Chan et al. 2008).

Strand Bias

Prior to the use of these signatures, it should be noted that for oligonucleotide frequencies, the feature vector requires correction for biases between strands (Tyson et al. 2004). This is often remedied by counting the number of n-mers on the original sequence, as well as on the reverse complement, and then taking the average of the two prior to normalization.

Normalization Techniques for Nucleotide Frequency

Given the observed nucleotide frequencies for each sequence, it is often necessary to normalize each observation prior to further analysis. (*Note:*

it is still possible to simply take the frequencies of the observed number of n -mers in a sequence.)

Markov Normalization

The dinucleotide odds ratio is a special case of Markov normalization. In the general case, the maximal-order Markov normalization of an observed nucleotide frequency vector is given by

$$\rho_i^k = \frac{f_{1\dots k} f_{2\dots k-1}}{f_{1\dots k-1} f_{2\dots k}},$$

where the appropriate statistic for tetranucleotide frequency, for example, is given when $k = 4$. This normalization scheme essentially aims to filter out lower/higher nucleotide frequencies. Lower-order normalization schemes are possible, but they have poorer correlation properties with phylogenetic distances (Mrazek 2009).

Z-Score Normalization

Another approach to normalization uses the Z-score transform to assess the statistical significance of observed n -mers (Tyson et al. 2004). For tetranucleotide frequency, the Z-score normalization is computed as follows: the expected value for a given tetramer is calculated by

$$E(n_1 n_2 n_3 n_4) = \frac{N(n_1 n_2 n_3) N(n_2 n_3 n_4)}{N(n_1 n_2)},$$

and the variance is calculated using

$$\sigma^2(n_1 n_2 n_3 n_4) = E(n_1 n_2 n_3 n_4) \times \frac{[N(n_2 n_3) - N(n_1 n_2 n_3)][N(n_2 n_3) - N(n_2 n_3 n_4)]}{N(n_2 n_3)^2},$$

which gives the required normalization for each tetramer:

$$Z(n_1 n_2 n_3 n_4) = \frac{N(n_1 n_2 n_3 n_4) - E(n_1 n_2 n_3 n_4)}{\sqrt{\sigma^2(n_1 n_2 n_3 n_4)}}.$$

The Oligonucleotide Frequency-Derived Error Gradient (OFDEG)

An extension to oligonucleotide frequency-based features is the oligonucleotide

frequency-derived error gradient (OFDEG) (Saeed and Halgamuge 2009). OFDEG was observed by exploring the relationship of nucleotide frequency between short fragments and the original DNA sequence from which they are sampled. The observation is based on a resampling method, where instead of using the entire sequence to estimate the maximum likelihood point estimate of nucleotide frequency, the OFDEG measure resamples the nucleotide base composition of varying length subsequences to capture the distribution of oligomeric frequencies.

For example, it is straightforward to compute the un-normalized nucleotide frequency of an entire genome (referred to in this definition as an occurrence vector). Similarly, the occurrence vector for a short subsequence sampled from anywhere along the genome can be easily computed. Intuitively, the error between these two occurrence vectors (defined in terms of Euclidean distance) would be large. Nevertheless, the error is recorded and another subsequence, of increased length, is sampled again from anywhere along the genome. Trivially, the error between the occurrence vector of this new subsequence and the occurrence vector of the genome would be reduced. This process is continued until the length of subsequences is equivalent to the length of the genome, while keeping track of the error at each sampling instance. The resulting error as a function of subsequence length is found to be linear (up to a given subsequence length). The rate of error reduction (or gradient) of this linear trend, within the bounds of the linear region, is referred to as the OFDEG. It has been found that this linear gradient is different for various genomes and is remarkably consistent within genomes as well as between fragments of a genome. The measure essentially captures the relative magnitude of biases in nucleotide base composition in a manner similar to entropic measures and has been used in combination with other complementary features to successfully group related sequences in various simulated and real-world metagenomes.

Combining Features to Increase the Taxonomic Resolution of Binning

It has recently been demonstrated that when two signatures capture the different characteristics of base composition, they can be used to group sequences differently, and in cases where these groups are mutually exclusive and at different taxonomic resolutions, such features can be arranged hierarchically to increase the taxonomic resolution at which sequences in a metagenome are grouped (Saeed et al. 2012).

This concept was demonstrated using the combination of GC content and OFDEG as a preliminary set of features to coarsely group a metagenome and then using tetranucleotide frequency to refine these coarse groups further. This is particularly important when the number of populations in a metagenome increases, as tetranucleotide frequency on its own is known to saturate in its discriminatory power when this occurs (Saeed et al. 2012). Since both OFDE and GC content generate coarse groups, the groups can then be processed using a high-resolution feature set for refinement. This has been found to improve the binning performance over existing methods on simulated as well as real metagenomic data sets (Saeed et al. 2012).

Grouping Related Metagenomic Fragments Based on Sequence Composition

Methods that operate on nucleotide composition for grouping related sequences can be classified in terms of two broad machine learning paradigms: those which construct supervised classifiers and those which rely on unsupervised exploratory clustering.

Supervised Classification

A classifier can be trained using existing knowledge of patterns based on the analysis of reference sequences in current databases. These methods consider the classification of each sequence in isolation, and their accuracy will be dependent on the representativeness of the

training set in relation to the metagenome under investigation.

Unsupervised Clustering

Unsupervised learning is not predicated on the availability of reference sequences for training. Instead, methods which operate in this paradigm group related DNA sequences by the inferred similarity of patterns, which is consistent with the exploratory nature of metagenomic studies. When using patterns it can often be advantageous to use this approach, particularly when the sampled community consists of microbes that are either underrepresented or not represented in existing databases. Moreover, these methods can be applied directly to a data set to reveal hidden patterns among related sequences in a metagenome without enforcing a priori knowledge of what phylotypes should be present.

However, for clustering methods which operate on density estimation, these methods require a sufficient number of sequences per population in order for it to be discovered as a cluster. As such, highly complex metagenomes with no dominant populations are difficult to analyze in this manner and are at present perhaps better suited to supervised methods (provided a suitable training set can be constructed).

The Effects of Various Forms of Noise in Grouping Metagenomic Sequences Using Nucleotide Base Composition

Category I: When unrelated, or distantly related, genomes have highly similar compositional signatures, the number of false positives in grouping sequences can increase and will consequently affect binning specificity.

Category II: Genomes that have large intra-genomic variation in base composition can often increase the number of false negatives (these can sometimes be observed as outliers) during binning and consequently affect binning sensitivity.

Category III: A more complex form of noise occurs when organisms partially share common characteristics in base composition, which will cause groups to overlap.

With the use of model-based clustering in an unsupervised setting with a hierarchical set of features, there is the potential to increase the accuracy of binning by removing these forms of noise (Saeed et al. 2012).

Limitations and Future Directions

Significant advances in compositional binning approaches have primarily looked at the issue of representing the composition of a sequence, rather than refining machine learning methods that operate in an unrepresentative feature space. Such is the case of the succession of GC content by higher-order oligonucleotide frequencies, for example. For instance, with an increase in the number of fully sequenced bacterial and archaeal genomes, it was observed that compositional features tend to saturate in their capacity to uniquely describe a microbial genome (or clade) (Teeling et al. 2004). The use of complementary features can address this limitation to a certain extent (Saeed et al. 2012).

Given that metagenomes only contain fragments, which in most cases can be quite short, the length of sequences often limits the representativeness of features based on nucleotide frequency (Mavromatis et al. 2007). This is because these signatures are statistical in nature and require sufficient sequence lengths with which to estimate a representative signature. The minimum sequence length has been argued to be 1 kbp (McHardy and Rigoutsos 2007), 40 kbp (Teeling et al. 2004), and even 50–100 kbp (Karlin et al. 1997); but in general caution is advised when applying such techniques to sequences less than 1 kbp as this may result in unrepresentative and sparse feature vectors.

This limitation of sequence length in metagenomes is not only due to currently achievable read lengths but also the complexity of assembling metagenomes (in comparison to single-genome studies). For complex communities (species rich), the required coverage for reasonable levels of assembly (N50 contig length greater than 1 kbp) translates into substantial sequencing requirements. In light of this,

methods that operate on nucleotide frequency alone can be seen to be at a disadvantage, but with the anticipated longer read lengths and higher throughput that future sequencing platforms are capable of generating, coupled with the development of a wide variety of novel tools for metagenomic data analysis, these issues may be largely alleviated and composition-based binning will be an important tool for metagenome analysis.

Summary

The analysis of nucleotide base composition in grouping related metagenomic sequences allows for more in-depth analysis of the functional potential of a sampled microbial community, in the context of constituent members (inferred populations), rather than simply observing the overall functional potential of a community. The features in current use are essentially based on nucleotide frequencies, which describe the relative abundance of *n*-mers in a sequence, and various extensions to these signatures have also been introduced, such as the oligonucleotide frequency-derived error gradient (OFDEG).

The performance of using composition-based features for binning can be improved when using complementary features in combination, which can result in an increase in the taxonomic resolution of the groups that result. In general, however, the accuracy of grouping sequences using nucleotide base composition is largely governed by the algorithm used to analyze the patterns (whether in a supervised or an unsupervised setting), the available sequence lengths, and the choice of compositional feature. On the other hand, the level of taxonomic resolution that can be achieved in such an analysis is more heavily influenced by the choice of compositional feature alone. In most cases, these can be alleviated with advances in sequencing technology. Nevertheless, there is much that can be unveiled when patterns are extracted from metagenomic sequences. It is, however, a matter of knowing what patterns to extract and how best to extract them before an attempt is made to group them.

References

- Chan C-K, Hsu A, Halgamuge SK, Tang S-L. Binning sequences using very sparse labels within a metagenome. *BMC Bioinforma.* 2008;9:215.
- Karlin S, Mrazek J, Campbell A. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol.* 1997;179(12):3899–913.
- Karlin S, Campbell A, Mrazek J. Comparative DNA analysis across diverse genomes. *Annu Rev Genet.* 1998;32:185–225.
- Mavromatis K, Ivanova N, Barry K, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4(6):495–500.
- McHardy A, Rigoutsos I. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol.* 2007;10:499–503.
- Mrazek J. Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol.* 2009;26(5):1163–9.
- Saeed I, Halgamuge SK. The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics.* 2009;10(S3):S10.
- Saeed I, Tang S-L, Halgamuge SK. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* 2012;40(5):e38.
- Teeling H, Meyerdierks A, Bauer M, et al. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol.* 2004;6(9):938–47.
- Tyson G, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004;428:37–43.

O

Open Resource Metagenomics

Trevor C. Charles and Josh D. Neufeld
Department of Biology, University of Waterloo,
Waterloo, ON, Canada

Synonyms

Open-access metagenomics; Open-source metagenomics; Shared metagenomic libraries

Definition

Open resource metagenomics encompasses the emerging resource sharing system that facilitates distribution of metagenomic libraries throughout the research community. These libraries are constructed from DNA isolated directly from environmental samples. Broad host range cosmids are ideal for open resource metagenomics, accommodating large DNA inserts for screening or selections in multiple prokaryotic or eukaryotic hosts.

The Challenges of Functional Metagenomics

By capturing DNA extracted directly from environmental samples, the first metagenomic libraries were constructed in the late 1990s (Handelsman et al. 1998; Rondon et al. 2000).

Metagenomic libraries are typically constructed for specific applications such as retrieving genes with desired functions. Functional metagenomics is the use of metagenomic libraries to isolate genes of interest based on associated activity of captured environmental genes (recently reviewed in Ekkers et al. 2012). Metagenomic libraries may be screened or selected in several potential host organisms (Wexler and Johnston 2010), commonly for functions of potential biotechnological value. This approach facilitates the discovery of novel genes, without requiring culture of the organisms that naturally carry those genes or sequence homology to known genes.

In recent years, the combination of high-capacity sequencing and advances in computational analysis of metagenomic sequence data has resulted in dramatic improvements in gene discovery in the absence of functional screening (Thomas et al. 2012). Despite these improvements, a fundamental limitation is that links between sequence and function tend to be substantially incomplete. This is not only a limitation of metagenomic library analysis, it is also an important caveat for the study of genomes from individual organisms. For example, it is often not possible to assign a function to a gene product of a characterized protein family, although that is precisely the limitation of computational methods for sequence-based analyses. Confident determination of specific functions, such as substrate specificity associated with sequence motifs, relies on the availability of experimental data. Arguably, the most interesting and valuable

metagenomic genes will be those whose function could not have been predicted by sequence alone; these genes would be more likely to encode products with truly novel properties.

A major advantage of metagenomic libraries is that once they are made, they can be a permanent resource, a snapshot of the microbial community that the DNA was extracted from. The same library, if stored properly, can be screened multiple times, indefinitely. Below we outline several methodological considerations for maximizing benefit from open resource metagenomic libraries.

Although they have sometimes been used, small-insert libraries are not optimal for functional metagenomics. The smaller the insert, the less chance that individual clones will contain full operons, including the regions required for control of gene expression. As a result, the use of bacterial artificial chromosome (BAC; Kikirde et al. 2012) and cosmid/fosmid (Aakvik et al. 2009; Neufeld et al. 2011; Taupp et al. 2011) vectors enables the cloning of fragments that are large enough to include multiple operons. Such large-insert libraries require fewer clones to ensure that they are representative.

Depending on DNA yields, quality, and size, metagenomic libraries of environmental microbial communities may yield several million clones. If such libraries are distributed into 384-well plates, this would represent over 2,500 plates per million clones. Plate storage would require extensive freezer space, and screening such libraries, one clone at a time, would be prohibitively laborious and costly, even with the use of robotic manipulation. An alternative strategy we recommend is to recover and maintain the libraries as pools of clones. This procedure involves physical harvesting and mixing of all individual colonies from all initial library plates, followed by the preparation of aliquot suspensions for cryopreservation and subsequent distribution.

Another important consideration is that different host backgrounds will selectively express only a subset of an environmental metagenome. For example, *Bacteroides* genes use specialized

promoters for their transcription (Mastropaolo et al. 2009). Host-specific limitations on gene expression include posttranscriptional controls, including translation initiation, codon usage, protein folding, enzyme activation, and transport. Also, wild-type and mutant strains that are most appropriate for a given screen might be available only in a host background that does not support replication of a given vector. This is especially true when using vectors that only replicate in *Escherichia coli* and other *Gammaproteobacteria*. For these reasons, it is advantageous to choose or design vectors that can be maintained in diverse host backgrounds.

Metagenomic libraries are often constructed for specific applications, such as to screen for a desired enzyme activity. Unlike the situation for single culture isolates, which must be deposited in accessible culture collections or otherwise made available as a requirement of publication of research results involving them, there is no such requirement or expectation for metagenomic libraries. This is unfortunate, as high-quality metagenomic libraries are technically challenging and costly to construct, and their full value is often not realized if their use is restricted to one or a few research groups.

Achieving Metagenomic Resource Sharing

We formally proposed that to ensure maximum value, metagenomic libraries should be made publicly available to members of the research community, without restriction (Neufeld et al. 2011). This is the concept of *open resource metagenomics* that libraries be pooled to ensure ease of archiving as frozen stocks and for subsequent distribution and handling. We also recommended that cosmid libraries be used, because they allow the efficient cloning of large inserts of >30 kb. To facilitate screening in a diversity of host backgrounds, cosmid vectors with broad host range origins of replication are recommended, as well as Gateway recombinational systems for easy transfer of inserts to other vectors. An example of such a resource is



the Canadian MetaMicroBiome Library project (CM²BL; <http://cm2bl.org>), which houses a collection of Canadian soil metagenomic libraries in an IncP cosmid Gateway vector. The largest library in this collection contains over eight million clones. To assist users in deciding which libraries to choose for a given application, extensive metadata and taxonomic sequence information is accessible in an online database.

Summary

The *open resource metagenomics* initiative aims to increase the availability of metagenomic libraries to the research community as a public and scientific resource. The principle of free and open sharing of metagenomic libraries is central to this initiative, including direct access to associated metadata and DNA sequences. Increased gene discovery as a result of the use of these libraries not only has the potential to provide novel, biotechnologically useful genetic material but should increase the overall understanding of gene functions and their relationship to DNA sequence.

References

- Aakvik T, Degnes KF, Dahlsrud R, Schmidt F, Dam R, Yu L, Völker U, Ellingsen TE, Valla S. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiol Lett.* 2009;296:149–58.
- Ekkers DM, Cretoiu MS, Kielak AM, Van Elsas JD. The great screen anomaly – a new frontier in product discovery through functional metagenomics. *Appl Microbiol Biotechnol.* 2012;93:1005–20.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:R245–9.
- Kakirde KS, Wild J, Godiska R, Mead DA, Wiggins AG, Goodman RM, Szybalski W, Liles MR. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene* 2012;475:57–62.
- Mastropaolo MD, Thorson ML, Stevens AM. Comparison of *Bacteroides thetaiotaomicron* and *Escherichia coli* 16S rRNA gene expression signals. *Microbiology.* 2009;155:2683–93.
- Neufeld JD, Engel K, Cheng J, Moreno-Hagelsieb G, Rose DR, Charles TC. Open resource metagenomics; a model for sharing metagenomic libraries. *Stand Genomic Sci.* 2011;5:203–10.
- Rondon MR, August PR, Betterman AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, Macneil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM. Cloning the oil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol.* 2000;66:2541–7.
- Taupp M, Mewis K, Hallam SJ. The art and design of functional metagenomic screens. *Curr Opin Biotechnol.* 2011;22:465–72.
- Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp.* 2012;2:3.
- Wexler M, Johnston AB. Wide host-range cloning for functional metagenomics. *Methods Mol Biol.* 2010;668:77–96.

P

Phylogenetics, Overview

Phylogenetics: A Root and Branch Analysis of the Tree of Life

Roy Sleator
Department of Biological Sciences, Cork
Institute of Technology, Cork, Co. Cork, Ireland

Synonyms

Evolutionary relatedness

Definition

Phylogenetics, derived from the Greek terms *phylon* (meaning “tribe”) and *genetikos* (meaning “genitive” or origin), is the study of the evolutionary history of species, organisms, genes, or proteins through the construction and analysis of mathematical entities known as trees or phylogenies.

Introduction

Darwin’s *The Origin of Species* marked the birth of phylogeny, a discipline whose primary aims are to classify all living organisms, grouping all extant descendants of a given ancestor within specific groups or clades; to provide insights into the shared properties of members within

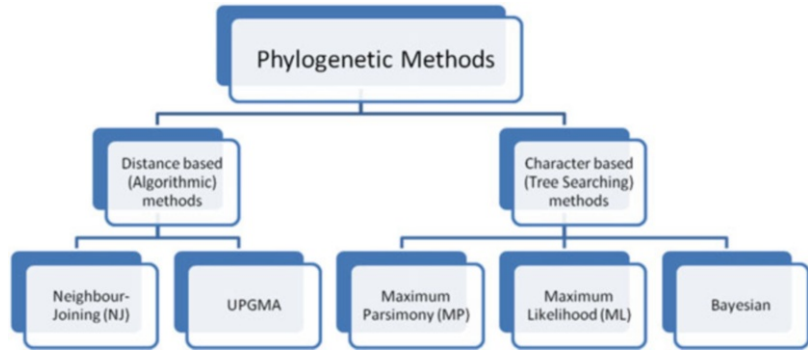
each clade; and to allow retro direction, i.e., the ability to infer ancestral properties based on observable characteristics of extant organisms.

A significant limitation of traditional morphology-based phylogeny approaches is the fact that reconstructing ancient evolutionary events requires a vast sum of character changes. Furthermore, many of these morphological characters are likely under selective pressure and subject to convergence (Sleator 2010). Based solely on this criterion, most organisms lack sufficient phenotypic characters to perform effective comparative analyses (Lopez and Baptiste 2009).

The development of modern DNA and protein sequence technologies has however effectively eliminated this limitation. Modern phylogenetic analysis involves the progressive alignment of nucleic acid and/or protein sequences between extant organisms. A hypothesis is then produced to explain the repartition of character states, and the results presented as a phylogenetic tree – which is simply a graphic representation of the computed output.

The accelerating accumulation of molecular sequence data arising from recent concerted large-scale genomic and metagenomic sequencing projects (Sleator et al. 2008) continues to afford new opportunities and perspectives for dissecting evolutionary relationships. Indeed, while early molecular phylogenetic approaches centered on individual DNA sequences coding for RNA or proteins, or the derived amino acid sequences of the latter, more recent analysis of

Phylogenetics, Overview, Fig. 1 Tree-building methods. Schematic overview of the major analytical approaches to phylogenetic tree building



whole genomes has led to the development of phylogenomics – a powerful approach to analyze complete genome sequences as a metasequence (Forte and Gadelles 2009).

Discussion

Tree-Building Methods: The Major Analytical Approaches

While several methods exist for inferring evolutionary relatedness, most can be classified as either *distance-* or *character-based methods* (outlined in Fig. 1). Distance (or *algorithmic*) methods employ an algorithm incorporating a model of evolution (e.g., amino acid substitution) to compute a distance matrix from which a phylogenetic tree is calculated by means of progressive clustering. Specifically, distances in the matrix relate to the number of differences between each pair of sequences (either DNA or protein). The model of evolution specifies how amino acid substitutions occurred in the protein sequence since they last shared a common ancestor. Finally, the tree is constructed from the numerical data in the matrix, with the most closely related sequences occupying a position on the tree which is distant from the less closely related sequences. Both the neighbor-joining (NJ) and the unweighted pair group method using arithmetic averages (UPGMA) approaches to tree building employ distance-based methods. Although fast and readily available in user-friendly software packages such as MEGA (Tamura et al. 2011), distance-based methods

have a number of significant limitations. NJ, for example, provides only a single tree as opposed to character-based methods which compute a consensus tree from several optimal or near optimal candidates. Furthermore, NJ may compute different trees depending on the order in which the constituent sequences are added. Finally, given that differences are presented as distance values, it is impossible to identify the specific character changes that support a branch (Soltis and Soltis 2003a).

Character-based methods (also referred to as tree-searching methods) search for the most probable tree for a specific sequence set based on characters at each position of the sequence alignment and a model of evolution. The most common character-based approaches include maximum parsimony (MP), maximum likelihood (ML), and to a lesser extent Bayesian methods.

MP seeks to find the tree or trees that are compatible with the minimum number of substitutions among sequences, i.e., the fewest evolutionary changes. An advantage of MP is that it provides diagnosable units (i.e., specific sets of characters) for each clade and branch lengths in terms of the number of changes on each branch of the tree. However, a significant limitation of the MP approach is that it requires strict assumptions of consistency across sites and among lineages. Thus, MP performance is significantly affected when mutational rates differ between conserved and hypervariable regions or if evolutionary rates are highly variable among evolutionary lineages. Finally, parsimony lacks an explicit model of evolution.

ML methods are based on specific probabilistic models of evolution and search for the tree with maximum likelihood under these models. The model of evolution may be empirical, derived from general assumptions about the evolution of sequences, or parametric, based on values estimated from the dataset. The major advantage of likelihood approaches is that they are based on powerful statistical theory which facilitates the application of robust statistical hypothesis testing and significant refinements to the resulting phylogenetic trees. However, while these strong statistical foundations make ML techniques arguably the most powerful approach in terms of phylogenetic reconstruction, paradoxically this strength is also a significant weakness, in that ML approaches are computationally intensive and, as a result, significantly slower than alternative approaches. As such, ML analysis can only be practically applied to a limited number of sequences (Soltis and Soltis 2003a).

In practice, both distance- and character-based methods tend to be used in tandem. An initial tree may be estimated by a distance-based method and used to test the parameters of the model of evolution. The most appropriate of these might then be used in a maximum likelihood tree search.

Testing the Reliability of a Tree

There are two approaches to finding the best tree: those that use optimality criteria that can be evaluated for any given tree (used for MP and ML) and those that involve the progressive clustering of sequence subsets (used for NJ and UPGMA). In the optimality methods trees are evaluated one by one by either exhaustive, branch and bound, or heuristic searches. Exhaustive searches evaluate all possible bifurcating trees to find a globally optimal topology; such an approach is only feasible for a relatively small number of taxa (<10). Rather than evaluating every possible tree, the branch and bound approach first chooses a local optimum value for tree length representing the total number of evolutionary changes on the tree; any tree length greater than the local optimum is automatically discarded, thus saving time and computational expense. Branch and bound

searches are effective up to ~20 taxa. Heuristic (or “best guess”) searches employ a “hill climbing” approach; an initial tree is chosen and subsequently modified; changes leading to an inferior tree descend the hill and the tree is rejected; changes leading to an improvement ascend the hill – when no further improvement is possible, the search is terminated. Although an extremely fast approach, there is no guarantee that the returned tree is the global optimal (the summit) or merely a local optimum (a foothill’s plateau).

Once an optimum tree is chosen, some statistical measure of internal support for clades must also be provided to prove that the tree is sufficiently robust and biologically meaningful. To this end a variety of methods have been proposed to verify the evolutionary reliability of trees of which the most commonly used is the bootstrap analysis. Bootstrapping can be divided into both parametric and nonparametric approaches (Wrobel 2008). Nonparametric bootstrapping is a numerical resampling approach in which a subset of sequence alignments referred to as bootstrap or pseudo-alignments are formed from the dataset by random sampling. This process is repeated several times (depending on the size of the dataset and the specifications of the analysis) usually with a default setting of 1,000 replicates. Bootstrap values are conservative measures of phylogenetic accuracy with values of 70 % or more representing “true” clades in experimental phylogenies. Parametric bootstrapping on the other hand creates replicate samples using numerical simulation as opposed to resampling. This approach is usually applied to test competing hypotheses.

Although generally effective, the bootstrap approach rests on a number of assumptions which are not optimal when applied to molecular sequence analysis (for an overview see [Box 1](#)). In addition to bootstrapping another measure of internal support which is often used in phylogenetic analyses is jackknifing. Although similar to bootstrapping, jackknifing involves one significant difference; rather than resampling the data, this approach uses only subsets of the available data (i.e., resampling without replacement to

create a smaller dataset). The purpose of which is to account for the presence of possible “outlier” characters which might have a disproportionate influence on the resulting tree.

Other less common approaches to measuring internal support include the decay index for parsimony analyses (Hernandez Fernandez and Vrba 2005) and the posterior probabilities generated in Bayesian inference (Wrobel 2008).

Box 1. Limitations of Bootstrap Analysis when Applied to Molecular Sequences

- The statistical bases of bootstrap analysis require that all positions of an alignment are independently identically distributed. However, this assumption fails to hold true for either nucleotide or amino acid sequences. For example, in proteins certain di-residues (in the primary structure) are either over- or underrepresented (Karlin et al. 1991), while strong correlations are observed between positions that interact within the 3D structure (Karlin et al. 1994).
- Bootstrap analysis is hampered by unequal evolutionary rates. If mutational rates are too high or uneven among lineages, the bootstrap proportion P is usually an overestimate (Soltis and Soltis 2003b).
- Molecular sequences are not representative of a homologous population, and as such resulting bootstrap values may not signify reliable clusters (Brocchieri 2001).

Difficulties Associated with Creating Reliable Phylogenetic Trees

Phylogenetic inferences are only as good as the alignments they are drawn from – “*Garbage in; garbage out.*” The majority of current alignment protocols are based on dynamic programming (DP) procedures which seek to identify the maximal alignment score, a value determined by the choice of scoring matrix (e.g., PAM or BLOSUM) and the assignment of gap penalties. Rather than searching for the optimal alignment of n sequences in an n -dimensional space, most DP methods employ fast heuristic or “greedy” approaches, progressively aligning pairs of

sequences. However, while effective this approach has a number of shortcomings in terms of phylogenetic analysis (for an overview of these shortcomings, see Box 2).

An alternative approach involves the application of motif finding algorithms which select common sequence motifs and align only these most conserved domains with no allowance for gaps or insertions (Lawrence et al. 1993).

In addition to alignment difficulties, two of the most significant problems associated with assessing tree reliability are long-branch attraction (LBA) associated with mutational saturation and lateral gene transfer (LGT) mediated, at least in part, by viruses and mobile genetic elements (Sapp 2007). As mutations cumulate during evolution, a point of mutational saturation is reached at which there is no further divergence between taxa (Brocchieri 2001). From this point on it becomes impossible to estimate evolutionary distance; furthermore very divergent sequences tend to be attracted together (Fig. 2) – hence the name – thus skewing their true position (Lopez and Baptiste 2009).

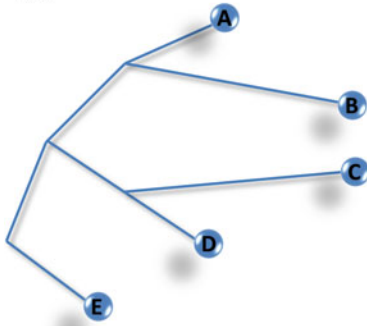
Box 2. Sequence Alignment Shortcomings

- Heuristic methods, although fast, only provide a best guess or estimate of the optimal alignment.
- Alignments are sensitive to the choice of similarity matrix (for amino acid sequence alignments) and gap penalty which are user adjustable – thus requiring human intervention.
- Hierarchically aligning pairs of sequence is prone to generate biases and dominance by the most similar sequences.

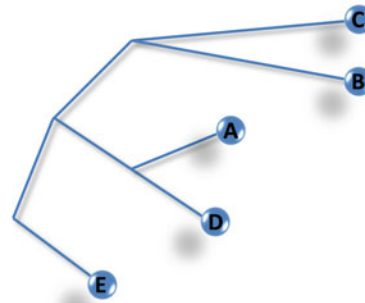
What Next...?

Phylogenomics – the merging of phylogenetics and genomics – is perhaps the most exciting recent development in the field of evolutionary mapping (Delsuc et al. 2005). Rather than concentrating on a single phylogenetic marker, whole-genome phylogenomic approaches involve comparisons of gene content: the

(i) Real Tree



(ii) Inferred Tree



Phylogenetics, Overview, Fig. 2 Long-branch attraction. A simulated example of long-branch attraction. (i) The real tree of the relationships among five taxa, with two taxa (*B* and *C*) having long evolutionary

branches. (ii) An inferred tree of the taxa in which *B* and *C* are artificially grouped together because of the phenomenon of long-branch attraction

presence or absence of orthologous genes (or gene families) and/or gene order. Genomic relationships based on genomic content and organization (representing the genomic profile) are inferred using the genomic signature which computes differences within, and between, species specific sequences based on dinucleotide relative abundance differences. The generality and robustness of the genomic signature gives it an advantage over traditional approaches which, based on individual sequences, are strongly influenced by mutational events such as LGT.

Finally, while it is tempting to consider only the Darwinian-Mendelian model of vertical gene transfer in phylogenetic analysis, recent evidence suggests that the role of LGT in shaping evolution can no longer be ignored. Indeed, in certain prokaryotes the LGT rate is comparable to and, in some instances, significantly higher than the rate of spontaneous mutation (Lawrence 2002). LGT has also been observed between eukaryotes (Andersson et al. 2007) as well as between organelles of the same cell (Archibald et al. 2003). A major consequence of LGT is that instead of focusing on the elusive “tree of life” (Puigbo et al. 2009), phylogenetic analysis must now consider the whole forest, corresponding to the integrated framework of vertical and lateral gene transfer (Lopez and Baptiste 2009). Slowly, but

surely, evolutionary biologists are beginning to “see the wood for the trees” (Sleator 2011).

Summary

Recent rapid expansions in the DNA and protein databases, arising from large-scale genomic and metagenomic sequence projects, have forced significant development in the field of phylogenetics, the study of the evolutionary relatedness of the planet’s inhabitants. Advances in phylogenetic analysis have greatly transformed our view of the landscape of evolutionary biology, transcending the view of the tree of life which has shaped evolutionary theory since Darwinian times. Indeed, modern phylogenetic analysis no longer focuses on the restricted Darwinian-Mendelian model of vertical gene transfer but must also consider the significant degree of lateral gene transfer which connects and shapes almost all living things.

Cross-References

- ▶ [DNA Methylation Analysis by Pyrosequencing](#)
- ▶ [Horizontal Gene Transfer and Bacterial Diversity](#)

References

- Andersson JO, Sjogren AM, Horner DS, Murphy CA, Dyal PL, Svard SG, Logsdon JR JM, Ragan MA, Hirt RP, Roger AJ. A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics*. 2007;8:51.
- Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloniella natans*. *Proc Natl Acad Sci U S A*. 2003;100:7678–83.
- Brocchieri L. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*. 2001;59:27–40.
- Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 2005;6:361–75.
- Forster P, Gabelle D. Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. *Nucl Acids Res*. 2009;37:679–92.
- Hernandez Fernandez M, Vrba ES. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. *Biol Rev Camb Philos Soc*. 2005;80:269–302.
- Karlin S, Bucher P, Brendel V, Altschul SF. Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Biophys Chem*. 1991;20:175–203.
- Karlin S, Zuker M, Brocchieri L. Measuring residue associations in protein structures. Possible implications for protein folding. *J Mol Biol*. 1994;239:227–48.
- Lawrence JG. Gene transfer in bacteria: speciation without species? *Theor Popul Biol*. 2002;61:449–60.
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*. 1993;262:208–14.
- Lopez P, Baptiste E. Molecular phylogeny: reconstructing the forest. *C R Biol*. 2009;332:171–82.
- Puigbo P, Wolf Y, Koonin E. Search for a ‘Tree of Life’ in the thicket of the phylogenetic forest. *J Biol*. 2009;8:59.
- Sapp J. The structure of microbial evolutionary theory. *Stud Hist Philos Biol Biomed Sci*. 2007;38:780–795.
- Sleator RD. An overview of the processes shaping protein evolution. *Sci Prog*. 2010;93:1–6.
- Sleator RD. Phylogenetics. *Arch Microbiol*. 2011;193:235–9.
- Sleator RD, Shortall C, Hill C. Metagenomics. *Lett Appl Microbiol*. 2008;47:361–6.
- Soltis DE, Soltis PS. The role of phylogenetics in comparative genetics. *Plant Physiol*. 2003a;132:1790–800.
- Soltis PS, Soltis DE. Applying the bootstrap in phylogeny reconstruction. *Stat Sci*. 2003b;18:256–67.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28:2731–9.
- Wrobel B. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. *J Appl Genet*. 2008;49:49–67.

PhyloPythia(S)

Alice C. McHardy

Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Definition

PhyloPythia and its successor *PhyloPythiaS* are fast and accurate oligomer signature-based classifiers for the taxonomic assignment of metagenome sequence fragments.

Introduction

Metagenomics uses random shotgun sequencing to recover genome sequence information from microbial communities without the need for cultivation of its member species. It thus gives access to the vast portion of the microbial world that cannot be cultured with standard techniques (Hugenholtz 2002). The sequencing of randomly sheared microbial community DNA initially generates a collection of short sequence fragments called reads. Depending on the sequencing technology used, the amount of generated data and read lengths vary (Metzker 2010; Droge and McHardy 2012): while traditional Sanger sequencing generates reads of around 800 bp, the commercially available “next-generation” sequencing technologies return reads of approximately 50–75 bp (SOLID sequencing by Applied Biosciences/Life Technologies), 75–300 bp (sequencing by synthesis technology by Solexa/Illumina), 100–200 bp (semiconductor chip sequencing by Ion Torrent/Life Technologies), and 550–1,000 bp (pyrosequencing by

454/Roche). The recently developed single-molecule sequencers produce read lengths of over 1 kb (PacBio SMRT) and of 5–10 kb (Oxford Nanopore technology). Currently, a single run of an Illumina HiSeq 2000 machine produces up to six billion paired-end reads or 600 Gb of sequence data (Illumina 2012).

Bioinformatics methods are subsequently applied to process the data. Assembly software such as MetaVelvet (Namiki et al. 2012) can be used to reconstruct longer contiguous sequence fragments, or contigs, based on overlaps in reads. For paired-end reads, the distances between reads originating from the two ends of an individual DNA fragment are approximately known. If paired-end reads are assembled into different contigs, the orientation of these contigs relative to each other and the size of the unassembled gap between them can be inferred. This ordering of contigs with gaps of known sizes is also referred to as a scaffold. The resulting sequence fragments, i.e., the contigs, scaffolds, and remaining unassembled reads, could principally originate from any member species of the microbial community.

In taxonomic assignment or “binning,” the fragments are assigned to individual species or higher-ranking clades (see Droge and McHardy (2012) for a recent review). The term “binning” was coined as a metaphor to describe the process of separating the fragment mixture by placing individual fragments into bins representing the different taxonomic origins. Besides variations caused by amplification bias of sequencing, the number of reads recovered for a community member should be approximately proportional to the product of its abundance and the size of its genome (Segata et al. 2012). Thus fragments are more likely to originate from the more abundant community members, which are more extensively covered by sequencing. Taxonomic assignment is different from taxonomic profiling for a metagenome. In profiling, the relative abundances of the different community members are estimated based on taxonomic assignment of either universal or clade-specific marker genes found on a subset of the sample fragments (Wu and Eisen 2008; Sharpton

et al. 2011; Segata et al. 2012; Wu and Scott 2012).

With the exception of highly complex communities, such as those found in soil, assembly and taxonomic assignment of metagenome samples sequenced to sufficient depth allows the reconstruction of draft genomes, corresponding to sets of contigs or scaffolds representing more than 50 % of a genome (Pope et al. 2010; Hess et al. 2011; Iverson et al. 2012). This enables a functional analysis and reconstruction of metabolic potential for individual community members. The annotation of assembled and unassembled metagenome fragments can be performed with publicly available servers such as MG-RAST, IMG/M, and CAMERA (Glass et al. 2010; Sun et al. 2011; Markowitz et al. 2012). In annotation, the presence and functionalities of genes and operons are identified and metabolic pathways reconstructed by comparing enzymes predicted to be encoded in these fragments with known reference pathways for model organisms.

In the following, the *PhyloPythia* and *PhyloPythiaS* software for the taxonomic assignment of metagenome sequence fragments are described.

Description

PhyloPythia and its successor *PhyloPythiaS* are oligomer signature-based classifiers for the taxonomic assignment of metagenome sequence fragments (McHardy et al. 2007; Patil et al. 2011). The methods are named after the *Pythia*, the priestess at Apollo’s oracle in ancient Delphi. They use the similarity in oligomer usage between a query sequence and a target clade as information. For prokaryotes, this allows to assign genome sequence fragments to species or higher-ranking taxonomic clades from which they originate. Oligomer- or composition-based taxonomic assignment differs from sequence similarity-based or phylogenetic methods in that global instead of local properties of the genome sequence are used as information. There is no requirement for homologous sequences of related

taxa to be known for every analyzed fragment. A fraction of a species' genome sequence, typically 100 kb or more, suffices as reference data. Reference data can be obtained by identifying contigs with conserved marker genes such as 16S rRNA from the sample itself or by additional sequencing of large insert libraries containing marker genes (Warnecke et al. 2007; Pope et al. 2010). Oligomer-based assignment therefore is advantageous for taxonomic assignment of metagenomes from microbial communities with few available sequenced genomes of its members or of related species. Oligomer-based taxonomic assignment is faster than alignment-based methods, as no sequence similarity searches in a large collection of reference sequences are required. This makes it well suited for the analysis of large next-generation sequence samples. For short fragments of less than 1 kb or for assignment over long taxonomic distances, homology-based methods tend to be more accurate (Patil et al. 2011). With *PhyloPythia*, the relative frequencies of 4–6 mer oligomer patterns with up to two wildcard characters in a sequence fragment are used as features to train ensembles of multi-class support vector machine classifiers with a Gaussian kernel for individual taxonomic ranks. These are subsequently combined for the assignment of variable length sequence fragments. *PhyloPythiaS* uses an ensemble of structured support vector machines with a linear kernel trained with the relative frequencies of 4–6 mer oligomers in sequence fragments. The structured output formulation allows to learn a classifier simultaneously for the entire taxonomy under consideration of commonalities of clades with partially shared evolutionary histories.

Summary

Metagenomics uses random shotgun sequencing to recover genome sequence information from microbial communities without the need for cultivation of its member species. It thus gives

access to the vast portion of the microbial world that cannot be cultured with standard techniques. Bioinformatics methods are subsequently applied to process the data. Assembly software is used to generate genomic sequence fragments, which could principally originate from any member species of the microbial community. In taxonomic assignment or “binning,” the fragments are assigned to individual species or higher-ranking clades from which they originate. *PhyloPythia* and its successor *PhyloPythiaS* are oligomer signature-based classifiers for the taxonomic assignment of metagenome sequence fragments. Oligomer signature-based taxonomic assignment is faster than alignment-based methods, as no sequence similarity searches in a large collection of reference sequences are required. Oligomer signature-based assignment is well suited for the taxonomic assignment of metagenomes from microbial communities with few available sequenced genomes of its members or of related species. For microbial community members with draft genomes reconstructed by taxonomic binning, a functional analysis based on gene content and reconstruction of metabolic potential can be performed.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Genome Portal, Joint Genome Institute](#)
- ▶ [KEGG and GenomeNet, New Developments, Metagenomic Analysis](#)

References

- Droge J, McHardy AC. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinforma.* 2012; 13(6):646–55.
- Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc.* 2010; 2010(1):pdb prot5368.
- Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, et al. Metagenomic discovery of

biomass-degrading genes and genomes from cow rumen. *Science*. 2011;331(6016):463–7.

Hugenholtz P. Exploring prokaryotic diversity in the genomic era. *Genome Biol*. 2002;3(2):REVIEWS0003.

Illumina. 2012. Available from: http://www.illumina.com/Documents/systems/hiseq/datasheet_hiseq_systems.pdf.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*. 2012;335(6068):587–90.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*. 2012;40(Database issue):D123–9.

McHardy AC, Garcia-Martin H, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*. 2007;4(1):63–72.

Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010;11(1):31–46.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*. 2012;40(20):e155.

Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, et al. Taxonomic metagenome sequence assignment with structured output models. *Nat Methods*. 2011;8(3):191–2.

Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, et al. Adaptation to herbivory by the tamar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci U S A*. 2010;107(33):14793–8.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9(8):811–4.

Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, et al. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput Biol*. 2011;7(1):e1001061.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res*. 2011;39(Database issue):D546–51.

Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. 2007;450(7169):560–5.

Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*. 2008;9(10):R151.

Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*. 2012;28(7):1033–4.

Plasmid Capture from Metagenomes

Brian V. Jones

Center for Biomedical and Health Science
Research, University of Brighton, School of
Pharmacy and Biomolecular Sciences, Brighton,
East Sussex, UK

Definitions

Metagenome: The collective genomes of all members of a bacterial community.

Mobile metagenome: The total pool of mobile genetic elements associated with a bacterial community.

Mobile genetic element (MGE): A discrete genetic unit capable of mediating its own transfer between distinct DNA molecules and/or between distinct host cells of the same or different species. Plasmids, transposons, insertion sequences, conjugative transposons, integrons, and bacteriophage are all examples of MGE.

Plasmid: Closed circular DNA molecule that replicates within host cells as an autonomous extrachromosomal element.

Plasmidome: Plasmid fraction of the mobile metagenome. May be defined as the total pool of plasmids associated with a microbial community and a component of the mobile metagenome as a whole.

Horizontal gene transfer: Transfer and acquisition of genetic material between distinct cells or species, outside of and in addition to the normal process of inheritance (vertical gene transfer).

Synonyms

Gut microbiota; Lateral gene transfer (LGT); Mobile microbiome

Introduction

Complex and diverse microbial ecosystems exist in a wide range of habitats ranging from aquatic and terrestrial environments, to those created on and within animals, plants, and other metazoans. The activities of these microbial consortia (microbiomes) contribute to important environmental processes such as nutrient cycling and bioremediation, while those associated with higher eukaryotic organisms are now widely recognized to be intimately involved in host health and aspects of host development (Ley et al. 2006; Jones 2010; Jones and Marchesi 2007a; Strom 2008).

However, members of both host-associated and “free-living” environmental microbiomes in turn play host to a wide range of mobile genetic elements (MGE) such as plasmids, transposons, and bacteriophage, which are now also being recognized as important components of these microbiomes (Jones and Marchesi 2007a; Jones et al. 2010; Jones 2010; Ogilvie et al. 2012; Kav et al. 2012; Reyes et al. 2010; Zhang et al. 2011). Collectively the total pool of MGE associated with a particular microbial ecosystem is referred to as its mobile metagenome (Jones and Marchesi 2007a, b), and there is increasing interest in understanding how this versatile, and dynamic reservoir of genes and genetic elements is involved in the development of these ecologies. For host-associated microbiomes there is also the added dimension of how the mobile metagenome of a particular ecosystem may impact on the health of the higher eukaryotic host, either through effects on the host microbiome or directly through the functions encoded by constituent MGE (Jones 2010; Ogilvie et al. 2012; Ley et al. 2006).

Moreover, MGE have also been proposed to facilitate the spread of beneficial functions within a bacterial community (Jones and Marchesi

2007a; Jones 2010; Lozupone et al. 2008; Heuer and Smalla 2012; Ley et al. 2006). MGE are capable of moving between distinct molecules of DNA and/or host cells and are also well documented to acquire new genetic material from host bacteria and subsequently disseminate this to other species. This feature of MGE facilitates the exchange and maintenance of genetic material between diverse species, a process termed horizontal gene transfer (HGT). HGT allows cells to rapidly acquire new genes and activities which facilitates adaptation to new environments, and the formation of new functional pathways, and is believed to be a pivotal factor in the evolution and diversification of bacteria (Ochman et al. 2000; Heuer and Smalla 2012; Jones and Marchesi; Jones 2010).

This is of particular relevance to host-associated ecosystems, such as the human microbiome, where HGT is proposed to have played a key role in stabilizing the functional output of such ecosystems. For example, in the human gut microbiome dissemination of key traits to multiple species in the community through HGT is thought to generate functional redundancy and protect against loss of important activities from the community as a whole (Ley et al. 2006; Lozupone et al. 2008; Jones and Marchesi 2007a; Jones 2010). In this context, it is notable that the human microbiome has now been shown to support an emergent and extensive network of gene exchange (with the highest rates of transfer observed in the gut microbiome) (Smillie et al. 2011), and it seems likely that MGE forge the majority of connections within this network.

Plasmids and Plasmidomes

Of the numerous types of MGE that will make up a particular mobile metagenome, those capable of autonomous cell to cell transfer are of special interest. Plasmids in particular are believed to be highly important in this regard and to be prevalent in many bacterial ecosystems. Not only are plasmids frequently capable of mediating their own transfer between distinct and diverse

bacterial species, but also act as vehicles for other MGE, and are known to encode a diverse array of accessory functions, including those relevant to health of higher eukaryotic organisms (Reviewed in Ogilvie et al. 2012; Ochman et al. 2000; Smalla et al. 2000a). Functions encoded by plasmids include virulence factors, antibiotic resistance determinants, bacteriocines, nutrient acquisition and utilization, and degradation of xenobiotic compounds, as well as factors that mediate tolerance of a wide range of physical parameters (reviewed in Ogilvie et al. 2012; Ochman et al. 2000; Smalla et al. 2000a; Heuer and Smalla 2012).

Plasmids are covalently closed circular molecules of DNA which replicate as extrachromosomal elements in the cytoplasm, independently of the host cell chromosome. The copy number of different plasmids can vary considerably, ranging from 1 to 2 copies per cell for some plasmids to several hundred copies per cell for others (Espinosa et al. 2000; Novick 1987). This variation in copy number contributes to gene dosage effects for plasmid encoded genes, potentially increasing the output from plasmid encoded activities.

The size and gene content of these elements is also highly variable and ranges from small cryptic plasmids encoding no obvious functions outside of those essential for replication and maintenance to large mega plasmids of several hundred kilobases, which encode a diverse array of activities (Espinosa et al. 2000; Novick 1987; Heuer and Smalla 2012). Typically, larger plasmids are present in low copy number as they present a greater metabolic burden to host cells and often also encode all machinery necessary to initiate their own transfer between host cells via conjugation.

Plasmids are classified into distinct families, generally distinguished based on their ability to coexist and replicate within the same host cell (incompatibility groups) and the sequence homology of their replication machinery (Espinosa et al. 2000; Novick 1987). However, from studies of the established and well-characterized plasmid families, it is clear that plasmid genomes are highly diverse in nature,

with plasmids in a particular family exhibiting a high degree of similarity around regions involved in basic replication and maintenance (the plasmid “backbone,” or core replicon) but considerable variation in overall size and gene content. Many plasmids have also been described as possessing a modular organization, with essential backbone functions and accessory genes organized as distinct gene clusters (Schlüter et al. 2007; Heuer and Smalla 2012). This modularized genome architecture affords plasmids a high degree of genetic flexibility in terms of gene loss or recruitment and is consistent with the diversity of plasmids and functions represented within a particular plasmid family.

Considering the diversity of the prokaryotic world and the relatively small numbers of plasmids characterized to date, it is clear that our knowledge of these elements remains limited. In conjunction with the insights into microbial ecology and diversity provided by the application of molecular genetic approaches (such as metagenomics) to the study of microbial communities, this has prompted many researchers to adopt a broader view of plasmids (and other MGE) associated with a particular microbiome (Jones et al. 2010; Ogilvie et al. 2012; Kav et al. 2012; Zhang et al. 2011). This shifts the emphasis to the global population of plasmids resident in a given ecosystem and the collective functions and activities they encode, giving rise to the concept of the plasmidome (Kav et al. 2012). The plasmidome refers to the total pool of plasmids associated with a particular mobile metagenome, and may be thought of as a distinct component of the mobile metagenome as a whole.

Accessing the Plasmidome

Plasmids are probably the best studied MGE, and a range of strategies exist to specifically recover and characterize these genetic elements (reviewed in Ogilvie et al. 2012). These include approaches that have been specifically designed to permit community-level analysis of microbial plasmidomes and to capture and analyze

plasmids from the non-cultivable fraction of microbial communities, which account for the vast majority of bacterial species in these ecosystems. The development of such tools has been, and will continue to be, a major challenge with current approaches each exhibiting distinct strengths and weaknesses when applied to community-level analysis of plasmids (Summarized in Table 1).

A particular issue faced by all approaches to survey microbial plasmidomes, as well as other facets of a given mobile metagenome, is the difficulty in evaluating the ability of any method to provide universal access to the plasmidome and identify any bias in the plasmids that may be identified and recovered (Ogilvie et al. 2012). Unlike analysis of the core chromosomal content of a microbiome, where detailed surveys of

Plasmid Capture from Metagenomes, Table 1 Relative merits of approaches available for analysis of microbial plasmidomes and plasmid capture from metagenomes (Modified from Ogilvie et al. 2012)

Plasmid isolation strategy	Advantages	Disadvantages	Reference
Endogenous isolation	<ul style="list-style-type: none"> • Original bacterial host is known • May be used for all cultivatable bacteria • Applicable to all plasmid types 	<ul style="list-style-type: none"> • Requires host cultivation restricting utility for study of natural communities • Reliance on plasmid encoded traits if surrogate host species required for plasmid characterization 	Reviewed in Smalla and Sobczyk (2002) Heuer and Smalla (2012) Ogilvie et al. (2012)
Exogenous isolation	<ul style="list-style-type: none"> • Culture independent • Selective isolation of self-transmissible or mobilizable elements • Potentially capable of isolating all plasmid types (circular and linear) and sizes • Can isolate plasmids irrespective of abundance in community 	<ul style="list-style-type: none"> • Relies on plasmid encoded traits for plasmid transfer, selection, and maintenance in surrogate host • Original bacterial host unknown • Range of plasmids isolated dependent on mating conditions used and dictated by numerous “unknown” environmental variables influencing host cell physiology and plasmid transfer kinetics 	Bale et al. (1988)
PCR-based detection	<ul style="list-style-type: none"> • Culture independent • High throughput • Sensitive • Scope for accurate quantitation of plasmids 	<ul style="list-style-type: none"> • Original bacterial host unknown • Complete characterization of plasmid detected generally impossible • Limited to detection of known and characterized plasmid lineages used for primer design 	Götz et al. (1996)
TRACA	<ul style="list-style-type: none"> • Culture independent • Suitable for development of high-throughput strategies • Can isolate plasmids irrespective of abundance in a community • Fully independent of plasmid encoded traits • Sequence-based characterization of plasmids facilitated by known Tn sequence in plasmids • Potentially applicable to all circular plasmids and bacterial communities • May permit capture of MGE other than plasmids when present as circular DNA molecules 	<ul style="list-style-type: none"> • Original bacterial host unknown • Transposon may inactivate genes of interest, impeding phenotypic characterization • Currently available Tn elements and surrogate host may limit range of plasmids isolated • Linear plasmids not captured • Transformation step may introduce size bias • Plasmids belonging to same incompatibility group as Tn origin may not be captured due to stability issues in surrogate host • Potential for bias towards numerically dominant plasmids 	Jones and Marchesi (2007b) Jones et al. (2010) Warburton et al. (2011) Zhang et al. (2011)

(continued)

Plasmid Capture from Metagenomes, Table 1 (continued)

Plasmid isolation strategy	Advantages	Disadvantages	Reference
Standard metagenomic libraries (BAC/Fosmid)	<ul style="list-style-type: none"> • Culture-independent • Suitable for development of high-throughput strategies • Initial capture independent of plasmid encoded traits • Sequence-based characterization facilitated 	<ul style="list-style-type: none"> • Original bacterial host unknown • Likely bias towards numerically dominant plasmids • Screening relies on plasmid encoded traits expressed in surrogate host species • Not specifically designed for plasmid capture, and non-plasmid sequences dominate libraries • Generally only incomplete, partial plasmids identified • General compatibility of library construction methods with plasmid capture unknown • Plasmids belonging to same incompatibility group as vector (BAC/Fosmid) may not be represented due to instability of clones in surrogate host • Plasmids belonging to same incompatibility group as vector may not be captured due to stability issues in surrogate host 	Kazmierczak et al. (2009)
Shotgun sequencing of plasmidomes	<ul style="list-style-type: none"> • Culture-independent • Suitable for development of high-throughput strategies • Independent of plasmid encoded traits • Potential for complete access to circular elements within a bacterial plasmidome 	<ul style="list-style-type: none"> • Original bacterial host unknown • Removal of contaminating chromosomal DNA potentially problematic • Not suitable for survey of linear plasmids with present strategies for removal of chromosomal DNA • Accurate assembly of complete plasmids will likely require a more comprehensive set of reference plasmid genomes than presently available • Pre-sequence processing of plasmid DNA (removal of chromosomal fragments and plasmid DNA amplification) likely to introduce bias into final dataset. Requires subsequent quantitative analysis to confirm relative abundance of particular plasmids 	Zhang et al. (2011) Kav et al. (2012)

population structure can be first undertaken using conserved housekeeping genes present in all bacterial chromosomes (such as genes encoding 16S rRNA), no such global survey is possible for plasmids (Ogilvie et al. 2012). As such, surveys of microbial plasmidomes are impeded by a fundamental lack of knowledge regarding the composition of these malleable gene pools, making the development and validation of methods which access a representative cross section of the plasmidome virtually impossible at present. Nevertheless, available strategies still offer the potential to provide much insight into microbial

plasmidomes and have been applied to study a range of microbial ecosystems yielding important fundamental insights into the composition and functional content of associated plasmid pools.

Endogenous isolation: The simplest and most widely used approach to study plasmids is the direct isolation of plasmid DNA from host bacteria. This approach classically involves the cultivation of host species, usually with selection for particular traits of interest believed to be plasmid encoded (reviewed in Smalla and Sobczyk 2002; Jones and Marchesi 2007a; Ogilvie et al. 2012).

Extracted plasmid DNA is subsequently transferred into a new host, ideally of the same species, with *E. coli* K12-type strains most commonly deployed. Plasmids are then typically characterized based on the phenotypes they confer upon host species but are increasingly examined at the nucleotide level, and plasmids sequenced as part of whole genome sequencing projects may also be considered as examples of endogenous isolation.

Aside from its simplicity and general applicability to all plasmid types (including linear plasmids), the major benefit of this approach is the identification of the natural hosts species of a particular plasmid. Conversely, the reliance on cultivation of host bacteria, as well as the reliance on plasmid encoded traits and their expression in surrogate hosts species (selectable markers and plasmid replication machinery), severely restricts the utility of this approach for access to the plasmidome. However, the general strategy of direct isolation of plasmid DNA from host cells can also be applied to the total community without prior cultivation, and when combined with high-throughput sequencing or other culture-independent approaches, this direct extraction method forms the basis for many “metagenomic” strategies for plasmidome analysis (discussed in detail below).

Exogenous isolation: Exogenous isolation approaches were the first to address some of the limitations inherent in endogenous approaches for community level analysis of plasmidomes (Bale et al. 1988; Hill et al. 1992). Exogenous methods rely on the natural ability of plasmids to initiate or participate in cell-cell transfer between distinct host species. This strategy accesses plasmids using a selectable surrogate host species (most typically *E. coli*) in biparental or tri-parental matings with the donor population, during which plasmids may be transferred from donor cells in the community to the selectable recipient (Fig. 1; Bale et al. 1988; Hill et al. 1992; Reviewed in Ogilvie et al. 2012; Smalla and Sobecky 2002; Heuer and Smalla 2012). Essentially this system utilizes the surrogate host as a “fishing net,” to pick up plasmids circulating within the donor community under study, and

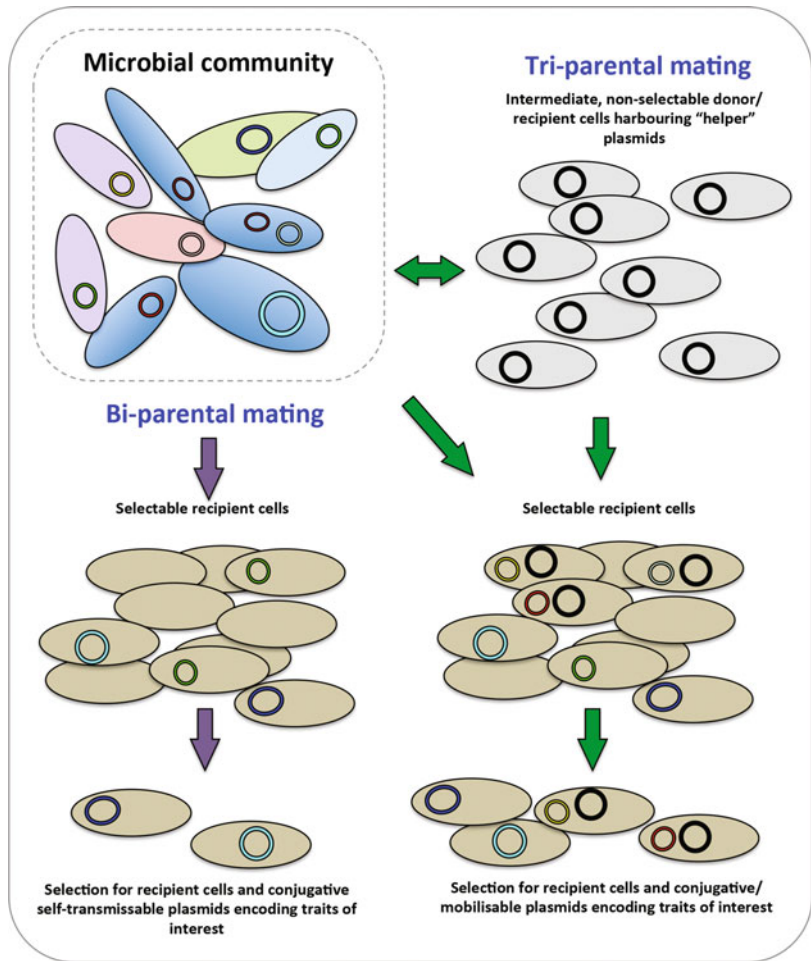
plasmid carrying recipient cells are subsequently identified by cultivation on media selectable for the recipient organism (often rifampicin resistance), as well as plasmid encoded traits.

Biparental matings, involving only the donor community and selectable recipient, can be used to retrieve self-transmissible plasmids capable of initiating autonomous conjugal transfer processes (Fig. 1). Alternatively donor cells carrying a “helper” plasmid may also be introduced along with the selectable recipient, in a tri-parental mating approach (Hill et al. 1992). In this case, the “helper” plasmid sets up plasmid conjugation apparatus, which can subsequently be exploited by plasmids that may be mobilized between cells, but are not capable of independent transfer (Fig. 1). In particular, the retrieval of self-transmissible elements may be seen as a strength of the exogenous isolation approach, since these elements are likely to be the most informative and important in understanding MGE-mediated prokaryotic gene flow both within and between microbiomes.

Although this method offers a number of significant advantages over endogenous approaches, the capture of plasmids is still reliant on plasmid encoded traits, including the presence of selectable markers, as well as the ability of plasmids to successfully replicate in the surrogate host species used (Ogilvie et al. 2012; Smalla and Sobecky 2002; Heuer and Smalla 2012). Plasmids lacking in traits selected for, or unable to replicate successfully in surrogate hosts, will not be captured using these approaches. In addition, the cell-cell transfer of plasmids is influenced by numerous environmental variables, as well as the physiological status of donor and recipient cells, with metabolically inactive community members unlikely to participate in conjugal transfer processes. These factors also impact on plasmid transfer rates, the types of plasmid that can be acquired and the portion of the plasmidome that may be accessed (Ogilvie et al. 2012; Smalla and Sobecky 2002; Heuer and Smalla 2012). Collectively, these factors restrict the range of plasmids that may be captured and limit the utility of this approach for studying microbial plasmidomes.

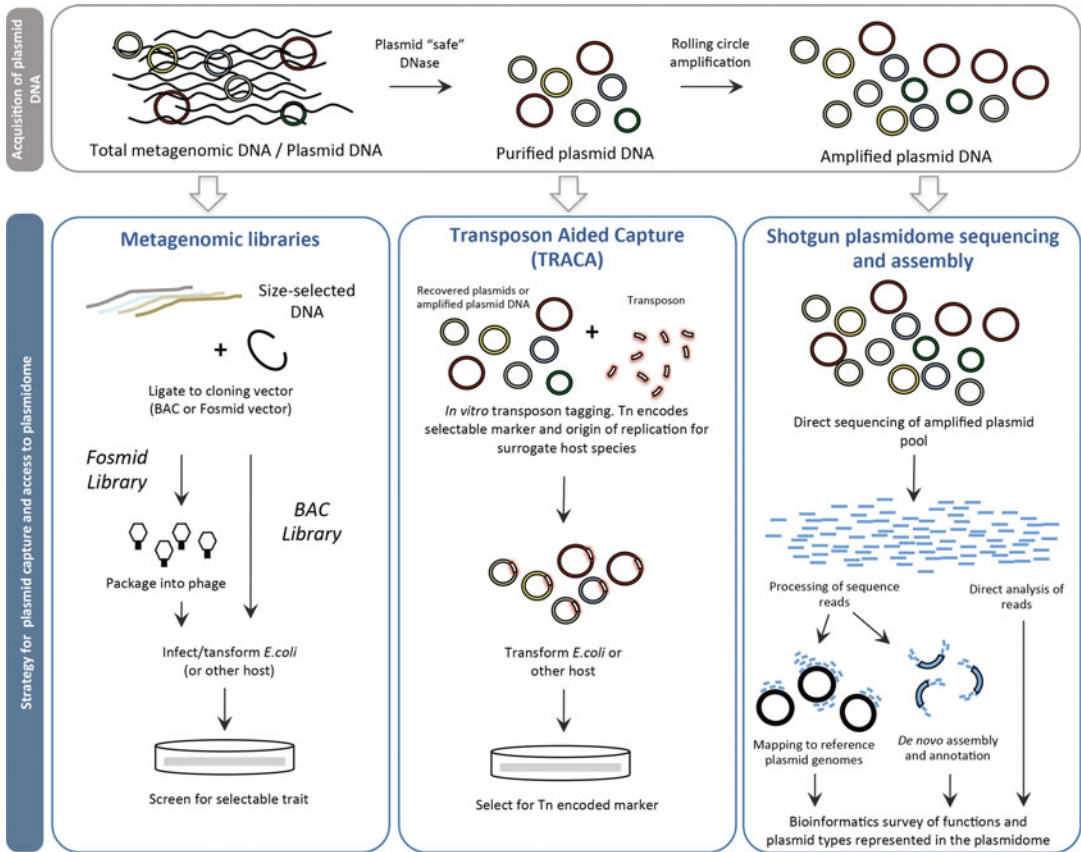
Plasmid Capture from Metagenomes,

Fig. 1 Overview of exogenous isolation approaches for the acquisition of plasmids from microbial communities. *Arrows* indicate plasmid transfer between donor (mixed microbial community), recipient, and “helper” populations. *Purple arrows* indicate plasmid transfer in biparental matings in which selectable recipient cells are used to acquire self-transmissible plasmids directly from the donor population. *Green arrows* indicate transfer events in tri-parental matings, in which cells harboring a self-transmissible “helper” plasmid are utilized to initiate conjugal transfer events with the donor population and the selectable recipient, in order to acquire mobilizable but non-self-transmissible plasmids



Direct plasmid detection by PCR: A range of PCR primers have been developed in order to distinguish between plasmids of different families based on backbone sequences, but these have also been employed as surveying tools to identify the presence of particular plasmid types in total community DNA extracts (Götz et al. 1996; Smalla et al. 2000b). While this approach is potentially useful in gaining an overview of the types of plasmids comprising a particular plasmidome and their relative abundance (if utilized with a quantitative PCR strategy), its usefulness is currently limited by the relatively small number of plasmid genomes available from which discriminatory primer sets may be established.

This limits the range of plasmids encompassed in such surveys to those families already isolated and characterized. A further disadvantage is that along with a lack of data on host range, no information on functional content of plasmids is offered by this method, and there is little or no scope to characterize detected plasmids in greater detail. As such, this approach does not at present constitute a viable strategy for in depth and comprehensive analysis of entire plasmidomes, but may be used to augment other strategies and provide further information on isolated plasmids. Despite the present limitations, the usefulness of this approach is likely to grow as more sequence information and associated data is generated, and greater numbers of habitat associated reference data sets become available in the future.



Plasmid Capture from Metagenomes,

Fig. 2 Overview of culture-independent metagenomic approaches for microbial plasmid analysis. Acquisition of plasmid DNA: Plasmid DNA may be harvested and processed in a number of ways before use in strategies to capture plasmids or access the plasmidome. Plasmid DNA may be acquired from either total metagenomic DNA extracts of the microbial community or specific plasmid extraction methods. Recovered pools of plasmids may subsequently be processed to remove contaminating

chromosomal sequences and to amplify the recovered plasmid DNA for certain plasmidome if necessary. Plasmid capture and plasmidome access: Recovered plasmidome extracts may then be used in conjunction with one or more culture-independent approaches for plasmid capture or general access to the plasmidome. Available culture-independent approaches include the generation of standard metagenomic libraries, the use of the TRACA plasmid capture approach, or direct shotgun sequencing of amplified plasmids

Transposon aided capture (TRACA): The culture-independent transposon-aided capture system (TRACA) has been specifically designed for the acquisition of plasmids from whole communities and to overcome some of the main limitations of endogenous and exogenous approaches (Jones and Marchesi 2007b). The basic premise of this system is to retrofit all plasmids with a suitable selectable marker and an origin of replication compatible with the surrogate host biomachinery, using an *in vitro*

transposon (Tn) system encoding this information (Fig. 2).

Following Tn integration, plasmids are subsequently transformed into a surrogate bacterial host and cells carrying plasmids selected for based on antibiotic resistance genes harbored by the inserted Tn. In this way, plasmids may be acquired independently of the traits they encode, and their replication in the surrogate host is facilitated (Jones and Marchesi 2007b). This provides access to plasmids in a bacterial community

regardless of functions encoded and has been successfully applied to study plasmids in a number of environments, including the human gut, the oral cavity, and activated sludge (Jones and Marchesi 2007b; Warburton et al. 2011; Zhang et al. 2011).

Although the TRACA system offers major advantages over other approaches, this method does not circumvent all issues and may be subject to a unique limitation in regard to the size of plasmids that can be captured when using this approach (reviewed in Jones and Marchesi 2007a; Ogilvie et al. 2012). Plasmids isolated by this system to date have all been in the smaller size range (~14 Kb and smaller), indicating the TRACA system may be biased towards the capture of small plasmids or even unable to acquire larger plasmids altogether. The reasons behind this potential size restriction are presently unclear, although the transformation step in which Tn-tagged plasmids are introduced into surrogate host cells is known to work more efficiently with smaller DNA molecules, and there is also potential for a size bias to be introduced during the purification of plasmid DNA (Jones and Marchesi 2007b).

It is also possible that the size range of plasmids captured by this system will be a function of the plasmidome composition and the predominance of smaller plasmids in the ecosystems that have been explored with this method to date (Ogilvie et al. 2012). Although there is presently no definitive data available on the average plasmid size in any given microbial ecosystem, initial evidence suggests that physical features of plasmids, such as size, are responsive to pervading environmental and ecological conditions in the same way as host chromosomes (Slater et al. 2008). Overall, it is most probable that both the composition of the plasmidome and inherent attributes of the TRACA system dictate the profile of plasmids captured by this approach. Regardless of these potential limitations, the TRACA method provides an additional and useful tool for the exploration of bacterial plasmidomes, overcoming some of the major disadvantages of other methods. There is also much scope to improve the existing TRACA approach

and expand the range of plasmids that may be acquired with this system.

Retrieval of plasmids from standard metagenomic libraries: Access to plasmid sequences contained in standard metagenomic libraries derived from total community DNA have also been described (Fig. 2; Kazimierczak et al. 2009). In particular, the isolation of plasmids or plasmid fragments, from such libraries of the organic pig gut microbiome, has been demonstrated and included those with the ability to replicate autonomously when liberated from the library vector and reconstructed by self-ligation (Kazimierczak et al. 2009). Despite the novelty of this approach, this strategy suffers from the same drawbacks as endogenous and exogenous methods in its reliance on plasmid encoded traits for initial plasmid identification and subsequent demonstration of autonomous replication in surrogate host species (Kazimierczak et al. 2009).

Furthermore, this approach is not at present designed to specifically retrieve plasmids, but rather total community DNA which is dominated by chromosomal sequences. As such this approach is not presently suitable for the specific analysis of microbial plasmidomes, and in the original study by Kazimierczak et al. (2009), libraries were analyzed for clones encoding antibiotic resistance genes, rather than plasmid sequences per se. However, there is clearly scope to utilize this method to further explore existing metagenomic data sets and enhance the interpretation of these valuable resources by illuminating mobile genetic elements captured in these repositories.

Shotgun sequencing of plasmidomes: More recently the first true applications of the metagenomic approach to study plasmidomes have been described (Fig. 2; Zhang et al. 2011; Kav et al. 2012). In these studies, plasmid DNA was extracted from the target community without any prior enrichment or cultivation, subjected to high-throughput sequencing, and fragments of plasmid genomes subsequently assembled from the resulting reads (Zhang et al. 2011; Kav et al. 2012). This permitted a global survey of plasmid-encoded functions present in the bovine plasmidome (Kav et al. 2012), as well as an

activated sludge microbial community (Zhang et al. 2011), demonstrating proof of principal for the shotgun sequencing approach to plasmidome analysis.

Although this approach should in theory be able to offer total and unbiased access to the entire plasmidome of a given microbial community, in practice limitations and potential biases remain. For example, in the study by Kav et al. 2012, sufficient plasmid DNA for sequencing was only obtained after amplification of the recovered plasmid DNA by rolling circle amplification. As such there is potential for some plasmids to be preferentially amplified over others, introducing bias into the resulting data set. In addition, the complete removal of contaminating chromosomal sequences is also challenging, and despite the availability of “plasmid safe” DNases which do not act on circular molecules, total elimination of chromosomal DNA from plasmid extracts appears to constitute a bottleneck in this strategy (Zhang et al. 2011; Kav et al. 2012), with linear plasmids also likely to be removed during this process. As such there is further potential to alter the composition of the plasmid pool obtained during this stage of plasmid DNA preparation.

There is also potential for errors in assembly due to the mosaic nature of these elements, a situation that may be exacerbated by the presence of any contaminating chromosomal sequences. In this regard, the availability of reference plasmid genomes captured by methods which acquire whole, intact plasmids (such as exogenous isolation and TRACA) will constitute a highly valuable resource that will significantly enhance the power and accuracy of the shotgun plasmidome approach (Fig. 2), and some researchers have already begun to combine these strategies (Zhang et al. 2011). Finally, extensive sequencing will likely be required for most plasmidomes, in order to move beyond representation of numerically dominant plasmids (particularly for assembly of complete replicons) and provide the depth of coverage required to access the full diversity of a given plasmidome.

Despite these potential issues, it is clear that the shotgun sequencing approach to plasmidome

analysis constitutes a major advance in accessing plasmids resident in microbial communities, in terms of both depth of coverage and the cross section of plasmids that may be covered. Further development of such approaches, in parallel with the development of more detailed and extensive reference data sets from plasmids captured through TRACA or exogenous approaches, for the first time places the comprehensive analysis of a microbial plasmidome within reach.

Retrieval of Host Range Data Following Plasmid Capture from Metagenomes

A major drawback of all culture-independent community-level approaches for investigation of microbial plasmidomes, and capture of plasmids from metagenomic data sets, is the loss of host range data inherent in these strategies (Table 1). All such strategies effectively divorce acquired plasmids or plasmid sequences of any phylogenetic affiliation, undermining a primary motivation for undertaking many such surveys: a fundamental understanding of gene flow in these communities. Despite this, several approaches may be used to supplement the initial culture-independent plasmid capture strategy and provide some indication of plasmid phylogenetic affiliation and long-term host range.

Plasmids captured through culture independent approaches may subsequently be utilized to develop fluorescent probes suitable for use in fluorescence associated cell sorting (FACS) applications (reviewed in Ogilvie et al. 2012). The development and use of such probes in FACS systems permits intact cells harboring target genes or sequences to be separated from the rest of the microbial community and subsequently identified through culture-independent molecular genetic approaches, such as 16S rDNA sequence analysis (Zwirgmaier et al. 2004). This strategy, termed Ring-FISH (recognition of individual genes by fluorescence in situ hybridization), has previously been implemented and demonstrated as a feasible approach for the recovery of cells encoding genes of interest, including those encoded by plasmids.

Alternatively a range of *in silico* approaches have been applied to plasmid host affiliation (reviewed in Ogilvie et al. 2012). Plasmid sequences may be compared directly to curated sequence databases where phylogenetic information on plasmid genomes and other genes is available. The homology of plasmid sequences to database entries may then be used to infer phylogeny of captured plasmids (Jones and Marchesi 2007b; Jones et al. 2010; Kav et al. 2012; Zhang et al. 2011). However, the mosaic nature of plasmids and the potential for a single element be composed of genetic material with highly diverse origins, coupled with inherent biases in public databases due to the paucity of available plasmid genomes, undermines the accuracy of this approach and particularly when applied to fragmentary data sets such as metagenomic libraries and shotgun plasmidomes.

Alternatively, strategies based on correlation of nucleotide usage patterns in plasmids with bacterial chromosomes have also been described (Campbell et al. 1999; Suzuki et al. 2010). These are based on the premise that over time, plasmids and other MGE that are long-term residents of a given host species adapt to their host at the nucleotide level and acquire a corresponding “genomic signature” in terms of nucleotide usage profiles (Campbell et al. 1999; Suzuki et al. 2010). As this underlying genomic signature has been shown to permit discrimination between chromosomal sequences of different bacterial species, there is also scope to employ plasmid nucleotide usage patterns to retrieve host range information. Dinucleotide and trinucleotide usage patterns, based on the abundance of all possible two-nucleotide or three-nucleotide combinations in a given DNA sequence, have been used in this way and shown to provide insight into plasmid host range, at least in terms of potential long-term bacterial host species to which plasmids are well adapted (Campbell et al. 1999; Suzuki et al. 2010). There is much scope to incorporate such analyses into culture-independent surveys of bacterial plasmidomes, as downstream processing steps that may provide some of the phylogenetic inference lacking in metagenomic approaches.

Summary

There is now much evidence to support the concept of distinct, community-associated plasmidomes and wider mobile metagenomes (reviewed in Jones 2010; Ogilvie et al. 2012). However, the mobile and promiscuous nature of many MGE (including many plasmids) makes this a much less clearly defined genetic reservoir, and membership of a particular mobile metagenome will be far less exclusive than for the core chromosomal complement of the associated microbiome (Jones 2010). A greater understanding of the composition and functional capacities of these mobile metagenomes, and key MGE such as plasmids, will be important for understanding and ultimately manipulating many important microbial ecosystems, as well as providing fundamental insight into the mechanisms of gene flow within and between distinct microbiomes. Although no available method for accessing microbial plasmidomes represents a panacea for the study of these dynamic gene pools, the application of tools currently available, particularly when used in combination, holds much potential for greatly expanding our knowledge of plasmid diversity, abundance, and functionality within microbial mobile metagenomes.

References

- Bale MJ, Day MJ, Fry JC. Novel method for studying plasmid transfer in undisturbed river epilithon. *Appl Environ Microbiol.* 1988;54(11):2756–8.
- Campbell A, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci U S A.* 1999;96: 9184–9.
- Espinosa M, Cohen S, Couturier M, et al. Plasmid replication and copy number control. In: CM Thomas (ed) *The horizontal gene pool, bacterial plasmids and gene spread.* Amsterdam: Harwood Academic Publishers; 2000. p. 207–48.
- Götz A, Pukall R, Smit E. Detection and characterization of broad-host-range plasmids in environmental bacteria by PCR. *Appl Environ Microbiol.* 1996;63:1980–6.
- Heuer H, Smalla K. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol Rev.* 2012. doi:10.1111/j.1574-6976.2012.00337.x.

- Hill K, Weightman AJ, Fry JC. Isolation and screening of plasmids from the epilithon which mobilise recombinant plasmid pD10. *Appl Environ Microbiol.* 1992; 58:1292–300.
- Jones BV. The human gut mobile metagenome: a metazoan perspective. *Gut Microbes.* 2010;1(6): 417–33.
- Jones BV, Marchesi JR. Accessing the mobile metagenome of the human gut microbiota. *Mol Biosyst.* 2007a;3:749–58.
- Jones BV, Marchesi JR. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat Methods.* 2007b;4:55–61.
- Jones BV, Sun F, Marchesi JR. Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. *BMC Genomics.* 2010; 11:46.
- Kav AB, Sasson G, Jami E, et al. Insights into the bovine rumen plasmidome. *Proc Natl Acad Sci U S A.* 2012; 109:5452–7.
- Kazimierczak KA, Scott KP, Kelly D, Aminov RI. Tetracycline resistome of the organic pig gut. *Appl Environ Microbiol.* 2009;75:1717–22.
- Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell.* 2006;124:837–48.
- Lozupone CA, Hamady M, Cantral BL, et al. The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci U S A.* 2008;105:15076–81.
- Novick RP. Plasmid incompatibility. *Microbiol Rev.* 1987;51:381–95.
- Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature.* 2000;405:299–304.
- Ogilvie LA, Firouzmand S, Jones BV. Evolutionary, ecological and biotechnological perspectives on plasmids resident in the human gut mobile metagenome. *Bioeng Bugs.* 2012;3(1):1–19.
- Reyes A, Haynes M, Hanson N, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature.* 2010;466:334–8.
- Schlüter A, Szczepanowski R, Pühler A, et al. Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance pool. *FEMS Microbiol Rev.* 2007;31:449–77.
- Slater FR, Bailey MJ, Tett AJ, Turner SL. Progress towards understanding the fate of plasmids in bacterial communities. *FEMS Microb Ecol.* 2008;66:3–13.
- Smalla K, Sobczyk PA. The prevalence and diversity of mobile genetic elements in bacterial communities of different environmental habitats: insights gained from different methodological approaches. *FEMS Microbiol Ecol.* 2002;42:165–75.
- Smalla K, Osborne AM, Wellington EMH. Isolation and characterisation of plasmids from bacteria In: CM Thomas (ed) *The horizontal gene pool, bacterial plasmids and gene spread.* Amsterdam: Harwood Academic Publishers; 2000a. p. 207–48.
- Smalla K, Krögerrecklenfort E, Heuer H, et al. PCR-based detection of mobile genetic elements in total community DNA. *Microbiology.* 2000;146:1256–7.
- Smillie CD, Smith MB, Friedman J, et al. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature.* 2011. doi:10.1038/nature10571.
- Strom SL. Microbial ecology of ocean biogeochemistry: a community perspective. *Science.* 2008;320:1043–5.
- Suzuki H, Yano H, Brown CJ, Top EM. Predicting plasmid promiscuity based on genomic signature. *J Bacteriol.* 2010;192(22):6045–55.
- Warburton P, Allan E, Hunter S, et al. Isolation of bacterial extra-chromosomal DNA from human dental plaque associated with periodontal disease, using transposon-aided capture (TRACA). *FEMS Microbiol Ecol.* 2011;78:349–54.
- Zhang T, Zhang X-X, Ye L. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PloS ONE.* 2011;6:e26041.
- Zwirgmaier K, Ludwig W, Schleifer KH. Recognition of individual genes in a single bacterial cell by fluorescence in situ hybridization – RING-FISH. *Mol Microbiol.* 2004;51(1):89–96.

Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies

Martin Wu
Department of Biology, University of Virginia,
Charlottesville, VA, USA

Synonyms

Automated Phylogenomic Inference Application (AMPHORA)

Introduction

The small ribosomal unit RNA (SSU rRNA or 16S rRNA) has been widely used in microbial systematic and diversity studies. The appeal of using 16S rRNA gene as a marker gene is numerous. First of all, it is distributed in every single

cellular organism. Secondly, because regions of 16S rRNA sequence are highly conserved, 16S rRNA gene can be PCR amplified from a wide diversity of taxa using “universal” primers and sequenced, bypassing the need to isolate and culture the organisms in question. Consequently, millions of 16S rRNA reference sequences are available for microbial classification and identification (Cole 2009).

Although 16S rRNA has been the “gold standard” in microbial diversity studies, it has several shortcomings. First, because 16S rRNA only makes up a tiny fraction of a genome (~0.1 %), its application as a marker gene in classifying metagenomic sequences is seriously limited. Secondly, the widely recognized bias in 16S rRNA PCR skews the estimation of the relative abundance of species in a population (Acinas et al. 2005). Thirdly, the 16S rRNA gene copy number varies substantially from species to species, further complicates the effort to accurately estimate microbial composition (Kembel et al. 2012). To circumvent these problems, protein-coding genes such as *rpoB*, *pyrG*, *recA*, and *HSP70* have been used as alternative phylogenetic markers to complement rRNA-based analyses (Ludwig and Klenk 2000; Santos and Ochman 2004). Because protein genes are conserved at the amino acid level and not at the nucleotide level, they evolve faster and thus have more power at resolving the relationships of closely related species than the 16S rRNA gene. Unfortunately for the same reason, it is extremely difficult to design “universal primers” that can be used to PCR amplify protein-coding genes from distantly related species (Santos and Ochman 2004). As a result, protein-coding genes have seen very limited use in broad-spectrum microbial surveys.

Recent explosive growth in genomic sequences has changed the landscape. Thousands of complete bacterial genomes are available and many more are on the way of being sequenced (Pagani et al. 2012). With each genome sequence come along thousands of protein-coding genes, vastly expanding the amount of data available for protein marker genes. In metagenomic studies, genomes of a mixed microbial population are

sequenced directly from environments without prior isolation, culturing, and PCR amplification. Metagenomics therefore overcomes a major hurdle for using protein genes for microbial diversity studies in that it makes the sequences of protein genes readily accessible. Because metagenomic sequencing is random in nature, microbial composition estimated based on metagenomic sequencing is less biased than the 16S rRNA PCR-based survey. When using single-copy protein-coding genes for relative species abundance estimation, it further eliminates the bias associated with the copy-number variations of the 16S rRNA gene.

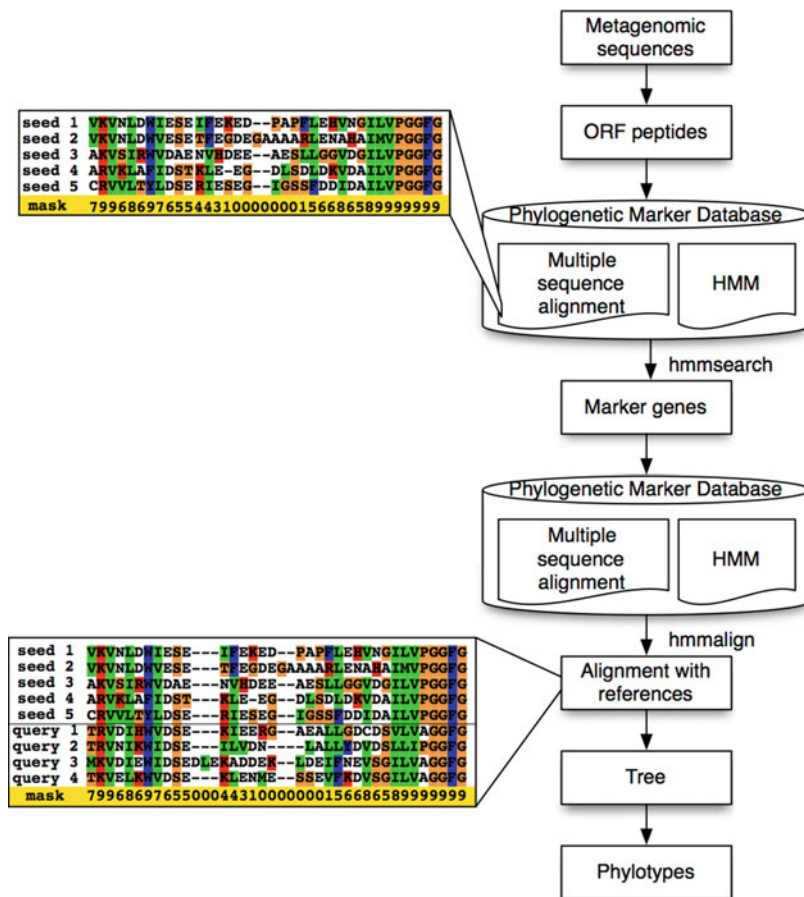
The rapid growth of genomic data also presents challenges for using protein-coding genes in microbial diversity studies. In order to answer the question of “who is there” in metagenomic studies, there is a pressing need for developing an automated high-throughput, high-quality application for metagenomic phylotyping. Several factors should be considered for such an application. First, because genes can be exchanged in bacteria and archaea, it is imperative to only use genes that are recalcitrant to lateral gene transfer for phylotyping. Secondly, for accurate estimation of the microbial composition, only single-copy protein genes should be used as the marker genes. Thirdly, tree-based phylotyping involves multiple steps including marker identification, sequence alignment, tree reconstruction, and taxonomy assignment. For large-scale phylogenetic analysis, several technical hurdles need to be overcome to make high-quality sequence alignments prior to the phylogenetic inference.

Description

AMPHORA is an automated phylogenomic inference application (Wu and Eisen 2008; Wu and Scott 2012). It offers speed, reliability, and high-quality analyses using protein-coding genes as alternative marker genes for microbial diversity studies. The main components of the AMPHORA are illustrated in Fig. 1 and are described in detail below.

Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies, Fig. 1 A

flowchart illustrating the major components of AMPHORA



Protein Phylogenetic Marker Database

AMPHORA relies on a core phylogenetic marker database to identify a set of protein marker genes from the input sequences. The phylogenetic marker database contained 31 bacterial markers initially (Wu and Eisen 2008) and was recently expanded to include 104 genes from the archaeal domain (Wu and Scott 2012). To limit potential complications from paralogy and lateral gene transfers, only single-copy genes that are “universally” distributed in bacteria or archaea were selected. As expected, most of the marker genes are housekeeping genes involved in DNA replication, transcription, translation, or central metabolism, which are thought to be less prone to lateral gene transfers (Jain 1999; Sorek et al. 2007). The use of single-copy genes provides the additional benefit by reducing the bias in the relative species abundance estimation.

AMPHORA uses the HMMER3 package to search for marker genes in the input sequences. Profile Hidden Markov Model (HMM)-based sequence similarity search is as fast as BLAST but is more sensitive (Eddy 2011). AMPHORA can take either protein or DNA sequences as input, which means that users can use AMPHORA to phylotype metagenomic reads directly without having to first annotate the DNA sequences. When DNA sequences are used, AMPHORA will first identify the open reading frames (ORFs) and then search the translated peptide sequences for marker genes.

High-Quality and Highly Reproducible Sequence Alignments

Molecular phylogenetic analysis assumes common ancestry, or homology, for every single column of a multiple sequence alignment. However,

this assumption is often violated when distantly related sequences are aligned. Low-quality alignment regions are noisy and can obscure the true phylogenetic signal contained elsewhere in the alignment. It has been shown that alignment quality can have greater impact on the accuracy of the tree than does the tree-building method employed (Lake 1991; Morrison and Ellis 1997; Hwang et al. 1998; Cammarano et al. 1999; Landan and Graur 2007). Therefore, preparing high-quality sequence alignments is the most critical part of tree-based phylotyping process. Quality of the sequence alignment at each column can be assessed (a step known as masking), and low-quality regions of the alignment can be deleted or down weighted (a step known as filtering) prior to making a tree. Masking and filtering improve the accuracy of phylogenetic analysis (Grundy and Naylor 1999; Castresana 2000; Loytynoja and Goldman 2008; Wu et al. 2012).

One great advantage of using AMPHORA is that it provides automated high-quality alignment masking and filtering. This is achieved by taking advantage of a unique feature of the profile HMM-based multiple sequence alignments. When using HMM to align sequences, new sequences can be mapped to the “seed” sequence alignment that is used to build the HMM, column by column. If the columns in the “seed” alignment have precomputed quality scores, they can then be transferred to the new alignment, thereby providing automated masking and filtering. Quality scores have been assigned to the “seed” alignments of the AMPHORA’s marker genes using a probability-based alignment masking program named Zorro (Wu et al. 2012). Incorporating Zorro makes it practical to quickly expand the phylogenetic marker database to include hundreds of marker genes. It also makes it much easier for users to add markers of their own choice and to build their personalized phylogenetic marker database to use with AMPHORA.

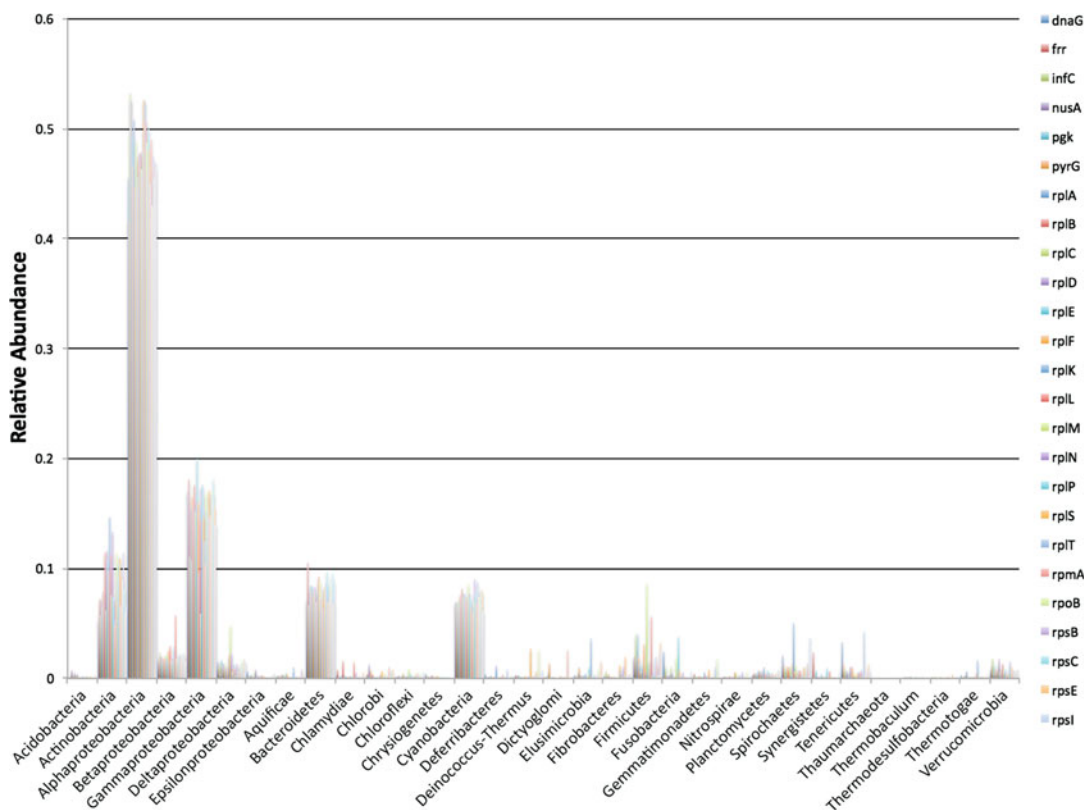
Tree-Based Phylotyping

By comparing to the reference sequences, metagenomic sequences can be classified and

assigned taxonomy. There are two approaches of phylotyping. Similarity-based phylotyping such as MEGAN works by BLAST searching the metagenomic sequence against a reference database such as NCBI nonredundant amino acid database and then assigning the common taxonomy of the top hits to the sequence (Huson et al. 2007). Similarity-based phylotyping is extremely fast. However, it requires the user to select an arbitrary cutoff to define the top hits. Since different microbial species and protein families evolve at different rates, there is no single universal cutoff that is applicable in all situations. Also because of the evolutionary rate variation, top hits are not guaranteed to be the closest relatives of the query sequence (Koski and Golding 2001). Therefore, taxonomy assigned using the top hits can be misleading, especially when no close relatives are available in the database.

Tree-based phylotyping works by placing the metagenomic sequences into a phylogeny of the reference sequences. The metagenomic sequence is assigned the taxonomy of its sister clade, the closest relative according to the phylogeny. Since evolutionary methods can account for the evolutionary rate variations, tree-based phylotyping is more robust than similarity-based phylotyping. In addition, there is no need to choose an arbitrary cutoff in tree-based phylotyping. It has been shown that tree-based phylotyping outperformed similarity-based phylotyping methods (Wu and Eisen 2008).

Insertion of the sequences into the reference tree has been one of the rate-limiting steps in tree-based phylotyping. However, new placement algorithms make it possible to insert thousands of sequences into a reference tree simultaneously, therefore dramatically speeding up the process (Matsen et al. 2010; Berger et al. 2011). AMPHORA takes advantage of RAXML’s evolutionary placement algorithm and can perform either parsimony or likelihood tree-based phylotyping. It places sequences into the NCBI’s taxonomic hierarchy and assigns a confidence score at each rank of the taxonomic classification.



Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies, Fig. 2 Bacterial composition of the GOS dataset analyzed using AMPHORA

AMPHORA Analysis of the Global Ocean Survey Dataset

AMPHORA was used to phylotype the environmental shotgun sequencing reads of the Global Ocean Survey (GOS) (Rusch et al. 2007). From the 41 million predicted peptides, 213,583 peptides were identified that corresponded to the 31 bacterial and 104 archaeal marker genes. Using the number of reads per marker, it was estimated that 95.4 % of the reads in GOS dataset belonged to bacteria while, 4.6 % of the reads were from archaea, indicating that the ocean surface water is dominated by bacteria. The relative abundance of major bacterial groups is shown in Fig. 2. Alphaproteobacteria is the most abundant group overall, making up 47.8 % of the bacterial population. This is mainly due to a single clade of *Pelagibacter ubique* that constituted 35.8 % of the bacterial population sampled in GOS.

Because all the marker genes in AMPHORA are single-copy genes, the relative abundance of

sequences in each marker gene can be used as approximation for the relative organismal abundance in the population. In agreement, the relative abundance of *Pelagibacter ubique* clade estimated by AMPHORA (35.8 %) is very close to previous quantitative estimations by fluorescence in situ hybridization showing that, on average, cells of the clade account for one-third of the ocean surface bacterioplankton communities (Morris et al. 2002). Also as expected, the bacterial diversity profiles are remarkably consistent between the different marker genes (Fig. 2).

Summary

Metagenomics has the potential to transform the way we study microbial diversity. To fully realize this potential, it is important to develop a set of well-curated protein-coding genes as alternative

marker genes. AMPHORA builds on a set of universally conserved, single-copy protein genes that are ideal for analyzing bacterial diversity. It facilitates the large-scale phylogenetic analysis of these marker genes and should be of broad application in the study of microbial evolution and ecology.

References

- Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, et al. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol.* 2005;71:8966–9.
- Berger SA, Krompass D, Stamatakis A. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol.* 2011;60:291–302.
- Cammarano P, Creti R, Sanangelantoni AM, et al. The archaea monophyly issue: a phylogeny of translational elongation factor G(2) sequences inferred from an optimized selection of alignment positions. *J Mol Evol.* 1999;49:524–37.
- Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000;17:540–52.
- Cole JR, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 2009;37:D141–5.
- Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7:e1002195.
- Grundy WN, Naylor GJ. Phylogenetic inference from conserved sites alignments. *J Exp Zool.* 1999; 285:128–39.
- Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17:377–86.
- Hwang UW, Kim W, Tautz D, et al. Molecular phylogenetics at the Felsenstein zone: approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. *Mol Phylogenet Evol.* 1998;9:470–80.
- Jain R. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci.* 1999;96:3801–6.
- Kembel SW, Wu M, Eisen JA, et al. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol.* 2012;8:e1002743.
- Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol.* 2001;52:540–2.
- Lake JA. The order of sequence alignment can bias the selection of tree topology. *Mol Biol Evol.* 1991;8:378–85.
- Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol.* 2007;24:1380–3.
- Loytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science.* 2008;320:1632–5.
- Ludwig W, Klenk H-P. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, Garrity GM, editors. *Bergey's manual of systematic bacteriology*, vol. 1. New York: Springer-Verlag; 2000. p. 49–65.
- Matsen FA, Kodner RB, Armbrus EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinforma.* 2010;11.
- Morris RM, Rappe MS, Connon SA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature.* 2002;420:806–10.
- Morrison DA, Ellis JT. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol Biol Evol.* 1997;14:428–41.
- Pagani I, Liolios K, Jansson J, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 2012;40:D571–9.
- Rusch DB, Halpern AL, Sutton G, et al. The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 2007;5:e77.
- Santos SR, Ochman H. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol.* 2004;6:754–9.
- Sorek R, Zhu Y, Creevey CJ, et al. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science.* 2007;318:1449–52.
- Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 2008;9:R151.
- Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics.* 2012;28:1033–4.
- Wu M, Chatterji S, Eisen JA. Accounting for alignment uncertainty in phylogenomics. *PLoS ONE.* 2012;7(1): e30288.

Proteomics and Metaproteomics

Rembert Pieper, Shih-Ting Huang and Moo-Jin Suh
J. Craig Venter Institute, Rockville, MD, USA

Synonyms

Global proteomics; Protein profiling of microbial communities; Proteomics of biological systems

Definition

Proteomics pertains to the comprehensive analysis of expressed proteins from a cell, a multicellular system, an extracellular environment, or a large set of recombinant clones. This is achieved using combinations of protein separation, identification, and/or assay techniques, such as liquid chromatography-mass spectrometry (LC-MS), two-dimensional gel electrophoresis-mass spectrometry (2DE-MS), affinity purification-mass spectrometry (AP-MS), and protein- or antibody-based microarrays. The objectives in proteomics research can be diverse; they include protein quantification on a global scale, highly parallel analysis of protein functions and interactions, structural characterization of protein complexes, unraveling trafficking of proteins and their distribution in different cellular compartments, and discovery of protein signatures for a disease state or other perturbation. Metaproteomics is a recent extension of proteomics where the biological systems under study are increased in complexity. This pertains to two or more coexisting organisms that may functionally interact with each other, with mutual benefits or to the advantage of some and detriment of other species.

Introduction

Proteomics is a relatively young scientific discipline at the interface of analytical biochemistry and molecular biology. Together with transcriptomics, the discipline emerged in part as a result of the “genomics revolution,” specifically the availability of databases derived from genome sequencing and annotation efforts that reliably predict potentially expressed proteins. Protein sequence information for all open reading frames (ORFs) is an important component of high-throughput mass spectrometry- and microarray-based proteomics. Proteins arrayed on microarray chips are usually derived from the expression of genes in recombinant systems. Expression requires sequence information to generate clones for the targeted genes. Proteins

analyzed by mass spectrometry (MS), the technology that advanced proteomics the most and resulted in Nobel Prize awards in Chemistry for K. Tanaka and J. B. Fenn in 2002, can be identified from complex mixtures on a global scale using computational methods that compare experimental mass spectra to the entirety of theoretical peptide masses and sequences derived from protein sequences annotated in a searchable database. Typically, peptides rather than proteins are analyzed in MS-based proteomic experiments because their mass range (length of 5–30 amino acids) makes them more suitable for ionization and accurate mass analysis, and fragmentation of peptides in tandem MS experiments allows sequence analysis. Peptide-spectrum matches (PSMs) require mathematical algorithms for probability-based assignment of peptides to their protein(s) of origin. In addition to increasingly powerful algorithms and the exponential growth of complete genomic databases (for thousands of species and subspecies), MS techniques regarding ionization, accurate measurement of mass-to-charge ratio of peptides, and proteins and their fragmentation have also dramatically advanced. Mass spectrometers now measure proteins with sensitivities in the attomole range, mass accuracies in the 1–3 ppm range, and a peak resolution of up to 60,000, at very high speeds and with considerable automation. Proteomes of prokaryotic and mammalian cells can now be profiled in a few days to a couple of weeks, including proteins present in less than 100 copies in a cell. For example, the proteomes of yeast and the human HeLa cell line have been exhaustively characterized (de Godoy et al. 2008; Nagaraj et al. 2011). MS-based proteomics requires high-resolution separation techniques to reduce the complexity of peptides or proteins in a sample. Two-dimensional gel electrophoresis (2DE) has been used for protein separation before MS emerged as the method of choice. In the last decade, 2DE has been gradually replaced by shotgun proteomics, a strategy that takes advantage of controlled enzymatic fragmentation of proteins into peptides prior to MS analysis. Shotgun proteomics has a superior dynamic range for proteome coverage compared to 2DE and is a more

sensitive detection method and less problematic as it pertains to the exclusion of proteins difficult to solubilize and separate in gels.

Protein microarrays allow immobilization of thousands of purified proteins and their interaction analysis with other proteins or small molecule ligands (Wolf-Yadlin et al. 2009). This technique does generally not require MS since the position of proteins on the array is predefined, and a highly parallel assay on the microarray facilitates detection of an activity or interaction with a ligand or substrate. Interactions of proteins with small molecules have more recently been studied with chemical probes that establish covalent bonds to proteins and thus characterize their functions (Speers and Cravatt 2009). Here, MS is typically used for protein identification. Finally, by the use of protein interaction screens, large protein networks (“interactomes”) have been established, e.g., for the bacterium *Mycoplasma pneumoniae* (Kuhner et al. 2009) and for baker’s yeast (Ho et al. 2002). MS is the only proteomic technique that permits comprehensive analyses of posttranslational modifications which play a key role in the modulation of protein functions, localizations, interactions, and control of turnover rates in the cell (Olsen et al. 2010; van Noort et al. 2012). It is also the leading technology for research in metaproteomics, where the expressed proteome is derived from more than one species, often a microbial community. Community dynamics and in some cases a host species influence the protein complement expressed by each participating species. Metagenomic data or concatenated genomes of multiple species are essential input as they deliver the databases necessary for proteomic analysis of a complex biological system.

Expression Proteomics

Protein/Peptide Sample Preparation and Separation

Various areas of proteomic research were already mentioned. This overview focuses on expression proteomics. First, it is more relevant and applicable to the questions of metagenomics (i.e., the

competitive and/or synergistic nature of interactions in multi-species communities). Second, functional analysis of uncharacterized proteins requires multiple methodological approaches not yet feasible on a metagenomic scale or methodologically not distinct. In expression proteomics, sample preparation is an essential component and usually needs to be adapted to a given scientific objective. Table 1 lists the examples of common sample types and approaches to recover the protein mixtures prior to their analysis or that of their digestion products.

Expression proteomics may focus solely on protein identifications from a given biological

Proteomics and Metaproteomics, Table 1 Proteomics and metaproteomics sample preparation methods

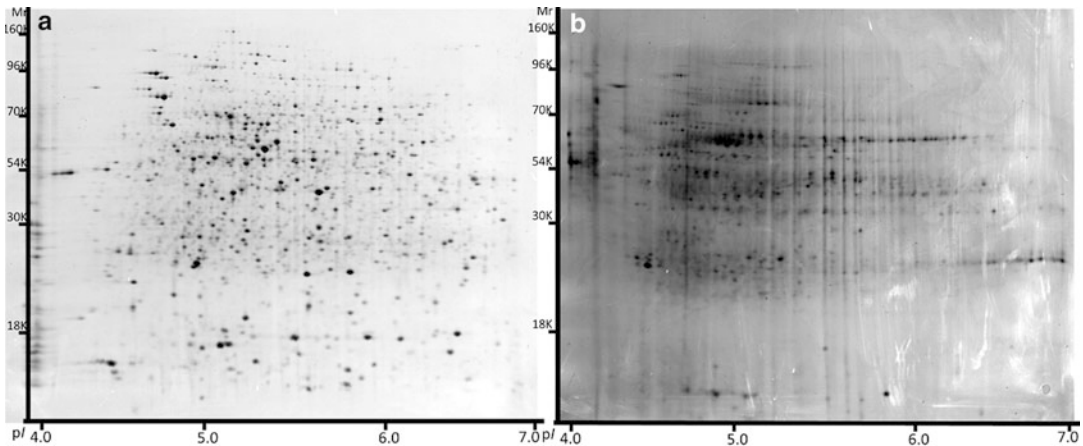
Sample group	Type of sample	Protein recovery method
Multiple-organism environmental sample	Soil, ocean water, stool/gut microbiome	Enrichment for cellular materials, cell lysis
Heterogeneous tissue	Liver, bladder	Isolation of cell types prior to cell lysis or tissue disruption
Cell culture	Bacterial, fungal, or mammalian cell culture	Concentration of extracellular fraction or cell lysis
Cell compartment or fraction	Mitochondria, nuclei, exosomes, chloroplasts, bacterial periplasm	Cell compartment isolation followed by its disintegration
Cellular complex	Proteasome, bacterial secretion, and secondary metabolite biosynthesis systems	Cell complex purification and disintegration
Protein-containing secretion fluid	Blood plasma, urine, hatch fluid of larvae	Removal of lipids, carbohydrates, cellular debris
Host-pathogen system	Intracellular viruses, bacteria, fungi, infected eukaryotic cells	Separation of host and pathogen cells; cell lysis
Life cycle stages of a species	Parasitic organism with a complex life cycle	Lysis of cells or cellular compartments

sample, but quantitative assessments of a subcellular or cellular proteome are often of interest (e.g., the comparative analysis of the *Escherichia coli* proteome isolated from exponential versus stationary-phase cultures or of the mouse liver proteome prior to, during, and after recovery from an infection with a hepatitis virus). Sample preparation may involve steps to remove nonorganic or other matter not of interest in a study (e.g., soil or digested foods in studies of soil and gut microbiomes, respectively), but typically it starts with the isolation of tissues, cells, cell organelles, or fluids followed by extraction and/or concentration of protein mixtures. An exception to this experimental sequence is a strategy that involves protein labeling with isotopes of a living cell/organism prior to cell and protein extraction. Stable isotope labeling of amino acids in culture (SILAC) is a frequently used method where cells are cultured with defined media containing amino acids (e.g., Lys, Arg) that contain carbon and nitrogen atom isotopes which alter the total mass of the amino acid. If two types of samples are to be compared in an experiment (e.g., two cell cycle time points or mammalian cells pre- and post-viral infection), they are cultured with different Lys and/or Arg isotopes. Following combination of the two cell populations and their lysis, otherwise identical proteins (peptides) can be compared quantitatively based on their isotope mass differences. Chemical isotope-labeling methods for proteins such as iTRAQ follow the same principles (differential quantification from a multiplexed sample), but the labeling step occurs after isolation of the protein digestion products.

Traditionally, the first step of proteomic analysis has been high-resolution separation of proteins in two-dimensional gels that permitted mapping of the most abundant proteins of a complex mixture and their relative quantification (O'Farrell 1975). Proteins are visualized in 2D gels using protein-binding dyes such as Coomassie Brilliant Blue, and more sensitive fluorescent dyes that stain most proteins resolved in a gel (up to 1,000 protein spots). Sample preparation for 2D gels includes solubilization or dilution of a protein mixture in a denaturing

solution (8 M urea, 2 M thiourea, detergents such as 4 % CHAPS or 1 % Nonidet-NP40, 0.1 % ampholytes, and DTT as a reducing agent). Prior to this step, the removal of salts and other macromolecules positively affect resolution and identification of proteins. Many different 2DE modification techniques have been introduced to improve spot resolution in alkaline and acidic pH ranges, in high and low M_r regions, and for lipid-associated and hydrophobic proteins (Gorg et al. 2004). Figure 1 displays a 2D gel profile of an *E. coli* O157:H7 cell lysate next to one from a mouse stool microbiome fraction. While more than 500 proteins were resolved in the *E. coli* gel, it is evident that the resolution limit of the more complex stool protein sample (hundreds of different gut microbial species and secreted human proteins) was reached. For such complex metaproteomic samples, 2D gels are not useful because few proteins are well resolved and identifiable as distinct spots. Differential quantification of proteins from 2D gels is performed with software tools that allow pixel-based spot intensity measurements, gel-to-gel spot matching and normalization, and generation of annotated spot maps that characterize the proteome under investigation. Subcellular proteomes of bacteria exposed to various environmental stress conditions have been analyzed in 2D gels (Pieper et al. 2008). Sample preparation of body fluids such as serum may include LC separations to remove highly abundant proteins and fractionate other proteins prior to 2DE (Pieper et al. 2003). Sample preparation for analysis of eukaryotic subcellular compartments often involves buoyant density-based centrifugal enrichment steps and differential display. Tagging of genes with reporter gene constructs to localize expressed proteins in subcellular compartments has also been used. An example is the comprehensive survey of the mitochondrial proteome in yeast (Prokisch et al. 2004). Methods used for the disintegration of cells, the isolation of subcellular organelles, and the subsequent protein extraction and solubilization require project- and cell type-specific optimization.

The shotgun proteomics workflow integrates a protein digestion step prior to analyte (peptide)



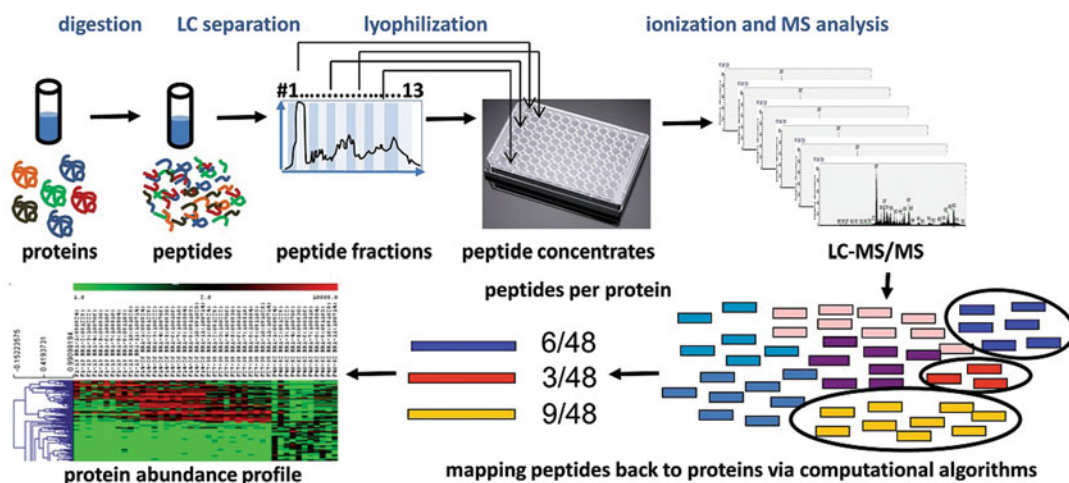
Proteomics and Metaproteomics, Fig. 1 Protein profiles of (a) Shiga toxin-producing *E. coli* (serotype H157:O7) and (b) a murine stool fraction enriched in bacteria displayed in 2D gels. Samples of ~150 μ g protein were loaded onto pH 4–7 25 cm immobilized pH gradient strips and isoelectrically focused applying 64 kVh. Following reduction and alkylation of proteins in the IPG

strips, proteins were separated according to size in second dimension 8–18 %T SDS-PAGE gels (25 \times 20 cm) for 1.8 kVh. Gels were stained with the dye Coomassie Brilliant Blue G250 (Courtesy of Christine Peterson and Prashanth Parmar for their contributions to the gel electrophoresis data depicted in the courtesy)

separation, identification, and quantification. Shotgun proteomics was developed 10 years ago (Wolters et al. 2001). Proteins are solubilized, denatured, and subjected to proteolysis with endoproteases such as trypsin, LysC, and/or GluC. Using a combination of these enzymes increases the coverage of a proteome. The mixture of digested proteins typically contains more than 100,000 peptide fragments in a wide abundance range. Digested proteins are sometimes applied to an LC column (e.g., with a reversed phase or ion exchange matrix) or an immobilized pH gradient gel strip to separate peptides further and reduce their complexity in the resulting fractions. Peptides are fractionated based on certain biophysical traits such as hydrophobicity, M_r , or net charge. Peptide eluates from LC columns may be directly spotted on plates for serial analysis via matrix-assisted laser desorption ionization MS (MALDI-MS). A far more time- and cost-effective and less tedious approach to obtain comprehensive proteome coverage, however, is to apply concentrated and desalted peptide fractions to online LC tandem mass spectrometry (LC-MS/MS). The shotgun proteomics workflow is displayed in the schematic of Fig. 2.

Protein Identification

MALDI-MS and LC-MS/MS have been the standard techniques for identifying proteins from 2D gel spots, often with considerable automation in the spot excision from gels and the enzymatic digestion to generate dissolved peptide mixtures. MALDI-time of flight (TOF) MS generates peptide mass fingerprints (PMFs) that are analyzed with an MS algorithm in which the protein of origin is identified based on the count of mass-matching peptides in the experimental spectrum compared to those predicted from the *in silico* enzymatic digest for a given protein (Fig. 3a). Nano-electrospray ionization (ESI) is the main ionization technique for LC-MS/MS experiments. LC-MS/MS not only provides one separation dimension for peptides, but also the data for protein identification, first on the MS level via generation of an accurate mass-to-charge ratio (m/z) for a peptide and then via data-dependent selection of an MS (peptide) peak for further fragmentation via gas-phase collision-induced dissociation (CID). MS peaks in fragment spectra, typically y - and b -ions resulting from the cleavage of peptide bonds, define the peptide sequence. Computational methods are available

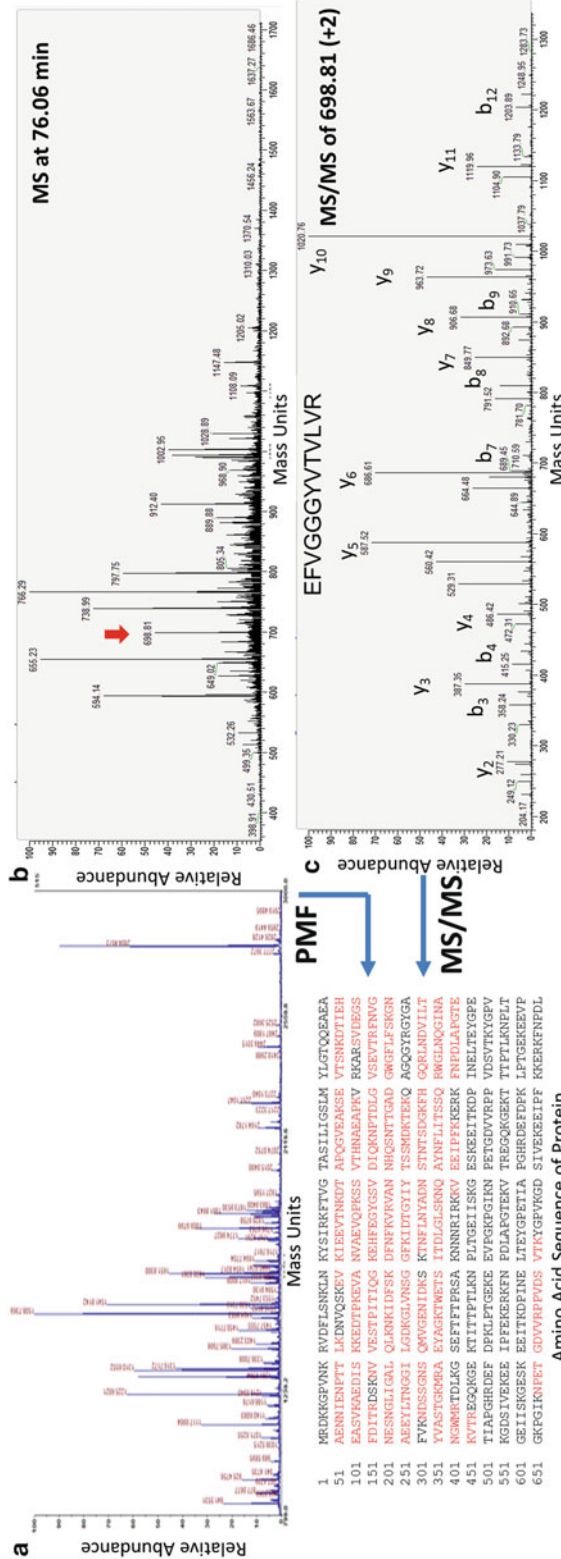


Proteomics and Metaproteomics, Fig. 2 Shotgun proteomics workflow. After generation of a cell lysate or protein extract, proteins are digested with an endoproteinase (e.g., trypsin). The peptide mixture is separated on a reversed phase C₁₈ or a strong anion exchange LC column. Peptide fractions are lyophilized and applied to nano LC-MS/MS sequentially. The mass spectra

combined from all fractions are collected and interpreted with an algorithm and a relevant protein sequence database. Identified peptides are assigned to proteins of origin and counted to obtain a protein quantity estimate. Abundance profiles from different samples can be displayed in form of heat maps

to assign a peptide sequence based on its original m/z value and tandem MS data that deliver a series of daughter m/z values for N- and C-terminal fragment ions (Fig. 3b). The MS and subsequent MS/MS analyses are performed in automated duty cycles defined by the LC-MS instrument software so that tens or even hundreds of thousands of MS and subsequent MS/MS scans are performed in series. The aggregate of data from these scans describes the proteome in the shotgun proteomics experiment. Due to the fact that the matching of theoretical MS/MS (peptide fragment) spectra and experimental spectra is performed with probabilistic models defined by a software and its inherent algorithm(s), shotgun proteomics data yield a number of peptide (and protein) identifications at a specific false discovery rate, often in conjunction with an MS score matrix that also attributes a measure of correct peptide identification. Algorithms integrated in such software tools are used to score PSMs and assign the highest scoring PSM to a peptide. Protein inferences are made by assigning peptides to a distinct protein sequence in the database (Fig. 3c). The larger

the database of theoretical PSMs, the more computationally challenging it is to determine the best peptide match for an experimental mass spectrum. Herein lies one of the fundamental challenges of metaproteomics: protein sequence databases to be searched not only contain sequences derived from one but numerous fully annotated genomes or large metagenomic read assemblies that are partially annotated. Their content of predicted protein sequences is substantially increased. MS platforms have recently moved towards ultra-high pressure LC for high-resolution peptide separations and high-resolution, high-mass accuracy MS, such as the Orbitrap and Quadrupole-TOF instruments. Excellent peptide separation has the benefit that more peptides derived from low-abundance proteins are enriched in fractions and more likely to be selected during the MS data-dependent duty cycle for MS/MS analysis. High-mass accuracy and resolution enhance the confidence in peptide (sequence) assignments via PSMs. For a detailed review of LC-MS platforms used for proteomics applications, see (Yates et al. 2009).



Proteomics and Metaproteomics, Fig. 3 Principles of peptide mass fingerprinting (PMF) and tandem mass spectrometry (MS/MS). (a) PMF of a purified 2D gel protein spot (e.g., from MALDI-TOF/TOF analysis). (b) Snapshot of MS and subsequent MS/MS scan in a shotgun proteomics dataset. *Top*: MS spectrum representing a peptide mixture derived from a variety of proteins, with one peptide peak at m/z 698.81 (+2 ion charge). *Bottom*: the peptide at m/z 698.81 was selected for CID fragmentation in a Velos Pro ion trap instrument and the sequence EFVGGGYVTLVLR assigned based on y- and b-ion series. For m/z values of the PMF and m/z values for a peptide and its MS/MS data, assignments to a protein in the searched database are made (c)

Protein Quantification

Relative protein quantification from 2D gel profiles involves correct spot matching based on their gel coordinates and MS data for each individual spot. 2DE has low dynamic range (two orders of magnitude) and high-abundance thresholds for accurate volumetric spot density measurements, resulting in quantitative analyses limited to the top 5–20 % of the actual proteome. These are a major reason as to why quantitative proteomics has moved towards the use of other techniques: (1) shotgun proteomics which allows quantification of a far larger proportion of the proteome with higher dynamic range and (2) targeted proteomics which allows high-precision peptide quantification in absolute terms in a wide dynamic range, but usually for a small number of proteins. The latter technique is moving towards larger scale, as demonstrated recently in an effort to monitor all yeast kinases and phosphatases (Picotti et al. 2010). Targeted proteomics is often associated with the term multiple reaction monitoring (MRM) proteomics where stable isotope-labeled peptide standards are used for quantitative comparisons. MRM proteomics is dominated by triple quadrupole/ion trap hybrid MS instruments. It requires preselection of “target” peptides, often tryptic peptides that are unique to a given protein of interest. Such peptides are generated in situ via enzymatic digestion of a sample in which the protein is to be quantified. Equivalent chemically synthesized peptide standards are spiked into the sample in known concentrations to allow absolute quantification. MRM experiments continue to advance in speed and complexity (multiplexing is possible) as shown in a recent study in which 63 urinary proteins were simultaneously measured in hundreds of samples (Chen et al. 2012).

A variety of methods for global quantification of proteins are available for shotgun proteomics as recently reviewed (Mueller et al. 2008; Elliott et al. 2009). In addition to isotope label-based approaches (e.g., SILAC, iTRAQ), spectral counting and MS¹ (peptide) intensity-based measurements are common. Unlike MS¹ peak

intensity measurements, spectral counting can be performed with low-resolution (ion trap) MS instruments. For both methods, software tools have been created allowing estimation of protein copies per cell (absolute quantification) for a large number of distinct proteins expressed in a cell. The schematic in Fig. 2 illustrates how spectral counting data are analyzed for proteome-wide quantification following assignment of peptides to different proteins of origin. Recently, more than 10,000 human proteins have been identified and quantified via shotgun proteomics using an Orbitrap mass analyzer, presumably representing the entire expressed proteome of a cancer cell line (Nagaraj et al. 2011). An additional layer of complexity is added when proteomic data are searched for specific post-translational modifications (PTMs). Among the PTMs are N-terminal truncation, phosphorylation, N-acetylation of Lys residues and N-termini, and glycosylation of various side chains, all of which can modify the protein's cellular function, localization, and trafficking and interaction with other proteins or ligands inside or outside of the cell. Ubiquitination of Lys residues often sends a protein into a degradation pathway. Comprehensive knowledge of all PTMs and their dynamics in a specific environment or cell state has not yet been achieved. Likewise, proteomic research is just beginning to provide information on distinct protein activities not functionally annotated in databases. One of the promising technologies is activity-based proteomic profiling that allows labeling of proteins in their active sites by the use of chemical probes (Speers and Cravatt 2009). The main limitation is generating a large number of specific chemical probes for high-throughput screens. A field more adapted to high-throughput analyses is interaction proteomics, where proteins of unknown function can be associated with protein complexes of known function, thus allowing assignment of new biological roles. This field includes AP-MS which, for example, has been utilized to study 178 soluble protein complexes of the *M. pneumoniae* proteome (Kuhner et al. 2009).

Metaproteomics

The term metaproteomics was first introduced in 2004 by Rodriguez-Valera to describe the concept of an expressed protein complement from environmental microbial communities (Rodriguez-Valera 2004). Metagenomics has been a driving force behind metaproteomic efforts; it essentially defines the protein sequence databases to be searched, and it also provides a biological context. Major interest has also arisen from the human microbiome project. Metaproteomics should take into consideration the host environment. This would expand the definition of metaproteomics to the study of biological systems consisting of two or more species that may interact in a mutualistic manner or to the detriment of some but the benefit of other species. This definition would include symbiotic relationships (e.g., N₂-fixing bacteria with legumes), host-pathogen relationships (infectious disease), and host-commensal relationships (e.g., the gut microbiome). The boundary of the latter two is fluid, as metagenomic and other studies have revealed. Metaproteomic data are of interest because they add a degree of function to the description of a complex community: microbes (and their hosts) live in the same environment, compete for the same resources, and send molecular signals to each other including quorum sensing, chemotaxis, and adhesion in response to the changing environment. The competition for resources implicates metabolism that is enabled by proteins. Likewise, inter- and intraspecies signaling implicates proteins and peptides or structures synthesized by proteins (e.g., LPS and secondary metabolites). Thus, quantitative analysis of proteins via metaproteomics promises to deliver new insights into the dynamics of complex biological communities. Studies may be highly experimental, considering the efforts to model microbial communities or a pathogen invading a macrophage *in vitro*. They may constitute a natural environment such as polybacterial biofilms growing in hydrothermal hot springs or on a urinary tract device (Hall-Stoodley et al. 2004).

While the first metaproteomic studies pertained to low complexity systems, they highlighted the ability to elucidate dynamic aspects of the adaptation of species to community living. Ram et al. investigated natural acid mine drainage microbial biofilms (Ram et al. 2005). More than 2,000 proteins from five different species were identified using shotgun proteomics, 48 % from *Leptospirillum* group II. Oxidative stress and refolding proteins were highly expressed, supporting the notion that their activities were critical in a challenging environment. Markert et al. investigated the proteome of an unculturable γ -proteobacterial endosymbiont of *Riftia pachyptila*, a deep-sea tube worm without a digestive system (Markert et al. 2007). The worm sustains a high growth rate using the symbiont's capacity for chemosynthesis of carbon compounds fixing CO₂ and oxidizing ambient H₂S. Using 2DE-MS proteomics, three abundant major sulfide oxidation proteins critical for energy metabolism in the endosymbiont were identified. It was determined that both the reductive tricarboxylic acid and Calvin cycles were used for CO₂ fixation. A more complex metaproteome, that of human distal gut microbiota, was examined by Verberkmoes et al. (2009). A particular challenge of analyzing such a complex metaproteome is the high number of species (and diverse subspecies and strains), most of which are not culturable and whose genomes remain to be sequenced. Therefore, databases for metaproteomic data searches, which are composed of only sequenced and fully annotated genomes of bacteria known to colonize the distal gut, are not truly representative of metagenomic (species) complexity. It is nonetheless useful to use such "imperfect" databases to gain insights into a human body-associated complex microbial metaproteome. Assessing quantitative estimates of protein counts representing distinct cellular function categories, it was reported that proteins linked to carbohydrate metabolism, energy generation, and ribosomal translation were most abundant in the distal gut metaproteome (Verberkmoes et al. 2009). Nearly 20 % of the mass spectra

matched protein sequences derived from *Bacteroides* and *Bifidobacterium*, confirming relatively high abundance of these genera in the human gut. Despite the application of a bacterial enrichment procedure, 30 % of the PSMs represented matches to human proteins, including a large proportion of those active in cell-cell adhesion and innate immunity. This finding supported the notion that the host immune system interacts extensively with its gut microbiome. Analysis of urinary tract metaproteomes linked to asymptomatic bacteriuria (Fouts et al. 2012) resulted in protein identifications from two to five different opportunistic pathogens and provided preliminary evidence for host-bacterial interactions, specifically a battle for iron. Human lactotransferrin, an iron sequestration protein, and iron acquisition proteins and receptors from *E. coli* and *Klebsiella pneumoniae* were identified in the same samples.

Summary and Outlook

In conclusion, proteomics is a highly advanced discipline that contributes to science at the biological systems level. Metaproteomics has clear potential to elucidate functional interactions of coexisting microbial species and, if applicable, those with their eukaryotic host environments. Major challenges to enable in-depth and accurate metaproteomic profiling efforts for highly diverse communities remain to be addressed. Only a fraction of the genomes represented in complex microbial communities have been sequenced. Comprehensive metagenomic sequence datasets are very promising resources for advanced proteomic data searches. However, such datasets can be incomplete and may have sequence inaccuracies and significant redundancy which, in turn, affects the reliability of assignments of peptides and proteins on the species level via PSMs derived from MS-based proteomic datasets. Further improvement of metagenomic assembly and computational methods will benefit the quality of metaproteomic datasets since their analysis depends on predicted protein sequence data.

A particular challenge pertains to the high amino acid sequence identities among highly conserved (housekeeping) proteins of related species in a microbiome. Since protein identification in shotgun proteomics relies on peptide sequence data followed by in silico assignment to proteins, it impedes taxonomic profiling on the species level analogous to the short reads of NextGen sequencing technologies. Nonetheless, metaproteomic data already contribute effectively to the elucidation of the metabolic capacity of complex biological systems and the cross-talk of such systems with their host environments. Robust computational algorithms and workflows will have a positive impact on the future of metaproteomics. Use of multiple “omics” technologies allows insights into complex intra- and extracellular biological processes and their cross-talk and integration into a biological system.

References

- Chen YT, Chen HW, et al. Multiplexed quantification of 63 proteins in human urine by multiple reaction monitoring-based mass spectrometry for discovery of potential bladder cancer biomarkers. *J Proteome*. 2012;75(12):3529-45
- de Godoy LM, Olsen JV, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*. 2008;455(7217):1251-4.
- Elliott MH, Smith DS, et al. Current trends in quantitative proteomics. *J Mass Spectrom*. 2009;44(12):1637-60.
- Fouts DE, Pieper R, et al. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med*. 2012;10(1):174.
- Gorg A, Weiss W, et al. Current two-dimensional electrophoresis technology for proteomics. *Proteomics*. 2004;4(12):3665-85.
- Hall-Stoodley L, Costerton JW, et al. Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol*. 2004;2(2):95-108.
- Ho Y, Gruhler A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*. 2002;415(6868):180-3.
- Kuhner S, van Noort V, et al. Proteome organization in a genome-reduced bacterium. *Science*. 2009;326(5957):1235-40.
- Markert S, Arndt C, et al. Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science*. 2007;315(5809):247-50.

- Mueller LN, Brusniak MY, et al. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res.* 2008;7(1):51–61.
- Nagaraj N, Wisniewski JR, et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol.* 2011;7:548.
- O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem.* 1975;250(10):4007–21.
- Olsen JV, Vermeulen M, et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal.* 2010;3(104):ra3.
- Picotti P, Rinner O, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat Methods.* 2010;7(1):43–6.
- Pieper R, Gatlin CL, et al. The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics.* 2003;3(7):1345–64.
- Pieper R, Huang ST, et al. Characterizing the dynamic nature of the *Yersinia pestis* periplasmic proteome in response to nutrient exhaustion and temperature change. *Proteomics.* 2008;8(7):1442–58.
- Prokisch H, Scharfe C, et al. Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol.* 2004;2(6):e160.
- Ram RJ, Verberkmoes NC, et al. Community proteomics of a natural microbial biofilm. *Science.* 2005;308(5730):1915–20.
- Rodriguez-Valera F. Environmental genomics, the big picture? *FEMS Microbiol Lett.* 2004;231(2):153–8.
- Speers AE, Cravatt BF. Activity-based protein profiling (ABPP) and click chemistry (CC)-ABPP by MudPIT mass spectrometry. *Curr Protoc Chem Biol.* 2009;1:29–41.
- van Noort V, Seebacher J, et al. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Mol Syst Biol.* 2012;8:571.
- Verberkmoes NC, Russell AL, et al. Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 2009;3(2):179–89.
- Wolf-Yadlin A, Sevecka M, et al. Dissecting protein function and signaling using protein microarrays. *Curr Opin Chem Biol.* 2009;13(4):398–405.
- Wolters DA, Washburn MP, et al. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem.* 2001;73(23):5683–90.
- Yates JR, Ruse CI, et al. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng.* 2009;11:49–79.

R

RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data

Norman J. MacDonald¹, Donovan H. Parks^{1,2} and Robert G. Beiko¹

¹Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

²Australian Centre for Ecogenomics, University of Queensland, Brisbane, QLD, Australia

Definition

Algorithm, software, and Web service for taxonomic classification of metagenome fragments using both homology and compositional information.

Introduction

A central task in many metagenomic studies is the inference of community function from sequence data. An additional challenge is the need to assign functional genes to particular members of the community, in order to determine which organisms are responsible for carrying out which molecular processes. While sequences derived from a given microorganism often carry a “signature” that reflects mutational bias and other processes in the genome of that organism (Campbell et al. 1999), these patterns are by no

means uniform and they often cannot distinguish closely related organisms. Further compounding the problem is the reliance on short-read-based approaches to metagenome sequencing, which can generate reads less than 200 nucleotides in length, and short or ambiguous assemblies in many cases. Successful classification methods use homology (e.g., BLAST comparisons against genes or proteins from a set of reference genomes) or composition (e.g., distribution of tetranucleotide sequences) for classification, with a newer generation of “hybrid” classifiers using both (e.g., PhymmBL; Brady and Salzberg 2009). We have developed RITA, a hybrid approach that uses streamlined approaches to rapidly generate homology and composition information and combines these sets of predictions in a supervised classification pipeline that sorts sequences into different classification groups based on the strength and agreement of the two types of predictions.

Requirements

Software: RITA is implemented in Python and can be used as a stand-alone program (<http://kiwi.cs.dal.ca/Software/RITA>) or via the Web service (<http://ratite.cs.dal.ca/rita>). Queries to the Web service are limited to 10,000 sequences at a time. For compositional classifications RITA uses the Fragment Classification Package FCP (Parks et al. 2011; <http://kiwi.cs.dal.ca/Software/FCP>).

For the stand-alone version, a locally installed copy of either the BLAST+ software suite or USEARCH (Edgar 2010) is necessary.

Reference Databases: Since RITA is a supervised classifier, it requires a reference database of sequenced genomes with associated taxonomic information. Genomic information is typically acquired from the NCBI database of sequenced genomes and can be performed automatically using the scripts provided with the FCP software package. From these sequenced genomes (and optionally, similarly formatted files provided by the user), RITA can build reference models for both composition and homology. If rank-flexible classification (described below) is to be performed, a set of 16S rRNA gene sequences corresponding to the reference sequenced genomes will be required as well. Instructions on acquiring and preparing these can be viewed at <http://kiwi.cs.dal.ca/Software/RITA>.

Input Data: The user must provide their metagenomic sequences in a FASTA-formatted file. The sequences can be of any length. If rank-flexible classification is desired, a list of sampled 16S rRNA gene sequences must also be provided.

The RITA Pipeline

The primary objective of RITA is to make taxonomic assignments that consider both the agreement between composition and homology and the strength of evidence from both types of classification technique. Homology-based classification is performed using local alignment-based comparative tools such as BLAST (Altschul et al. 1997). Many variants of BLAST have been developed which differ in the type of sequence information being compared (e.g., nucleotide, 6-way translated nucleotide, amino acid), as well as sensitivity and speed. Although RITA can be configured to run in a number of different ways, the default approach is based on the sequential use of three different algorithms: Discontiguous MEGABLAST for fast but low-sensitivity comparisons between a nucleotide query and nucleotide database, BLASTN for slower but more sensitive nucleotide-nucleotide comparisons, and BLASTX for sensitive

comparisons between a translated query sequence and reference database of protein sequences. The objective in using this ordering is to place the fastest algorithms first, which removes the need to run the slower algorithms on all query sequences. The stand-alone version of RITA also includes the option to use UBLAST (Edgar 2010), which aims to prioritize searches against a reference database in order to avoid searching the entire database. Approaches such as LCA (Huson et al. 2007), MetaPhyler (Liu et al. 2011), and CARMA (Gerlach and Stoye 2011) use phylogenetic information for taxonomic classification, but our trials of RITA showed no additional benefit to the use of phylogenetic trees in the classification scheme we describe below.

For compositional classifications, we encode each reference genome as a series of nucleotide words (i.e., k -mers) of a fixed length to generate frequency distributions of each word. These frequency profiles are then used to train a naïve Bayes (NB) classifier (Parks et al. 2011), which assigns likelihoods to each query fragment based on the match between its k -mer profile and those representing the different genomes in the reference database. The genome with the largest likelihood for a given fragment is the best compositional match to that fragment. The crucial assumption of the NB classifier is of independence among input k -mers: while this assumption is clearly violated by k -mer decompositions of DNA sequences (for instance, the frequency of the 6-mer AAAAAA will be closely tied to that of AAAAAC), in practice this does not impact on the performance of the classifier. Phymm is a compositional classifier that uses more-sophisticated Markov models of sequence composition: while these are better at describing the compositional profile of a genome, in practice they are much slower and no more accurate than our NB approach.

The RITA pipeline combines homology and composition information by first assessing whether the predictions of composition and homology agree. While homology alone outperforms composition alone in most classification tasks, the genomic patterns reflected in compositional profiles provide complementary information, and agreement between the two types of

data is not trivially obtained. If agreement is found for a given fragment and the first BLAST algorithm considered, then the fragment will be classified with the predicted taxonomic label and assigned to group 1, the highest-confidence group. If the predictions of composition and homology disagree, then classification using homology alone will be attempted in the following manner. When running RITA, the user specifies a minimum margin for homology-based classification based on e-values: the default value is 20 orders of magnitude. If the globally best e-value is greater than the best e-value from a different taxonomic group by an amount greater than or equal to this margin, then the result is considered as strong evidence for assignment to the best-matching group, and the fragment is assigned to group 2. If the fragment remains unclassified, the same procedure is followed for subsequent BLAST algorithms with potential classifications to group 3, group 4, etc. If all homology-based options have been exhausted, classification is made to one of two groups based on the NB classifier alone. Similar to the homology margin described above, the globally best NB likelihood is compared to the best likelihood from a different taxonomic group. If this ratio exceeds a user-specified amount, then the fragment is assigned to the higher-confidence composition-only group. If the ratio does not exceed this amount, then the fragment is assigned to the last and lowest-confidence group.

The procedure above describes rank-specific classification, where all fragments are classified at a given taxonomic rank, for instance, phylum or genus. However, different groups of microbes may be more or less represented by sequenced genomes, and there may be more evidence to make precise assignments to some groups than to others. In the extreme case, some bacterial phyla are represented by a single sequenced individual, making it impossible to distinguish between genera and other groups within this phylum. One solution to this problem is to classify all fragments to a very high rank such as phylum or class, but this discards precision in cases where it may be available. Our solution is to use a *rank-flexible* version of RITA that assigns an

appropriate taxonomic group *and* rank based on the strength of available evidence. To perform rank-flexible classification of a metagenome sample using RITA, the user must provide a list of 16S rRNA genes that were identified from the sample. These genes are used to limit the taxonomic scope of the RITA predictions. The provided 16S rRNA genes are mapped into a tree of all 16S genes from the reference database of sequenced genomes. All genomes represented within a minimal clade containing one of the sampled 16S genes will be flagged as assignable to a taxonomic rank that is no more precise than the rank covering all members of that clade. For example, if a sampled 16S gene maps to the reference tree such that all of its sister taxa are from the same order, then RITA will consider matches to those taxa to be equivalent at the rank of order. In this manner, the level of classification is determined by the density of reference genome sampling around the observed 16S rRNA gene sequences from the environmental sample.

Interpreting RITA Results and Factors Affecting Prediction Accuracy

RITA Output: RITA returns detailed results of both composition- and homology-based models. Most critical in the RITA output is a tab-separated file that lists the predictions associated with each DNA sequence. Examples of RITA output are given in Table 1, with some taxonomic ranks omitted to fit each result on a single line:

The first column contains the name of the sequence as obtained from the sequence file. The second and third columns give the confidence group associated with the prediction, first by number and then by name. Group 1, “NB_DCMGABLAST,” indicates agreement between the first homology prediction method used (Discontiguous MEGABLAST) and the NB classifier, while group 2 corresponds to a prediction made based on a strong separation between the best and second-best groups according to homology. The fourth column shows the taxonomic rank at which the prediction was made, and the remaining columns give the

RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data, Table 1 Examples of RITA output

seq1	1	NB_DCMGABLAST	CLASS	Actinobacteria	<i>Nocardioides_sp._JS614</i>
seq2	2	DCMEGABLAST_RATIO	CLASS	Deltaproteobacteria	<i>Syntrophus aciditrophicus</i>
seq3	5	NB_RATIO	CLASS	Alphaproteobacteria	<i>Phenylobacterium zucineum</i>
seq4	6	NB_ML	CLASS	Sphingobacteria	<i>Pedobacter saltans</i>

labels associated with that prediction, with the final column showing the actual genome that yielded the best prediction.

Summarizing Results: In most cases, we do not recommend using all classes when building taxonomic summaries of the contents of a metagenome. In particular, the accuracy of the final two classes (which are based on composition only) tends to be very poor when sequences are short. If high precision is desired, then a user can focus their attention on either group 1 alone or those groups in which homology is a factor in the prediction. In the example above, this would include groups 1–4 and exclude the last two groups, 5 and 6. However, when sequences are long (>2,500 nt in length) due to assembly, predictions based on composition alone are more reliable and can be included in the final set of predictions. Also, if the user has a reasonable expectation of “who is there” based on, e.g., taxonomic assignment of marker genes, this knowledge can be used as the basis for accepting a subset of predictions from the last two groups.

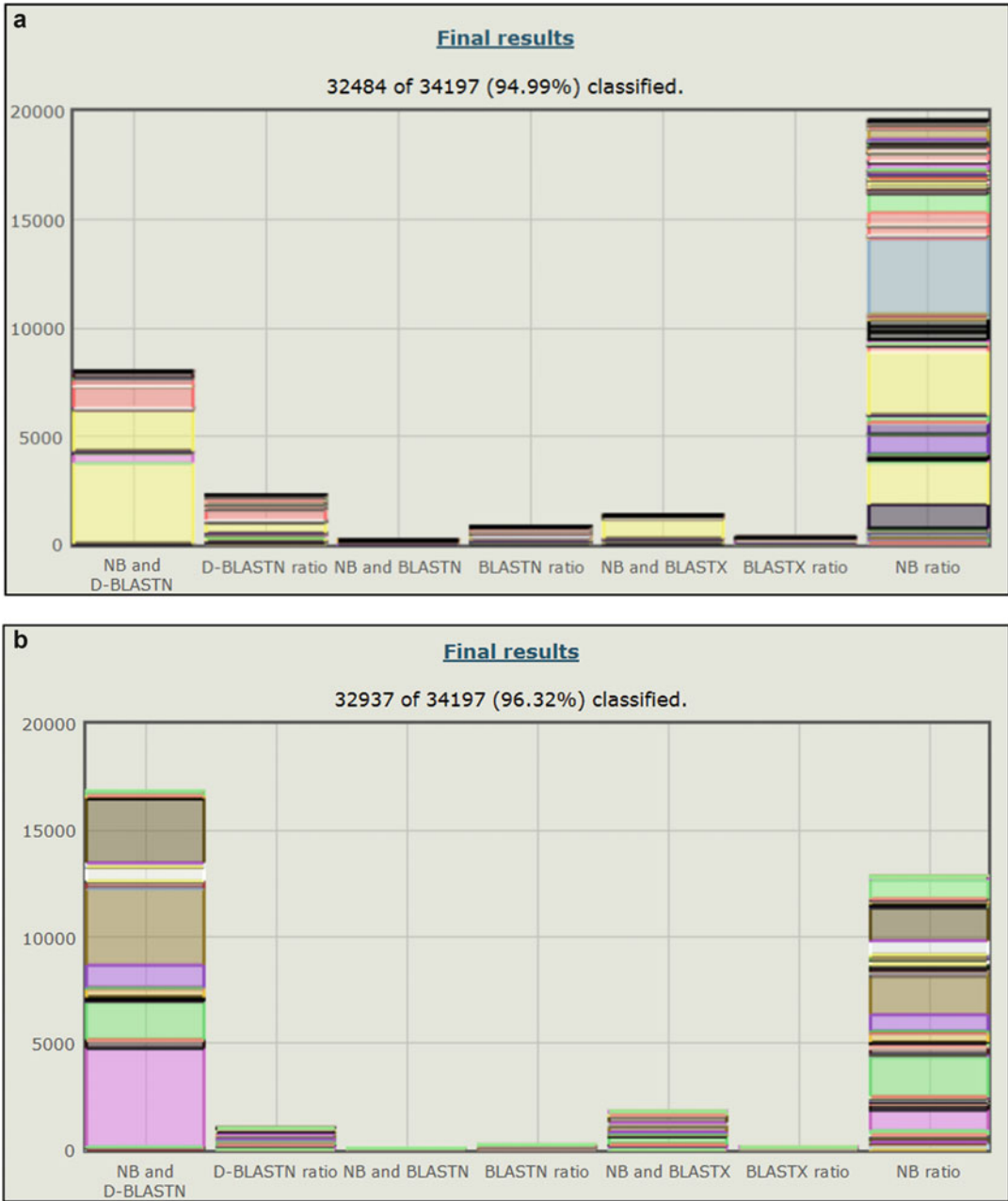
Factors Affecting the Accuracy of RITA Predictions: Several factors have been tested and shown to impact on the accuracy of RITA predictions. Among the most notable are:

Reference genome availability. Classification of a fragment to a taxonomic group at a given rank obviously depends on the existence of at least one sequenced genome from this group in the reference database. Even at the level of genus, inclusion of multiple reference genomes is desirable to adequately map out the pan-genome for homology-based predictions and to capture compositional variation within the group. Compositional signal is highly variable within order, class, and phylum, and best matching to homologs is difficult as well due to the confounding effects of gene loss and lateral gene transfer.

Consequently the classification accuracy on fragments from genomes that are taxonomically novel (i.e., that have no relatives in the reference database at ranks such as order or class) will be extremely poor. This presents a significant challenge when samples are known to be enriched in poorly represented phyla such as Verrucomicrobia, Acidobacteria, or the many candidate phyla that lack sequenced representatives. If human microbiome samples are being processed using RITA, it is highly desirable to add the draft genomes sequenced by the Human Microbiome Consortium (Markowitz et al. 2012) to increase the coverage of common human-associated taxonomic groups: the effects of including or excluding these genomes are shown in Fig. 1.

Short fragments. The effect of fragment length on classification accuracy has been extensively characterized (McHardy and Rigoutsos 2007; Brady and Salzberg 2009; MacDonald et al. 2012). While hybrid classifiers such as RITA can give accuracy in excess of 50 % even on metagenomic fragments ~50 nt in length, a high degree of misclassification is likely and many false-positive predictions can be anticipated. Restricting predictions to the “agreement” groups such as group 1 is highly desirable in this case.

Long fragments. A different problem is seen when applying RITA to long, assembled metagenomic fragments. Since RITA considers only the best BLAST match for a given fragment, the homology prediction for a long assembly will be based on one of many genes. If the prediction associated with this gene is incorrect (for instance, if it was recently transferred into the sequenced organism from a different genome), then homology and composition will likely disagree, and the entire fragment will likely be assigned to a



RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data, Fig. 1 RITA classifications of ~33,000 metagenomic fragments from obese twin gut metagenomes (Turbaugh et al. 2009). D-BLASTN = Discontiguous MEGABLAST. The number of assignments to each RITA group is shown, with different colors indicating assignments to different genera. (a) Assignments made to a set of reference genomes, excluding the draft genomes

sequenced by the HMP. A majority of sequences are assigned to the low-confidence “NB ratio” category. (b) Classifications of the same data set with inclusion of the HMP reference genomes, showing a doubling of the number of assignments to the highest-confidence group (NB and D-BLASTN) and a near-halving of assignments to the NB ratio group. Plots were generated by the RITA Web server

composition-only bin. In this case, the NB classification will likely be correct due to the length of the fragment, and inspection of the homology affinities of other genes in this fragment will likely confirm the correct classification.

“Sticky” taxa. Some genera are both extremely diverse in their gene content and well represented in the set of sequenced genomes. Notable examples of this include genera *Clostridia*, *Streptococcus*, and *Bacillus*. Since these genera have large pan-genomes and appear to share frequently with other lineages, many query fragments may be incorrectly assigned to these large groups. Care should be taken when results include a large number of these genera. Also, genera such as *Buchnera* and *Sulcia* that are dominated by small genomes tend to have low genomic G + C contents; as a consequence, fragments from low-G+C regions of other genomes may tend to be incorrectly assigned to these organisms. Since many of these organisms are restricted to highly specific settings such as insect bacteriomes, spurious matches to these groups can readily be identified.

Summary

RITA is a hybrid supervised classification system for metagenomic reads that has been shown to give useful accuracy on fragments as small as 50 nt in length. Accurate classification depends on the criteria listed above, in particular the availability of good reference databases. The key to RITA’s speed is the use of the very fast NB classifier and prioritizing the slower homology search approaches. BLASTX in particular is very slow and can increase running time by an order of magnitude, so avoiding translated nucleotide queries or using the UBLASTX algorithm of Edgar (2010) is crucial to rapid execution. Predictions based on composition alone are less reliable than those that include homology as a criterion, but composition-based predictions become more accurate with increasing sequence length.

Cross-References

- ▶ [MEtaGenome Analyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3](#)

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402.
- Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods.* 2009;6:673–6.
- Campbell A, Mrázek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci U S A.* 1999;96:9184–9.
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010;26:2460–1.
- Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.* 2011;39:e91.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007;17:377–86.
- Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics.* 2011;12 Suppl 2:S4.
- MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.* 2012;40:e111.
- Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Jacob B, Ratner A, Liolios K, Pagani I, Huntemann M, Mavromatis K, Ivanova NN, Kyrpides NC. IMG/M-HMP: a metagenome comparative analysis system for the human microbiome project. *PLoS One.* 2012;7:e40151.
- McHardy AC, Rigoutsos I. What’s in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol.* 2007;10:499–503.
- Parks DH, MacDonald NJ, Beiko RG. Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinforma.* 2011;12:328.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. *Nature.* 2009;457:480–4.

S

SATe-Enabled Phylogenetic Placement

Tandy Warnow
Institute for Genomic Biology, University of
Illinois, IL, USA

Synonyms

Evolutionary tree; Phylogenetic tree; Phylogeny;
Tree

Introduction

Taxonomic identification of DNA fragments produced in a shotgun sequencing analysis is a basic problem in metagenomic data analysis. One approach for this problem operates as follows: the fragmentary sequences are assigned to genes and then these fragmentary sequences are inserted into a calculated or precomputed taxonomy based on the same gene. The insertion of fragments of a gene sequence into a tree on full-length sequences is called “phylogenetic placement.”

The input to the phylogenetic placement problem is generally assumed to be a reference alignment *A* on full-length sequences for a gene and its maximum-likelihood tree *T* (Felsenstein 2003; Price et al. 2010; Stamatakis 2006; Swofford 2003), as well as a set *Q* of “query sequences.” The set *Q* thus represents the fragmentary

sequences, whose taxon identification is uncertain but for which the gene assignment is assumed correct. In general, the reference alignment *A* and tree *T* are also assumed correct, and so the objective is to place the fragments in *Q* into *T* as close as possible to their correct position.

Methods for phylogenetic placement include EPA (Berger et al. 2011), pplacer (Matsen 2010), PaPaRa (Berger and Stamatakis 2011), and others. Of these, EPA and pplacer are essentially identical in performance and technique: first, a Profile Hidden Markov Model (HMM, Eddy 1998) is computed for the reference alignment, and then it is used to align each of the query sequences, one at a time. Thus, |Q|-extended alignments are computed, each containing the reference sequences and one query sequence, and inducing *A* on the reference sequences. Then, maximum-likelihood methods are used to insert the query sequence into the tree *T*. The calculation of the extended alignment and the placement of a single query sequence into the tree itself is also reasonably fast; however, because there can be many query sequences, this approach can be computationally intensive. However, the analyses of different query sequences are independent, and so this process can be easily parallelized. Furthermore, this approach has good accuracy when the reference alignments and trees are correct.

In Mirarab et al. (2012), PaPaRa and pplacer were studied on a range of datasets, varying the rate of evolution and the number of sequences. This study showed that both PaPaRa and pplacer

had good accuracy for genes that evolve under low rates of evolution, but when the rate of evolution increased, then their accuracy dropped substantially. Two observations resulted: first, under a high rate of evolution, the reference alignment and tree would be difficult to estimate, an observation that has been made elsewhere. However, more surprisingly, when the sequences evolved under a high rate of evolution, even with a good alignment and tree, the technique for computing the extended alignment did not have good accuracy.

Mirarab et al. developed a divide-and-conquer technique for improving this approach to phylogenetic placement, which they termed SEPP. This technique operates by using the reference tree to divide the dataset into subsets (as used in SATé-II (Liu et al. 2012)) and then uses HMMER (Eddy 1998) to compute an HMM on each subset using the induced alignment from the reference alignment. Thus, SEPP stands for SATé-enabled phylogenetic placement. Thus, instead of using a single HMM to represent the reference alignment, a collection of HMMs is used, each on a different subset of the taxa. The calculation of the extended alignment for each query sequence is then made by using HMMER to score the fit between the query sequence and each of the subset HMMs, and the one that has the best score is used to align the query sequence to the alignment on that subset. Because the subset alignments are all in agreement with the reference alignment on the full dataset, transitivity then provides the alignment of the query sequence to the full dataset. In this way, the extended alignment of each query sequence can be computed. Once the extended alignment is calculated, the query sequence can be inserted into the reference tree using maximum likelihood, just as in EPA and pplacer.

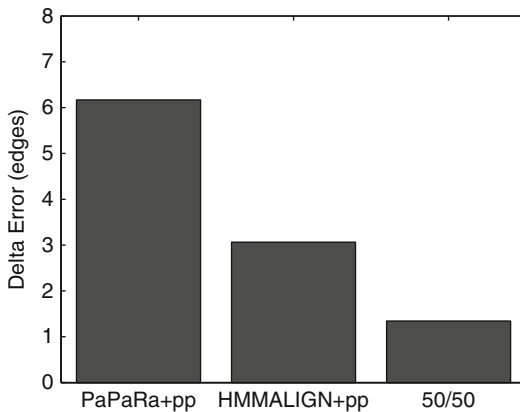
SEPP also allows the user to limit the subtree of the reference tree into which the query sequence will be placed through an additional parameter. Thus, SEPP takes two parameters: the number of leaves in the subtree on which SEPP builds an HMM (based on the induced alignment) and the number of leaves in the (perhaps larger) subtree that contains the

alignment subset, into which the query sequence is then placed. Both parameters influence the accuracy and running time of SEPP.

Thus, the most important difference between SEPP and EPA and pplacer is just how the extended alignment is computed. The technique in SEPP for calculating the extended alignment is based on decomposing the taxon set into subsets using the reference tree, and so the important issue is how the taxon set is decomposed. They used the centroid edge decomposition strategy first employed in the SATE multiple sequence alignment method (Liu et al. 2012). This strategy removes an edge that breaks the taxon set roughly in half and then repeats the process on each subtree until the desired number of subtrees is computed. Thus, SEPP is SATE-enabled phylogenetic placement.

SEPP takes two parameters – the size of the “alignment subsets” that the reference tree is decomposed into and the size of the larger subsets (called “placement subsets”) into which the query sequences can be placed after their extended alignments are computed. Both parameters impact accuracy and speed. For example, smaller alignment subsets result in better accuracy but increase the running time. Similarly, larger placement subsets improve accuracy but increase the running time. The experimental study showed that setting both parameters identically and decomposing to ten subsets gave a good trade-off between accuracy and running time.

The experimental study in Mirarab et al. (2012) showed that this default setting for SEPP gave improved accuracy compared to pplacer and PaPaRa; results from this study are reproduced below in Fig. 1. The test datasets have 500 query sequences (half “long” and half “short,” where long sequences have a length on average of 250 and short sequences have a length on average of 100), and the placement methods insert these query sequences into a reference tree and alignment on 500 full-length sequences (average length 1,000 nt). Mirarab et al. (2012) also showed that SEPP provided improved computational performance over these methods with respect to both time and peak memory usage for very large datasets.



SATe-Enabled Phylogenetic Placement, Fig. 1 Delta tree error of three phylogenetic placement methods on simulated datasets with 500 query sequences and 500 reference sequences, given the true alignment and the true tree. The delta error is the average number of additional distance from the correct placement produced by using the extended alignment rather than the true alignment in placing each query sequence. We show results obtained using PaPaRa or HMMALIGN to compute the extended alignments followed by pplacer to place each query sequence. We also show SEPP(50/50) (the default setting), which uses alignment subsets of size 50 (10% of the reference tree) to compute the extended alignment, and then places the query sequences into the same subtree. Note that SEPP(50/50) has less than half the error of HMMALIGN+pplacer, and that PaPaRa+pplacer has about double that of HMMALIGN+pplacer (reproduced from Mirarab et al. 2012, with permission from the publisher)

Summary

SEPP is a technique for performing phylogenetic placement, a basic algorithmic problem in large-scale phylogeny estimation and also in taxonomic classification of fragmentary sequences that are often produced in a shotgun sequencing analysis of metagenomic data. Phylogenetic placement methods such as EPA and pplacer use a reference tree and alignment on full-length sequences for a given gene, represent the reference alignment using a HMM, and then align each fragmentary sequence to the reference alignment using the HMM. This extended alignment for the given fragmentary sequence is then used to find the best placement in the reference tree. SEPP produces more accurate placements

than both EPA and pplacer, because instead of using a single HMM to represent the entire reference alignment, it uses multiple HMMs, each on a different subset of the taxa. Although formulated for use specifically with HMMER tools for computing HMMs and aligning sequences to the reference alignment, it could also be used to boost other phylogenetic placement methods. Finally, the divide-and-conquer technique employed in SEPP is a general technique for boosting machine learning methods that could be extended to other classification problems in bioinformatics.

Funding

This work was supported by NSF grant DEB 0733029 to T.W.

References

- Berger SA, Stamatakis A. Aligning short reads to reference alignments and trees. *Bioinformatics*. 2011; 27(15):2068–75.
- Berger SA, Krompass D, Stamatakis A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol*. 2011;60(3):291–302.
- Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14:755–63.
- Felsenstein J. *Inferring phylogenies*. Sunderland: Sinauer Associates; 2003.
- Liu K, Warnow T, Holder M, Nelesen S, Yu J, Stamatakis A, Linder CR. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol*. 2012;61(1):90–106.
- Matsen F, Kodner R, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*. 2010;11:538.
- Mirarab S, Nguyen N, Warnow T. SEPP: SATe-enabled phylogenetic placement. *Pacific Symposium on Biocomputing*. 2012.
- Price M, Dehal P, Arkin A. FastTree 2 - approximately maximum likelihood trees for large alignments. *PLoS ONE*. 2010;5:e9490.
- Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
- Swofford DL. *PAUP*: phylogenetic analysis using parsimony (*and other methods)*, version 4. Sunderland: Sinauer Associates; 2003.

Serial Analysis of V1 Ribosomal Sequence Tags

Zhongtang Yu¹ and Mark Morrison²

¹Department of Animal Sciences, Environmental Science Graduate Program, The Ohio State University, Columbus, OH, USA

²Diamantina Institute, The University of Queensland, Woolloongabba, Brisbane, QLD, Australia

Synonyms

Serial analysis of ribosomal sequence tags (SARST); Serial analysis of V6 ribosomal sequence tags (SARST-V6)

Definition

By this technique, the V1 hypervariable region of 16S rRNA genes is amplified as a ribosomal sequence tag (RST) by PCR using universal primers, concatenated head to tail, cloned, and sequenced. By enabling multiple RSTs to be sequenced from each RST concatemer, SARST-V1 substantially increases the number of sequences on either the Sanger or next-generation sequencing platforms, thus, increasing the depth of coverage of microbiome analysis.

Introduction

Sufficient characterization of actual diversity is a prerequisite to understanding the function of microbiomes and to exploring and manipulating them for beneficial applications. Sequencing and phylogenetic analysis of 16S rRNA genes have been the primary approaches in such a pursuit. Detailed characterization of microbiomes, however, requires a large number of 16S rRNA genes to be sequenced from each microbiome sample, especially when members present at low abundance need to be identified. Before next-generation sequencing (NGS) technologies for

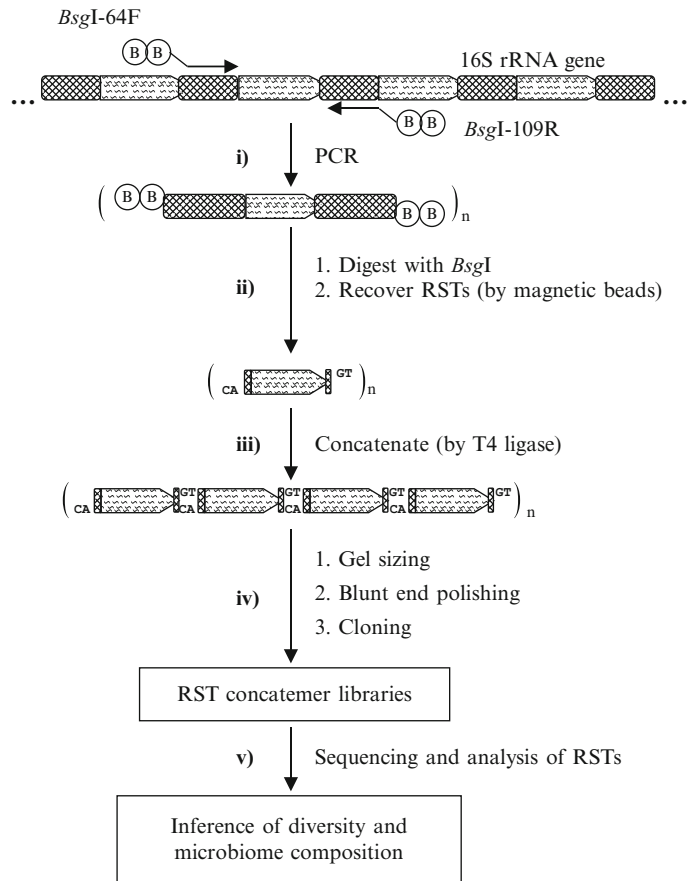
DNA sequencing became available, it was not feasible to generate adequate sequences of 16S rRNA genes from any complex microbiome because 16S rRNA genes first need to be cloned, and then individual clones need to be sequenced in a one-clone-one sequence fashion, which is a costly and labor-intensive process. Indeed, all the 16S rRNA gene datasets produced by the Sanger sequencing technology are too small to capture the full diversity (Bent and Forney 2008; Tiedje et al. 1999). One strategy to reduce the cost of DNA sequencing-based analysis of microbial diversity is to sequence concatemers of a sequence tag of 16S rRNA genes using the serial analysis of ribosomal sequence tags (SARST) (Kysela et al. 2005; Neufeld et al. 2004; Yu et al. 2006). SARST was adapted from the serial analysis of gene expression (SAGE) (Velculescu et al. 1995), an approach first developed to substantially improve analysis of gene expression in eukaryotes (Carulli et al. 1998). In SARST, one of the hypervariable regions of 16S rRNA genes is used as a sequence tag. By far, SARST has been developed based on either hypervariable V1 (referred to as SARST-V1) (Yu et al. 2006) or V6 (SARST-V6) (Kysela et al. 2005). Except for the two different hypervariable regions as the sequence tags, SARST-V1 and SARST-V6 have similar procedures.

Overview of SARST-V1 Procedures

The entire process of SARST-V1 (Fig. 1) consists of (i) amplification of the V1 region of 16S rRNA genes using a pair of bacterial primers; (ii) digestion of the PCR amplicons to cut off the primers; (iii) purification and concatenation of individual ribosomal sequence tags (RSTs); (iv) gel sizing, end repair, and cloning of the concatemers; and (v) sequencing of cloned RST concatemers and phylogenetic analysis of individual RSTs. The detailed procedures have been described elsewhere (Yu and Morrison 2011; Yu et al. 2006). Here we describe the major steps, alternatives, and cautions when warranted.

Serial Analysis of V1 Ribosomal Sequence Tags, Fig. 1

Schematic of the SARST-V1 process. BB, dual biotin label conjugated to the 5' end of the primers. *BsgI*-Bact64F and *BsgI*-Bact109R, bacterial forward primer and reverse primer, each with an extension containing a *BsgI* recognition site (the figure was modified from reference 22 with permission from Wiley-Blackwell)



(i) **PCR amplification.** The V1 region is amplified using *BsgI*-Bact64F (5'-dual biotin-TTT GAC CGT GCA GCY TAA YRC ATG CAA GTCG-3') and *BsgI*-Bact109R (5'-dual biotin-TTT GAC CGT GCA GYY CAC GYG TTA CKC ACC CGT-3'). Each primer has an extension region that contains a recognition site for *BsgI* (bolded and underlined), and the most 5' nucleotide of this extension region is labeled with at least one biotin or biotin-tetra-ethyleneglycol (biotin-TEG) molecule.

The quality and quantity of the PCR products are evaluated using PAGE (8 %T, 19:1) mini gel. Then, the PCR products are purified using the QIAquick PCR Purification Kit (QIAGEN, Valencia, CA) or by ethanol precipitation following extraction with phenol/chloroform. Hot-start PCR using a hot-start

Taq DNA polymerase or a hot-start dNTP mix is recommended in the PCR amplification to reduce formation of primer dimers, which can contaminate the RSTs.

(ii) **Digestion of PCR products and primer removal.** The purified PCR products are digested with *BsgI*, a type II restriction endonuclease that cuts 16 base pairs (bp) downstream from the recognition site. The released RSTs are separated from the primers using streptavidin-coated magnetic beads, such as Dyna 280 beads (Dyna, Oslo, Norway), which immobilize the primers that have a biotin label at the 5' end.

(iii) **Concatenation of individual RSTs.** Each of the freed RSTs has one 2-nt overhang at both 3' termini, and these overhangs facilitate annealing of individual RSTs in hand-to-tail orientation in series (Fig. 1).

Consequently, individual RSTs are ligated head-to-tail to form concatemers in 5'–3' orientation. Because of the short overhangs and the desire to form long (>0.5 kb) concatemers, the concatenation is often performed overnight using a DNA ligase, such as T4 DNA ligase. It should be noted that *BsgI* is the most suitable type II endonuclease available. New type II endonucleases that produce 3- or 4-nt overhangs will improve concatenation. Additionally, when new primers are designed to target other V regions, the recognition site of *BsgI* (or the type II restriction endonuclease used) should be at such a distance from the 3' end that digestion of the PCR products leaves at least five base pairs of the primers at each end of the freed RSTs. These conserved base pairs allow delineation of individual RSTs from the sequenced concatemers.

- (iv) **Gel sizing, end repair, and cloning of concatemers.** Concatenation of individual RSTs produces concatemers of varying lengths. The concatemers of 0.5–2.0 kb need to be size selected, typically using gel (either agarose or polyacrylamide) electrophoresis, and recovered from the gel slice using commercial kits, such as the MinElute Gel Extraction Kit (QIAGEN). Following end repair by T4 DNA polymerase, the concatemers are then cloned by ligation into a cloning vector (e.g., pZero-2.1 from Invitrogen or pSmartLCKan from Lucigen) that has been digested with a blunt-end restriction endonuclease. Alternatively, an adenine overhang can be added to each end of each concatemer so that the concatemers can be cloned using the TOPO TA cloning kit (Invitrogen). Direct cloning of the blunt-ended concatemers might be preferred because it increases cloning efficiency of SAGE concatemers (Koehl et al. 2003), which have similar length as RST concatemers.
- (v) **Sequencing and phylogenetic analysis of cloned RST concatemers.** The cloned concatemers are sequenced using the Sanger

DNA sequencing technology. A typical Sanger sequencing read (greater than 500 bp) can determine the sequence of 19 individual RSTs (Yu et al. 2006). Individual RSTs are then delineated using the conserved base pairs that flank individual RSTs. The individual RSTs first can be grouped into OTUs and then compared to databases (Neufeld et al. 2004; Poitelon et al. 2009; Yu et al. 2006), or they can be compared to databases without grouping (Kysela et al. 2005). BLASTn and SEQ MATCH are two programs that can be used to compare RSTs to the sequences archived in GenBank (<http://www.ncbi.nlm.nih.gov/>) and RDP (<http://rdp.cme.msu.edu/>), respectively. Other programs, such as ESPRIT (Sun et al. 2009), Mothur (Schloss et al. 2009), Qiime (Caporaso et al. 2010), CD-HIT (Li and Godzik 2006), and UniFrac (Lozupone et al. 2006), can also be used in RST analysis. Most of the RST datasets produced in previous studies can be found either in the NCBI Gene Expression Omnibus (GEO) database (Ashby et al. 2007; Neufeld et al. 2004; Yu et al. 2006) or in the Sequence Read Archive (SRA) (Huber et al. 2007).

Full-length 16S rRNA gene sequences can be grouped or assigned to species and genera using 97 % and 95 % sequencing similarity as the cutoff values, respectively (Ludwig et al. 1998; Stackebrandt and Goebel 1994). However, most researchers also use these cutoff values in analyzing partial sequences. Because sequence divergence is not evenly distributed along the 16S rRNA gene (particularly among the nine V regions), different cutoff values are needed when different regions of 16S rRNA genes are analyzed (Kim et al. 2011). As such, different cutoff values of sequence similarity are needed to group and assign individual RSTs to RST-based OTUs. Alternatively, individual RSTs can be compared to rRNA gene sequence databases to identify longer sequences, which can then be used to characterize the microbiomes.

SARST Based on Other V Regions

Besides the V1 region, the V6 region (987–1,045 nt) (Kysela et al. 2005; Poitelon et al. 2009) has also been used in SARST. As demonstrated in analyzing soil, rumen, and marine samples, both V1 and V6 appeared to have sufficient phylogenetic information to allow taxonomic assignments of the recovered RSTs (Neufeld and Mohn 2005; Neufeld et al. 2004; Pinloche et al. 2013; Poitelon et al. 2009; Yu et al. 2006). The V5 region has also been used in a SAGE-like analysis (referred to as serial analysis of rRNA genes, SARD) of the soil microbiome (Ashby et al. 2007). Since it only generates 14-bp RSTs, SARD may not provide enough phylogenetic information for reliable taxonomic assignments of the RSTs. Other V regions can be targeted in SARST, but when choosing a V region for SARST, the following need to be considered: the length and divergence of sequence, the availability of universal primers, and the frequency of the recognition site of the type II restriction endonuclease chosen within the V region.

Summary

SARST was developed before NGS technologies became available, and it significantly improved upon the traditional one-clone-one-sequence approach with respect to both cost and coverage. SARST will still be useful when NGS technologies is affordable. First, SARST-V1 can generate as much phylogenetic information as longer 454 pyrosequencing reads (Pinloche et al. 2013). Second, deep coverage is not compromised when multiple bar-coded microbiome samples are analyzed simultaneously in a single NGS run. Additionally, as the read length of NGS continues to increase, concatemers of RSTs can be sequenced without cloning of RST concatemers.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)

References

- Ashby MN, Rine J, Mongodin EF, Nelson KE, Dimster-Denk D. Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. *Appl Environ Microbiol.* 2007;73:4532–42.
- Bent SJ, Forney LJ. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J.* 2008;2:689–95.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7:335–6.
- Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, Guerin J, Osborne M, Stein G, Lian J, Lomedico PT. High throughput analysis of differential gene expression. *J Cell Biochem Suppl.* 1998;30–31(Suppl):286–96.
- Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. Microbial population structures in the deep marine biosphere. *Science.* 2007;318:97–100.
- Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods.* 2011;84:81–7.
- Koehl A, Friauf E, Nothwang HG. Efficient cloning of SAGE tags by blunt-end ligation of polished concatemers. *Biotechniques.* 2003;34:692–4.
- Kysela DT, Palacios C, Sogin ML. Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. *Environ Microbiol.* 2005;7:356–64.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.

- Lozupone C, Hamady M, Knight R. UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*. 2006;7:371.
- Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M, Neumaier J, Bachleitner M, Schleifer KH. Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis*. 1998;19:554–68.
- Neufeld JD, Mohn WW. Unexpectedly high bacterial diversity in Arctic Tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. *Appl Environ Microbiol*. 2005;71:5710–8.
- Neufeld JD, Yu Z, Lam W, Mohn WW. Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities. *Environ Microbiol*. 2004;6:131–44.
- Pinloche E, McEwan N, Marden JP, Bayourthe C, Auclair E, Newbold CJ. The effects of a probiotic yeast on the bacterial diversity and population structure in the rumen of cattle. *PLoS ONE*. 2013;8:e67824.
- Poitelon JB, Joyeux M, Welte B, Duguet JP, Prestel E, Lespinet O, DuBow MS. Assessment of phylogenetic diversity of bacterial microflora in drinking water using serial analysis of ribosomal sequence tags. *Water Res*. 2009;43:4197–206.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75:7537–41.
- Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in Bacteriology. *Int J Syst Bacteriol*. 1994;44:846–9.
- Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W. ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res*. 2009;37:e76.
- Tiedje JM, Asuming-Brempong S, Nüsslein K, Marsh TL, Flynn SJ. Opening the black box of soil microbial diversity. *Appl Soil Ecol*. 1999;13:109–22.
- Velculescu V, Zhang L, Vogelstein B, Kinzler K. Serial analysis of gene expression. *Science*. 1995;270:484–7.
- Yu Z, Morrison M. Sequence-based characterization of microbiomes by Serial Analysis of Ribosomal Sequence Tags (SARST). *Handbook of molecular microbial ecology I*. Wiley; 2011. p. 265–73.
- Yu Z, Yu M, Morrison M. Improved serial analysis of V1 ribosomal sequence tags (SARST-V1) provides a rapid, comprehensive, sequence-based characterization of bacterial diversity and community composition. *Environ Microbiol*. 2006;8:603–11.

SILVA Databases

Christian Quast¹, Elmar Pruesse¹, Jan Gerken¹, Timmy Schweer¹, Pelin Yilmaz¹, Jörg Peplies² and Frank Oliver Glöckner^{3,4}

¹Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

²Ribocon GmbH, Bremen, Germany

³Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

⁴Jacobs University Bremen gGmbH, Bremen, Germany

Synonyms

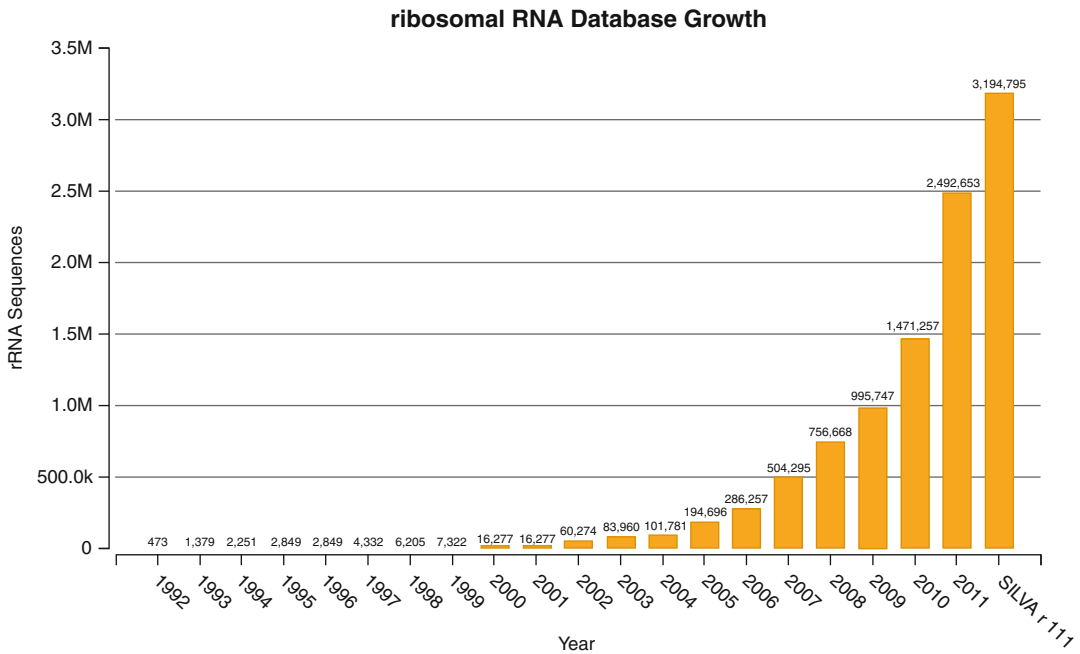
Alignment; Classification; Probe and primer evaluation; Quality assessment; Ribosomal RNA gene datasets; Taxonomy

Definition

SILVA (from Latin *silva*, forest) is a comprehensive web resource (<http://www.arb-silva.de>) for up-to-date, quality-controlled databases of aligned ribosomal RNA gene (rDNA) sequences from the *Bacteria*, *Archaea*, and *Eukarya* domains.

Introduction

Sequencing the ribosomal RNA gene (rDNA) is the method of choice for nucleic acid-based detection and identification of microbes, their taxonomic assignment, phylogenetic analysis, and investigation of microbial diversity. Today (July 2012), more than 3.5 million small and large subunit (SSU and LSU) rDNA sequences are publicly available and their analysis demands for appropriate software tools and specialized, quality-controlled databases. The SILVA datasets, established in 2007, provide high-quality,



SILVA Databases, Fig. 1 Growth of the ribosomal RNA databases since 1992

comprehensive rDNA datasets comprising sequences from the *Bacteria*, *Archaea*, and *Eukarya* domains. All sequences are checked for anomalies, carry a rich set of sequence-associated contextual information, and have multiple taxonomic classifications and the latest validly described nomenclature. The SILVA datasets are based on the EMBL/EBI Nucleotide Sequence Database (EMBL-Bank), a member of the International Nucleotide Sequence Database Collaboration (INSDC) comprising all publicly available DNA sequences. They are generated by an automatic software pipeline for the extraction of SSU and LSU rDNA sequences as well as quality control. The alignment is based on the latest comprehensive ARB (Ludwig et al. 2004) alignments. The datasets are extensively annotated by third-party data integration. Substantial manual curation of the alignment and taxonomy is performed on each public release. SILVA dataset updates and new online features are continuously released on the SILVA web portal (<http://www.arb-silva.de>)

which provides detailed statistics and documentation of the resource.

SILVA Datasets

The SILVA project provides datasets for all SSU and LSU rDNA sequences found in EMBL-Bank that fulfill the SILVA quality criteria. Since their first public release in February 2007, based on EMBL-Bank release 89, these datasets have increased in size by a factor of 10 and 5 for the SSU Parc and LSU Parc datasets, respectively. Moreover, the growth is clearly exponential (Fig. 1) as is the growth of the general DNA sequence databases. Detailed information on the current SILVA database content can be found in the documentation section of the SILVA web portal.

The SILVA SSU and LSU rDNA datasets each consist of two subsets: (1) the “Parc” datasets comprising the complete SILVA

database content and (2) the “Ref” datasets comprising high-quality subsets of sequences in the Parc datasets. For the SSU dataset, additionally, two subsets are provided, (3) the “Ref NR” dataset, a nonredundant version of the Ref subset, and (4) a type strain dataset provided by the “All-Species Living Tree Project” (LTP) (Munoz et al. 2011) which is also available for the LSU rDNA. All datasets and individual subsets can be downloaded as ARB files for direct use with the ARB software package (Ludwig et al. 2004). In the future, only SSU Ref and SSU Ref NR datasets will be offered in the ARB format to avoid unmatched hardware demands on the user side.

SILVA Parc

All sequences in the Parc datasets have a minimum length of 300 aligned nucleotides within the boundaries of the rRNA genes. Sequences are only accepted if they have less than 2 % ambiguities or homopolymers or vector contamination. Additionally, after the alignment, minimal quality requirements for sequence quality and base-pair score, as well as alignment identity and quality, are applied. For details, please refer to the section “Quality Control” and the respective dataset documentation page (e.g., <http://www.arb-silva.de/documentation/background/release-111/>).

SILVA Ref

The SILVA Ref datasets represent subsets of the corresponding SILVA SSU and LSU Parc datasets. They comprise only “full-length” or nearly “full-length” sequences. An SSU sequence is considered to be of “full length” if it contains at least 1,200 aligned bases within the rRNA gene boundaries. For sequences classified as *Archaea*, this threshold has been lowered to 900 aligned bases to avoid losing the majority of sequences. LSU sequences are considered “full length” if they are at least 1,900 bases long.

More stringent thresholds for alignment quality and identity are applied for the Ref datasets. Consequently, the Ref datasets contain considerably less sequences than the Parc datasets,

particularly about one-fourth in case of the SSU Parc database and about one-tenth for the LSU Parc database (ratios for the SILVA 111 release).

Sequences originating from the “Human Skin Microbiome” (HSM) (Grice et al. 2009), the “Mouse Wound Microbiota” (MWM) (Grice et al. 2010), and the “Guerrero Negro Hypersaline Microbial Mat” (GNHM) large-scale sequencing projects are excluded from the SSU Ref dataset. Instead, these sequences, with more than 490,000 (SILVA 111) long sequence reads in total, are provided in a dedicated dataset. This is done to further restrict the size of the SILVA SSU Ref dataset and to avoid overrepresentation of sequences of a specific origin.

For both SILVA Ref datasets, the ARB files are supplemented with a manually classified “guide tree,” incrementally built using the ARB parsimony tool with filters to remove highly variable positions and followed by removal of sequence entries represented by anomalous tree branch lengths. These trees also represent the basis for the SILVA taxonomy (see section “SILVA Taxonomy” below).

SILVA Ref NR (Nonredundant)

For users interested in a representative SSU rDNA sequence collection, the SILVA project offers a nonredundant (NR) version of the SSU Ref subset. This dataset is created by applying clustering at 99 % (up to SILVA 108) and 98 % (from SILVA 111 on) sequence identity. Of each cluster, only the longest sequence is kept. This reduces the size of the dataset to less than 50 % of its original size, even though the sequences omitted in the SSU Ref dataset from the HSM, MWM, and GNHM projects (see above) are included for clustering. Sequences from cultivated species are preserved in all cases to lead as an anchor for taxonomy. The resulting SSU Ref NR dataset with its manually curated “guide tree” can be used as a representative dataset for classification, phylogenetic analysis, and probe design. It is the recommended dataset to be used as a starting point for all users interested in environmental rDNA sequence analysis.

SILVA Taxonomy

A substantial revision of the classification of all prokaryotic sequences in the Ref datasets was first published with SILVA release 100. Based on the “guide trees,” all phylogenetic assignments are manually curated, taking into account taxonomic information provided by Bergey’s Taxonomic Outline of the Prokaryotes (Garrity et al. 2004); the taxonomic outlines for volumes 3, 4, and 5 of Bergey’s Manual; and the List of Prokaryotic names with Standing in Nomenclature (Euzéby 1997). Furthermore, extensive effort is spent to represent prominent uncultured and not validly published environmental clades, groups, and taxa, respectively. The majority of these clades and groups are annotated in the “guide tree” based on literature surveys and personal communications. Taxonomic groups consisting only of sequences from uncultured organisms are named after the clone sequence submitted earliest. Due to this exhaustive manual approach, SILVA currently contains the most up-to-date and detailed bacterial and archaeal taxonomic classification.

To create also an improved and unified taxonomy for *Eukarya* based on 18S rDNA sequences, the Eukaryotic Taxonomy Working Group (ETWG) has been founded in October 2011. The first version of these efforts is deployed with SILVA release 111.

SILVA SEED Datasets

SILVA uses customized and specialized reference datasets for specific tasks within its software pipeline. Such internal reference datasets are called SEEDs.

SEEDs for Alignment

As of July 2012, the SEED used for SSU rRNA gene sequence alignment has 50,000 alignment positions including all gaps and consists of about 57,000 high-quality, aligned SSU rDNA reference sequences. The alignment SEED of the LSU rRNA gene comprises 150,000 positions but includes only about 3,000 aligned sequences. Both SEEDs contain representative sequences from the *Bacteria*, *Archaea*, and *Eukarya*

domains and are manually curated and continuously enhanced.

SEEDs for Quality Control

The SEED used for the detection of sequence anomalies in SSU sequences is based on the corresponding alignment SEED with all sequences removed if any indication of an anomaly was found. This reduces the size of the SEED by a factor of 6. The detection of anomalies is not done for LSU rDNA sequences because none of the available tools can be applied.

For identification of vector contaminations, a SEED based on the EMVEC (EBI) and UniVec (NCBI) reference datasets is used with all sequences removed resembling an rDNA sequence.

Data Retrieval

Three strategies are applied to retrieve SSU and LSU rDNA sequences from EMBL-Bank:

- A keyword search is used to extract annotated SSU and LSU rDNA sequences. Additionally, a set of relaxed keywords is applied to account for sequences with spelling mistakes in the annotation.
- A whitelist taken from the Ribosomal Database Project (RDP) (Cole et al. 2005) is used to retrieve sequences that are not covered by the keyword search.
- HMMs (one for each of the three domains of life for both LSU and SSU) taken from the RNAmmer tool (Lagesen et al. 2007) are searched against the complete EMBL-Bank. Sequences that match one of the HMMs and were not already imported by one of the two previous approaches are added.

In all cases, the entries in the datasets are flagged by its origin of retrieval.

Alignment

After import, sequences are aligned using the SINA software (SILVA Incremental Aligner)

(Pruesse et al. 2012). Similar to the ARB project, the tool follows the concept of an incremental alignment. Briefly, no de novo multiple sequence alignment is created; instead, the highly accurate manual alignment of closely related sequences found in the corresponding alignment SEED is used as a template to align each sequence included in the SILVA datasets. This approach guarantees a high-quality alignment of rDNA sequences.

Quality Control

Every imported and aligned SSU and LSU gene sequence has to pass a multistage quality inspection to assure the high quality of the SILVA datasets. Sequences are checked for sequence and alignment quality using various parameters. Sequences are excluded from the SILVA releases in case they fail any of the applied tests or show reduced quality based on combined quality values. Additionally, sequences are tested for anomalies but no filtering is done by the SILVA project based on these results. The information is provided to the users for individual filtering of the datasets, if required.

Detailed statistics on the SILVA quality control can be found on the SILVA web portal for all SILVA releases, e.g., <http://www.arb-silva.de/documentation/background/release-111/>.

Sequence Quality

The SILVA sequence quality checks test for ambiguous bases, extended homopolymeric stretches, and vector contaminations.

Ambiguous bases are nucleotides representing valid characters according to the International Union of Pure and Applied Chemistry (IUPAC) DNA encoding but do not resolve to “A”, “C”, “G”, or “T”. A maximum of two percent of ambiguous nucleotides within the rRNA gene boundaries is allowed by SILVA.

Homopolymers are stretches of identical nucleotides that commonly appear with a maximum of up to four nucleotide repetitions in native rDNAs. In contrast, extended stretches within a sequence represent an indication of

reduced sequence quality caused by the sequencing process. As a consequence, if homopolymers of five or more nucleotides are found within a sequence and these stretches count for more than 2 % of the sequence within the rRNA gene boundaries, the sequence is excluded from the SILVA datasets.

Unaligned overhangs of a sequence are checked against the vector SEED using BLAST (Altschul et al. 1998) to identify cloning artifacts. If it is likely that the unaligned part of a sequence is a vector sequence and the unaligned part is longer than the aligned part, the sequence is excluded from SILVA. Sequences in SILVA are not allowed to contain more than 2 % vector contamination.

The three parameters are combined into an overall “sequence quality” value. This score represents the mean of the three individual parameters. It is normalized to values in the range of 0–100, such that 100 represents the best possible quality of a sequence.

All thresholds to reject a sequence were defined based on statistical analysis of the retrieved SSU and LSU rDNA sequences.

Alignment Quality

Four characteristics of the alignment process are evaluated in the pipeline and a sequence is rejected if it fails to pass one of these: the base-pair score, the alignment quality, the alignment identity, and the alignment length within the boundaries of the rRNA gene.

The base-pair score is calculated from the number of bases involved in helix binding according to the secondary structure model of Gutell et al. (1994).

The alignment quality score is a measure of the identity of the query sequence to the reference sequences that are used as a template for the alignment. High values (>90) indicate that closely related sequences have been found in the alignment SEED and that the resulting alignment is likely to be accurate. Low values suggest that further manual inspection of the particular sequence is needed.

Additionally, the alignment identity of the query sequence to its closest relative in the

alignment SEED is considered to guarantee the specificity of the alignment. Two positions in the alignment are considered identical if both positions have the same unambiguous nucleotide according to the IUPAC encoding.

To fit the SILVA unified scoring scheme, the base-pair and alignment quality scores are normalized to values between 0 and 100, such that 100 represents the maximum score.

Chimera/Anomaly Detection

To detect sequence anomalies, a customized version of the Pintail software (Ashelford et al. 2005) is used. This software checks whether a pair of sequences is mutually anomalous (e.g., chimeric) by computing a distance profile and comparing it to a predicted distance profile. The result is “yes,” “likely,” or “no,” depending on the amount of measured deviation from expectation. From this operation, the SILVA Pintail score is constructed by running each sequence against the ten most similar sequences retrieved from the chimera SEED. Sequences that have passed all tests with “no” (not anomalous) get a score of “100 %,” whereas all tests returning “likely” would yield a 50 % score. Only SSU sequences are checked for anomalies because the Pintail software does not contain profiles for sequences other than 16S rDNAs.

Third-Party (Meta) Data

One of the unique features of the SILVA datasets is extensive data integration based on various third-party resources and manifold linkage of the SILVA database entries to external data sources.

Taxonomies

Every sequence in the SILVA databases carries the EMBL-Bank taxonomy assignment. Where available, the greengenes (DeSantis et al. 2006) and RDP (Cole et al. 2005) taxonomies are added for comparison. All entries of the SILVA Ref datasets are also assigned to the taxonomy of the SILVA project (see section “[SILVA Taxonomy](#)”).

For LSU rDNA sequences, only the EMBL-Bank and SILVA taxonomy are available due to a lack of additional resources.

Nomenclature

With every SILVA release, all organism names are updated according to the “Nomenclature Up-to-Date” website of the “Deutsche Sammlung für Mikroorganismen und Zellkulturen” (DSMZ). All synonyms and name replacements are recorded.

Strain Annotation

The strain field of an entry in the SILVA datasets is annotated using SILVA-specific labels if an entry matches one or more of the following criteria:

- The label “e[G]” is added if an entry is part of the list of genomes offered by the EBI.
- The label “I[T]” is added if the entry is part of the type strain datasets of “The All-Species Living Tree” Project (Munoz et al. 2011).
- The label “s[T]” is added if an entry is listed as a type strain by the StrainInfo project (Dawyndt et al. 2005).
- The label “s[C]” is added if an entry is a cultured strain according to the StrainInfo project.
- The label “r[T]” is added if an entry is listed as a type strain by the RDP project.

Furthermore, manually curated habitat information and GPS coordinates are assigned to each entry based on information provided by the megx.net project (Kottmann et al. 2010).

SILVA Website/Online: Service

One of the problems associated with the ever increasing amount of sequences is the hardware resources required to store and analyze the data. As a response to allow users to still work with these datasets, features requesting comprehensive reference datasets such as probe and primer evaluation for testing the *in silico* accuracy of oligonucleotide signatures are now offered by the SILVA web portal. Additionally, the SILVA website offers extensive data retrieval

The screenshot displays the SILVA Taxonomy Browser interface. At the top, the SILVA logo and navigation links (Home, Browser, Search, Aligner, Download, Documentation, Projects, FISH & Probes, Shop, Contact) are visible. The main content area shows a taxonomy tree on the left and a detailed view of the selected sequence on the right. The taxonomy tree is expanded to show the following structure:

- SILVA > Bacteria > Proteobacteria > Alphaproteobacteria > Caulobacterales > Caulobacteraceae > Amorphus > DQ097300
- Caulobacterales (0.14%)
 - Caulobacteraceae (0.18%)
 - Amorphus (100%)
 - Amorphus coralli
 - Amorphus sp. YIM D10
 - uncultured bacterium

The detailed view on the right for accession number DQ097300 includes the following information:

- Accession Nr: DQ097300
- Description: Amorphus coralli strain R5.Sph.026 16S ribosomal RNA gene, partial sequence.
- Regions: 1
- Length: 1438
- Quality:
 - Sequence: [Progress bar]
 - Alignment: [Progress bar]
 - Pintail: [Progress bar]
- Links:
 - Details
 - Link to EMBL
 - This page (permalink)

SILVA Databases, Fig. 2 The entry of *Amorphus coralli* (DQ097300) within the genus *Amorphus* displayed in the SILVA “Taxonomy Browser”

functions for the compilation of individual sequence subsets from the comprehensive online database as well as preconfigured, quality-constrained subsets for direct download.

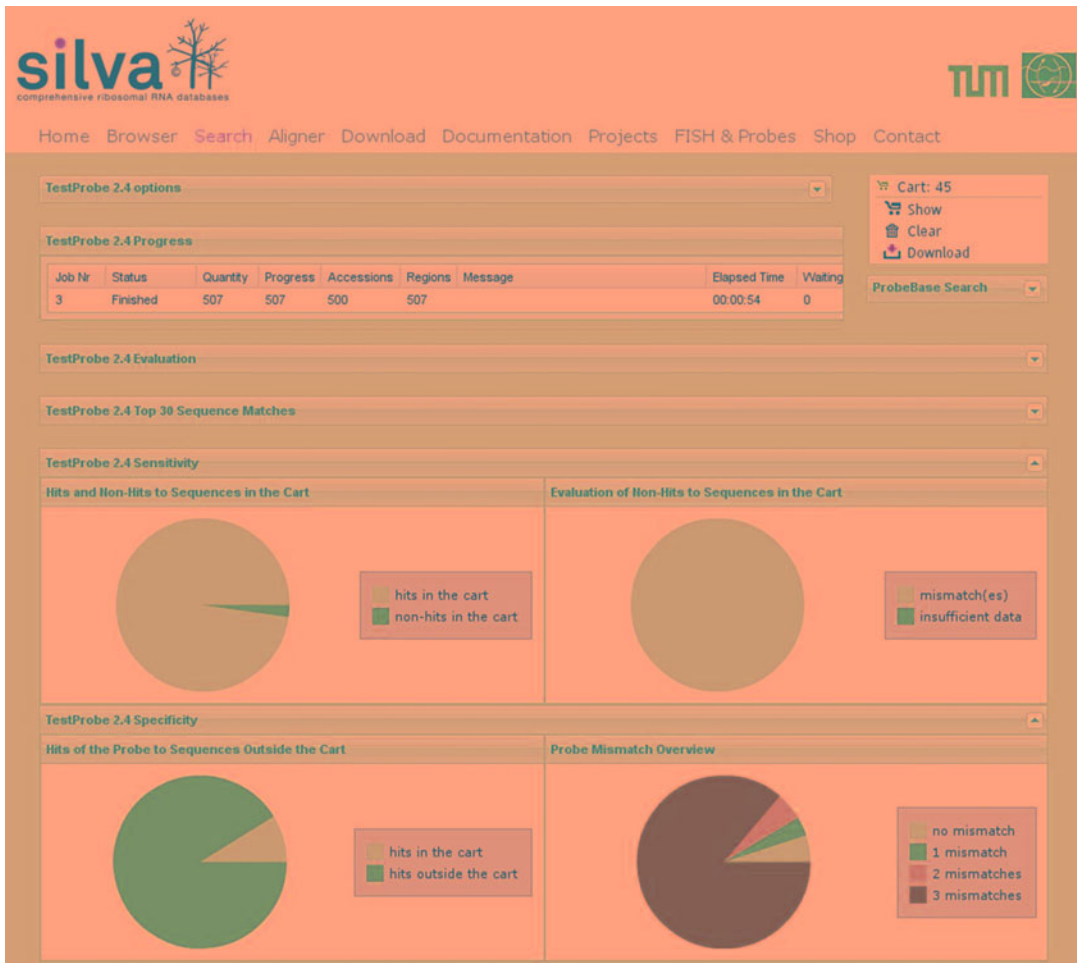
Taxonomy Browsing, Searching, and Download Cart

The SILVA “Taxonomy Browser” allows navigation through a selected taxonomy by clicking on the respective nodes. The browser starts with showing all taxonomic groups of the highest level of the selected taxonomy. By selecting one of these groups, a new list view appears with all subgroups, preserving the former levels within a horizontal scroll bar layout. If a sequence entry is selected, a detailed summary will be opened. This summary shows full annotation of an entry and a traffic light like view of the main quality parameters (Fig. 2).

The browser can also be used to create customized subsets of the SILVA databases and to display the results of the online services provided by SILVA. For each taxonomic group in the browser, the fraction of corresponding sequences in the cart can be highlighted (Fig. 2).

The advanced search functionality offered on the SILVA website allows the user to easily compile custom subsets of sequences. Besides simple searches, e.g., for accession numbers, organism names, taxonomic entities, or publication DOI/PubMed IDs, complex queries including several database fields are also possible. Constraints such as the sequence length or quality values can be used to further filter the sequences.

Customized sequence subsets compiled by the user including the results of the SILVA online services can be collected in the SILVA cart system and downloaded in various formats.



SILVA Databases, Fig. 3 The web interface and results of the SILVA “TestProbe” service

Alignment, Sequence-Based Searches, and Classification

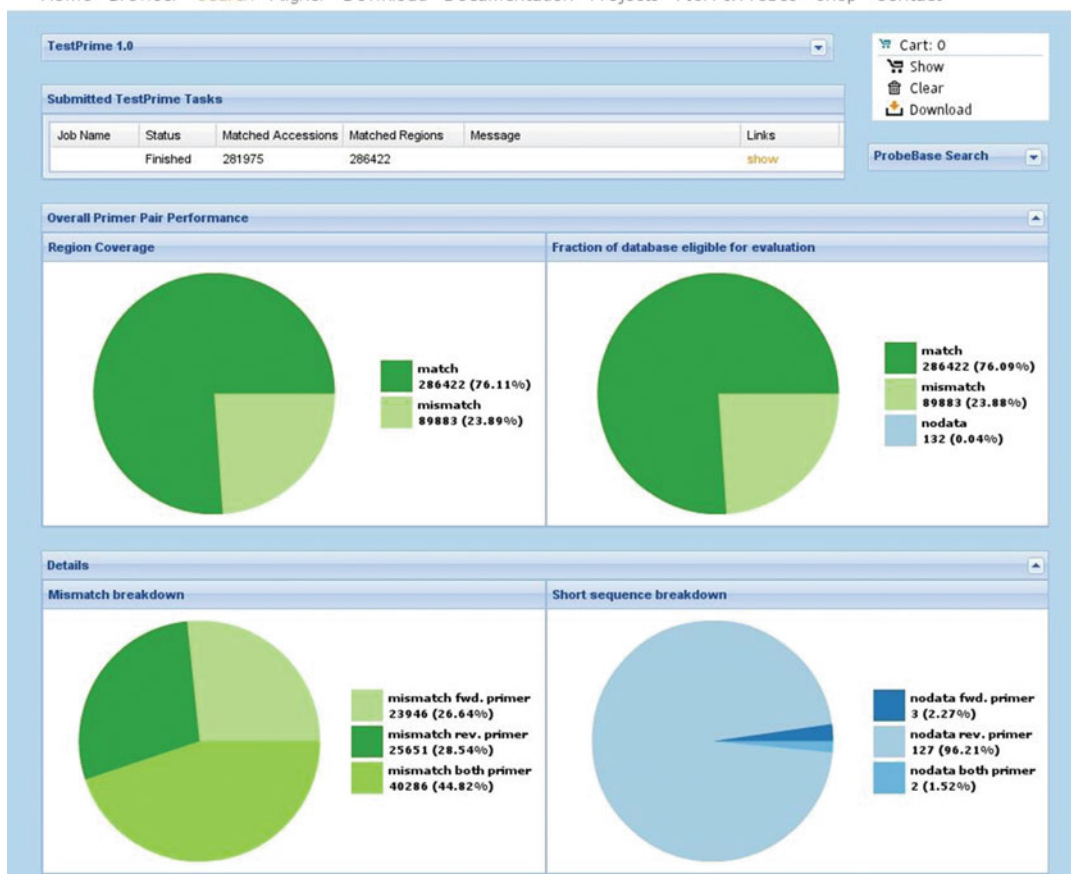
Users can align their own sequences using the SILVA SSU and LSU SEEDs with a fully configurable online version of the SILVA aligner (SINA). The aligned sequences can be downloaded in either ARB or FASTA file formats.

Submitted sequences can also be searched against one of the predefined datasets (Parc, Ref, or NR). This function will return a list of closely related sequences which can be added to the cart system for building and downloading customized datasets.

Finally, the “Least Common Ancestor” feature of the aligner can be used to classify sequences against any of the taxonomies provided by the SILVA project.

TestProbe

The SILVA probe match and evaluation tool called “TestProbe” detects and displays all occurrences of a given probe or primer sequence within any specified SILVA datasets or subsets thereof. It is offered to test and visualize *in silico* specificity and target group coverage (sensitivity) of rDNA-targeting probes and single primers against the SILVA datasets. The tool can be



SILVA Databases, Fig. 4 The web interface and results of the SILVA “TestPrime” service

configured to allow up to five mismatches between probe and target sequences and mismatches can be weighted. The resulting number of matches and non-matches is shown as a set of pie charts (Fig. 3), and an additional list provides sequence names, accession numbers, and a graphical representation of the probe’s binding site within all matches. Sequences in this list can be added to the cart system for subsequent download.

TestPrime

Similar to the SILVA “TestProbe” tool, “TestPrime” allows searching for all sequences

within the SILVA datasets or subsets thereof which are targeted by a given pair of primers. The number of allowed mismatches can be configured and results are shown in overview pie charts (Fig. 4) and the corresponding sequences can be selected for download.

Summary

The SILVA project provides comprehensive, quality-controlled, richly annotated, and aligned reference rDNA datasets to support the molecular assessment of biodiversity, as well as

investigations of the evolution of organisms. Applications of these datasets range from basic research in microbiology and molecular ecology to the detection of contaminants and pathogens in biotechnology and medicine. The taxonomically fully classified Ref and Ref NR datasets are perfectly suited for the classification of metagenomic or amplicon-based next-generation sequencing data.

The combination of SILVA datasets with the ARB software suite provides an easy to use workbench for researchers to perform in-depth sequence analysis and phylogenetic reconstructions as well as manual curation of rDNA datasets. Furthermore, the SILVA datasets have become an integral part of the MOTHUR (Schloss et al. 2009), QIIME (Caporaso et al. 2010), and MG-RAST (Meyer et al. 2008) analysis tools and pipelines.

Cross-References

- ▶ [A 123 of Metagenomics](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)

References

- Altschul S, Madden T, et al. BLAST and PSI-BLAST: a new generation of protein database search programs. *FASEB J*. 1998;12:A1326.
- Ashelford KE, Chuzhanova NA, et al. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol*. 2005;71:7724–36.
- Caporaso JG, Kuczynski J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7:335–6. Nature Publishing Group.
- Cole JR, Chai B, et al. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*. 2005;33:D294–6.
- Dawyndt P, Vancanneyt M, et al. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Trans Knowl Data Eng*. 2005;17(8):1111–26.
- DeSantis TZ, Hugenholtz P, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72:5069–72.
- Euzéby JP. List of bacterial names with standing in nomenclature: a folder available on the internet. *Int J Syst Bacteriol*. 1997;47(2):590–2.
- Garrity GM, Bell JA, et al. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology. 2nd ed. New York: Springer; 2004. Release 5.0.
- Grice EA, Kong HH, et al. Topographical and temporal diversity of the human skin microbiome. *Science*. 2009;324:1190–2.
- Grice EA, Snitkin ES, et al. Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proc Natl Acad Sci U S A*. 2010;107:14799–804.
- Gutell RR, Larsen N, et al. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev*. 1994;58:10–26.
- Kottmann R, Kostadinov I, et al. Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res*. 2010;38:D391–5.
- Lagesen K, Hallin P, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35:3100–8.
- Ludwig W, Strunk O, et al. ARB: a software environment for sequence data. *Nucleic Acids Res*. 2004;32(4):1363–71.
- Meyer F, Paarmann D, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 2008;9:386.
- Munoz R, Yarza P, et al. Release LTPs104 of the all-species living tree. *Syst Appl Microbiol*. 2011;34:169–70.
- Pruesse E, Peplies J, et al. SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*. 2012;28:1823–9.
- Schloss PD, Westcott SL, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75:7537–41.

Simultaneous Quantification of Multiple Bacteria

Annalisa Ballarini and Olivier Jousson
Laboratory of Microbial Genomics, Centre for Integrative Biology (CIBIO), University of Trento, Trento, Italy

Synonyms

Composition assessment; Abundance determination

Definition

The term “quantification” derives from the Latin terms *quant* (meaning “how much”) and *facere* (meaning “to make”). Quantifying is the act of determining the quantity of, measure. Simultaneous quantification of multiple bacteria stands for medium- to high-throughput detection and abundance determination of a bacterial community.

Introduction

Bacteria are widespread and abundant in the environment and comprise bacterial strains which are pathogenic for plants, animals, or human beings. Historically, the threat represented by pathogenic bacteria lead to the development of cultural-, serological-, and molecular-based methods to identify and possibly quantify the causative agent of an occurring infectious disease. These methods commonly target with high sensitivity and specificity one or few bacterial strains or species, but are not suitable for collecting comprehensive information on a microbial community. Recently, the advent of high-throughput sequencing technologies and the extensive metagenomic studies performed on human microbiomes have shown how shifts in bacterial composition and quantity may also correlate locally or systemically with the health status of the human host. These high-throughput metagenomic technologies can provide comprehensive information on microbial composition, functions, and dynamics, accelerating the development of complementary or alternative methods for environmental studies, clinically oriented studies, and routine diagnostics. The methods for simultaneous quantification of multiple bacteria are based on the following techniques: PCR, microarray, or high-throughput sequencing.

End-point and real-time quantitative PCRs are widely used techniques for identification and/or quantification of bacteria. The PCR is based on a primer extension reaction catalyzed by DNA polymerase, thus requiring a priori knowledge of the potential target bacteria.

These techniques amplify a few copies of the target DNA to millions of copies after 30–40 cycles, ensuring high sensitivity of detection. Being also user-friendly, fast, and cost-effective, these techniques are broadly used for clinical detection of potential pathogens. Multiplexing can also be applied by means of multiple primer pairs hybridizing to different target sequences. However, the multiplexing capacity is limited, as simultaneous detection of a high number of bacteria leads to decreased sensitivity, increased costs, and bacterial misidentifications. Thus, these techniques cannot reach the throughput required for bacterial community assessment.

High-throughput sequencing (see entries “► [Approaches in Metagenome Research: Progress and Challenges](#)”; and “► [Metagenomic Research: Methods and Ecological Applications](#)”) is the most used method for metagenomic studies, allowing the highest throughput and in-depth determination of bacterial communities’ composition without the requirement of a priori knowledge. The advantage of this technology is to be able to detect not only well-characterized microbes but also variant strains (Roh et al. 2010). It can be applied either for deciphering all genetic information including functional classes’ composition via whole-genome shotgun sequencing or for defining exclusively the taxonomic composition of bacterial communities by targeted sequencing of phylogenetic markers. Despite the continuous cost reductions, whole-genome shotgun sequencing still requires high expense, time, and complex computational resources (see entry “► [Computational Approaches for Metagenomic Datasets](#)”). For these reasons, this method is prohibited for prolonged clinical studies or routine diagnostics. On the other hand, targeted sequencing is less costly and complex in data processing but commonly addresses the 16S rRNA gene, which was reported to have a phylogenetic resolution limited to the family or genus level for several clades.

Microarray-based methods allow parallel detection of a large number of sequences in a single hybridization. As PCR-based assays, they require the a priori knowledge of the potential targets. Their feature is to combine a medium

to high throughput with a small format, a rapid and automated processing, and a low cost per sample. Thus, they span in between PCR and sequencing and are well suitable for cost-effective clinically oriented studies and informative routine diagnostics. Moreover, applied to metagenomic samples, they can provide a useful complementary approach to high-throughput sequencing, determining the appropriate sequencing depth according to the complexity of the bacterial communities.

Here below, the overview is focused on DNA microarrays as tools for simultaneous quantification of multiple bacteria and more extensively on well-known 16S microarrays and the recently developed BactoChip, a multi-marker phylogenetic microarray.

DNA Microarrays for Simultaneous Detection of Multiple Bacteria

DNA microarrays technology allows high-throughput screening of nucleic acid sequences for complementary binding. The sequences bound to the solid surface of the microarray may be synthetic oligonucleotides or DNA fragments either synthesized directly or spotted on the surface. In the presence of the target DNA sample, nucleic acid hybridization can occur. The stringency and rate of hybridization can be controlled by varying temperature, salt concentration, and washes (passive hybridization) or by applying electric fields on a microelectronic device (active hybridization). After hybridization, unbound template DNA is washed away and the bound template is detected, using dedicated scanners, mostly by means of fluorescent label or enzymatically active moieties previously incorporated in the DNA sample. The high-affinity binding of template nucleic acids to their complementary target can be used for the identification of microorganisms and relative abundance determination. Key steps for highly accurate bacterial detection are target DNA region selection and probe design. According to the target regions chosen, microarrays for bacterial profiling can be classified in two main

categories: microbial function or phylogenetically targeted.

Functional gene microarrays target mostly a combination of functional classes' genes. The feature of these microarrays is to define the capabilities of the bacterial community under investigation rather than its composition. For clinical purposes, arrays commonly target functional genes belonging to virulence and antibiotic resistance gene families (Jaing et al. 2008). Environmental applications focused instead on the known functions of the specific bacterial niche under study.

An example of this microarray type is the **GeoChip**, which was developed for characterizing microbial communities isolated from the environment both at structural and functional level. It proved successful in association with high-throughput approaches to provide in-depth information of defined environmental niches, such as sulfate-reducing bacterial communities important to environmental cleanup [see entry "[► GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities](#)"].

The second category is represented by **phylogenetic oligonucleotide microarrays** (see entry "[► Phylogenetics, Overview](#)"), which target instead phylogenetic marker genes, and can be based on a single or multiple marker approach. In contrast to functional gene arrays, they aim at discriminating bacteria by defining their identity. Single marker phylogenetic arrays target variants of universally conserved rRNA sequences. Some of the publicly described chips targeting single phylogenetic markers are listed here.

An example of a panmicrobial 16S-based microarray is the **GreenChipPm** (Palacios et al. 2007). This single marker detection array was designed to target respiratory pathogens (vertebrate viruses, fungi, bacteria, and protozoa). It includes, among others, probes designed to specifically bind variable segments of the 16S rRNA gene, the most well-known "universal" bacterial marker for phylogenetic determination.

The PhyloChipTM (Second Genome, San Francisco) is an oligonucleotide microarray targeting the segments of the single 16S rRNA

gene for high-throughput detection of microbial communities both in the environment and clinical samples (Brodie et al. 2007; Ghosh et al. 2009; Wu et al. 2010). Several versions were developed, the latest being the G3 version (Hazen et al. 2010). The G3 comprises 1.1 million DNA probes and covers nearly 60,000 operational taxonomic units. In order to increase the reliability, some microarray designs, including the PhyloChips, define multiple target regions within the marker adopting a so-called multiple probe concept to increase the overall detection accuracy.

HOMIM (Preza et al. 2009) is the acronym for Human Microbial Identification Microarray, a tool developed to detect simultaneously 300 bacterial species from the oral microbiome, including non-cultivable ones. The target bacteria were selected among the ones identified by 16S rRNA sequencing in health roots and root caries in elderly. Experiments performed with this array showed a general agreement in the results with 16S RNA gene sequencing analysis. Since 2008, a core facility at the Forsyth Institute (Cambridge, Massachusetts) provides a service, based on this platform, to rapidly screen clinical samples from the oral cavity, esophagus, and lungs.

The HITChip (human intestinal tract chip) (Rajilic-Stojanovic et al. 2009) is a microarray-based metagenomic tool designed for profiling the human gastrointestinal microbiota. This phylogenetic microarray comprised 4,809 oligonucleotide probes and discriminate 1,140 species via two hypervariable regions of the small subunit ribosomal RNA (SSU rRNA) gene. The validation performed with SSU rRNA clones and clinical samples proved that this microarray provides a highly reproducible fingerprint and has also quantification potential. In particular, tests performed with synthetic mixtures showed it can detect 40 different amplicons and also those with relative abundance of 0.1 %. The HITChip showed to correctly identify a universal microbiota at genus-level resolution.

Overall, the most used phylogenetic marker is the 16S rRNA gene. The presence of highly conserved regions flanking the variable 16S rRNA

target sequence allows introducing a PCR-based amplification step for bacterial target enrichment (e.g., GreenChipPm). This pre-amplification step ensures an increased sensitivity in detection but does increase the processing time and may introduce biases in relative abundance quantitation.

Besides that, being universally conserved within the bacterial kingdom, the 16S RNA gene may not be sufficient for specific and reproducible bacterial identification, especially in complex systems. In fact, the high conservation score of this gene across taxa has been reported to cause cross-hybridization events, affecting both resolution and abundance determination, and to fail to discriminate below the genus level for many clades.

Recently, as alternative to 16S or single marker array for microbial profiling, a multiple marker phylogenetic microarray has been designed, the **BactoChip** (Ballarini et al. 2013). The array design was based on the notion that metagenomic sequencing data offer a powerful view on the microbial diversity of the sampled communities and an increasingly higher number of complete and annotated bacterial genomes are publicly available.

The BactoChip: A Multi-marker Phylogenetic Microarray for Species-Level Resolution

The BactoChip (Ballarini et al. 2013) was designed with the aim to overcome the issues of resolution and abundance determination of 16S-based microarrays and thus approach the throughput and specificity of sequence-based techniques. Up to date, one version of the BactoChip has been described, detecting via a PCR-independent approach a set of 54 bacterial species belonging to multiple genera of clinical interest. The number of target bacteria was limited by the availability of typed strains for experimental validation and of complete bacterial genome sequences for computational microarray design. However, the developed method for marker selection may be extended to the whole microbial world, thus allowing high accuracy of

microbial composition assessment even in complex samples.

Computational and Experimental Design.

The BactoChip *in silico* design is based on the knowledge deriving from metagenomic datasets and complete bacterial genome sequences. The computational tool for DNA marker identification employed a pairwise identity threshold above 99 % to define core genes for most species, where core genes are those shared by all available sequenced strains of the same species. Unique genes (i.e., core genes unique for each bacterial species) were then selected by removing all core genes with *blastn* hits outside the target species. Probes targeting an average of 10 markers per bacterial species were designed to have similar physicochemical parameters and were directly synthesized on “custom high-definition Agilent DNA Comparative Genomic Hybridization arrays 8x15K” (Agilent Technologies, Santa Clara, CA, USA). Besides internal control and other probes, the BactoChip includes 2,094 marker gene probes targeting 54 bacterial species.

Testing on Pure Isolates, Synthetic Communities, and Clinical Samples. The BactoChip was validated by performing hybridization experiments with 37 bacterial species singularly, multiple congeneric species, and synthetic bacterial communities of up to 15 microorganisms. Also, it was tested with oral microbiomes from two healthy subjects spiked with 5 different species at known relative abundance. Single reference strains used for validation were collected from the LGC Standards ATCC, the Leibniz Institute DSMZ, or university hospitals. Synthetic communities were obtained by mixing single strains in known DNA quantities. Oral microbiomes were collected from saliva, DNA was extracted with standard protocols, and the bacterial load was determined by real-time PCR. The BactoChip identified univocally almost all tested species (97.3 %) from 19 genera with near-perfect accuracy (AUC > 0.99). In case of malfunctioning probes (false negative or false positive), the presence of multiple probes per marker genes and multiple genes per species prevented species misidentification. Testing

performed with multiple congeneric bacterial species from the *Staphylococcus* genus showed how this microarray design can resolve to the species level even genera known to be poorly resolved by the 16S marker genes. The performance of the BactoChip in identifying bacteria and determining relative abundances was tested by means of synthetic bacterial communities comprising 9 and 15 different species at even and staggered concentrations. The species-level specificity was confirmed also in this experimental setting. The microarray quantified both bacterial communities with high accuracy with an overall high correlation (0.97, $p < 10^{-10}$) between reference relative abundance values and estimated ones. Experiments performed on saliva microbiomes isolated from healthy volunteers, spiked in with reference species in known amounts, proved the feasibility of this approach for microbiome profiling, and detected the native and spiked-in species within clinical samples over a 100-fold dynamic range.

Summary and Conclusions

High-throughput metagenomic technologies have provided an extensive amount of data on microbial composition, functions, and dynamics, accelerating the development of complementary or alternative methods for environmental studies, clinically oriented studies, and routine diagnostics. Definitely, next-generation sequencing technology leads, without the need of a priori knowledge, to the maximum amount of information on the genomic sequences' composition of a microbial sample. However, this technology requires complex computational analyses to extrapolate information of interest and still requires high costs and processing times.

Among the alternative molecular-based techniques currently available (multiplex, real-time PCR, or array-based assays), microarrays represent the most promising technique for parallel detection and relative abundance quantitation of bacteria with complex microbial samples, combining a high-throughput with a user-friendly rapid protocol and a low cost per sample. Besides

functional microarrays, which commonly target defined environmental niches and aim at functional classes' classification (e.g., GeoChip), microarrays for microbial identification are phylogenetic based. Up to date, ribosomal genes (in particular the 16S) are the most used phylogenetic markers for microbial profiling through microarray (e.g., GreenChipPm, PhyloChipTM, HOMIM, HITChip). However, being the 16S gene highly conserved throughout the bacterial kingdom, it is difficult to resolve bacteria below the family or genus level for some clades.

Metagenomic data available on human body sites samples has shown how defining the bacteria profile at species level may generate a more in-depth understanding of the relation between bacterial composition and health. Recently, a multi-marker phylogenetic microarray was described (the BactoChip) which proved to be highly specific in bacterial species identification, feasible for microbial profiling, and reliable for relative abundance quantification over a 100-fold dynamic range, even within complex ecosystems. Being based on complete genomic sequences, the BactoChip array design stands on a lower number of reference sequences available, in public sequence databases, in comparison to the historically used 16S rRNA phylogenetic marker sequences. However, the exponentially increasing amount of complete bacterial genome sequences will soon fill this gap allowing an optimized marker selection for accurate microbial profiling both for the ecosystem and the human body.

Cross-References

- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes](#)
- ▶ [GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)
- ▶ [Phylogenetics, Overview](#)

References

- Ballarini A, Segata N, Huttenhower C, Jousson O. Simultaneous quantification of multiple bacteria by the bactoChip microarray designed to target species-specific marker genes. *PLoS One*. 2013;8(2): e55764.
- Brodie EL, DeSantis TZ, Parker JPM, Zubietta IX, Piceno YM, Andersen GL. Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci U S A*. 2007;104:299–304.
- Ghosh D, Roy K, Williamson KE, Srinivasiah S, Wommack KE, Radoosevich M. Acyl-homoserine lactones can induce virus production in lysogenic bacteria: an alternative paradigm for prophage induction. *Appl Environ Microbiol*. 2009;75:7142–52.
- Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Zhou J, Mason OU. Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science*. 2010;330: 204–8.
- Jaing C, Gardner S, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P, Williams P, Gu P, Wagner M, Manohar C, et al. A functional gene array for detection of bacterial virulence elements. *PLoS One*. 2008;3(5): e2163.
- Palacios G, Quan P-L, Jabado O, Conlan S, Hirschberg D, Liu Y. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis*. 2007;13(1):73–81.
- Preza D, Olsen I, Willumpson T, Boches SK, Cotton SL, Grinde B, Paster BJ. Microarray analysis of microflora in root caries of the elderly. *Eur J Clin Microbiol Infect Dis*. 2009;28(5):509–517.
- Rajilic-Stojanovic M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H, de Vos WM. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundance microbiota of young and elderly adults. *Environ Microbiol*. 2009;11(7):1736–51.
- Roh SW, Abell G CJ, Kim K-H, Nam YD, Bae JW. Comparing microarrays and next-generation sequencing technologies for microbial ecology research (review). *Trends Biotechnol*. 2010; 28(6):291–299.
- Wu CH, Sercu B, Van De Werfhorst LC, Wong J, DeSantis TZ, Brodie EL, Hazen TC, Holden PA, Andersen GL. Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. *PLoS One*. 2010;5: e11285.

STAMP: Statistical Analysis of Metagenomic Profiles

Donovan H. Parks^{1,2} and Robert G. Beiko¹

¹Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

²Australian Centre for Ecogenomics, University of Queensland, Brisbane QLD, Australia

Definition

Cross-platform software providing statistical analyses and plots of taxonomic and functional profiles.

Introduction

Comparative metagenomic studies aim to understand differences in the structure and function of microbial communities from different habitats. Statistical approaches can be used to highlight differences between pairs of metagenomic samples or defined groups of samples (e.g., samples from sick and healthy individuals). STAMP (Statistical Analysis of Metagenomic Profiles) is a software platform for analyzing metagenomic profiles (Parks and Beiko 2010), such as taxonomic profiles indicating the number of marker genes assigned to different taxonomic units or functional profiles indicating the number of sequences contributing to a specific subsystem or pathway. It aims to promote best practices in reporting statistical results by encouraging the use of effect sizes and confidence intervals when assessing biological importance. A user-friendly, graphical interface permits easy exploration of statistical results and generation of publication-quality plots for inferring the biological relevance of features in a metagenomic profile. STAMP is open-source, extensible via a plug-in framework and available for all major platforms.

Defining Metagenomic Profiles and Sample Metadata

STAMP requires metagenomic profiles to be specified in a tab-separated text file. The first

row of the file contains the header for each column. Columns indicating the hierarchical structure of a feature must be organized from the highest to lowest level in the hierarchy. There are no restrictions on the depth of a hierarchy and hierarchies may be multifurcating. However, hierarchies must form a strict tree structure (i.e., a child can have only one parent). The number of sequences or reads assigned to each leaf node in the hierarchy must be specified for each metagenomic sample. To allow for different normalization methods, counts may be integers or real numbers. An example STAMP profile is given in Table 1.

Several methods have been proposed for generating taxonomic or functional profiles from metagenomic data. STAMP supports analyzing profiles generated by MG-RAST (Meyer et al. 2008), IMG/M (Markowitz et al. 2008), mothur (Schloss et al. 2009), CoMet (Lingner et al. 2011), and RITA (MacDonald et al. 2012). Profiles generated using these software platforms can be converted to STAMP-compatible profiles using functionality provided within STAMP. The simple format of STAMP profiles helps ensure that results from other software platforms can be converted for processing by STAMP.

Additional data associated with each metagenomic sample can be defined through an optional tab-separated metadata file. The first column of this file indicates the name of each sample and should correspond to an entry in the

STAMP: Statistical Analysis of Metagenomic Profiles, Table 1 Example STAMP profile

Hierarchical level 1	Hierarchical level 2	Sample 1	Sample 2	Sample 3
Category A	Subcategory A1	0	4.4	4
Category A	Subcategory A1	3	5	5
Category A	Subcategory A2	4.8	3.5	2
Category B	Subcategory B1	2	32	6.5
Category C	Subcategory C1	1	2	2
Category C	Subcategory C1	7.2	6	4

STAMP: Statistical Analysis of Metagenomic Profiles, Table 2

Example metadata file

Sample Id	Location	Phenotype	Gender	Sample size
Sample 1	Canada	Obese	Female	4,000
Sample 2	Canada	Lean	Male	2,000
Sample 3	Italy	Lean	Female	3,000

corresponding STAMP profile. Additional columns may specify any other data relevant to the samples being considered. Within STAMP, these additional columns can be used to define groups (i.e., collections of one or more samples) over which statistical tests and plots can be calculated. An example metadata file is given in Table 2.

Statistical Analysis of Metagenomic Profiles

STAMP provides statistics for assessing biologically relevant differences between pairs of metagenomic samples or treatment groups. Two-sample (e.g., Fisher's exact test, G-test), two-group (Welch's t-test, White's nonparametric t-test), and multigroup (ANOVA, Kruskal-Wallis H-test) statistical hypothesis tests are provided for identifying statistically significant features. Features with p-values below a nominally chosen threshold (e.g., 0.05) can reasonably be assumed to be enriched or depleted due to ecological differences between samples or treatment groups as opposed to representing a sampling artifact. STAMP also reports effect size statistics such as the difference or ratio between proportions in order to aid in determining if a statistically significant feature is of biological relevance. Consideration of effect sizes is essential as small, biologically uninteresting differences may be statistically significant when sample sizes are large. Confidence intervals are computed for all effect size statistics. These indicate the range of effect size values that have a specified probability (typically 95 %) of being compatible with the observed data and are an important additional statistic for reasoning about biological relevance.

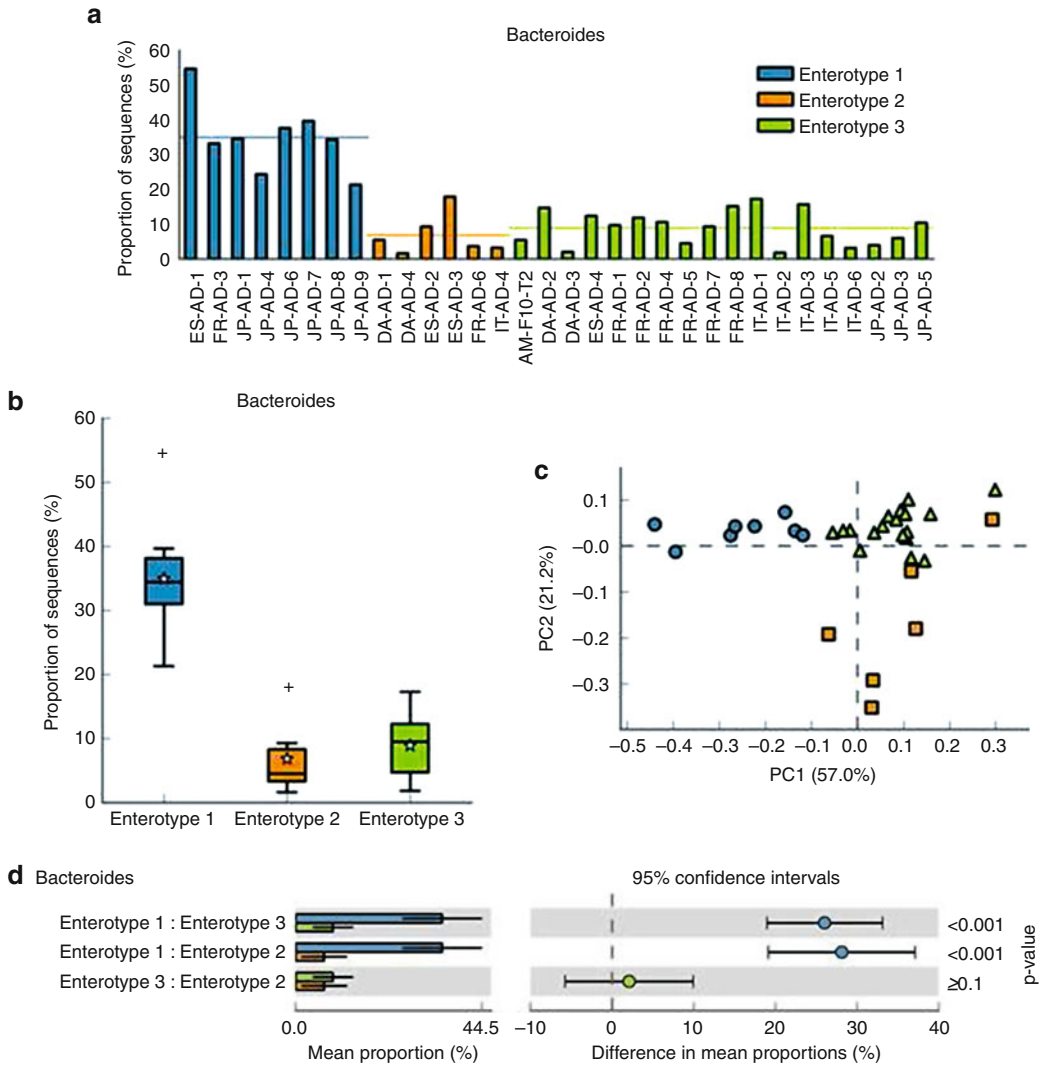
Metagenomic profiles typically consist of several hundred or thousand features. Care must be taken when performing multiple hypothesis tests. For example, a profile consisting of 1,000 features will have 50 features with a p-value less than 0.05 simply due to chance variation. STAMP provides two techniques for correcting p-values when multiple hypothesis tests are being performed. The first controls the *familywise error rate* using a correction method such as Bonferroni, Holm-Bonferroni, or Šidák. This adjusts the reported p-values so that the probability of observing one or more false positives is less than a specified probability. During data exploration, this approach can be too conservative and it may be beneficial to adjust the p-values using a *false discovery rate* procedure. Under this approach, a q-value is calculated for each feature that indicates the expected proportion of false positives within the set of features with a smaller q-value (Benjamini and Hochberg 1995). Additionally, STAMP can filter features using a number of criteria in addition to p- or q-values in order to focus on biologically interesting features, e.g., those with a large effect size or consisting of a substantial number of reads.

Exploration of Metagenomic Profiles

STAMP provides the following interactive, publication-quality plots for exploring metagenomic profiles:

Bar plots indicate the proportion of sequences of each feature within a pair of samples or the proportion of sequences of a single feature across all samples (Fig. 1a).

Box plots illustrate how the proportion of sequences of a single feature is distributed within different treatment groups using a box-and-whiskers graphic (Fig. 1b). Box-and-whiskers graphics show the median of the data as a line, the mean of the data as a star, the 25th and 75th percentiles of the data as the top and bottom of the box, and use whiskers to indicate the most extreme data point within $1.5 \times (75\text{th} - 25\text{th percentile})$ of



STAMP: Statistical Analysis of Metagenomic Profiles, Fig. 1 Exploration of the gut microbiota of 32 individuals reported by Arumugam et al. (2011) to form three distinct clusters or enterotypes. (a) Bar plot showing the relative proportion of *Bacteroides*. Samples are colored according to the enterotype to which they have been assigned. (b) Box plot showing the distribution in the proportion of *Bacteroides* from samples assigned to each

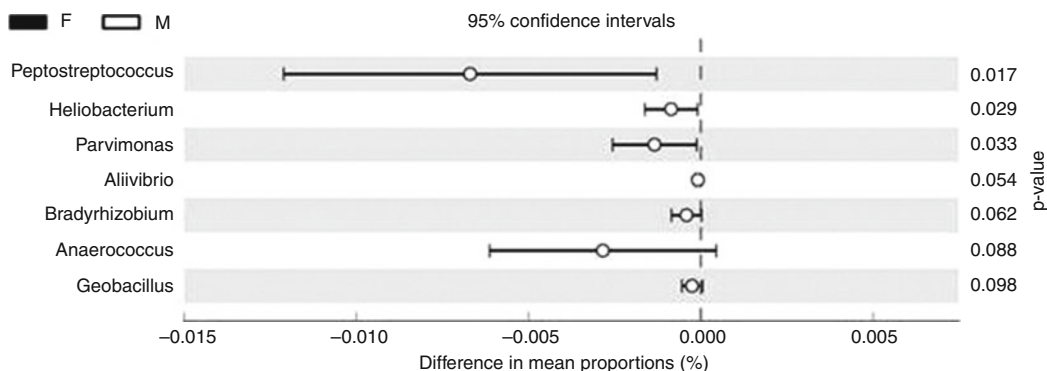
enterotype. (c) Principal coordinate analysis plot determined from the proportion of reads assigned to each genera within a sample. (d) Post hoc plot for *Bacteroides* indicating (1) the mean proportion and standard deviation within each enterotype, (2) the difference in mean proportions between each pair of enterotypes along with 95% confidence intervals, and (3) a p-value indicating if the mean proportion is equal for a given pair

the median. Data points outside of the whiskers are shown as crosses.

PCA plots give the first three principal components of a metagenomic profile as determined by applying principal component analysis (Fig. 1c). Clicking on a marker within the

plot indicates the sample represented by the marker.

Post hoc plots contrast each pair of groups considered in a multigroup statistical hypothesis test (Fig. 1d). It indicates the mean proportion of sequences within each group, the difference



STAMP: Statistical Analysis of Metagenomic Profiles, Fig. 2 Exploration of compositional differences in the gut microbiota of males and females sampled by Arumugam et al. (2011). The extended error bar plot

indicates all genera where Welch's t-test produces an uncorrected p -value < 0.1 . All genera are overabundant within the gut microbiota of males (M) compared to females (F)

in mean proportions for each pair of groups along with the confidence interval of this effect size statistic, and a p -value indicating if the mean proportion is equal for a given pair.

Extended error barplots display the p -value, effect size, and associated confidence interval for all unfiltered features in a metagenomic profile (Fig. 2). In addition, a bar plot indicates the proportion of sequences assigned to a feature in each sample or group. This provides all information required to reason about the biological relevance of a feature in a single plot.

Scatter plots indicate either the proportion of sequences or mean proportion of sequences assigned to each feature within a pair of samples or a pair of treatment groups, respectively. This plot is useful for identifying features that are clearly enriched in one of the two samples or groups. When considering a pair of samples, confidence intervals calculated with the Wilson score method can be shown. For a pair of treatment groups, different statistics indicating the spread of the data can be displayed (e.g., standard deviation, minimum and maximum proportions).

All plots provide a range of customization options. For example, PCA plots can be restricted

to the first two principal components, and individual panels of the extended error bar plot can be selectively hidden. Plots can be saved in either vector (PDF, PS, EPS, SVG) or raster (PNG) formats. The resolution of raster files can be set to allow for generation of plots suitable for printed publication or display on posters.

Tabular views of statistical results are also provided and columns can be sorted to help identify interesting patterns. Tables can be saved as tab-separated value files for subsequent display in any text editor or spreadsheet program or for inclusion as supplemental information in publications.

Summary

Statistics can greatly aid in the comparison of metagenomic profiles. STAMP provides a simple graphical environment for performing statistical analyses that are tailored to the needs of comparative metagenomic studies. It provides a range of statistical hypothesis test and can identify statistically significant features between pairs of samples or defined treatment groups. Different multiple test correction methods are provided in order to account for the large number of features

typical of metagenomic profiles and to aid in data exploration. The biological relevance of significant features can be assessed through a range of publication-quality plots that provide key statistics such as effect sizes and confidence intervals. Interactive filtering allows the most biologically interesting features to be quickly identified and plots of specific features to be generated. STAMP's wide range of statistics and simple interactive interface makes it a valuable tool in comparative metagenomic studies.

Cross-References

- ▶ [MEtaGenome ANALyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3](#)

References

- Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473:174–80.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57:289–300.
- Lingner T, ABhauer KP, Schreiber F, Meinicke P. CoMet – a web server for comparative functional profiling of metagenomes. *Nucleic Acids Res*. 2011;39 Suppl 2:W518–23.
- MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res*. 2012;40:e111.
- Markowitz VM, Ivanona NN, Sveto E, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*. 2008;36(Database issue):D534–8.
- Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma*. 2008;9:386.
- Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. 2010;26:715–21.
- Schloss PD, Westcott SL, Ryabin T, et al. Introducing mother: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75:7537–41.

Subtractive Hybridization Magnetic Bead Capture: Molecular Technique for Recovery of Full-Length ORFs from Metagenomes

Don Cowan, Sandra Ronca and Jean-Baptiste Ramond
Centre for Microbial Ecology and Genomics (CMEG), Genome Research Institute (GRI), University of Pretoria, Hatfield, Pretoria, South Africa

Synonyms

Recovery of full-length ORFs from metagenomic DNA

Definition

Subtractive hybridization magnetic bead capture (SHBMC) is a sequence-based metagenomic technique for the recovery of full-length ORFs from heterogeneous metagenomic DNA samples.

Introduction

It is widely acknowledged that the vast majority (~99 %) of microorganisms present in the environment are resistant to culture using classical microbiological methods. Approximately half of the total estimated bacterial phyla (61) are still to be cultured (Vartoukian et al. 2010). However, environmental microbial communities constitute a valuable resource for biotechnology and are a valid target for identification of novel genes and/or biological compounds such as biocatalysts or secondary metabolites (Sharma et al. 2005). In order to bypass the limitations of microbial culturing and to discover new microbial genes and functions, two approaches have been implemented, either culture-based, through the development of innovative strategies and media

to “culture the unculturables” (Vartoukian et al. 2010), or culture-independent, by using “meta-omics” technologies (Riesenfeld et al. 2004; Cowan et al. 2005).

Metagenomics enable to investigate in depth the totality of (microbial) genomes present in any given environments (Riesenfeld et al. 2004), including extreme habitats which constitute fields of choice for the discovery of robust enzymes and biological compounds suitable for industrial processes (Cowan et al. 2005). In practice, metagenomic studies can either be function- or sequence-driven (Riesenfeld et al. 2004; Schmeisser et al. 2007). While the former is widely used, it remains limited (i) by the choice and number of substrates available, (ii) by the difficulty of designing novel substrates, (iii) by the fact that heterologous expression systems and hosts are required, and (iv) by the need to clone full-length open reading frames (ORFs) or gene clusters to enable activities to be detected. Contrastingly, the latter allows access to all the sequences from any given environment and thus to its complete metabolic/catabolic potential providing that similar sequences have previously been annotated and their encoded activity(ies) characterized (Schmeisser et al. 2007). In metagenomic gene-mining studies (whether function- or sequence-based), one of the main experimental challenges is therefore to recover complete ORFs. The recent advent of high-throughput second-generation “meta”sequencing technologies potentially provides access to all the sequences present in a metagenome, and while it facilitates the isolation of the targeted sequence(s) from metagenomic samples, it does not avoid laborious experimental procedures. Here, we report on a novel molecular biology technique for the recovery of full-length ORFs from environmental metagenomes, termed subtractive hybridization magnetic bead capture (Meyer et al. 2007).

Method

SHMBC is a sequence-based technique developed for the retrieval of complete ORFs from

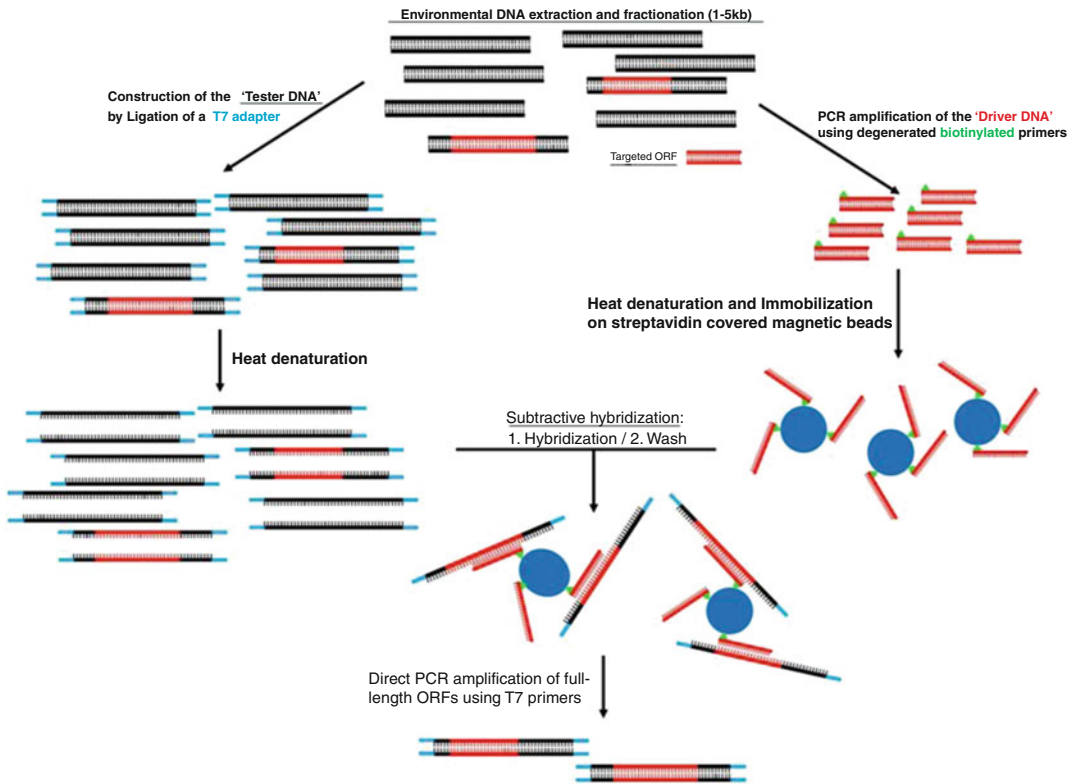
metagenomic DNA from environmental samples. The three core elements of the technology include high-quality metagenomic DNA and the production of “tester DNA” and “driver DNA,” where the latter is immobilized on magnetic beads (Fig. 1).

Metagenomic DNA Extraction and Fragmentation

Prior to performing SHMBC, it is essential to obtain high-quality and high molecular weight metagenomic DNA, by chemical (e.g., cell lysis using detergents) and/or mechanical (e.g., bead-beating) extraction protocols (Roh et al. 2006). Due to variable physical and chemical compositions, extracting high-quality metagenomic DNA from environmental samples can be challenging, most notably in attaining a complete representation of the microbial (functional) diversity, including rare phyla/sequences.

For functional investigation, the isolation of complete ORFs and/or gene clusters is crucial. For SHMBC, metagenomic DNA fragment sizes of 1–5 kb are ideal as they are short enough to permit relatively easy PCR amplification following subtractive hybridization (Fig. 1). The detergent-based metagenomic DNA extraction method developed by Zhou is recommended as it typically yields high-quality metagenomic DNA (>23 kb; Zhou et al. 1996). The whole extraction process requires ~6 h of labor. Alternatively, various column-based metagenomic DNA extraction kits are commercially available and are efficient in extracting high-quality DNA (~10 kb in 1 h; Knauth et al. 2013).

The fragmentation of metagenomic DNA to the appropriate size can be performed by physical disruption methods (e.g., by freeze/saw or freeze-boiling cycles or bead-mill homogenization) or enzymatic digestion (using restriction enzymes). Metagenomic DNAs contaminated by co-extracted compounds (notably humic acids or heavy metals), which can hamper downstream restriction and/or PCR amplification reactions, can be diluted or purified. These procedures generally lead to the reduction of metagenomic DNA yields.



Subtractive Hybridization Magnetic Bead Capture: Molecular Technique for Recovery of Full-Length ORFs from Metagenomes, Fig. 1 Schematic subtractive hybridization magnetic bead capture (SHMBC) protocol

“Tester DNA” Construction

The “tester DNA” is a modified metagenomic DNA sample to be probed by SHMBC for full-length ORF recovery (Fig. 1). Once fractionated to an appropriate size (1–5 kb), the metagenomic DNA is manipulated to generate the “tester DNA” by ligating T7 adapters at the 3′ and 5′ ends (in blue, in Fig. 1). Recommended protocols are given in Meiring et al. (2010). The T7 adapters contain T7 priming sites which enable a direct PCR amplification of the 1–5 kb DNA fragments following subtractive hybridization.

“Driver DNA” Production

The “driver DNA” is the hybridization probe used for the recovery of full-length ORFs from the “tester DNA.” Its production is thus crucial for the success of SHMBC. The “driver DNA” can be PCR-amplified directly from the purified metagenomic DNA using gene-specific primers

(Fig. 1). However, environmental samples are characterized by composite bacterial communities, potentially with polymorphic sequences coding for multiple related genes. To PCR-amplify heterologous gene sequences, the use of degenerate primers is necessary. For the production of valid “driver DNA,” a compromise must therefore be made between primer degeneracy (the number of degenerate bases) and primer coverage (the number of matched homologous gene). Highly specific primers may only target limited numbers of organisms/genes, while excessive degeneracy often leads to high levels of nonspecific binding and to the amplification of untargeted sequences.

In general, in order to design degenerate PCR primers, homologous nucleotide sequences from different microorganisms are retrieved from databases (e.g., GenBank) and aligned to such that conserved ~20 mer long sequences can be

identified. An alternative is to identify conserved amino acid sequences in homologous proteins, as they usually constitute key components of active sites and/or are necessary for protein stability. Moreover, the genetic code being degenerate (or redundant), synonymous codons (i.e., triplet of nucleotide coding for a single amino acid) generally differs by their last base, which may be replaced in degenerate primers by the nucleotide “inosine.” Amplicons obtained with degenerate primer sets must initially be cloned and sequenced to verify their specificity.

Biotinylated ORF-/gene-specific “driver DNAs” are produced using 5′-biotinylated gene-specific forward degenerate primers (the reverse remaining unlabelled; Meyer et al. 2007, Meiring et al. 2010). Single-stranded “driver DNAs” are immobilized to streptavidin-coated magnetic beads which have high affinity for biotin (Fig. 1). The “driver DNA”-magnetic bead complex constitutes the hybridization probe for SHMBC.

Subtractive Hybridization of “Tester DNA” and Full-Length ORF Amplification

“Tester DNAs” and “driver DNAs” are generally hybridized overnight. To modify SHMBC selectivity, hybridization temperatures and hybridization buffer salt concentrations can be adjusted. To ensure specific post-subtractive hybridization PCR amplifications, unbound “tester DNAs” are eliminated with successive SDS washes. Recovered magnetic beads with hybridized “tester DNA” can be used directly to amplify target ORFs using T7 primers (Fig. 1). To amplify full-length ORFs, high-fidelity and reading Taq polymerases are recommended.

Comparison with Other Techniques

The function-driven isolation of ORFs has most commonly been used with pure isolates or clones from metagenomic libraries. However, this strategy is limited because of significant technical and methodological challenges. Notably, less than 1 % of environmental bacteria can be cultured. Obtaining axenic cultures, a prerequisite for any

physiological characterization, is a fastidious and unreliable process as numerous variables potentially influence microbial growth (e.g., amounts of various but specific nutrients, pH, temperature, atmospheric gas composition, etc.). Function-based screening of clones is dependent on the expression of genes in foreign hosts, and only few model microorganisms are widely used as transformation hosts (e.g., *Escherichia coli*, *Bacillus subtilis*, *Geobacillus* sp., *Streptococcus pneumoniae*, *Neisseria gonorrhoeae*, *Haemophilus influenzae*, *Helicobacter pylori*, *Acinetobacter baylyi*, and some cyanobacteria). In addition, a successful transformation guarantees neither the expression of heterologous genes nor the production of functional proteins/enzymes in the foreign host. Finally, the detection of an enzymatic activity is dependent on the existence or design of suitable media and/or of an assay to detect the specific enzyme activities (Waschkowitz et al. 2009).

Current advances in second-generation sequencing technology (454 pyrosequencing, Illumina, and SOLiD™) have increased the effectiveness of sequence-based screening as hundreds of millions of sequencing reads can be acquired in a single run, enabling the detection of rare ORFs in metagenomic samples. However, isolation of full-length sequences still necessary for functional studies, and the short sequence read lengths (which do not cover entire ORFs), may become problematic (Morales and Holben 2011). In consequence, complex computational gene assemblage strategies must be implemented in order to recover full-length ORFs (Liu et al. 2012). When working with genes or ORFs with multiple homologous present in a database, the latter can be used as reference sequences during gene annotation and assembly processes. However, in the absence of such a reference sequence, de novo approaches can only provide a probability of sequence fragments belonging to a specific gene cluster (Thomas et al. 2012). New gene families can be annotated from next-generation sequencing datasets by de novo gene assembly methods (i) by reconstructing novel sequences/genes based on nucleotide frequency, (ii) by implementation of a conventional Overlap

Layout Consensus (OLC) strategy, or (iii) by probabilistic De Bruijn graphs (Paszkiwicz and Studholme 2010). However, gene size variations due to large insertions, deletions, or polymorphisms can lead to complicated de novo assemblies even within closely related taxa. Finally, computational assembly of next-generation sequencing data is limited by the fact that short-read assemblies rely on data reduction algorithms, in which reads from low-abundance organisms may be discarded, ultimately leading to the disappearance of rare ORFs from the datasets.

Applications and Improvements

The sequence-based SHMBC technique constitutes a valuable prescreening method, as it collects ORFs of interest from complex metagenomic mixtures; and the latter can further be functionally analyzed. Such an approach increases the chance of obtaining positive hits in post-functional screening protocols.

Recently, a comparable subtractive hybridization approach in combination with a pre-enrichment microsatellite strategy was performed to isolate novel *phaC* gene sequences from the marine bacteria *Paracoccus homiensis* (Latisnere-Barragan and Lopez-Cortes 2012). This methodology allowed the efficient isolation of full-length *phaC* ORFs and the construction of enriched plasmid libraries of *phaC* genes, thereby reducing the experimental costs of genome sequencing (Latisnere-Barragan and Lopez-Cortes 2012). In this study, a single genome was screened but the authors suggest that SHMBC coupled with microsatellite enrichments could be used to retrieve *phaC* sequences from complex microbial community metagenomes.

In forensic studies, SHMBC was developed (i) to extract and pre-concentrate STRs (Short Tandem Repeat) from degraded DNA samples, which is a common problem in crime scene analysis, and (ii) to compensate for STR allele imbalance, allele dropout, and sequence-specific inhibitions generally encountered in such samples (Wang and McCord 2011). The authors

conclude that SHMBC using the specifically designed probes (i.e., “driver DNA”) significantly improved the recovery of STR alleles from degraded DNA samples.

SHMBC was recently amended by combining the DNA fractionation and linker ligation steps (Harris et al. 2009). Such a procedure increased the efficiency in constructing the “tester DNA.” In this particular study, SHMBC was applied as an enrichment tool to identify and isolate sex-specific regions in the complex genome of Australian python (*Morelia spilota imbricata*). The authors stressed the critical importance of nondegraded DNA.

SHMBC as a pre-enrichment tool could also be used in comparative meta-transcriptomic studies. In such studies, the analysis of full-length ORFs may identify the most frequently represented functional genes in different ecosystems and thus potentially unravel differential trophic structures and functions.

To conclude, we strongly believe that this technique should be considered by the scientific community for use prior to any “meta-functional” or functional genomic studies. SHMBC can readily be automated and routinely used as a full-length ORFs pre-enrichment tool to detect functions of interest in metagenomic samples (Latisnere-Barragan and Lopez-Cortes 2012), for disease diagnosis (Wang et al. 2011), and could further be implemented to test processed food products for the presence of genetically modified organisms (GMOs) and/or other adulterations (such as horse meat contamination in beef lasagne!) or GMO cross-contaminations in crop fields.

Summary

A wide range of function-based and/or sequence-based screening techniques have been developed to study/isolate ORFs in metagenomes. The subtractive hybridization magnetic bead capture (SHMBC) technique is potentially a cost-effective and efficient method for the isolation of full-length ORFs from metagenomic DNA preparations. This approach could be widely

used as a pre-enrichment tool prior to performing post-functional studies or sequence-based functional analyses.

Cross-References

- ▶ [Approaches in Metagenome Research: Progress and Challenges](#)
- ▶ [Biological Treasure Metagenome](#)
- ▶ [Metagenomic Research: Methods and Ecological Applications](#)
- ▶ [Mining Metagenomic Datasets for Antibiotic Resistance Genes](#)
- ▶ [Mining Metagenomic Datasets for Cellulases](#)
- ▶ [Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies](#)

References

- Cowan D, Meyer Q, Stafford W, Muyanga S, Rory Cameron R, Wittwer P. Metagenomics gene discovery: past, present and future. *TRENDS Biotechnol.* 2005;23:321–9.
- Harris RP, Groth DM, Ledger J, Lee CY. Identification of sex specific DNA regions in the snake genome using a subtractive hybridization technique. *Proc Assoc Advmt Anim Breed Genet.* 2009;18:572–5.
- Knauth S, Schmidt H, Tippkotter R. Comparison of commercial kits for the extraction of DNA from paddy soils. *Lett Appl Microbiol.* 2013;56:222–8.
- Latisnere-Barragan H, Lopez-Cortes A. Isolation of *phaC* gene from marine bacteria *Paracoccus homiensis* strain E33 by magnetic beads subtractive hybridization. *Ann Microbiol.* 2012;62:1691–5.
- Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012. doi:10.1155/2012/251364.
- Meiring T, Mulako I, Tuffin MI, Meyer Q, Cowan DA. Retrieval of full-length functional genes using subtractive hybridization magnetic bead capture. *Methods Mol Biol.* 2010;668:287–97. Clifton, NJ.
- Meyer QC, Burton SG, Cowan DA. Subtractive hybridization magnetic bead capture: a new technique for the recovery of full-length ORFs from metagenome. *Biotechnol J.* 2007;2:36–40.
- Morales SE, Holben WE. Linking bacterial identities and ecosystem processes: can ‘omic’ analyses be more than the sum of their parts? *FEMS Microbiol Ecol.* 2011;75:2–16.
- Paszkievicz K, Studholme DJ. *De novo* assembly of short sequence reads. *Brief Bioinform.* 2010;11:457–72.
- Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet.* 2004;38:525–52.
- Roh C, Villatte F, Kim B-G, Schmid RD. Comparative study of methods for extraction and purification of environmental DNA from soil and sludge samples. *Appl Biochem Biotechnol.* 2006;134:97–112.
- Schmeisser C, Steele H, Streit WR. Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol.* 2007;75:955–62.
- Sharma R, Ranjan R, Kapardar RK, Grover A. ‘Unculturable’ bacterial diversity: an untapped resource. *Curr Sci.* 2005;89:72–7.
- Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microbiol Inform Exp.* 2012;2:3–3.
- Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of ‘unculturable’ bacteria. *FEMS Microbiol Lett.* 2010;309:1–7.
- Wang J, McCord B. The application of magnetic bead hybridization for the recovery and STR amplification of degraded and inhibited forensic DNA. *Electrophoresis.* 2011;32:1631–8.
- Wang HH, Zhao CY, Li F. Rapid identification of mycobacterium tuberculosis complex by a novel hybridization signal amplification method based on self-assembly of dna-streptavidin nanoparticles. *Braz J Microbiol.* 2011;42:964–72.
- Waschkowitz T, Rockstroh S, Daniel R. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Appl Environ Microbiol.* 2009;75:2506–16.
- Zhou JZ, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. *Appl Environ Microbiol.* 1996;62:316–22.

T

Taxa Counting Using Specific Peptides of Aminoacyl tRNA Synthetases

David Horn
School of Physics and Astronomy, Tel Aviv
University, Tel Aviv, Israel

Synonyms

Short Read Analysis; Specific Peptides; Taxa Counting

Definition

Motifs that appear on Aminoacyl tRNA Synthetases can serve as specific peptides (SP) whose presence in a metagenome indicates which taxa it contains. This is used to devise a method, based on gene fragments rather than on 16S rRNA sequences, which allows for taxa counting from short read metagenomic data. It is exemplified on human gut microbial data.

Introduction: The SP Approach

Specific peptides (SPs) are short deterministic motifs whose presence in the protein sequence is a good predictor of an enzymatic activity of the

protein. SPs were introduced by Kunik et al. (2007), and their predictive powers on full protein sequences were established by Weingart et al. (2009). Their results are the basis of the webtool <http://horn.tau.ac.il/DME11.html> which supplies enzymatic assignments for queried protein sequences. This methodology has been applied directly to short reads, obtaining enzymatic and taxonomic signatures of data, by Weingart et al. (2010). These authors have extracted a set of SPs that are associated with single proteins of the aaRS families, known as the S61 set (because the EC numbers of these enzymes, indicating their 4-level enzymatic classification, start with 6.1.1.). The application of SPs to taxa counting in metagenomic data has been developed by Persi et al. (2012). To ensure high precision of the prediction process, it is required that the length of the SPs in the S61 set is at least nine amino acids. The resulting list contains 3,949 SPs.

The Taxa Counting Algorithm

For short read data one first converts all genomic reads to amino acid strings in the six possible reading frames. One then identifies all reads that share a single SP. Choosing the largest group of such reads, one tries to group the short reads into sets such that all reads within a set are consistent with one another (i.e., can be fused

with each other) and every set is inconsistent with the other ones. Although this mathematical problem is NP complete, one may devise simple algorithms that carry it out efficiently (Persi et al. 2012). The strong consistency conditions can be relaxed to allow for errors, such that reads within a set may differ from each other by one amino acid, and different sets have to differ from each other by at least two amino acids. The number of different sets becomes a lower bound on the number of different taxa. For short reads, distinguishing between species belonging to the same genus is impossible. Depending on the length of the short reads, chances are high however for distinguishing between different families, classes, and phyla. For the case of long sequences or extensive contigs, one can resort to searching for sequences that share several SPs of the same aaRS enzyme. This allows one to address the question of counting different species or even different strains of the same species.

Tests and Applications

Persi et al. (2012) have compared the S61 SP approach with the 16S rRNA analysis on an artificial metagenome composed of 64 genomes of different species that represent bacterial taxonomic diversity. For some of the principal phyla, they selected pairs of strains of the same species, such that the resolutions of the taxonomic delineation of the two methods can be tested and compared. The SP approach has been proved to match the accuracy provided by the 16S analysis and sometimes even to surpass it. The novel method has then been applied to species counting in the human gut microbiome employing the data of Qin et al. (2010). These data were based on samples taken from 124 individuals. In addition to raw short read data, the authors have presented genomic contigs, as well as a nonredundant set of 3.3 M ORFs derived from full genomic analysis (also called “prevalent genes”). The analysis of the prevalent genes has led Qin et al. (2010) to conclude that there

exist more than 1,000 different species in their metagenomic data. Persi et al. (2012) has argued that the prevalent genes, when analyzed using the S61 approach, display only half this count. If, however, the full set of contigs is analyzed, an estimate of over 1,000 different species and strains is obtained. The number of different genera has been estimated to be relatively small, presumably of the order of a few tens.

Of particular interest is the application of the novel method to short read data. Here this method is quite unique. It allows for a quick estimate of species count directly from raw data. Short read singletons that are often discarded from metagenomic analysis, because they cannot combine with other short reads to form longer contigs, can be readily included in this analysis. Moreover, one can test the sensitivity of the results to sample size, to the minimal distance d allowed between reads that are classified in the same taxa, and to noise in the data.

The raw data contain errors, and every misidentification of an amino acid will affect taxa counts. The probability of such errors was estimated to be below 1 %. This was then tested by inserting artificial random errors at the level of 1 % into analyzed reads. The results showed that the $d \geq 2$ counts of the set with artificial errors are similar to the $d \geq 1$ estimates drawn from the raw data. One may therefore conclude that limiting oneself to $d \geq 2$ analysis of the raw data suffices to eliminate the majority of errors in the data. Sample sizes of order 1,000 short reads of the Qin et al. (2010) data lead to counts of 200 or more taxa. The counts keep increasing linearly with sample size, indicating that greater depth unravels larger numbers of strains and species. Focusing on large distances between reads, such as $d \geq 7$, the taxa counts in the analysis of Persi et al. (2012) saturate at about 60, providing a stable bound on the number of species that are expected to have quite large Hamming distances (over 150) between their relevant protein sequences. Finally it is interesting to note that an analysis of Persi et al. (2012) carried out for all short reads of one of the subjects has shown

10 % novel species with respect to the contigs of Qin et al. (2010), and about 45 % novelties when compared to all Uniprot enzymes.

Discussion

The richness of microbiomes has become a widely recognized topic. It is often being analyzed by using 16S rRNA definitions of OTUs, whose direct contact with observed and analyzed organisms may be lacking. The alternative method provided by the S61 SP approach is based on peptides that have been extracted from an analysis of enzymes recorded in Swiss-Prot. This is the only bias of this method. It is conceivable that some genes of new species will be so far removed from the known ones that no SP match will occur and they may thus avoid detection. Therefore the list of SPs should be updated from time to time (Weingart et al. 2009) as the database grows.

A major advantage of the S61 SP methodology is its simplicity: its straightforward implementation does not require any further choice of parameters or comparisons with additional databases. Furthermore, it is satisfying to realize that it can be applied to short reads. Even those short reads that cannot be combined into contigs may lead to informative conclusions on taxa counting using the S61 approach.

Summary

Taxonomic deciphering of metagenomic data usually relies on 16S rRNA analysis. Alternatively one may use genomic information, in particular genes related to single proteins, i.e., those known to appear only once in a genome. Some of the Aminoacyl tRNA Synthetases (aaRS) fit this description. Employing specific peptides, whose occurrence is restricted to these protein families, one can devise algorithms of taxa counting. The latter turn out to be informative even for short read metagenomic data.

Cross-References

- ▶ [Computational Approaches for Metagenomic Datasets](#)
- ▶ [Human Gut Microbial Genes by Metagenomic Sequencing](#)

References

- Kunik V, Meroz Y, Solan Z, et al. Functional representation of enzymes by specific peptides. *PLoS Comput Biol.* 2007;3(8):e167.
- Persi E, Weingart U, Freilich S, Horn D. Peptide markers of aminoacyl tRNA synthetases facilitate taxa counting in metagenomic data. *BMC Genomics.* 2012;13:65.
- Qin J, Li R, Raes J, Arumugam M, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:08821.
- Weingart U, Lavi Y, Horn D. Data mining of enzymes using specific peptides. *BMC Bioinformatics.* 2009;10:446.
- Weingart U, Persi E, Gophna U, Horn D. Deriving enzymatic and taxonomic signatures of metagenomes from short read data. *BMC Bioinformatics.* 2010;11:390.

Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3

Wolfgang Gerlach¹ and Jens Stoye²

¹Institute for Genomics and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

²Faculty of Technology, Bielefeld University, Bielefeld, Germany

Synonyms

Classification; Metagenome; Taxonomy

Definition

CARMA3 is a program to assign taxonomic identifiers to metagenomic sequences of unknown taxonomic origin.

Introduction

The vast majority of microbes cannot be cultivated in a monoculture and thus cannot be sequenced by means of traditional methods. To explore these microbes, they have to be analyzed within their natural microbial communities. High-throughput sequencing (HTS) technologies like Roche's 454 sequencing, ABI's SOLiD, or Illumina's Genome Analyzer make it possible to sequence microbial DNA samples of such communities, called *metagenomes*. Due to the restricted read lengths produced by these technologies, reconstruction of complete genomic sequences from a metagenome is impossible. However, by comparing the metagenomic fragments with sequences of known function, it is possible to analyze the biological diversity and the underlying metabolic pathways in microbial communities.

To infer the taxonomic origin of metagenomic reads, two kinds of methods, composition-based and comparison-based, can be distinguished. The composition-based methods extract sequence features like GC content or k-mer frequencies and compare them with features computed from reference sequences with known taxonomic origin (Abe et al. 2005; Diaz et al. 2009; Karlin et al. 1997; McHardy et al. 2007). A disadvantage is that short reads are not suited for this method as rather long reads are required to obtain a reasonable classification accuracy. The comparison-based methods, in contrast, rely on homology information obtained by database searches. They can be further subdivided into methods that are based on hidden Markov model (HMM) homology searches (Eddy 1998) and those that are based on BLAST homology searches (Altschul 1990, 1997; Gish and States 1993). CARMA version 1 (Krause et al. 2008) and CARMA version 2 (Gerlach et al. 2009) belong to the HMM-based methods. CARMA version 3 (Gerlach and Stoye 2011) has been implemented in two variants, one of which is HMMER3-based and therefore also belongs to the HMM-based methods.

For the taxonomic classification of metagenomic reads based on BLAST, different

methods have been developed. Probably the most basic method is to use BLAST to search for the best hit in a database of sequences with known origin. Since the evolutionary distance between the source organisms of the metagenomic fragment and the database sequence is unknown, a classification result solely based on a best BLAST hit has to be interpreted carefully. In general, such a classification is more reliable on higher taxonomic levels (e.g., superkingdom or phylum) than on lower taxonomic levels (e.g., genus or species), but it is difficult to decide which taxonomic level is reliable enough, as this strongly varies for each metagenomic fragment.

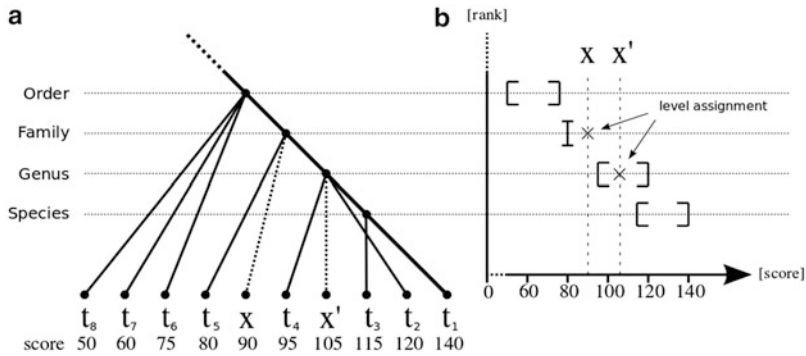
The program MEGAN (Huson et al. 2007, 2011) is based on the lowest common ancestor (LCA) approach. A BLAST search is performed, and all BLAST hits that have a bit score close to the bit score of the best hit are collected. The metagenomic fragment is then classified by computing the LCA of all species in this set. One of the reasons for the improved classification accuracy of this approach is that fragments with ambiguous hits are assigned at higher taxonomic levels.

The SORT-ITEMS method (Haque et al. 2009) extends the LCA approach and uses additional techniques to reduce the number of false-positive predictions. One approach is the reduction of the number of hits by using a reciprocal BLAST search step. Another technique used is the adaptation of the taxonomic assignment level for all hits, based on different alignment parameters like sequence similarity between the metagenomic fragment and the aligned database sequence.

Inspired by these techniques, in particular the reciprocal search step of SORT-ITEMS, CARMA3 (Gerlach and Stoye 2011) was developed to further improve the accuracy of the taxonomic classification. It makes explicit use of the assumption of a model of evolution where different gene families have different rates of mutation, but within each family this rate does not change too much.

Methods

The first step in CARMA3 consists of a BLASTx search of the metagenomic DNA sequence



Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3, Fig. 1 (a) Projections of BLAST hits obtained from reciprocal search onto the lineage of t_1 . The dashed edges represent projections of

unknown phylogenetic affiliations x and x' of metagenomic sequences q and q' , respectively. (b) Intervals given by reciprocal bit scores for each taxonomic rank and level assignments of x and x' based on their score

against the NCBI NR protein database. All protein fragments in the database that have an alignment with the metagenomic sequence are extracted. These sequences, as well as the protein translation of the metagenomic DNA sequence, as given by the BLAST alignment between the metagenomic query and the best database hit, are used to create a small protein BLAST database. In the second step of CARMA3, the reciprocal BLAST search, the extracted protein fragment that corresponds to the best BLAST hit is searched against this database using BLASTp. Since the protein fragment that is searched against the database is included in the database, this database sequence produces a perfect alignment and yields the best BLAST bit score.

Let t_i be the taxonomic affiliation of the i th best BLAST hit in the reciprocal search, and let x be the (unknown) species of the metagenomic sequence. Clearly t_1 , which is also the taxonomic affiliation of the best BLAST hit in the first BLAST search, is the phylogenetically closest known relative of x . Since the taxonomy assignment t_1 is usually located at taxonomic rank species, strain, or substrain, and metagenomic sequences mostly come from species that are phylogenetically more distantly related, using t_1 as taxonomic classification for x would be an overprediction. Therefore, the purpose of this method is to approximate the lowest common ancestor of t_1 and x , which would be the best possible taxonomic classification.

The reciprocal search provides similarity scores in terms of BLAST bit scores between t_1 and all other database sequences. Since the taxonomic affiliations of the other database sequences, except the metagenomic sequence, are known, the reciprocal search provides means to correlate BLAST bit scores with phylogenetic distances. Database sequences that are more closely related to t_1 tend also to have higher reciprocal bit scores than the less closely related sequences. A toy example for this is given in Fig. 1a.

Each t_i is projected onto the lowest common ancestor of t_i and t_1 , a taxon within the lineage of t_1 . For each taxon in the lineage of t_1 that gets projections from a subset of t_i , an interval is defined by the minimum and the maximal reciprocal bit scores from the BLAST hits in this subset. Intervals for the reciprocal search example are depicted in Fig. 1b. These intervals can be used to assign a metagenomic sequence to a taxon in the lineage of t_1 based on its reciprocal score. In general (case a) this method tries to assign the metagenomic sequence to the lowest taxonomic rank at which its reciprocal score is still within the borders of the interval at that rank. If such an interval does not exist (case b), the lowest taxonomic rank is chosen for which all bit scores are still lower than the bit score of the metagenomic sequence.

Two examples for the taxonomic classification are given in Fig. 1b. Metagenomic read

q with unknown phylogenetic affiliation x has a reciprocal score of 90. Since the bit score of q is higher than the bit score of the single hit in the interval at rank family ($t_5 = 80$), but smaller than any hit in the interval at taxonomic rank genus ($t_4 = 95$ and $t_2 = 120$), x gets assigned to the rank family (case b). The second metagenomic read q' with reciprocal score of 105 is within the borders of the interval at rank genus and thus x' gets assigned to the rank genus (case a).

Real data often does not show the properties as assumed in this model, and sometimes reciprocal scores are missing for a taxonomic rank. To accommodate for this, CARMA3 additionally employs techniques like polishing, linear interpolation, and a fallback method, described in detail in the original publication (Gerlach and Stoye 2011). CARMA3 is also available in a variant that is based on HMMER3 homology searches against the Pfam (Finn et al. 2010) database. In this variant the metagenomic sequences are aligned against Pfam family alignments from which reciprocal scores can be computed that are required for the taxonomic classification. Both the BLAST and the HMMER variants of CARMA3 can also be used for the taxonomic classification of amino acid sequences.

Results and Discussion

CARMA3 is available via the WebCARMA pipeline that takes metagenomic reads as input and output taxonomic and functional classifications. The pipeline runs on the compute cluster of the Bielefeld University Bioinformatics Resource Facility at the Center for Biotechnology (CeBiTec) and is freely accessible at <http://webcarma.cebitec.uni-bielefeld.de>. The complete source code of CARMA3 (C/C++) has been released under the GPL and is available for download from the WebCARMA homepage.

CARMA3 has been evaluated in various experiments including simulated and real metagenomes. In the following the results of two of these experiments are shown. The first experiment is a qualitative comparison of CARMA3 with

Sort-ITEMS and MEGAN using simulated data. The simulated metagenome consists of 25 randomly chosen bacterial genomes from the NCBI ftp site (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>). $N = 25\,000$ metagenomic reads were simulated using MetaSim (Richter et al. 2008) with the default 454 sequencing error model resulting in an average read length of 265 bp. The second experiment is an example of the applicability of CARMA3 in the case of very large metagenomes that can be produced, for example, by the Illumina sequencing technology. In this experiment the real data set consists of 3.3 million nonredundant microbial genes of the gene catalogue of the human gut microbiome (Qin et al. 2010). Fecal samples from different individuals were sequenced with the Illumina Genome Analyzer (GA) which yielded in 576.7 Gb of sequence. The reads were assembled into longer contigs, and a gene finder was used to detect open reading frames (ORFs). Similar ORFs were clustered to obtain the final nonredundant gene set. This gene set was downloaded and the ORFs were translated into protein sequences using the NCBI Genetic Code 11.

Comparison with Other Methods Using Simulated Data

To evaluate the different BLAST-based methods regarding their ability to classify sequences of unknown source organism, three BLAST NR protein databases were created: “order-filtered,” without sequences from species that share the same order as any of the species from the simulated metagenome; “species-filtered,” without sequences from species in the simulated metagenome; and “All,” the complete NR database.

The BLASTx runs for CARMA3, Sort-ITEMS, and MEGAN against these three databases were performed with default E-value threshold ($-e\ 10$), soft sequence masking ($-F\ "mS"$), and frameshift penalty 15 ($-w\ 15$). To ensure comparability, CARMA3 used the same thresholds as Sort-ITEMS regarding the BLASTx hits, a minimal bit score of 35, and a minimal alignment length of 25. The parameter

Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3, Table 1 Comparison of the taxonomic classification accuracy of the different BLASTx-based methods CARMA3, SOrt-ITEMS, and MEGAN using the order-filtered database

	CARMA3		SOrt-ITEMS		MEGAN	
	TP	FP	TP	FP	TP	FP
Superkingdom	12,696	861	12,576	786	12,626	1,849
Phylum	8,989	1,224	9,254	1,736	8,079	1,985
Class	4,066	1,495	4,062	1,937	3,649	2,479
Order	–	2,507	–	4,011	–	4,975
Family	–	1,186	–	2,565	–	4,087
Genus	–	210	–	798	–	4,041
Species	–	23	–	0	–	3,544

for the minimal number of reads that are required to report a taxon in SOrt-ITEMS and MEGAN was set to 1 in all experiments. To ensure comparability of MEGAN with the other two BLAST-based methods, the top percent parameter was increased from ten (default) to 15 resulting in more conservative predictions. Although CARMA3 is a parameter-free method, an artificial parameter p was introduced to slightly increase the sensitivity at the cost of decreased specificity, in order to yield a sensitivity comparable to that of the other two methods. This allowed for evaluating each of the methods based on their number of false positives. The values of p were 1.024 for order-filtered, 1.033 for species-filtered, and 1.15 for the unfiltered database.

The taxonomic classification methods assign to a metagenomic read one taxon and therefore also one taxonomic rank. This taxon implicitly provides a taxonomic classification also for the higher taxonomic ranks. For example, the taxon Gammaproteobacteria at the taxonomic rank class implicitly provides the taxonomic classification Bacteria at the taxonomic rank superkingdom. The taxonomic ranks below the predicted taxon can be considered to be classified as “unknown.” Therefore, for each taxonomic rank, a metagenomic read can either be correctly classified and counts as a true positive (TP), can be wrongly classified and counts as a false positive (FP), or it is not classified and counts as unknown (U). As for each taxonomic rank the

numbers TP, FP, and U sum up to the total number N of reads used in the evaluation and U equals $N - TP - FP$, U is not explicitly given in the results.

The complete table for all results can be found in the original publication. Table 1 below shows the results for the evaluation on the order-filtered database. While CARMA3 performs better than SOrt-ITEMS at rank class, since it has the same number of true positives but fewer false positives, for the ranks superkingdom and phylum, it is not clear which method is better. At the taxonomic ranks order to genus, where the metagenomic sequences have been filtered away, CARMA3 has much fewer ($\sim 37\text{--}74\%$) false positives than SOrt-ITEMS. CARMA3 has better results than MEGAN at all taxonomic ranks, while SOrt-ITEMS has better results than MEGAN at all taxonomic ranks below superkingdom. The results for the species-filtered and the complete NR database, where closely related or identical reference species are available in the database, show that in such a setting CARMA3 performs similar to the other two methods.

Taxonomic Classification of the Human Gut Microbiome with CARMA3

A taxonomic classification based on BLAST has the advantage of a high sensitivity, which is in particular important if no closely related reference species are available. The main bottleneck of this approach is the computation time required for the BLAST search. Over 98 % of the total

running time of a CARMA3 analysis is due to the initial BLASTx search against the NR database. While a BLASTx analysis of a complete 454 run is feasible on a compute cluster in the order of hours or a few days, this approach seems to be less practical for the analysis of all unassembled reads produced by a complete run of an Illumina sequencing machine that produces one to two orders of magnitude more bases in total than a 454 sequencing machine in a single run.

One way to overcome this limitation is the usage of data reduction techniques. This is a common strategy to handle the amount of data produced by Illumina sequencing machines (Qin et al. 2010; Hess et al. 2011). Typical steps involve the assembly of reads into longer fragments, gene detection with a gene finder to detect open reading frames (ORFs), clustering of highly similar ORFs, and translation of the nonredundant ORFs into protein sequences. Such a metaproteome has, in contrast to the full set of unassembled Illumina reads, a size that makes the analysis with the BLASTp variant of CARMA3 possible on a compute cluster in the order of hours or a few days. To evaluate the applicability of CARMA3 on amino acid sequences derived from assembled Illumina reads, the BLASTp variant of CARMA3 was used to analyze the gene catalogue of the human gut microbiome (Qin et al. 2010). The results were compared to the taxonomic classification of another study of the human intestinal microbial flora based on 13,355 prokaryotic 16S ribosomal RNA gene sequences (Eckburg et al. 2005).

Both methods, the 16S rDNA analysis and CARMA3, identify Firmicutes and Bacteroidetes as the most abundant phyla, followed by Proteobacteria, Actinobacteria, Verrucomicrobia, and Fusobacteria. Also, in both analyses, the phylum Firmicutes consists mainly of the class Clostridia. Nearly all genera of the Clostridia that have been predicted by the 16S rDNA analysis, like *Eubacterium*, *Ruminococcus*, *Dorea*, *Butyrivibrio*, and *Coprococcus*, have also been predicted by CARMA3. Also most of the species of Clostridia like *E. rectale*, *E. hallii*, *R. torques*, *R. gnavus*, *F. prausnitzii*, *D. formicigenerans*, and

D. longicatena that are found by the 16S rDNA analysis could be confirmed by CARMA3. However, the species *E. hadrum* and *R. callidus* that have been found by 16S rDNA were not found by CARMA3. The genus *Clostridium* which is the taxon found by CARMA3 to have the highest abundance in the class Clostridia is not reported by the 16S rDNA analysis. The reason for this might be that the 16S rDNA sequence of *Clostridium bartlettii*, which mostly contributes to the genus *Clostridium* and is known to be found in human feces, might not have been available at the time of the 16S rDNA analysis (Song et al. 2004). Also the species *R. inulinivorans* and *R. intestinalis* of the genus *Roseburia*, which are found by CARMA3 but not by the 16S rDNA analysis, are known to occur in human feces (Duncan et al. 2002; Scott et al. 2011). For the second most abundant phylum, the Bacteroidetes, the authors of the 16S rDNA analysis report a high variability in the distribution of phylotypes in samples from different subjects. Nevertheless, all phylotypes reported by the authors of the 16S rDNA analysis, *B. vulgatus*, Prevotellaceae, *B. thetaiotaomicron*, *B. caccae*, and *B. fragilis*, were among the 11 or, in case of *B. putredinis*, among the 22 most abundant taxa predicted by CARMA3 (Gerlach et al. 2011, Supplementary Figs. S22–S25).

The comparison of the taxonomic predictions of the 16S rDNA analysis and CARMA3 has revealed a high consistency in the results of both methods. This shows that CARMA3 can also be used for the taxonomic classification of amino acid sequences obtained from assembled Illumina reads.

Summary

CARMA3 is a method for the taxonomic classification of assembled and unassembled metagenomic sequences that can be used in combination with BLAST- and HMMER-based homology searches. Except for the homology search and the fallback scenario, this method is parameter-free. In addition, for the HMMER-based variant, it also provides a functional

classification of the metagenomic sequence. Typically, a metagenomic sample contains many novel species that have not been sequenced before. Such a scenario has been simulated with the order-filtered database, and it also has been shown that in most cases CARMA3 not only performs better than existing BLAST-based methods, but most strikingly, it is better at avoiding FP predictions on lower taxonomic ranks when only remote homologues are available for the classification of novel species.

One reason for the high accuracy of CARMA3 is because reciprocal hits provide a reasonable estimation of the last common ancestor of the metagenomic sequence and its best hit in the sequence database. In contrast to the other BLAST-based methods, this method is not based on the LCA and therefore does not discard reciprocal hits that can provide valuable information for the taxonomic classification.

A drawback of using BLASTx is its running time. The computational bottleneck of the CARMA3 pipeline is the homology search, in particular the BLAST search. In the evaluation the initial BLAST search accounted for over 98 % of the total running time. However, this is a problem shared with all BLAST-based approaches. Furthermore, it has been shown in the evaluation that this problem can be dealt with by the use of data reduction strategies which include assembly and gene detection steps.

Currently available biological sequence databases are known to be biased because they mainly contain sequences of species that are culturable. Although the authors have tried to minimize the effect of this bias on the results of their evaluation by creating the order-filtered database, this bias has to be kept in mind when generalizing the evaluation results to metagenomic reads from unculturable species.

Cross-References

- ▶ [MEtaGenome ANalyzer \(MEGAN\): Metagenomic Expert Resource](#)
- ▶ [PhyloPythia\(S\)](#)

References

- Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res*, Center for Information Biology, National Institute of Genetics, The Graduate University for Advanced Studies (Sokendai) Mishima, Shizuoka, Japan. 2005;12:281–290.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–402.
- Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinforma*. 2009;10:56.
- Duncan SH, Hold GL, Barcenilla A, Stewart CS, Flint HJ. *Roseburia intestinalis* sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *Int J Syst Evol Microbiol*. 2002;52(Pt 5):1615–20.
- Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the human intestinal microbial flora. *Science*, Division of Infectious Diseases and Geographic Medicine, Stanford University School of Medicine, Room S-169, 300 Pasteur Drive, Stanford CA 94305-5107, USA. 2005;308:1635–1638.
- Eddy SR. Profile hidden Markov models (review). *Bioinformatics*. 1998;14(9):755–63.
- Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2010;38(Database issue):D211–22.
- Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res*. 2011;39(14):e91.
- Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinforma*. 2009;10:430.
- Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet*. 1993;3(3):266–72.
- Haque MM, Ghosh TS, Komanduri D, Mande SS. SORT-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*. 2009;25(14):1722–30.
- Hess M, Sczyrba A, Egan R, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*. 2011;331(6016):463–7. New York, N.Y.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17:377–86.

- Huson DH, Mitra S, Weber N, Ruscheweyh H, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 2011;21:1552–60.
- Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol.* 1997;179:3899–913.
- Krause L, Diaz NN, Edwards RA, et al. Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *J Biotechnol.* 2008;136(1–2):91–101.
- McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4(1):63–72.
- Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464(7285):59–65.
- Richter DC, Ott F, Auch AF, Schmid R, Huson DH. Metasim: a sequencing simulator for genomics and metagenomics. *PLoS One.* 2008;3(10):e3373.
- Scott KP, Martin JC, Chassard C, Clerget M, Potrykus J, Campbell G, Mayer C-D, Young P, Rucklidge G, Ramsay AG, Flint HJ. Substrate-driven gene expression in *Roseburia inulinivorans*: importance of inducible enzymes in the utilization of inulin and starch. *Proc Natl Acad Sci U S A*, Rowett Institute of Nutrition and Health, University of Aberdeen, Bucksburn, Aberdeen AB21 9SB, United Kingdom. 2011;108(1):4672–4679.
- Song YL, Liu CX, McTeague M, Summanen P, Finegold SM. *Clostridium bartlettii* sp. nov., isolated from human feces. *Anaerobe.* 2004;10(3):179–84.

The Vaginal Microbiome in Health and Disease

Ronald F. Lamont

Department of Gynecology and Obstetrics,
Clinical Institute, University of Southern
Denmark, Odense University Hospital, Odense,
Denmark
Division of Surgery, University College London,
Northwick Park Institute of Medical Research
Campus, London, UK

Definition of the Human Vaginal Microbiome

The full collection of microbial genomes (bacterial, viral, fungal, etc.) in the human vagina.

Introduction

The resident microbial flora of the healthy vagina provides protection from infection by a number of different mechanisms. Until recently, our knowledge of the composition of the vaginal microbial flora came from qualitative/semiquantitative descriptive studies using culture-dependent techniques. Following the development and introduction of culture-independent molecular-based techniques, new information with respect to the composition of normal vaginal flora in health and disease has expanded our knowledge (Lamont et al. 2011). Most studies, whether culture dependent or independent, give the impression that the composition of vaginal flora is static and do not reflect the fact that such communities undergo shifts in their relative representation, abundance, and virulence between individuals and over time (Costello et al. 2009), all of which are affected by many factors. In this way, there may be a relatively stable “core” vaginal microbiome together with a “variable” microbiome that is affected inter alia by transient members of the community as well as by host factors such as environment, lifestyle, genotype, and immune response (Turnbaugh et al. 2007).

Normal Vaginal Flora

Culture and microscopy of “normal” vaginal flora typically demonstrates a predominance of *Lactobacillus* species, which are believed to promote a healthy vaginal milieu by providing numerical dominance but also by producing lactic acid to maintain an acid environment that is inhospitable to many bacteria. Lactobacilli also produce hydrogen peroxide (H₂O₂), antibiotic hydroxyl radicals, bacteriocins, and probiotics. Most of the data on the vaginal microbiome published to date have been derived from healthy asymptomatic women of reproductive age (Zhou et al. 2007; Srinivasan et al. 2010; Ravel et al. 2011; Gajer et al. 2012). Using culture-independent techniques, it can be demonstrated that a significant proportion (7–33 %) of healthy women lack appreciable numbers of *Lactobacillus* species in

the vagina that may be replaced by other lactic acid-producing bacteria such as *Atopobium vaginae*, *Megasphaera*, and *Leptotrichia* species. Although the structure of the communities may differ between populations, this demonstrates that vaginal health can be maintained, provided the function of these communities (lactic acid production) continues. Consequently, the absence of lactobacilli or the presence of certain organisms such as *Gardnerella vaginalis* or species of *Peptostreptococcus*, *Prevotella*, *Pseudomonas*, and/or *Streptococcus* does not constitute an abnormal state.

The Role of Lactobacilli from Culture-Independent Studies

Culture-based techniques, because they fail to detect fastidious organisms, underestimate the diversity of vaginal microbial flora, but because of deficiencies in the phenotypic identification of lactobacilli, they overestimate the diversity of *Lactobacillus* species in the vagina. Some 20 years ago, using culture-based phenotypic techniques, Redondo-Lopez et al. concluded that no two women were colonized by the same two *Lactobacillus* species (Redondo-Lopez et al. 1990). Using culture-independent techniques, we now know this is inaccurate, and because of their significant role in health and disease, much attention has been given to the identification of lactobacilli using genotypic means. Culture-independent studies using molecular-based techniques have been carried out in different populations from different geographic locations (Lamont et al. 2011). Racial variation and geographical area are important, and different racial groups within the same geographical region have significant differences in what is the dominant vaginal organism. In most populations, *Lactobacillus crispatus* is the most common dominant isolate, and White women are more likely to be dominated by *L. crispatus* and/or *Lactobacillus jensenii* than any other species of *Lactobacillus*. A number of genetic as well as environmental factors might explain at least part of this observation. Alternatively, diet might influence the *Lactobacillus* species resident in the gastrointestinal tract and hence the

vagina, as the lactobacilli of the gut vary between Japanese and Western women.

Over 120 species of *Lactobacillus* have been identified, and more than 20 species have been detected in the vagina. Using molecular-based techniques and in contrast to the assertion of Redondo-Lopez et al., outlined above (Redondo-Lopez et al. 1990), we now know that healthy vaginal flora does not contain high numbers of many different species of *Lactobacillus*. At least six subtypes or community state types (CSTs) of vaginal microbiome exist (Zhou et al. 2007; Zhou et al. 2010; Ravel et al. 2011; Gajer et al. 2012). Four of these CSTs are mainly dominated by one or two lactobacilli from a range of four species (*L. crispatus*, *L. jensenii*, *Lactobacillus iners*, and *Lactobacillus gasseri*). The remaining two CSTs lack substantial numbers of different species of lactobacilli and are composed of a diverse array of anaerobic bacteria including species associated with bacterial vaginosis (BV) such as *Prevotella*, *Megasphaera*, *G. vaginalis*, *Sneathia*, and *A. vaginae* (Fredricks et al. 2005). In *Lactobacillus*-dominated CSTs, other species are rare, are lower in titer, and tend to be novel phylotypes. The exclusion of other species is in keeping with the theory of “competitive exclusion” and the superior ability of microorganisms such as *L. crispatus* to compete with other bacteria for vaginal resources, a survival strategy known as “bacterial interference.” Alternatively, the rare coexistence of multiple dominant species of *Lactobacillus* could result from preemptive colonization by a particular species or from host factors that strongly influence the choice of species to colonize the vagina.

Lactobacillus iners: Under-detected and Underappreciated

The existence of *L. iners* was unknown prior to 1999, but due to molecular-based studies, it is now known to play a significant role in the vaginal microbial flora. Culture-independent methods have identified *L. iners*, a lactic acid-producing bacterium, as one of the organisms most frequently isolated from the vagina of healthy women. In contrast to *L. crispatus*,

which is rarely dominant in bacterial vaginosis (BV), *L. iners* can be detected at high levels in most subjects with and without BV, and in many studies it is the only *Lactobacillus* species detected in women with BV. It has been postulated that this may be because *L. iners* may be better adapted to the conditions associated with BV, i.e., the polymicrobial state of the vaginal flora and elevated pH. Alternatively, it could be the relative resistance of *L. iners* to unknown factors that led to the demise of other *Lactobacillus* species during the onset of BV or to a relative lack of antagonism of *L. iners* to BV-associated anaerobes, so that their dominance predisposes the individual to the acquisition of BV.

Community Group Variations Among Different Ethnic Groups

The vaginal microbiome and pH in asymptomatic, sexually active women who were fairly equally represented according to self-reported ethnic group (Hispanic, Black, Asian, White) has been studied (Ravel et al. 2011). The proportion of each community group and pH among the four ethnic groups varied significantly. Bacterial communities dominated by lactobacilli were found significantly more commonly in Asian and White women (80.2 % and 89.7 %, respectively) compared to only 59.6 % and 61.9 % in Hispanic and Black women, respectively. Similarly, median pH values were significantly higher in Black and Hispanic women compared to Asian and White women.

Abnormal Vaginal Flora

Abnormal vaginal flora may occur because of a sexually transmitted infection (STI), e.g., trichomoniasis, or through colonization by an organism that is not part of the normal vaginal community. Alternatively, abnormal vaginal flora may result from overgrowth or increased virulence of an organism that is a constituent part of normal vaginal flora such as *Escherichia coli*. Alterations in vaginal flora do not necessarily imply disease or result in symptoms. Disease

results from the interplay between microbial virulence, numerical dominance, and the innate and adaptive immune response of the host (Smith 1934). The most common disorder of vaginal flora is BV, which is a polymicrobial condition characterized by a decrease in the quality or quantity of lactobacilli and by a 1000-fold increase in the number of other organisms, determined by culture-dependent techniques, particularly anaerobes such as *Mycoplasma hominis*, *G. vaginalis*, and *Mobiluncus* species. BV is increasingly associated with adverse outcomes in gynecology such as pelvic inflammatory disease, postabortal sepsis, infertility, post-hysterectomy vaginal cuff infections, and the acquisition of STIs such as gonorrhea, *Chlamydia*, trichomoniasis, and HIV. In pregnancy, BV has been associated with early and late miscarriage, recurrent abortion, postpartum endometritis, and preterm birth.

Atopobium vaginae: Under-detected and Underappreciated

The genus *Atopobium* is a member of the family *Coriobacteriaceae* and forms a distinct branch within the phylum *Actinobacteria*. Following sequence analysis, three species formerly designated *Lactobacillus minutus*, *Lactobacillus rimae*, and *Streptococcus parvulus*, within the lactic acid-producing group of bacteria, have been reclassified as the genus *Atopobium*. In 1999, an organism similar but not identical to these three species was isolated from the vagina of a healthy woman in Sweden, and the organism was named *Atopobium vaginae* (Rodriguez et al. 1999). Since that time, using molecular-based techniques, *A. vaginae* has frequently been detected in the vagina and is found much more commonly in women with BV than in those with normal flora (Lamont et al. 2011). *A. vaginae* is strictly anaerobic and is very sensitive to clindamycin in vitro, but is highly resistant to nitroimidazoles such as metronidazole and secnidazole.

High Diversity of Flora in Bacterial Vaginosis Compared with Normal Flora

Using various molecular-based techniques and the Amsel clinical criteria, or Nugent score to

classify normal or abnormal flora, a number of studies have demonstrated a high diversity of organisms in women with BV compared to women with normal flora. Collectively, these studies demonstrate the presence of species such as *A. vaginae*, *Porphyromonas asaccharolytica*, bacterial vaginosis-associated bacteria (BVAB)-1, BVAB-2, and BVAB-3 in the order *Clostridiales* and species of *Megasphaera*, *Leptotrichia*, *Dialister*, *Chloroflexi*, *Eggerthella*, *Olsenella*, *Streptobacillus*, and *Shuttleworthia* which are either novel or unfamiliar to clinicians (Lamont et al. 2011). For many of these undetected or under-detected organisms, there is evidence of disease association. The renamed *Atopobium parvulum*, *Atopobium minutum*, and *Atopobium rimae* have been associated with oral infections, dental and tubo-ovarian abscesses, and abdominal wound infections, supporting the view that these organisms can be pathogenic to the host. *Leptotrichia sanguinegens/ammionii* has been reported in association with postpartum endometritis, adnexal masses, and fetal death and has been detected in the amniotic fluid of women with preterm labor, preterm prelabor rupture of the membranes, and preeclampsia. Also, in a study of 45 women with salpingitis and 44 controls (women seeking tubal ligation), bacterial 16S rDNA sequences were found in the fallopian tube specimens of 24 % of cases, but in none of the controls. Bacterial phylotypes closely related to *Leptotrichia* species and *A. vaginae* were among those identified in the cases. In addition, *Dialister pneumosintes* was found as the sole agent in the blood culture from a woman with suppurative postpartum ovarian thrombosis.

It has also been demonstrated that many of these organisms have specificity for BV and that the number of phylotypes found in association with BV is statistically significantly greater than the number detected in the presence of intermediate flora (a distinct entity in its own right) (Taylor-Robinson et al. 2003) or normal flora. This statistic largely results from the extreme dominance of lactobacilli in healthy women, which makes detection of other species unlikely, even when they are present at levels of 100,000 or

more cells/sample. In summary, these studies have demonstrated that different subjects with BV have different microbial profiles, indicating heterogeneity in the composition of bacterial taxa in women with BV. Women without BV had bacterial communities dominated by *Lactobacillus* species, accounting for 86 % of all sequences. In contrast, women with BV did not possess a single dominant phylotype, but instead had a diverse array of vaginal bacteria, often at relatively low abundances.

The Diagnosis of Bacterial Vaginosis

Bacterial vaginosis can be diagnosed clinically, microscopically, enzymatically, and chromatographically, using qualitative or semiquantitative culture methods or using composite clinical criteria. Currently, the gold standard is the Nugent score (Nugent et al. 1991), but the number of diagnostic methods testifies to the fact that no single test is ideal and that they can all provide false-positive and false-negative results.

Confounding Factors

Findings from molecular-based studies are now highlighting possible explanations for why diagnosis by microscopy may be inconsistent and why molecular methods may replace them:

1. *Mobiluncus*: One of the three organisms quantified as part of the Nugent score is *Mobiluncus*. Several cloning and sequencing studies have only rarely identified *Mobiluncus*. Fluorescence in situ hybridization (FISH) technology has demonstrated that BVAB-1 has curved-rod morphology, similar to *Mobiluncus* morphotypes, and it is possible that during microscopic examination of vaginal smears, *Mobiluncus* species may have been overrepresented and mistaken for BVAB-1. Alternatively, as species-specific PCR agrees with the Nugent score, *Mobiluncus* may be missed in universal PCR studies because it frequently falls below a threshold titer where it can be detected.
2. *Atopobium*: The urea produced by *Atopobium* species is associated with halitosis, and similarly, species of *Megasphaera* cause beer spoilage by producing turbidity, off-flavors

and off-colors. Accordingly, if two genera associated with malodorous metabolites can be found in the vagina of healthy women and amines can be found in women without BV, then diagnostic techniques to diagnose BV, based on amine production and odor formation, may need to be reconsidered. Microscopically, *Atopobium* species are gram-positive, elliptical cocci, or rod-shaped organisms that occur singly, in pairs, or in short chains. The variable cell morphology of *Atopobium* renders it well camouflaged among the mixture of other species present in bacterial communities where the Nugent score is ≥ 4 . *A. vaginae* is fastidious, grows anaerobically, and forms small pinhead colonies on culture that are easily missed. Although phylogenetically different from other lactic acid-producing bacteria, they are not phenotypically exceptional, and it is not difficult to see why the significance of this organism based on culture, microscopy, and phenotype may be overlooked and underappreciated.

3. *Symptomatic relationships*: Using species-specific primers, the relationships between five fastidious organisms associated with BV were compared with BV diagnosed by Amsel and/or Nugent scores, and also with the individual Amsel clinical criteria (Haggerty et al. 2009). The two biovars of *Ureaplasma urealyticum* (*Ureaplasma parvum* and *Ureaplasma urealyticum* – biovar 2) were associated with vaginal discharge and raised pH, but not with BV by either Amsel or Nugent criteria or any of the individual Amsel clinical criteria. In contrast, with *Leptotrichia sanguinegens/amnionii*, *A. vaginae*, and BVAB-1, an elevated pH >4.5 was a universal feature, and they were all associated with BV by both Amsel and Nugent criteria and with the finding of $>20\%$ of epithelial cells as clue cells, a feature that has already been reported. A positive test for amine odor upon the addition of 10% solution of potassium hydroxide was significantly more likely in women testing positive for BVAB-1. Douching is a recognized risk factor

for BV, and the detection of *Leptotrichia* and *A. vaginae* was three times more likely, and BVAB-1 twice as likely, when women reported douching.

Diagnosis of BV Using Qualitative and Quantitative Molecular Techniques

Some organisms or combinations of organisms have high sensitivities or specificities for the diagnosis of BV using the Amsel criteria and the Nugent score (Fredricks et al. 2005; Fredricks et al. 2007). Using quantitative real-time PCR, the association of individual organisms with BV diagnosed by Nugent score was examined qualitatively. At a threshold of $\geq 10^8$ DNA copies/ml, *Lactobacillus* species was predictive of normal flora (sensitivity 44%; specificity 100%). BVAB-1, BVAB-2, and BVAB-3 alone, or in combination, had high specificity for BV diagnosed by Amsel criteria.

Since *A. vaginae* and *G. vaginalis* are frequently detected in association with BV, a number of authors using molecular-based techniques have examined the possibility of combining these two organisms as a means of diagnosing BV. Using DNA quantitation, 19 out of 20 BV samples had either a DNA level for *A. vaginae* $\geq 10^8$ copies/ml or *G. vaginalis* $\geq 10^9$ copies/ml, and nine out of 20 had both. The combination of an *A. vaginae* DNA level $\geq 10^8$ copies/ml and a *G. vaginalis* DNA level $\geq 10^9$ copies/ml demonstrated the best predictive criteria for the diagnosis of BV with excellent sensitivity (95%), specificity (99%), negative predictive value (NPV, 99%), and positive predictive value (PPV, 95%) (Menard et al. 2008).

Culture-Independent Studies in Pregnancy

Culture-independent techniques have been used to measure prevalence, diversity, and abundance of organisms, particularly ureaplasmas in amniotic fluid, in association with suspected cervical insufficiency, preterm labor, preterm prelabor rupture of membranes (PPROM),

small-for-gestational-age babies, preeclampsia, and the potential for bacteria from the oral cavity to colonize amniotic fluid. However, apart from combining pregnant women with nonpregnant women to increase sample numbers, the information with respect to the vaginal microbiome in pregnant women is limited, particularly with respect to the outcome of pregnancy, especially preterm birth. Using species-specific primers, Wilks et al. quantified the production of H₂O₂ by lactobacilli from swabs taken at 20 weeks of gestation from the vagina of 73 women considered to be at high risk of preterm birth (Wilks et al. 2004). The levels of H₂O₂ production varied between species of *Lactobacillus*. The presence of lactobacilli producing high levels of H₂O₂ was associated with a reduced incidence of BV at 20 weeks of gestation and subsequent chorioamnionitis. The authors postulated that H₂O₂-producing lactobacilli reduced the incidence of ascending genital tract colonization in pregnancy, which leads to infection and preterm birth. In a longitudinal study of 100 pregnant women, vaginal swabs were obtained at mean gestational ages of 8.6, 21.2, and 32.4 weeks, respectively (Verstraelen et al. 2009). In the first trimester, 77 women had normal or *Lactobacillus*-dominated flora, 13 of whom developed abnormal flora in the second or third trimester. When the first-trimester normal flora was dominated by *L. gasseri* or *L. iners*, there was a tenfold risk of conversion to abnormal flora. In contrast, normal flora comprising *L. crispatus* had a fivefold decreased risk of conversion to abnormal flora. This may be because only a small percentage of *L. gasseri* and *L. iners* strains produce H₂O₂.

Knowledge of the vaginal microbiome in pregnancy is limited to only a few studies (Verstraelen et al. 2009; Hernández-Rodríguez et al. 2011; Aagaard et al. 2012), none of which analyzed samples collected longitudinally. Recently, using 16S rDNA sequencing in normal pregnant women sampled longitudinally, the vaginal microbiome was found to be different from that of nonpregnant women; also the vaginal microbiome during pregnancy is more stable than in the nonpregnant state (Romero et al. 2014).

Conclusions

Stability and resilience of the vaginal ecosystem is now recognized to be of importance in the health of a bacterial community as well as the response to perturbations. The relative abundance of certain phylotypes correlates well with low or high Nugent scores, which is used for the diagnosis of normal flora or BV. The inherent difference within and between women in different ethnic groups strongly argues for a more refined definition of the subtypes of bacterial communities normally found in healthy women and the need to appreciate differences between individuals so they can be taken into account in risk assessment and diagnosis of disease.

References

- Aagaard K, Riehle K, Ma J, Segata N, Mistretta TA, Coarfa C, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS ONE*. 2012;7:e36466.
- Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science*. 2009;326:1694–7.
- Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. *N Engl J Med*. 2005;353:1899–911.
- Fredricks DN, Fiedler TL, Thomas KK, Oakley BB, Marrazzo JM. Targeted PCR for detection of vaginal bacteria associated with bacterial vaginosis. *J Clin Microbiol*. 2007;45:3270–6.
- Gajer P, Brotman R, Bai G, Sakamoto J, Schütte U, Zhong X, et al. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med*. 2012;4:132ra152.
- Haggerty CL, Totten PA, Ferris M, Martin DH, Hoferka S, Astete SG, et al. Clinical characteristics of bacterial vaginosis among women testing positive for fastidious bacteria. *Sex Transm Infect*. 2009;85:242–8.
- Hernández-Rodríguez C, Romero-González R, Albani-Campanario M, Figueroa-Damián R, Meraz-Cruz N, Hernández-Guerrero C. Vaginal microbiota of healthy pregnant Mexican women is constituted by four *Lactobacillus* species and several vaginosis-associated bacteria. *Infect Dis Obstet Gynecol*. 2011;2011: 851485.
- Lamont R, Sobel J, Akins R, Hassan S, Chaiworapongsa T, Kusanovic J, et al. The vaginal microbiome: new information about genital tract flora using molecular based techniques. *BJOG*. 2011;118:533–49.

- Menard JP, Fenollar F, Henry M, Bretelle F, Raoult D. Molecular quantification of *Gardnerella vaginalis* and *Atopobium vaginae* loads to predict bacterial vaginosis. *Clin Infect Dis*. 2008;47:33–43.
- Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J Clin Microbiol*. 1991;29:297–301.
- Ravel J, Gajer P, Abdo Z, Schneider G, Koenig S, McCulle S, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A*. 2011;108(Suppl1):4680–7.
- Redondo-Lopez V, Cook RL, Sobel JD. Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora. *Rev Infect Dis*. 1990;12:856–72.
- Rodriguez J, Collins MD, Sjoden B, Falsen E. Characterization of a novel *Atopobium* isolate from the human vagina: description of *Atopobium vaginae* sp. nov. *Int J Sys Bacteriol*. 1999;49:1573–6.
- Romero, R, S Hassan, P Gajer, A Tarca, D Fadrosch, L Nikita, et al. The composition and stability of the vaginal microbiota of normal pregnant is different from that of non-pregnant women: results of a longitudinal study using culture-independent techniques. *Microbiome* 2014. In press.
- Smith T. Parasitism and disease. Princeton: Princeton University Press; 1934.
- Srinivasan S, Liu C, Mitchell C, Fiedler T, Thomas K, Agnew K, et al. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. *PLoS ONE*. 2010;5:e10197.
- Taylor-Robinson D, Morgan DJ, Sheehan M, Rosenstein IJ, Lamont RF. Relation between Gram-stain and clinical criteria for diagnosing bacterial vaginosis with special reference to Gram grade II evaluation. *Int J STD AIDS*. 2003;14:6–10.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JL. The human microbiome project. *Nature*. 2007;449:804–10.
- Verstraelen H, Verhelst R, Claeys G, De Backer E, Temmema M, Vaneechoutte M. Longitudinal analysis of the vaginal microflora in pregnancy suggests that *L. crispatus* promotes the stability of the normal vaginal microflora and that *L. gasseri* and/or *L. iners* are more conducive to the occurrence of abnormal vaginal microflora. *BMC Microbiol*. 2009;9:116.
- Wilks M, Wiggins R, Whiley A, Hennessy E, Warwick S, Porter H, et al. Identification and H(2)O(2) production of vaginal lactobacilli from pregnant women at high risk of preterm birth and relation with outcome. *J Clin Microbiol*. 2004;42:713–17.
- Zhou X, Brown C, Abdo Z, Davis C, Hansmann M, Joyce P, et al. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J*. 2007;1:121–33.
- Zhou X, Hansmann M, Davis C, Suzuki H, Brown C, Schütte U, et al. The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunol Med Microbiol*. 2010;58:169–81.

tRNA Gene Database Curated Manually by Experts

tRNADB-CE and Use of tRNAs as Phylogenetic Markers for Metagenomic Sequences

Takashi Abe¹, Hachiro Inokuchi², Yuko Yamada², Akira Muto³, Yuki Iwasaki² and Toshimichi Ikemura²

¹Graduate School of Science and Technology, Niigata University, Niigata, Japan

²Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga, Japan

³Faculty of Agriculture and Life Science, Hirosaki University, Hirosaki, Aomori, Japan

Synonyms

tRNA; tRNADB-CE; Taxonomic assignment using tRNA genes

Definition

The tRNA gene database curated manually by experts (“tRNADB-CE”) (<http://trna.ie.niigata-u.ac.jp>) has been constructed and annually updated by analyzing all available complete and draft genomes of Bacteria and Archaea, virus genomes, chloroplast genomes, and eukaryote genomes plus fragment sequences obtained from metagenome analyses of environmental samples. By compiling tRNAs from known prokaryotes that had identical sequences, we found high phylogenetic preservation of tRNA sequences, especially at the phylum level. Furthermore, a large number of tRNAs obtained by metagenome analyses of environmental samples had sequences identical to those found in known prokaryotes. The identical sequence group, therefore, can be used as molecular phylogenetic markers to clarify microbial community structures in environmental ecosystems as well as in clinical samples.

Introduction

In accord with the remarkable progress of DNA sequencing technology, a vast quantity of metagenomic sequences obtained from a wide variety of environmental and clinical samples have been decoded and released by DDBJ/EMBL/GenBank. A massive number of metagenomic sequences, including short sequences obtained with new-generation sequencers, should contain a large number of complete tRNA sequences because the lengths of tRNA sequences are short. However, practically no information on tRNA genes has been annotated for metagenomic sequences in DDBJ/EMBL/GenBank. The search for tRNA genes in metagenomic sequences can provide a new strategy to clarify microbial community structures in environmental samples. Thus, we included a vast number of tRNA genes found in metagenomic sequences in the tRNADB-CE (Abe et al. 2009, 2011).

When we focused on a group of tRNAs with an identical sequence, we found tRNAs only in a particular lineage of phylogenetic groups. Notably, such phylotype-specific tRNA sequences were also found in many species-unknown genomic fragments obtained by metagenome analyses. This fact shows that tRNA is a good phylogenetic marker for discovering the phylotype composition and microbial community structure in an environmental sample.

Search for tRNA Genes

In order to enhance the completeness and accuracy of searching for tRNA genes, three computer programs, tRNAscan-SE (Lowe and Eddy 1997), ARAGORN (Laslett and Canback 2004), and tRNAfinder (Kinouchi and Kurokawa 2006), were used in combination since their algorithms were partially different and rendered somewhat different results. First, we checked to what degree the predicted regions and the anticodons of individual tRNA genes were consistent with each other. The tRNA genes concordantly found by the three programs were stored in the database.

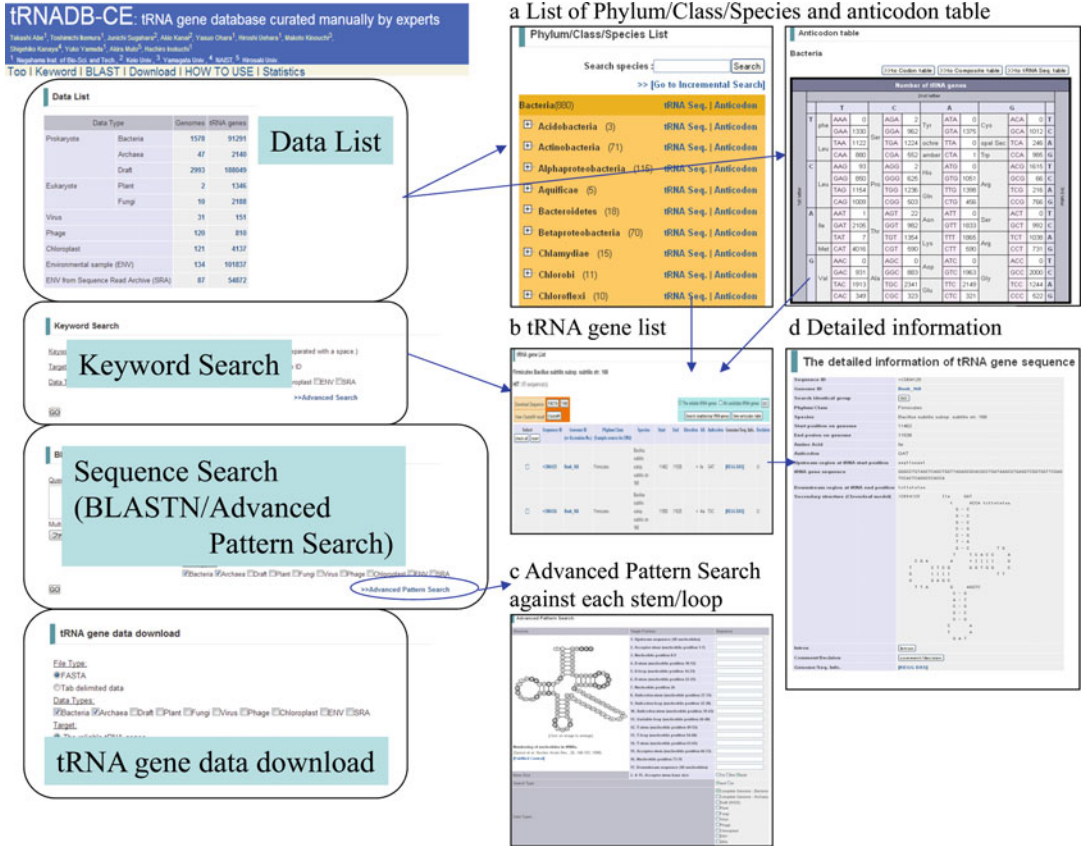
Second, three experts in the tRNA experimental field manually checked the discordant cases (approximately 3 % of the total of bacterial gene candidates) independently and included reliable cases in the database.

For fragment sequences obtained from metagenome analyses, only tRNA genes concordantly found by the three programs and those that had sequences identical to tRNAs already included in the database were stored. A large number of tRNA genes were detected in various environmental samples, and their numbers were separately listed by category of environment. This enabled us to clarify microbial community structures in environmental samples using tRNAs as phylogenetic markers. Because a significant portion of environmental DNA sequences are thought to be from unculturable microbes, tRNA genes of novel species should be included.

Functions of tRNADB-CE and Data Access

The tRNADB-CE allows browsing of the stored data and search for the database with user-specified input as described previously in detail (Abe et al. 2009, 2011). A browse page is presented in Fig. 1. First, a list of tRNA genes and anticodons can be browsed depending on the numbering of genomes (i.e., genome ID) or DNA fragments of environmental samples stored in the database. The statistical information for copy numbers of tRNA genes in each phylotype/species and the anticodon type in each amino acid group can also be browsed (Fig. 1a). By clicking the sequence ID of each tRNA gene, detailed information on the selected tRNA genes can be browsed, including tRNA gene sequences, their upstream and downstream sequences (10 nt), information on the secondary structures of the tRNA, and curation comments on the tRNA.

A “keyword search” can also be conducted using retrieved items such as species name, amino acid, anticodon, sequence ID, and genome ID. This function can be performed by using multiple keywords in combination. The database



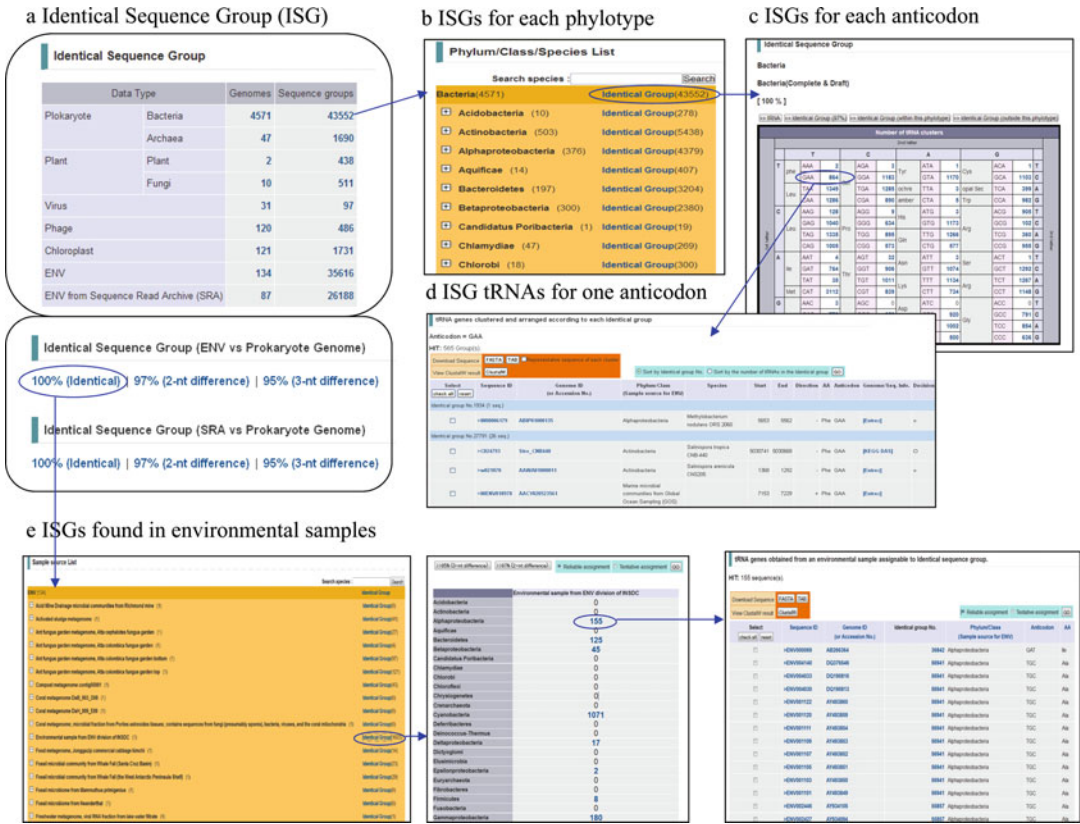
tRNA Gene Database Curated Manually by Experts, Fig. 1 Basic functions of tRNADB-CE

supports two types of sequence search: sequence similarity search “BLASTN” and pattern search. In pattern search (i.e., oligonucleotide sequence search), we can focus the search area on the stems/loops of cloverleaf structures and combine the areas in various patterns. After selecting tRNA genes of interest using the sequence search procedures, multiple alignments with ClustalW and downloads of aligned sequences and obtained dendrograms are available (Fig. 1b).

Identical Sequence Groups and Their Use as Phylogenetic Markers for Environmental Metagenomic Sequences

When we conducted the clustering of tRNA gene sequences, except the 3' CCA terminal sequence, from complete and draft genomes of Bacteria and

Archaea by sequence alignment using the CD-HIT (Li and Godzik 2006), we found high phylogenetic preservation of tRNA genes; a particular tRNA sequence was found only in a particular lineage of phylogenetic groups. We designated here the tRNA group with an identical sequence as “identical sequence group: ISG” (Fig. 2a) and listed the numbers of ISGs for each phylotype (Fig. 2b) and for each anticodon (Fig. 2c). tRNAs with one anticodon type were classified and listed according to the ISG along with the phylotype information of each tRNA (Fig. 2d), and thus, the range of phylotypes found for each ISG could be examined. If we focused on ISGs composed of more than five sequences, approximately 95 % of ISGs were conserved at a phylum level, showing most tRNAs to be good phylogenetic markers at least at the phylum level. The ISGs could provide



tRNA Gene Database Curated Manually by Experts, Fig. 2 List and search for identical sequence group (ISG)

a strategy for selecting reliable phylogenetic markers. In addition, approximately 65 % of ISGs were conserved even at the genus level, showing the possible existence of good genus-specific markers. By combining the data provided by this database with other detailed knowledge of a particular tRNA obtained by experiments or from literature, users may obtain useful phylogenetic markers (e.g., genus-specific markers) by themselves.

Interestingly, among tRNA genes found in metagenomic sequences derived from environmental samples, approximately 25 % of tRNA genes were identical in sequence to genes from species-known prokaryotes. Using tRNAs found in an environment sample that were assigned to ISGs, we could predict the microbial community structure in an environmental ecosystem at least at the phylum level (Fig. 2e). The database also has

a function for searching for sequences with 97 % or 95 % sequence identity (2- or 3-nt difference, respectively) (Fig. 2a). By using tools in the database and specific markers found by users (e.g., genus-specific markers), users can clarify microbial populations in an ecosystem by themselves.

The present strategy can be applied even to data of short sequences obtained with new-generation sequencers, such as Sequence Read Archive (SRA), in NCBI. In metagenomic analyses using new-generation sequencers, the phylogenetic characterization of short sequences with existing bioinformatics methods was particularly difficult, except for sequences unambiguously mapped on a known sequenced genome. Because complete tRNA genes can be found even from short genomic fragments of around 100 bases, tRNA genes should become one of the most effective means for identifying

microbial populations in an ecosystem in the case of metagenome studies conducted with next-generation sequencers.

When we consider the rapid growth of genomic and metagenomic sequences accumulated in DDBJ/EMBL/GenBank, our present strategy to search for reliable tRNAs, including manual curation by experts, may be inadequate. Our group previously developed BLSOM (Batch-Learning Self-Organizing Map) for oligonucleotide composition, which clustered (self-organized) genomic sequence fragments according to the phylogenetic group (Abe et al. 2003). The oligonucleotide BLSOM was successfully applied to the phylogenetic classification of a large quantity of metagenomic sequences (Abe et al. 2005). When we conducted BLSOM for the tetra- and pentanucleotide compositions of bacterial tRNAs, tRNAs were accurately separated according to the amino acid, showing the BLSOM to be an additional informatics strategy for the assignment of reliable tRNAs. When we focused on tRNAs with the same anticodon, tRNAs were separated according to the phylotype on BLSOM, showing that the BLSOM is also applicable to the phylogenetic assignment of tRNAs present in metagenomic sequences.

Summary

By compiling the tRNAs of known prokaryotes with identical sequences, we found high phylogenetic preservation of tRNA sequences, especially at the phylum level. Furthermore, a large number of tRNAs obtained by metagenome analyses had sequences identical to those found for known prokaryotes. The identical sequence

group, therefore, can be used as molecular phylogenetic markers to clarify microbial community structures of environmental ecosystems. The tRNADB-CE allows users to obtain phylotype-specific markers (e.g., genus-specific markers) by themselves and to clarify microbial community structures in ecosystems in detail.

tRNADB-CE can be accessed freely at <http://trna.ie.niigata-u.ac.jp>.

References

- Abe T, Kanaya S, Kinouchi M, Ichiba M, Kozuki T, Ikemura T. Informatics for unveiling hidden genome signatures. *Genome Res.* 2003;13:693–702.
- Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.* 2005;12:281–90.
- Abe T, Ikemura T, Ohara Y, Uehara H, Kinouchi M, Kanaya S, Yamada Y, Muto A, Inokuchi H. tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res.* 2009;37:D163–8.
- Abe T, Ikemura T, Sugahara J, Kanai A, Ohara Y, Uehara H, Kinouchi M, Kanaya S, Yamada Y, Muto A, Inokuchi H. tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Res.* 2011;39:D210–3.
- Kinouchi M, Kurokawa K. tRNAfinder: a software system to find all tRNA genes in the DNA sequence based on the cloverleaf secondary structure. *J Comp Aided Chem.* 2006;7:116–26.
- Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 2004;32:11–6.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.
- Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25:955–64.

U

Use of Bacterial Artificial Chromosomes in Metagenomics Studies, Overview

Lingling Wang¹, Shamima Nasrin², Mark Liles² and Zhongtang Yu³

¹Department of Animal Sciences, The Ohio State University, Columbus, OH, USA

²Department of Biological Sciences, Auburn University, Auburn, AL, USA

³Department of Animal Sciences, Environmental Science Graduate Program, The Ohio State University, Columbus, OH, USA

Synonyms

Environmental genomic libraries; Metagenomic libraries; Metagenomic studies

Definition

Cloning of large metagenomic DNA fragments using bacterial artificial chromosome (BAC) vectors provides an opportunity to study the functional diversity and to harness the metabolic potential of diverse microorganisms in various microbiomes. This technology is especially relevant to the study of biosynthetic pathways encoded by large gene clusters that would not be cloned as contiguous regions using other cloning vectors.

Introduction

Despite progress in culturing a greater diversity of the members of microbial assemblages, the vast majority of prokaryotic taxa is not readily cultured in laboratory and remains largely unknown. Because of this “great plate count anomaly,” microbiologists are forced to use cultivation-independent alternative approaches to access and study the functional diversity in microbiomes. Direct cloning of metagenomic DNA fragments into BAC clone libraries and subsequent analyses provide opportunities to investigate the genetic makeup and potential of a microbiome. Metagenomic DNA fragments can be cloned into plasmid, cosmid, or fosmid (a low-copy-number cosmid that is based on the F-factor replicon of *Escherichia coli*) vectors, but these cloning vectors can only carry relatively small DNA fragments: plasmids, <20 kb; cosmids, 37–52 kb; and fosmids, <42 kb. On the other hand, a BAC vector can carry a much longer DNA fragment (up to 300 kb). Although BAC libraries are technically more difficult to construct than other types of clone libraries, BAC clone libraries have several advantages. First, to clone a certain amount of metagenomic DNA, a smaller number of BAC clones are needed compared to libraries constructed using other cloning vectors. Second, the production of many bioactive compounds (e.g., antibiotics, multimodular polyketide, or nonribosomal peptide) is encoded by a gene cluster whose length typically exceeds what can be carried by a plasmid, cosmid, or fosmid vector

(Piel 2011). Third, a *cis*-regulatory element is often required for the expression of a gene or operon. However, the inserts cloned into a plasmid, cosmid, or fosmid vector might not allow for the cloning of both a gene or operon and its *cis*-regulatory element into the same clone. The lack of a *cis*-regulatory element can prevent the metabolic phenotype of interest from being detected during activity-based screening of clone libraries. Fourth, cloned metagenomic DNA fragments are less stable in a plasmid or a cosmid vector than in a BAC vector. Therefore, BAC libraries have unique utility in metagenomic studies of microbiomes. In this entry, the construction, screening, bioinformatic and biochemical analysis, and utilization of BAC clone libraries are overviewed.

Isolation of High Molecular Weight DNA

Because the purpose of BAC cloning is to recover contiguous regions of microbial genomes, the DNA fragments recovered from a microbiome sample should be significantly longer than that required for fosmid or other cloning vectors. Many studies have compared DNA extraction methods suitable for metagenomic applications (Delmont et al. 2011), and this section will review those methods that are appropriate for BAC cloning.

Extraction of high molecular weight (HMW) DNA from microbiome is always a significant issue due to the inherent conflict between the need to recover DNA from diverse microorganisms while preserving DNA integrity. Direct DNA extraction methods, by which DNA is recovered directly from an environmental sample, provide high DNA yields from phylogenetically diverse microorganisms; yet the vast majority of the DNA is sheared and likely less than 100 kb. In contrast, indirect DNA extraction methods, in which microbial cells are first isolated from sample matrices prior to DNA extraction, result in a lower DNA yield, but the resultant DNA is of significantly greater molecular weight compared to direct extraction methods.

However, indirect DNA extraction can produce less representative metagenomic DNA because some microbial cells can be difficult to isolate from the sample matrices. The choice of a direct or an indirect extraction method depends on the nature of the environmental sample. For example, for a sample with high levels of contaminants, such as soils and sediments, there may be an advantage in using indirect extraction to significantly reduce humic acid levels co-extracted with the DNA. However, indirect extraction methods may not yield sufficient DNA from samples with lower cellular abundance, such as sediments and aquifer. Therefore, it is important to use an empirical approach and evaluate multiple methods for HMW DNA extraction.

Several protocols have been developed to overcome the difficulties in extracting HMW DNA for metagenomic studies. Stein et al. (1996) introduced an innovative technique that embeds microbial cells in agarose gel matrix. These embedded cells are lysed in situ to prevent mechanical shearing of the microbial DNA. The Nycodenz extraction technique is another technique that is used to avoid mechanical shearing of HMW DNA (Berry et al. 2003). This technique prevents physical damage to bacterial cells by cushioning them during the high-speed centrifugation step. Some environmental contaminants are also removed during the centrifugation.

The use of multiple extraction methods for a single sample may increase the ultimate yield and phylogenetic representation of the recovered metagenomic DNA. The adoption of a particular DNA extraction method should be evaluated by the DNA yield and the representation of diverse microbial genomes within the recovered DNA. The diversity represented in the recovered DNA can be assessed by molecular phylogenetic analysis based on sequencing of the universally conserved 16S rRNA gene. Although phylogenetic diversity analysis can be influenced by many factors, including biases inherent in PCR, it is a rapid analysis to assess the phylogenetic composition of samples and libraries. Next-generation sequencing (NGS) of 16S rRNA gene amplicons permits a greater depth of

sequencing coverage compared to traditional Sanger sequencing technology.

Subsequent to extraction, the metagenomic DNA may need to be purified to achieve efficient cloning. Purification is especially needed to remove contaminants from metagenomic DNA extracted from soil samples because soil samples can contain high levels of humic acids and other phenolic compounds that tend to be co-extracted with the DNA and hamper downstream processes (e.g., restriction digestion and cloning of the DNA). Multiple approaches have been developed to purify metagenomic DNA, including the use of CTAB, hydroxyapatite column purification, or formamide denaturation. Formamide can be more effective in removing some contaminants from HMW DNA due to its inherent capability to denature DNA and remove contaminants that are tightly intercalated between the DNA bases and strands (Liles et al. 2008). Formamide or polyvinylpyrrolidone (PVPP) can also help remove nuclease contaminant. Hydroxyapatite chromatography has advantages over other purification approaches as this method can efficiently fractionate nucleic acids with different conformations (i.e., dsDNA, ssDNA, dsRNA, and ssRNA) while helping remove sample contaminants (Andrews-Pfannkoch et al. 2010). This differential elution of nucleic acids can easily be accomplished by changing the phosphate concentrations at a constant temperature or a combination of increasing temperatures and phosphate concentrations of the elution buffers.

Fragmentation and Size Selection of Metagenomic DNA

Fragments of metagenomic DNA with uniform length about 150 kb are prepared either enzymatically or mechanically. Enzymatic fragmentation relies upon partial restriction digestion. However, the extent of partial digestion is difficult to control. Additionally, partial restriction digestion can result in nonrandom DNA fragmentation and a significant reduction in DNA size. An alternative to partial restriction digestion is mechanical shearing, which

results in random fragmentation of metagenomic DNA. This method has been demonstrated with multiple eukaryotic genomes and recently applied to construction of soil BAC libraries studies (Kakirde and Nasrin et al., unpublished data).

A major challenge in constructing high-quality BAC libraries is to retain large DNA fragments while removing small ones that can be preferentially cloned. Multiple strategies are available to recover and clone large metagenomic DNA fragments. Pulsed field gel electrophoresis (PFGE) is the most frequently used method for size selection of partially digested or sheared metagenomic DNA. Alternatively, agarose gel electrophoresis can be used as it can provide better resolution of HMW DNA. Because sucrose gradient centrifugation can only resolve DNA fragments of 5–60 kb, it is not a suitable method to separate HMW DNA fragments for BAC library construction.

Construction of BAC Clone Libraries

Several BAC vectors have been developed for metagenomic cloning that enable transfer and expression of cloned DNA in multiple heterologous hosts. The initial development of the pBELOBAC11 vector (Kim et al. 1996) was instrumental in permitting BAC cloning of environmental DNA (Rondon et al. 2000). However, the inherent restriction of the pBELOBAC11 vector to single-copy within an *E. coli* host cell was a severe limitation for downstream analysis of the resultant BAC libraries. The development of inducible-copy control BAC vectors by inclusion of an RK2 origin of replication enabled more facile BAC library construction (Wild et al. 2002), and different derivatives of these vectors were constructed and commercially available (Lucigen, Middleton, WI; Epicentre, Madison, WI). The inducible-copy control BAC vectors were further modified by including the complete mini-RK2 replicon within the BAC vector to enable shuttling of BAC clones into multiple heterologous hosts (Kakirde et al. 2011), greatly expanding the host range for heterologous expression analysis of BAC libraries.

The size-selected metagenomic DNA fragments are ligated into an appropriate BAC vector using a DNA ligase. DNA fragments prepared by partial restriction digestion can be directly ligated to the chosen BAC vector that has been linearized with the same restriction enzyme. If randomly shared DNA fragments are cloned, however, both ends of each fragment need to be repaired to blunt ends. To increase cloning efficiency, adaptors of an appropriate restriction enzyme can also be ligated to the repaired ends. After ligation, the insert-carrying BAC vector is transformed into highly competent *E. coli* cells typically using electroporation. It should be noted that even if a shuttle vector capable of transfer to other hosts is used, it is advisable to first use *E. coli* for BAC library construction to take advantage of high transformation efficiency, even if the ultimate expression host is not *E. coli*. It should also be noted that this is the most difficult step in BAC cloning and that the larger number of fosmid libraries reported in the literature compared to BAC libraries is merely a reflection of the more facile fosmid cloning. Once transformants are isolated on respective antibiotic selection plates, a representative number of colonies should be evaluated for the percentage of BAC clones with insert the average insert size.

If the library statistics are satisfactory for further analysis, then colonies can be archived. The archiving of a BAC library, which typically consists of a vast number of clones, is an important step since researchers frequently need to access clones for confirmation, verification, screening, and other analyses. BAC clones are usually suspended in a cryoprotectant medium, which is usually 10–15 % glycerol or 8 % dimethyl sulfoxide (DMSO) in the original growth medium of the bacterial host. The BAC clones are usually grown in 96-well or 384-well format for high throughput handling and screening.

Screening of BAC Libraries

Unlike shotgun metagenomic sequencing, BAC libraries provide the opportunity to identify and

access the functional diversity that can be determined phenotypically. The strategies for activity-based screening of BAC libraries depend on the nature of the compounds or enzymes of interest and should be designed carefully (Taupp et al. 2011). The advent of new technologies has enabled the application of high throughput screening (HTS) approaches to identify clones with the desired phenotypes (producing certain compounds or enzymes) from a large number of BAC clones. The analysis of cell lysates, DNA, or supernatants from BAC clones can be performed with a great diversity of screening targets to identify the clones carrying the desired activities or genetic targets (Lakhdari et al. 2010). For example, BAC clones expressing the desired activity can be identified by applying an indicator substrate of the enzymes of interest into the growth medium. Depending on the nature of the assay, the active clones may be detected by visual inspection of an indicator agar plate, flow cytometry, a spectrophotometer, or fluorescent microtiter plate reader (Taupp et al. 2011).

One of the first proof-of-concept studies for identifying a functional natural product from a BAC library was accomplished via sequence-based screening. In the seminal paper of Béjà et al. (2000), the first bacterial proteorhodopsin, a light-driven proton pump, was discovered by identifying specific BAC clones that contained a 16S rRNA gene sequence and then identifying the other linked functional genes contained within the same clone. While this study used a fosmid library, this approach is equally applicable to BAC libraries and has been used to describe some of the functional diversity associated with as-yet-uncultured bacteria (Liles et al. 2008). This approach is inherently limited by the metagenomic DNA sequences that are immediately adjacent to an rRNA operons, but was nonetheless a useful method in the initial exploration of BAC libraries.

Enzymatic activities expressed from BAC clones may be identified via many different methods, including colorimetric or fluorescent assays, as well as indicator media (Taupp et al. 2011). For example, lipase-producing BAC clones can be detected on LB

agar plates supplemented with 1.0 % tributyrin by formation of a halo around individual clones due to tributyrin hydrolysis. Cellulases and xylanases can be detected using agar plates supplemented with carboxymethyl cellulose (CMC) and soluble xylan, respectively. Other enzymatic activities identified using a BAC approach include but are not limited to esterase, alcohol dehydrogenase, amidase, amylase, protease, chitinase, dehydratase, and β -lactamase (Lorenz and Eck 2005).

Identification of secondary metabolites expressed from a heterologous host is dependent upon having large-insert BAC clones, along with suitable transcriptional and translational machinery. The best examples of a functional metagenomic approach to identify antimicrobial activities were the isolation of turbomycin A and B (Gillespie et al. 2002), the identification of antibacterial activities expressed in cosmid libraries in different proteobacterial hosts (Craig et al. 2010), and identification of gene clusters involved in synthesis of antifungal activities (Chung et al. 2008).

In addition to activity-based screening, sequence-based screening is a widely used approach to find genes or gene clusters involved in particular functions within a BAC library. For example, an alternative to functional expression of libraries to identify antibacterial-active clones is to first identify clones that contain known pathways involved in secondary metabolite synthesis and then to express these pathways in a related host, permitting isolation of novel analogs of known metabolites. This has been demonstrated previously in the case of polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Feng et al. first identified the type II PKS biosynthetic system in two different cosmid clones by sequence-based homology screening of a cosmid library (Feng et al. 2010). Their sequence-based screening followed by heterologous expression of a type II PKS biosynthetic gene cluster identified three new fluostatins that were previously uncharacterized in cultured species (Feng et al. 2010). This approach can be equally applicable to screening BAC libraries.

Sequencing and Bioinformatic Analysis of BAC Libraries

The inserts of BAC clones that exhibit certain metabolic activities can be sequenced to determine the coding sequence, structural and regulatory features of the gene(s), and potential phylogenetic markers. Three different common sequencing strategies are typically used.

Subcloning and Sequencing of Individual BAC Clones

The insert of a BAC clone is first fragmented mechanically or enzymatically using a restriction enzyme. The resultant smaller inserts can be cloned into a plasmid vector (e.g., pUC vector) and then sequenced individually using the Sanger sequencing technology (see the “[Subcloning](#)” section below). The full length of the BAC insert can be assembled from the sequenced subclones (see “[Sequence Assembly](#)” section below). Because it is time-consuming to subclone and sequence a large number of BAC clones, this approach is primarily used to sequence one or a few of BAC clones of interest.

End Sequencing

Both ends of a BAC clone insert can be sequenced using the Sanger sequencing technology and the primers that specifically anneal to the vector regions that flank the insert (Pope and Patel 2008). This approach only allows sequencing a short region at both ends of a BAC clone, and thus only limited sequence information can be determined. In contemporary studies, end sequencing is primarily used to match a BAC clone with its corresponding sequence that is determined using shotgun sequencing of pooled BAC clones (see the “[Shotgun Sequencing](#)” section below). The genetic information of each BAC clone can then be analyzed with respect to its phenotypic activities observed during activity screening.

Shotgun Sequencing of Pooled Select BAC Clones

Recent advancement in DNA sequencing technologies, especially the NGS technologies, made

it more cost-effective and efficient to sequence pooled BAC clones of interest in a shotgun manner. The 454 FLX Platinum (Roche) is the most suitable NGS currently available to achieve effective shotgun sequencing of pools of BAC clones because it can generate relatively long sequence reads (average 500 bp). The Illumina systems (Illumina) have also been used even though they produce shorter (150 bp) sequence reads. Ion Torrent (Life Technologies) is a new NGS that can generate 260 bp reads and should be another suitable NGS technology for the aforementioned shotgun sequencing.

Pooled BAC clones are first randomly fragmented to small fragments, with the length of the fragments depending on the specific NGS technology used. Adapters (short oligonucleotides) may need to be ligated to the ends of each fragment to facilitate sequencing. Unique barcodes (short oligonucleotides) can be incorporated into the adaptor for each BAC clone. In that case, sequence reads for individual BAC clones can be separated based on the unique barcodes. However, it is often cost-prohibitive to barcode individual BAC clones when a large number of them are sequenced. Therefore, pooled BAC clones are typically sequenced in a shotgun manner. However, be aware that assembly of complete BAC inserts may be problematic using a pooled BAC clone format depending on the degree of coverage and sequence similarity among the BAC clone DNA inserts. Details on these NGS can be found in respective entries of this encyclopedia. It should be pointed out that except for the 454 FLX Platinum system, the other aforementioned NGS technologies produce short read. Very high (50× or greater) coverage is needed to assemble the individual NGS reads into long contigs and the full-length inserts. Alignment analysis of the end sequences (see [End sequencing](#) above) and the assembled BAC insert sequences can help match individual BAC clones with their respective sequences. To ensure high-quality sequences, BAC clones need to be prepared free of *E. coli* chromosomal DNA.

Sequence Assembly

Individual sequence reads from the same BAC insert are linked together using *de novo* sequence assembly, a bioinformatic process, to form contigs and to reconstruct the BAC insert sequence without reference to any genome sequence. Most software tools used in *de novo* sequence assembly (referred to sequence assemblers) seek overlapping sequences among individual sequence reads and then merge them together based on the overlapping sequences. A number of sequence assemblers are available that use different strategies. Sequence reads generated by subcloning and the Sanger technology are relatively long (500 bp or longer) and only low coverage ($\leq 10\times$) is needed to assemble complete BAC inserts. Such long reads can be assembled by alignment against each other and merged based on overlapping sequences using programs such as Phrap (Gordon 2004) and CAP3 (Huang and Madan 1999). Assembling short sequence reads generated by NGS technologies requires different strategies because it is computing intensive to align large numbers of short sequence reads required to assemble long contigs or full-length BAC inserts. Moreover, high sequencing coverage (at least 50×, depending on the read length) is needed to compensate for the short sequence reads. One alternative approach is to use a graph-based algorithm (e.g., the K-mers for de Bruijn graph) to detect certain short fragments to facilitate assembling short sequence reads. Velvet (Zerbino and Birney 2008) is one bioinformatic program that uses the de Bruijn graph to assemble short sequence reads. Other popular sequence assemblers used to assemble NGS reads include SSAKE, SHARCGS, VCAKE, Euler, SOAPdenovo, ABySS, ALLPATHS (Miller et al. 2010).

Sequence reads generated by the 454 FLX Titanium system on average reach 500 bp in length. Such a read length approaches that of Sanger sequencing reads. Two bioinformatic programs, Newbler (www.454.com) and Arachne (Batzoglou et al. 2002), are designed to assemble sequence reads generated by 454 systems.

Because each of these two sequence assemblers has its own preference and the assembly is somewhat different, contigs generated from each of them can be reassembled using an overlap-based algorithm, such as Minimus of the AMOS package (Sommer et al. 2007), to further improve the accuracy and extend the lengths of the contigs assembled.

Comparative sequence assembly involves alignment of sequence reads against reference genome(s). It is rarely applied to assembling of shotgun BAC sequence reads because few reference genomes are available in most cases. However, as more and more microbial genomes and metagenomes are sequenced in some habitats, such as human gut, comparative sequence assembly may be used to facilitate assembling shotgun BAC sequence reads. The Reference Mapper from Roche (www.roche.com), Eland from Illumina (www.illumina.com), Corona from ABI (www.appliedbiosystems.com), and some shareware bioinformatic programs (e.g., SOAP, MAQ, and segemehl) can be used in comparative sequence assembling (Kunin et al. 2008).

Prediction and Annotation of Open Reading Frame (ORF)

The contigs assembled provide the opportunity to determine the putative genes, their structure and organization, and putative function. Putative genes are defined by open reading frames (ORF) that encode proteins of minimal molecular weight (>50 amino acids). The early bioinformatic programs developed were intended to predict ORFs from sequenced genomes by either recognizing specific genome signals or comparing them to protein or cDNA databases. These specific genome signals are usually species specific. Hence, these bioinformatic programs have limited utility when applied to metagenomic sequence data. In recent years, several programs and database environments have been developed to specifically predict ORFs from metagenomic sequence data.

The commonly used ones include GeneMark.hmm, Metagene annotator (MGA, <http://metagenomics.anl.gov/>), and Orphelia (Yok and Rosen 2011). Orphelia differs from the other two *ab initio* bioinformatic programs in combining a similarity-based algorithm with a composition-based method. However, the specificity and sensitivity of these programs remain to be improved. Furthermore, sequencing and assembly errors significantly affect the accuracy of gene prediction. For instance, GeneMark.hmm is very sensitive to insertion and deletion, which are the main types of sequencing error of the 454 FLX system, thus producing false-positive and false-negative predictions caused by frame shifts. As the logarithms of ORF-finding programs continue to improve, future bioinformatic tools should improve in sensitivity and accuracy in finding ORFs in BAC clone libraries and other metagenomic sequence data.

A predicted ORF can then be annotated by comparing to databases, most of which maintain a publically accessible online server. The tentative function of an ORF is typically inferred using *in silico* bioinformatic analysis by comparing to comprehensive databases, such as GenBank (www.ncbi.nlm.nih.gov/), the EMBL Nucleotide Sequence Database (EMBL-bank, www.ebi.ac.uk/embl/), and the DNA Database of Japan (DDBJ, www.ddbj.nig.ac.jp/). Certain functional and structural features of ORFs can also be identified using specialty databases, such as KEGG (www.genome.jp/kegg/) for prediction of metabolic functions, SignalP (www.cbs.dtu.dk/services/SignalP/) for prediction of presence of signal peptides, and TransMembrane Protein DataBase (pdbtm.enzim.hu/) for presence of transmembrane domains. It should be noted that although annotation of ORFs has significantly improved over the past 10 years owing to the development of software tools and databases and the accumulation of sequenced and annotated genomes and metagenomes, the gene function predicted by *in silico* analysis sometimes does not really represent its actual biological characteristics. Moreover, inaccurate annotations can

be cascaded and amplified in databases. Researchers should always keep in mind such potential discrepancy when annotating ORFs identified in BAC libraries and other metagenomic sequence data.

Phylogenetic Analysis

Unlike BAC libraries constructed from pure cultures, BAC libraries from microbiome samples are constructed from hundreds or even thousands of species of diverse microbes. Thus, one of the primary analyses to be performed on BAC libraries is evaluating the microbiome composition. It is straightforward to infer the taxon from which a BAC sequence is derived if a phylogenetic marker, such as a SSU or LSU rRNA gene, can be found. However, most BAC clones lack such a phylogenetic marker. Alternative methods are available to taxonomically predict the origin of BAC clones based on a number of features, such as sequence composition or homology (Kunin et al. 2008). Composition-based software focuses on the sequence composition signatures, primarily oligonucleotide frequencies to distinguish contigs from each other. Phymm (www.cbcb.umd.edu/software/phymm/) and TETRA (www.megx.net/tetra/) are commonly used bioinformatic programs of this category. On the other hand, homology-based methods predict the taxonomic origin of BAC clones by searching for homologous sequence available in databases. Representative homology-based bioinformatic programs include BLAST (blast.ncbi.nlm.nih.gov/), MEGAN (ab.inf.uni-tuebingen.de/software/megan/), and SIGNATURE (www.cmbi.ru.nl/signature/). Some hybrid classifiers, such as PhymmBL that is a combination of Phymm and BLAST (Brady and Salzberg 2009), are also available that can improve taxonomic assignment accuracy. It should be cautioned that although short sequences can be accurately classified, accurate and reliable prediction requires long reads or contigs. Therefore, precise and accurate taxonomic classification of BAC clones depends on a delicate selection of sequences and assembly strategy.

Biochemical Characterization of Gene Functions of Interest

Subcloning

The proteins or enzymes encoded by a gene or gene cluster of a particular BAC clone can be biochemically characterized following subcloning and overexpression. Subcloning entails cloning the gene(s) of interest from the selected BAC clone into another vector, mostly an expression vector (e.g., a pET vector). The gene(s) can either be excised out of the BAC clone using an appropriate restriction enzyme or amplified using PCR amplification. In the latter case, a pair of primers are needed that anneal to the sequences flanking the target gene(s). However, neither method may not be suitable in some cases, for example, when no appropriate restriction enzyme is available that does not cut within the target gene(s) or when the target gene(s) is too long to be PCR amplified. Recently, a homologous recombination-based subcloning approach, referred to as BAC recombineering (Warming et al. 2005), is used to overcome the aforementioned subcloning obstacles.

Over Expression and Characterization of Expressed Gene Products

The enzyme encoded by the gene of interest can be biochemically characterized if overexpressed and purified. Briefly, a subcloned gene is transformed into an appropriate host, which is typically *E. coli*, or another expression system, depending on the characteristics of the protein to be expressed. It should be noted that some genes in BAC clones might not be expressed successfully because the chosen expression system lacks the suitable transcription or translation systems. Another expression host may be evaluated for its ability to express the gene(s) of interest using a shuttle vector. In addition, the expressed protein may be toxic to the expression host. Recently, fusion proteins and co-expression systems have been incorporated into the expression strategies to overcome some of the limitations mentioned above. A fusion protein can aid in solubilization

and/or purification of the overexpressed proteins. Commonly used fusion proteins include glutathione S-transferase (GST), maltose-binding protein (MBP), and histidine tags.

In some cases, a protein can be expressed but cannot be correctly folded due to the lack of an appropriate chaperon protein in the heterologous host. To overcome this obstacle, a co-expression vector that allows coexistence of multiple expression vectors and expression of a chaperon protein can be used to help the correct folding of heterologous proteins. The Duet vectors (Novagen) are among the systems used to co-express genes of interest from BAC clones. Upon induction, sufficient quantities of the expressed protein can be obtained to determine its characteristics, such as substrate range, product, optimum of temperature and pH, kinetics, and stability.

Prediction of Protein Structure

The three-dimensional (3D) structure of a protein provides invaluable insights into the molecular basis of its functions. Additionally, the detailed knowledge of the spatial arrangement of key amino acid residues within the overall 3D structure also helps design experiments to characterize the protein and understand the molecular mechanisms of functions. The structure of a purified protein can be determined experimentally using X-ray crystallography, high-resolution electron microscopy, or nuclear magnetic resonance (NMR) spectroscopy. Detailed information on these types of characterization is beyond the scope of this entry but can be found in other entries of this encyclopedia. Although experimental methods can help determine the actual 3D structure of a protein, they are expensive, time-consuming, and not always applicable. Thus, alternative bioinformatic programs are often used to predict the structure of a protein from its amino acid sequence. Some of the commonly used programs/servers include I-TASSER (zhanglab.ccmb.med.umich.edu/I-TASSER/), Modeller (salilab.org/modeller/), and Phyre2 (www.sbg.bio.ic.ac.uk/phyre2/).

Functions and Bioactive Compounds Identified by BAC

A variety of enzymes and other bioactive compounds have been identified through BAC libraries. These include xylanases, cellulases, lipases, proteases, amylases, esterases, and type II polyketide synthases. Examples of bioactive compounds discovered from BAC libraries include antibiotics, patellamide D, and ascidiacyclamide. New antibiotic resistance genes have also been found from BAC libraries. Future applications of BAC libraries will probably lead to discovery of novel compounds or enzymes useful to medical or technological purposes.

Summary

Metagenomic studies using BAC clone libraries allow access to metabolic activities and biocatalysts from uncultured microbes. Unlike shotgun deep sequencing, BAC libraries provide a unique opportunity to gain access to metabolic activities and the underpinning enzymes involved in synthesis or biodegradation of many useful compounds and biocatalysts. Functional diversity archived in BAC libraries can also be accessed repeatedly for various studies including detailed characterization of the enzymes and metabolic activities. Furthermore, BAC libraries also enable access and capture of large gene clusters that exceeds the capacity of fosmid vectors. Thus, BAC libraries complement both shotgun deep DNA sequencing and fosmid libraries in metagenomic studies of microbiomes.

Cross-References

- ▶ [A De Novo Metagenomic Assembly Program for Shotgun DNA Reads](#)
- ▶ [Fosmid System](#)
- ▶ [KEGG and GenomeNet, New Developments, Metagenomic Analysis](#)
- ▶ [Phylogenetics, Overview](#)

References

- Andrews-Pfannkoch C, Fadrosch DW, Thorpe J, Williamson SJ. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl Environ Microbiol.* 2010;76:5039–45.
- Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: a whole-genome shotgun assembler. *Genome Res.* 2002;12:177–89.
- Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science.* 2000;289:1902–6.
- Berry AE, Chiocchini C, Selby T, Sosio M, Wellington EMH. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol Lett.* 2003;223:15–20.
- Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods.* 2009;6:673–6.
- Chung EJ, Lim HK, Kim JC, Choi GJ, Park EJ, Lee MH, Chung YR, Lee SW. Forest soil metagenome gene cluster involved in antifungal activity expression in *Escherichia coli*. *Appl Environ Microbiol.* 2008;74:723–30.
- Craig JW, Chang FY, Kim JH, Obiajulu SC, Brady SF. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol.* 2010;76:1633–41.
- Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods.* 2011;86:397–400.
- Feng Z, Kim JH, Brady SF. Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *J Am Chem Soc.* 2010;132:11902–3.
- Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM, Handelsman J. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol.* 2002;68:4301–6.
- Gordon D. Viewing and editing assembled sequences using consed. In: Baxevanis AD, Davison DB, editors. *Current protocols in bioinformatics.* New York: Wiley; 2004. p. 11.12.11–43.
- Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res.* 1999;9:868–77.
- Kakirde KS, Wild J, Godiska R, Mead DA, Wiggins AG, Goodman RM, Szybalski W, Liles MR. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene.* 2011;475:57–62.
- Kim UJ, Birren BW, Slepak T, Mancino V, Boysen C, Kang HL, Simon MI, Shizuya H. Construction and characterization of a human bacterial artificial chromosome library. *Genomics.* 1996;34:213–8.
- Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev.* 2008;72:557–78. Table of Contents.
- Lakhdari O, Cultrone A, Tap J, Gloux K, Bernard F, Ehrlich SD, Lefevre F, Dore J, Blottiere HM. Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-kappaB modulation in the human gut. *PLoS One.* 2010;5.
- Liles MR, Williamson LL, Rodbumrer J, Torsvik V, Goodman RM, Handelsman J. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Appl Environ Microbiol.* 2008;74:3302–5.
- Lorenz P, Eck J. Metagenomics and industrial applications. *Nature reviews. Microbiology.* 2005;3:510–6.
- Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010;95:315–27.
- Piel J. Approaches to capturing and designing biologically active small molecules produced by uncultured microbes. *Annu Rev Microbiol.* 2011;65:431–53.
- Pope PB, Patel BK. Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol Ecol.* 2008;64:9–27.
- Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol.* 2000;66:2541–7.
- Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics.* 2007;8:64.
- Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol.* 1996;178:591–9.
- Taupp M, Mewis K, Hallam SJ. The art and design of functional metagenomic screens. *Curr Opin Biotechnol.* 2011;22:465–72.
- Warming S, Costantino N, Court DL, Jenkins NA, Copeland NG. Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Res.* 2005;33:e36.
- Wild J, Hradecna Z, Szybalski W. Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res.* 2002;12:1434–44.
- Yok NG, Rosen GL. Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics.* 2011;12:20.
- Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18:821–9.

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution

Thomas W. Schoenfeld and David Mead
Lucigen Corporation, Middleton, WI, USA

Introduction

High-temperature subterrestrial aquifers are vast ecosystems fueled solely by chemical reducing potential rather than solar radiation as is the case for surface life (Fournier 2005). The volume of the global thermal aquifer has been estimated as high as 10^{19} L (Gold 1992), with microbial and viral abundances approaching those of the oceans (Breitbart et al. 2004b). This study, previously reported in Pride and Schoenfeld (2008), Schoenfeld et al. (2008), and Heidelberg et al. (2009), examined planktonic viruses directly isolated from two mildly alkaline siliceous hot springs in Yellowstone National Park (YNP). With temperatures of 74 °C and 93 °C, life in these springs is comprised exclusively of bacterial and archaeal cells and viruses, all uniquely adapted to the temperature and chemistry extremes of the environment (Reysenbach et al. 2002). The springs in these water-driven systems are direct outflows of the thermal aquifer and not secondarily heated surface water, as is the case for vapor-driven systems (Fournier 2005). In this respect they are distinct from acidic springs, mud pots, and other thermal features that have provided many of the published thermophilic virus samples. Because the springs are direct outflows of the aquifers, conceivably, viruses in these springs may proliferate not only at the surface but deeper in the vent as well, where increased pressures and temperatures as high as 180–270 °C are found at depths of 100–550 m throughout the caldera of YNP (Fournier 2005). If viruses proliferate in the subsurface aquifer, hot springs separated by kilometer distances that share common water sources may also share viral populations.

The roles of viruses in the ecology of hydrothermal environments have not been studied in detail, although they appear to play a role in host mortality and carbon cycling (Breitbart et al. 2004b) and are probably the only predators. In better studied marine environments, an estimated 10^{30} viruses in the world's oceans (Suttle 2007) may comprise several hundred thousand different types (Angly et al. 2006) and are responsible for a significant proportion of microbial mortality and thus have a profound influence on carbon and other nutrient cycles (Suttle 2007). Viruses also may be important vehicles for lateral gene transfer via lysogeny and transduction and probably promote diversity by preferentially lysing the most abundant species (Weinbauer and Rassoulzadegan 2004). Analyses of viral metagenomes (Cann et al. 2005; Angly 2006; Bench et al. 2007) and cultured viral genomes (Pedulla et al. 2003; Kwan et al. 2005) have consistently shown that a minority of these sequences have detectable similarity to sequences in GenBank and very few are similar to known viruses. In spite of extensive sequencing of marine virus metapopulations, only a few small RNA genomes of 5–10 kb have been assembled (Culley et al. 2007), presumably due to the extreme viral diversity that confounds the assembly of viral genomes (see Chaps. 2–10, Vol. I).

Enrichment cultivation has provided most of the knowledge of thermophilic viruses (defined here as those growing at >70 °C). Since the first reports of thermophilic viruses (Sakaki and Oshima 1975; Martin et al. 1984), hundreds of bacteriophages (Yu et al. 2006), dozens of crenarchaeal viruses (reviewed in Snyder et al. 2003; Prangishvili and Garrett 2005), and one euryarchaeal virus (Geslin et al. 2003) have been isolated from thermal springs and vents around the world. Cultivated *Thermus* bacteriophages belong to four morphological families: *Myoviridae*, *Siphoviridae*, *Tectiviridae*, and *Inoviridae* (Yu 2006). Their morphologies and the available genomic sequences (Naryshkina et al. 2006) suggest similarity to mesophilic bacteriophages. Most known thermophilic bacteriophages appear to be lytic, although this could be

biased by the method of their discovery (Yu 2006). Cultivated thermophilic crenarchaeal viruses infect the genera *Sulfolobus*, *Acidianus*, *Pyrobaculum*, and *Thermoproteus*. Morphologies and genome content suggest crenarchaeal viruses are unrelated to viruses of *Euryarchaeota*, *Bacteria*, or *Eukarya* (Prangishvili et al. 2006a). All of the cultivated crenarchaeal viruses proliferate as chronic, nonlytic infections.

While enrichment cultures have been highly informative in the study of thermophilic viruses, important contextual information such as relative abundance, diversity, and distribution is lost. Furthermore, these analyses exclude the majority of viruses that are not readily cultivated (Snyder et al. 2004). No viral cultivation study fully replicates the temperature and pressure extremes and the chemistries that characterize the subsurface vents, which limits cultivation of not only viruses but hosts, as well. Unlike cellular life, no universal genetic marker (e.g., rDNA) exists for viruses. Direct metagenomic analysis of viruses from environmental samples circumvents these limitations and provides insight into biology, evolution, and adaptations to the environment and composition of viral assemblages through studies of their genomic sequences. No metagenomic analysis of waterborne viral populations in geothermal environments has been reported. In fact, planktonic life in thermal environments is under-explored in general, with microbial diversity studies of hot spring environments focused almost exclusively on sediments (Barns et al. 1994; Hugenholtz et al. 1998; Blank et al. 2002), adherent filaments (Reysenbach et al. 1994), or mats (Ward et al. 1998). The goal of this study was to profile the diversity, composition, and adaptations of viral assemblages in two hot springs of YNP based on metagenomic analysis of viruses inhabiting these environments.

Materials and Methods

Site Description and Sampling

Viral particles were isolated from Bear Paw (an unofficial name for LRNN374)

(N 44.5560955 W110.8347866) and Octopus (N44.5340836 W110.7978895) hot springs (Stoner et al. 2001). The temperatures of the hot springs are based on direct measurement on the day of the sampling. The pH values were determined by the USGS (McCleskey et al. 2004). Thermal water (400–600 L) was filtered using a 100 kD MWCO tangential flow filter (GE Healthcare). Viral particles were concentrated to 2 L, filtered through a 0.2 μm filter and further concentrated to 100 ml using a 100 kD filter. Viral concentrates were imaged by transmission electron microscopy (TEM) (Leo 912AB operating at 80KV). Direct viral enumeration was performed by epifluorescence microscopy (Noble and Fuhrman 1998). As recommended (Wen et al. 2004), samples were unfixed and were stained with SYBR Gold. The samples were stored at 4 °C for no more than 24 h before counting. Immediate freezing of samples in liquid nitrogen was not possible, so viral abundances may be somewhat underestimated.

Viral DNA Processing and Extraction

Viral concentrates were centrifuged at 12 K rpm for 20 min, syringe-filtered using a 0.2 μm Acrodisc filter (Gelman), and further concentrated to 400 μl by filtration using a 30 kD MWCO Centricon spin filter (Millipore). Those judged by epifluorescence microscopy to be substantially free of microbial cells were used for library construction. Viral concentrates were transferred to SM buffer (0.1 M NaCl, 8 mM MgSO_4 , 50 mM Tris-HCl pH 7.5) using a 30 kD MWCO spin filter. Benzonase endonuclease (Sigma, 10 U) was added, and the reactions were incubated for 30 min. at 23 °C. EDTA (20 mM), SDS (0.5 %), and Proteinase K (100 U) were added, and the reactions were incubated for 3 h at 56 °C. NaCl (0.7 M) and CTAB (1 %) were added, and DNA was extracted with phenol/chloroform and ethanol precipitated.

Library Construction and Sequencing

Viral DNA was physically sheared to 3–6 kb using a HydroShear device (Genomic Solutions, MI). The ends were made blunt using the DNATerminator end repair kit (Lucigen, WI),

and the fragments were ligated to a double-stranded asymmetrical linker comprised of one phosphorylated blunt end (5'-GATGCGGCCGCTTGTATCTGATACTGCT-3', Linker 1) and one non-phosphorylated staggered end (5'-GGAGCAGTATCAGATA CAAGCGGCCGCATC-3', Linker 2) to fix the primer in a defined orientation relative to the genomic DNA. Gel fractionation was used to remove unligated linkers and to isolate 3–6 kb fragments. These fragments were PCR amplified using Vent DNA polymerase (New England Biolabs, MA) and a primer targeted to Linker 1 (5'-AGCAGTATCAGATACAAGCGGCCGCA TC-3'). Amplification products were gel purified again, inserted into the cloning site of the transcription-free pSMART vector (Lucigen), and used to transform *E. coli* 10G cells (Lucigen). Libraries were sequenced by the Department of Energy's Joint Genome Institute (Walnut Creek, CA). The sequences were deposited in the GenBank trace archive and are retrievable using CENTER_NAME = "JGI" and SEQ_LIB_ID = "AOIX" for Bear Paw sequences and SEQ_LIB_ID = "APNO" and SEQ_LIB_ID = "ATYB" for octopus sequences.

Bioinformatics

Viral metagenome sequencing reads were compared to the nonredundant (nr) protein database (GenBank) using BLASTx (Altschul et al. 1997). The 50 most significant BLASTx scores ($E < 10^{-3}$) were recorded. The first occurrences of keywords in the output of the BLASTx were counted using PERL scripts written for this project, and the sequences were categorized by function. Sequences were assembled using the SeqMan[®] program (DNASTAR, WI) at a minimum of 50 % or 95 % identity over a minimum of 20 nt. Metagenome sequence libraries were compared to each other and to all the sequences in GenBank using tBLASTx (NCBI) with a cutoff of $E < 10^{-3}$. Where indicated, the apparent open reading frames were identified and translated using the Gene Mark program (Lukashin and Borodovsky 1998). These translations were compared to the nr protein database

using the BLASTp program. The rank abundances were calculated using the PHAge Community from Contig Spectrum (PHACCS) web utility located at <http://phage.sdsu.edu/research/tools/phaccs/> (Angly et al. 2005) based on an average genome length of 50 kb.

Results and Discussion

Sampling Sites, Viral Abundance, and Morphologies

The two hot springs that provided samples are listed in Table 1. Bear Paw hot spring is in the river group of the lower geyser basin of YNP, while Octopus is about 5 km away in the White Creek area. Although the pH values of these hot springs are both circumneutral, the temperatures and apparent microflora differ widely. Bear Paw is significantly cooler and is characterized by orange sedentary microbial growth in the pool. Octopus water emerges at the boiling point at the local elevation of 2,300 m, with none of the orange growth. Octopus hot spring is well documented to support prolific microbial life (Brock and Brock 1968), and its geochemistry (McCleskey 2004) is suitable for chemotrophic metabolism. Reported analyses based on rDNA sequences from filaments and sediments (Reysenbach et al. 1994; Blank et al. 2002) show that microbial diversity is relatively limited compared to moderate-temperature environments. These studies and others comparing lipid and isotope composition (Jahnke et al. 2001) suggest the microbes in the filaments and the sediments, close in proximity and temperature to the sample site in this study, are primarily *Bacteria*, with *Aquificales* and *Thermotogales* most highly represented. No detailed study of the planktonic life from Octopus or the chemical composition or life in Bear Paw has been published.

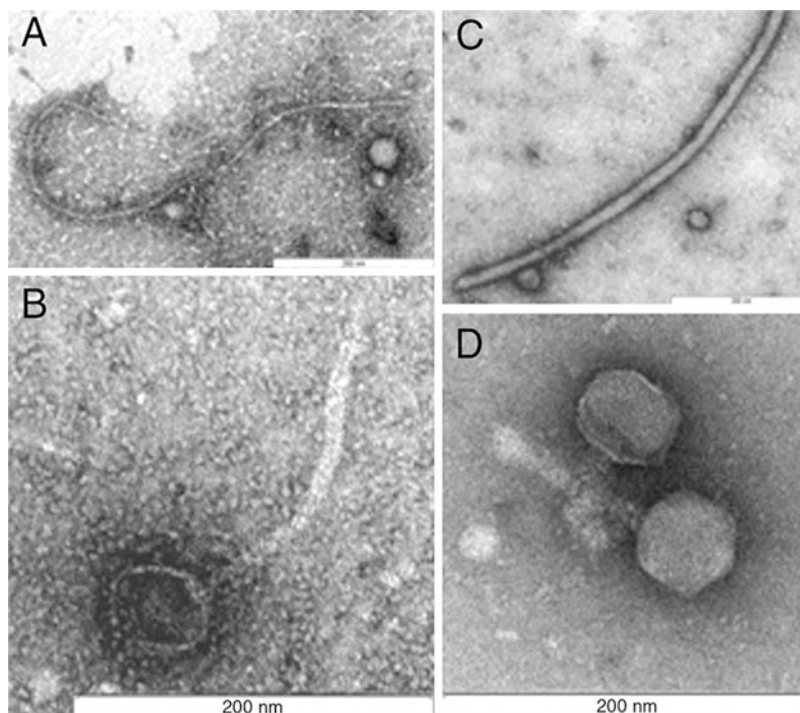
Virus-enriched fractions were isolated from 400 to 600 L of hot spring water for library construction and sequence analysis. Viral abundances (Table 1) were at the lower end of the range of 10^4 – 10^9 reported for thermal springs in Long Valley, California (Breitbart et al. 2004b), and moderate-temperature aquatic environments

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 1 Sample sites and abundance of viral and microbial counts

Hot spring	Temp	pH	Cells/mL	Viruses/mL	Virus: microbe ratio	Virus/mL in concentrate	Virus/mL theoretical ^a	Efficiency
Bear paw	74	7.34	4.3×10^6	1.44×10^6	0.33	1.48×10^8	7.21×10^9	2.1 %
Octopus	93	8.14	9.0×10^5	3.07×10^5	0.34	2.18×10^8	1.53×10^9	14.2 %

^aBased on a concentration factor of $5,000 \times$ (500 L to 100 mL)

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Fig. 1 TEM images of viruslike particles directly isolated from YNP hot springs. Images from Bear Paw (Panels **a** and **b**) and Octopus (Panels **c** and **d**) hot springs are shown. The bar in each figure is 200 nm (Images are courtesy of Sue Brumfield and Mark Young, Montana State University. Reproduced with permission from Schoenfeld et al. (2008))



(Wommack and Colwell 2000). The virus/microbe ratios (VMRs) in the hot springs were much lower than in moderate-temperature environments (typically 3–10). These low VMRs may be related to the observation that none of the cultured thermophilic crenarchaeal viruses proliferate via lytic infections, a lifestyle that would result in large burst sizes at the same time as the microbial population is reduced. Actual yields of viruses were significantly below theoretical yields (Table 1) for both two hot springs. It is not known if this loss was systematic and, therefore, biased the metagenomic analysis. Tailed, rod-shaped, and filamentous morphologies were observed in the concentrates (Fig. 1).

Morphologies of viral particles in the concentrates represent most morphological families of known thermophilic viruses. Tailed morphologies are commonly associated with bacteriophages and euryarchaeotal viruses (Geslin et al. 2003; Yu et al. 2006); rod-shaped and filamentous morphologies are more commonly associated with crenarchaeal viruses (Prangishvili and Garrett 2004).

Library Construction and Sequencing

Advances in sequencing capacity make analyses of large numbers of clones feasible; however,

challenges in sampling and library construction have prevented the widespread use of metagenomic shotgun sequencing for studying viral populations. At around 50 ag of DNA per virus, abundances of 10^5 – 10^6 viruses per ml correspond to 5–50 ng of viral DNA per liter. In practice, processing of hundreds of liters of spring water generally yielded no more than 100 ng of DNA, much lower than is normally required for library construction. This low yield of virus precluded cesium chloride purification of the viral particles, as is commonly used for marine viral metagenomic library construction. Viral DNA also contains cytotoxic genes and modified nucleotides that induce host restriction systems. A linker-dependent, anonymous method of DNA amplification was used to access this diversity, allowing construction of 3–8 kb insert libraries with none of the potential modified nucleotides. This library construction method has been used in the analysis of several cultivated and uncultivated viral genomes (Breitbart et al. 2003, 2004a; Seguritan et al. 2003; Lindell et al. 2004; Paul et al. 2005; Bench et al. 2007) but never fully described. Viral DNA was physically sheared, and short (20 bp) linkers were ligated to the DNA fragments to serve as priming sites for PCR. Amplified fragments were cloned into a transcription-free pSMART vector to minimize cloning bias due to cytotoxic sequences (Godiska et al. 2005). The use of flanking synthetic linkers provides identical primer annealing sites for each viral template in the mixture, which significantly limits amplification bias. A noteworthy characteristic of this approach is that it selects exclusively for dsDNA viruses. All cultivated thermophilic bacteriophage and archaeal viruses have dsDNA genomes except certain *Thermus*-specific *Inoviruses*, which have ssDNA genomes (Yu et al. 2006). Notably, several viral nucleic acid preparations from these and other springs sampled as part of this study had RNase-digestible material (data not shown), suggesting that RNA viruses inhabit these hot spring environments.

A total of 28,883 Sanger sequence reads were determined from Bear Paw (7,685 reads) and

Octopus (21,198 reads) hot springs. Paired-end reads averaged 981 nucleotides each or nearly 30 Mb total. Assuming an average genome size of 50 kb, which is supported by agarose gel electrophoresis of the viral genomic DNA (data not shown), this sequencing depth represents about 600 viral genomic equivalents. The quality of the libraries is highly dependent on the amount of DNA used in their construction. The sequence reads of the Octopus library contained very few anomalies that would suggest amplification bias or cloning artifacts. Some of the reads from the Bear Paw library were less random than the Octopus library, as demonstrated by several cases of sequence stacking.

Contaminating cellular DNA in viral DNA preparations was greatly reduced by filtration and nuclease treatment. Only viral preparations substantially free of microbial cells as judged by epifluorescence microscopy were used for library construction. Detection of rDNA sequences (5S, 16S, and 23S) in the libraries was used to identify contaminating cellular DNA. These sequences are absent in known viral genomes but highly conserved in microbial cells. A typical bacterial genome contains 15 rRNA genes (Coenye 2003). Most hyperthermophilic archaeal and bacterial genomes contain three to six rRNA genes, although the genomes of thermophilic *Geobacillus* that grow in the temperature range of Bear Paw contain up to 30 rRNA genes (Feng et al. 2007). BLASTn analysis identified only four rDNA sequences in the 10.4 microbial genome equivalents sequenced from the Octopus library (two 23S and two 16S) and eight in the 3.8 microbial genome equivalents from the Bear Paw library, suggesting viral enrichment was quite high, particularly for the Octopus library. This inference is supported by a high similarity to sequences of cultivated viruses (shown below) and a large number of BLASTx similarities to genes associated with viral functions. In particular, the hundreds of presumptive genes for viral functions, such as replication, transcription, translation, lysogeny, recombination, lysis, and structural proteins (Table 2), are consistent only with a predominately viral origin of the sequences.

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 2 Functional grouping of predicted genes in the viral metagenomes

	Bear paw	Octopus	Bear paw	Octopus
Total reads	7,685	21,198		
No BLASTx similarity	2,545	8,469		
COGs functional category	Number of reads matching a keyword		Percent with a keyword match	
F. Nucleotide transport and metabolism	1,445	2,130	35.09 %	37.81 %
J. Translation, ribosomal structure, and biogenesis	221	336	5.37 %	5.96 %
K. Transcription	278	325	6.75 %	5.77 %
L. Replication, recombination and repair	688	989	16.71 %	17.55 %
O. Posttranslational modification, protein turnover, chaperones	181	213	4.40 %	3.78 %
None virus specific	350	596	8.50 %	10.58 %
No match to a keyword	955	1,045	23.19 %	18.55 %

Identification of Likely Gene Products and Viral Lifestyles

BLASTx analysis of the individual reads was used to identify coding sequences in the libraries. While most reads revealed no significant similarity to known proteins (i.e., no BLASTx similarity; Table 2), a significant portion of the sequences could be assigned an apparent function based on BLASTx analysis. The majority of these predicted functions fall into five of the 23 NCBI Clusters of Orthologous Groups (COG) functional categories (Tatusov et al. 1997) or are virus-specific functions that have no assigned COG function, e.g., lysin, packaging, capsid, tail, or tape measure protein (Table 2). The five COG categories are all nucleic acid metabolism-, information processing-, and translation-related functions, which are commonly associated with phages and viruses.

Certain similarities were particularly informative. The 532 lysin-like genes among 600 viral equivalents suggest lytic viruses are quite common in the hot springs, in contrast to the cultured thermophilic crenarchaeal viruses, all of which are nonlytic. Although lysin genes were highly abundant and are typically proximal to holin genes, no homologs for holins were seen, probably reflecting the high molecular diversity observed in known holin genes (Young 1992). The 86 apparent integrase genes imply that

lysogeny is also common in thermal aquifers, consistent with previous studies that show integrase homologs in six crenarchaeal viral genomes (ATV, STSV1, and four SSV isolates) (Wiedenheft et al. 2004; Xiang et al. 2005; Prangishvili et al. 2006b), and induction of prophage by mitomycin C in 1–9 % of hot spring microbial cells (Breitbart et al. 2004b).

Viruses and Lateral Gene Transfer in Thermal Environments

Viruses have been implicated in lateral gene transfer and nonorthologous gene replacement in cellular genomes (Villarreal and DeFilippis 2000; Daubin and Ochman 2004). Viruses also may have played critical roles in the evolution of DNA as a genetic material, DNA replication mechanisms, the separation of the three domains of life, and the origin of the eukaryotic nucleus, reviewed in Forterre (2006). Gene similarities seen in the metagenomic libraries support the role of viruses in cellular evolution. The 13 apparent reverse transcriptases were almost exclusively related to the intron-associated maturase/reverse transcriptases and retrotransposon reverse transcriptases. These genes and the recombinase, integrase, and transposase genes represent 5.1 % and 3.4 % of the identifiable reads in the Bear Paw and Octopus libraries,

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 3 Sources of superfamily II helicase similarities to Octopus contig 158 and strength of similarity by BLASTx

Source of similarity	Domain	E-value
<i>Staphylococcus</i> phage Twort	Bacteriophage	2E-16
<i>Myxococcus xanthus</i>	Bacteria	1E-15
<i>Sulfolobus islandicus</i> filamentous virus	Archaeal virus	8E-15
<i>Lactobacillus plantarum</i> bacteriophage	Bacteriophage	3E-14
<i>Pyrococcus abyssi</i>	Archaea	4E-08
<i>Sulfolobus solfataricus</i>	Archaea	1.E-06
<i>Eremothecium gossypii</i> (a fungus)	Eukarya	9.E-05
<i>Tribolium castaneum</i> (an insect)	Eukarya	4.E-04
<i>Homo sapiens</i>	Eukarya	6.E-03

respectively, suggesting that the appropriate machinery for lateral gene transfer exists in hot spring viral genomes (Canchaya et al. 2003).

Other sequence similarities provide evidence of ongoing gene transfer within these populations. Helicase genes shared among viruses and cells from all domains have been considered examples of nonorthologous replacement of cellular genes by viral genes (Filee et al. 2003). Hundreds of reads showed sequence similarity to the superfamily II helicases of a wide range of cells and viruses. For example, the 2 kb Octopus contig 158 had significant similarity to helicases of bacterial, archaeal, and eukaryotic cells as well as to phage and archaeal viruses (Table 3).

Also common in the metagenomic libraries are presumptive ribonucleotide reductases (14 and 50 in Bear Paw and Octopus springs, respectively) and thymidylate synthase (seven and 51, respectively) genes. The conservation of these genes between viral and cellular genomes of all domains and the biochemical activities of the gene products imply that viral genes played a key role in the transition from RNA-based to DNA-based genomes (Forterre 2005). DNA polymerase (*pol*) genes have also been proposed

as likely examples of nonorthologous replacement by viral genes (Filee et al. 2002). 156 *pol* gene homologs were identified in the two metagenomic libraries, with all the polymerase families represented. In contrast, only one *pol* gene has been identified by BLASTx analysis of the known crenarchaeal viral genomes (ABV), and three *pol* genes are found in thermophilic bacteriophage genomes (Hjörleifsdottir et al. 2002); Naryshkina 2006). The high abundance of both *pol* and *lys* genes in the metagenomic libraries compared to cultured genomes suggests that the current view of diversity may be biased by the difficulty in culturing certain types of viruses.

Sequence Assembly and Estimation of Viral Diversity

The degree to which metagenomic reads assemble has been used to assess the diversity of the viral populations. Previous studies have used >95 % identity over 20 nucleotides as the assembly stringency (Breitbart et al. 2002, 2004a; Breitbart 2003; Angly et al. 2006). Using this criteria, the power law rank-abundance model built into the Phages Communities from Contig Spectrum tool (PHACCS, 5) predicted 1,400 and 1,310 viral types in Bear Paw and Octopus hot springs, respectively, with no one viral type representing more than about 2 % of the population (Table 4). For reference, 1,650, 3,350, 7,180, 7,340, and 2,390 viral genotypes were reported in estuarine, nearshore marine, open ocean, marine sediments, and fecal viral assemblages, respectively (Breitbart et al. 2002, 2003, 2004a; Angly et al. 2006; Bench et al. 2007), with no single viral species representing more than 2–3 % in any case.

There are several limitations in assessing actual numbers of viral species from metagenomic libraries. First, these models assume viral genomes evolve uniformly. However, different regions of viral genomes are clearly more conserved than others (Lindell et al. 2004). Genetic diversity outside the

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 4 Sequence assembly data and estimation of viral diversity

	Bear paw	Octopus	Totals	
Sequence reads	7,685	21,198	28,883	
	Bear paw 95 %	Octopus 95 %	Bear paw 50 %	Octopus 50 %
Contigs assembled	6,191	13,543	4,850	4,788
Avg. reads per contig	1.239	3.129	1.587	4.427
Largest contig (nt)	3,503	4,554	8,007	35,089
Power law richness	1,440	1,310	548	283
Evenness score	0.946	0.954	0.933	0.936
Most abundant virus	2.14 %	1.88 %	3.93 %	4.88 %
Shannon-Wiener score	6.88	6.85	5.88	5.29

conserved regions is probably higher than these models indicate. Second, the generation of new viral species by mosaicism, modular evolution, or lateral gene transfer (Villarreal and DeFilippis 2000; Canchaya et al. 2003; Weinbauer and Rassoulzadegan 2004) would not be detected using assembly of <1 kb sequence reads. On the other hand, given the dynamic nature of viral genomes, this approach is well suited to a view of the diversity and evolution of viruses that considers genes or groups of genes rather than whole genomes. Finally, assembly at >95 % nucleotide identity fails to account for molecular diversity among related viral types, which is higher than that of cellular species. In fact such stringency would fail to associate viruses that, based on classical criteria (host range, morphology, replication lineages, and physicochemical and antigenic properties), are considered to be related (Lucchini and Brussow 1999; Hatfull et al. 2006; Kwan et al. 2006) although they may share as little as 50 % nucleotide identity over much of their genomes.

Lower Stringency Assemblies Reveal Population Heterogeneity

To accommodate genomic heterogeneity inherent to viral populations, sequences were also assembled at 50 % identity (Table 4). As expected, the numbers of viral types decreased to 548 and 283 in Bear Paw and Octopus,

respectively. These lower stringency assemblies proved quite useful for associating sequences of related, but not identical, viral types and for studying diversity among these related viruses. At 95 % identity, the largest contigs were 3.5 and 4.6 kb for Bear Paw and Octopus, respectively (Table 4). At 50 % identity, Octopus reads assembled into 17 contigs of greater than 10 kb, including contigs of 35 kb and 19 kb, comprised of >1,000 reads each. In each case, reads were evenly distributed across the contigs. The 17 > 10 kb contigs comprise a total of 7.04 Mbp (33 % of total metagenomic sequence) or about 140 viral equivalents. The four strongest BLASTx hits to the 35 kb contig belonged to thermophilic crenarchaeal viruses *Acidianus Rod-shaped virus* (ARV), *Sulfolobus islandicus rod-shaped viruses* 1 (SIRV1) and 2 (SIRV2), and *Sulfolobus islandicus filamentous viruses* (SIFV) (Table 5). The only significant similarity for the 19 kb contig was to the thermophilic crenarchaeal virus, *Pyrobaculum spherical virus* (PSV). In the Bear Paw library, with roughly one third as many reads, the largest contig that assembled at 50 % identity was 8 kb. Five hundred thirty four (7 %) of the reads assembled into 19 contigs >4 kb. These include 0.5 Mbp of reads or ten viral equivalents.

The larger composite contigs allow associations that were impossible at standard stringency. More than 200 million bases have been sequenced from marine viral metagenomic libraries, but only one small phage genome has been

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 5 Numbers of 95 % contigs with tBLASTx similarities ($E < 0.001$) to the respective cellular genomes

	Bear Paw	Octopus
<i>Pyrobaculum</i>	124	684
Archaea		
<i>Aeropyrum</i>	62	626
<i>Sulfolobus</i>	38	326
<i>Acidianus</i>	25	185
Bacteria		
<i>Aquifex</i>	474	1,138

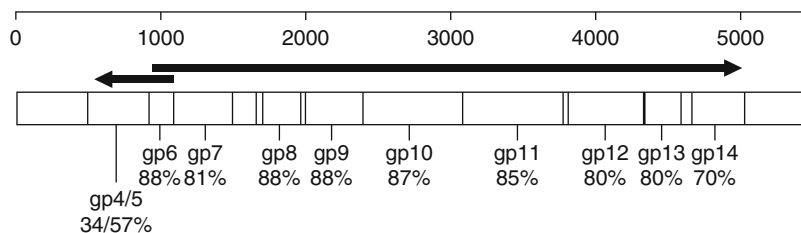
reconstructed (Angly et al. 2006). To validate the low-stringency assemblies and to further study the molecular biology of the viruses, the 4 kb cognates of one contig of four reads that assembled at 50 % NAID were PCR amplified, cloned, and sequenced (Schoenfeld 2014). This confirms that at least this assembly accurately reflects the virome sequence. Furthermore, this contig includes an apparent replisome, and amplification based on the low-stringency assembly allows study of an operon that, due to its size, could not otherwise be recovered from the fragmentary metagenomic data.

Certain contigs provide compelling evidence that the 50 % assemblies associate genuine orthologous sequences. An example is Bear Paw contig 327 (Fig. 2). Eleven open reading frames (ORFs) were identified by the GeneMark algorithm (Lukashin 1998). BLASTp analysis of each shows strongest similarity to the putative coding sequences of PSV (Haring et al. 2004). Nucleotide identities were as high as 88 %, gene order is perfectly preserved relative to the cultured virus, and gene overlap is identical between the composite contig and the cultivated virus. Interestingly, two different ORFs of the PSV genome, gp 4 and 5, are apparently related to each other, since both had significant similarity to the same region of the consensus contig. In both the cultured viral genome and the consensus contig, the gp7 PSV gene overlaps gp6 in the opposite orientation.

Contig 722 from the Octopus spring library provided a unique opportunity to associate

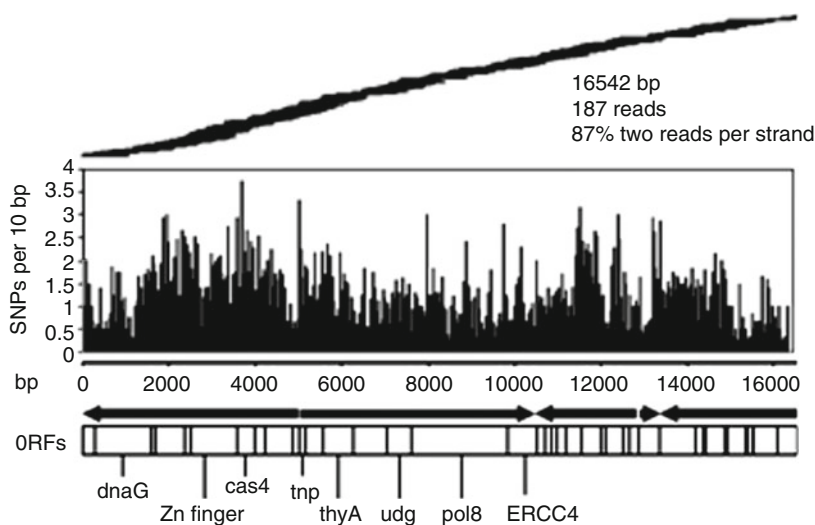
population diversity of an assembled metagenome with the biochemistry of the gene products (Fig. 3). This 16.5 kb contig, assembled at 50 % identity, includes 187 reads (average coverage of 11 reads per nucleotide position). GeneMark predicted 26 ORFs of greater than 100 nucleotides, including an apparent replication operon. The genes with the strongest similarity to four of these ORFs encode primase, uracil DNA glycosylase, family B DNA polymerase, and nucleotide excision repair nuclease (*dnaG*, *udg*, *polB*, and ERCC4 genes, respectively). Homologs of these ORFs belong to crenarchaeal DNA replication/repair complexes (Roberts and White 2003; Dionne and Bell 2005; Barry and Bell 2006). The predicted *polB* gene showed 28 % identity to *Pyrobaculum islandicus* polB2 (Kahler and Antranikian 2000) and has an archaeal codon profile (data not shown). Sequences from three of the discreet clones that comprise the *polB* gene in this contig have been expressed in *E. coli* to produce a functional thermostable DNA polymerase (data not shown). This contig also contains apparent homologs to a zinc fingerlike protein and a transposon-like integrase/resolvase (*tnp*), functions commonly associated with viruses and phages. Another ORF with highest similarity to the CRISPR-associated sequence *cas4* (Haft et al. 2005) is unlikely to be part of a functional CRISPR system. Unlike authentic Cas sequences, this one is virus-derived and is not proximal to a CRISPR sequence or other typically associated sequences. More likely this gene is a separate member of the Cas4 COG, presumably a RecB-like exonuclease (Haft et al. 2005).

To correlate the level of sequence divergence with predicted gene function, SNP frequency was aligned to the 50 % assembly consensus sequence of the contig. Overall distribution of SNPs in the contig was 0.705 per 10 bp. Replication-associated genes showed noticeably lower molecular diversity than the other ORFs. SNP distribution in the *dnaG*, *udg*, *polB*, and ERCC homologs was 0.565, 0.617, 0.569, and 0.548 per 10 bp, respectively, while the distribution in the Zn finger, *cas4*, and *thyA* homologs was 0.979, 1.31, and 0.728, respectively.



Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Fig. 2 Genes and gene order are highly conserved between a cultured crenarchaeal virus and a consensus contig from the Bear Paw library. Contig 372 (5,492 bp, 71 reads) was assembled at 50 % identity from the Bear Paw library. Open reading frames identified by GeneMark algorithm were compared by BLASTp to

proteins in GenBank. Similarities to *Pyrobaculum spherical virus* proteins are shown with percent coding identity. The gene names are based on the annotation in GenBank and are named in order of their location on the viral chromosome. Direction of transcription is indicated by the arrows (Reproduced with permission from Schoenfeld et al. (2008))



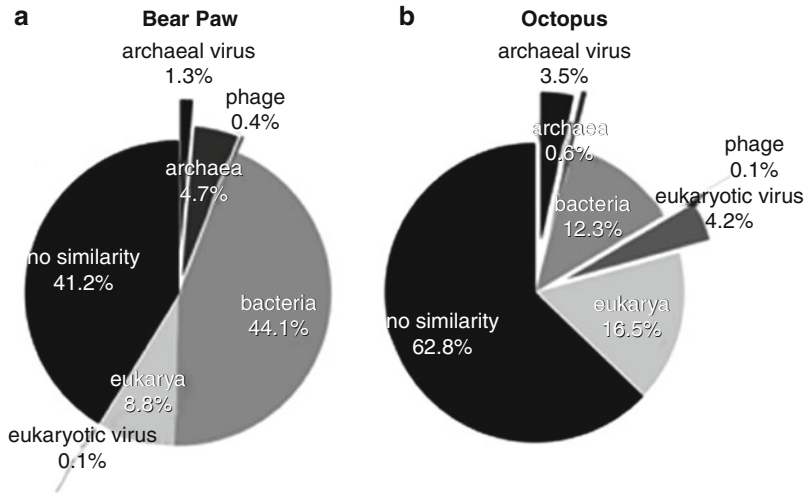
Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Fig. 3 Alignment of nucleotide polymorphisms with coding sequences in a 16.5 kb consensus contig from Octopus hot spring. Contig 722 was assembled at ≥ 50 % identity from the Octopus library. Sequence coverage is shown on the top, with each line representing a separate read. Single-nucleotide polymorphisms per ten

base pairs were normalized to the number of reads covering the respective nucleotide (middle) and are aligned with predicted open reading frames from the consensus sequence in the contig and the gene name of the strongest BLASTx similarity (bottom). Direction of transcription is shown by the arrows. Similarities to known genes were identified by BLASTp (Reproduced with permission from Schoenfeld et al. (2008))

Similarities to Known Viral and Microbial Genomes Imply Phylogeny

tBLASTx analysis was used to infer phylogenetic origin of the 95 % assembled contig sequences. A majority of the contigs (41 % from Bear Paw

and 63 % from Octopus) had no tBLASTx similarity ($E < 0.001$) to any sequence in GenBank (Fig. 4). Although it is typical for viral metagenomic libraries analyzed in this way to have a high proportion of sequences without identifiable homologs, these libraries contained



Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Fig. 4 Broad classification of viral metagenomic contigs based on tBLASTx similarities. Contigs assembled at 95 % identity from Bear Paw and Octopus reads (Panel a and b, respectively) were

compared to sequences in GenBank to infer phylogeny. Shown are frequencies of contigs with no significant sequence similarity in GenBank ($E < 0.001$) and those with sequence similarity to *Bacteria*, *Archaea*, *Eukarya*, and their respective viruses (Reproduced with permission from Schoenfeld et al. (2008))

the highest frequency of novel sequence reported to date using long read Sanger chemistry. This trend likely reflects the lack of sequence data from microorganisms in high-temperature environments as well as high diversity.

Interestingly, the libraries contained a sizable number of sequences with homology to eukaryotic genes, 16.5 % for Octopus Spring and 8.3 % for Bear Paw, which may reflect the commonly observed overlap in gene sequence homology between *Archaea* and *Eukarya* (Brown and Doolittle 1997). Almost all known crenarchaeal viruses were cultivated on three archaeal genera, *Pyrobaculum*, *Sulfolobus*, and *Acidianus*. Interestingly, these genera were three of the four most common archaeal sources of the sequence similarities to the two libraries, the other being *Aeropyrum* (Table 5). Genetic similarities to *Sulfolobus* and *Acidianus* are surprising because these two genera have been found exclusively in highly acidic environments. Nearly half the bacterial similarities were to *Aquifex*. Apparently no attempts have been made to cultivate phage on any strain in the *Aquificales* order.

Genome Signature Sequences to Associate Host/Virus Sequences

The ability to determine phylogenetic relationships in viral metapopulations is important to the current understanding of their community composition and function. In the absence of universal signature genes like 16S sequences, BLASTx and tBLASTx alignments have been the primary tools to determine phylogeny of viral metagenomic sequences and to correlate them with their respective hosts. BLASTx and tBLASTx focus on amino acid sequence similarities and ignore differences in codon usage and other patterns of nucleotide content, which can be highly informative.

Sequence signature-based methods, independent of nucleotide or amino acid alignment, are being developed to classify the phylogenies of viral metagenomes and their hosts. Phylopythia is an approach designed for cellular metagenomes (McHardy et al. 2007; see also Chap. 47, Vol. I) that classifies based on oligonucleotide composition differences. Alternative

approaches use differences in codon usage, which are generally conserved between hosts and viruses (Lucks et al. 2008). Genome signature-based phylogenetic classification (GSPC) analyzes differences in di-, tri-, and tetranucleotide utilization patterns to associate phylogenetic relationships, which are influenced by codon usage bias, as a basis for correlating hosts and viruses (Pride et al. 2006; Yooseph and Sutton 2008).

A GSPC study based on tetranucleotide utilization in the Yellowstone viral metagenomes from Bear Paw and Octopus hot springs was reported in Pride and Schoenfeld (2008), which includes the details of the analysis and the statistical support. To be statistically significant, the analysis used only contigs >1.9 kb (3.8 kb when analyzing both strands) assembled at 95 % identity. Contigs of this size should include 95 % of tetranucleotide combinations at least 7.5 times. Approximately 19.3 % and 39.0 % of the Bear Paw and Octopus metagenomic contigs, respectively, representing the more abundant viruses, conformed to these criteria. The GSPC analysis classified 20 of 22 Bear Paw contigs and 69 of 70 Octopus contigs, a much higher proportion of the reads than either BLASTx or Phylopythia with significantly stronger statistical support (see Pride and Schoenfeld 2008). The method is useful to group contigs by relatedness, which might assist assembly, and to infer phylogenies and hosts. The GSPC analysis suggests that Octopus viruses belong primarily to archaeal families *Globuloviridae* and *Fuselloviridae* (56 of 69) while Bear Paw members belong primarily to the bacteriophage family *Caudoviridae* (includes *Myoviridae*, *Podoviridae*, and *Siphoviridae*) (17 of 20). The analysis also estimates that 80 % of the Octopus contigs have archaeal signatures, while 77 % of Bear Paw contigs had bacterial signatures, a finding consistent with BLASTx analysis.

The apparent predominance of archaeal viruses seems inconsistent with the reported dominance of Octopus sediments and filaments by *Bacteria* (Blank et al. 2002; Rachel et al. 2002). Furthermore, the viral populations appear much more diverse than would be

predicted based on the low diversity of microbes in the sediments and filaments. The BLASTx, GSPC, and diversity data all suggest that the viruses are infecting hosts other than the sedentary surface bacteria, implying significant proliferation either in the pool or in the vent. The viruses used in this study were planktonic isolates collected close to the outflow source immediately after emergence, making it more unlikely that the hosts were surface microbes in the filament, sediments, or water column.

Alignment of the Metagenome to Cultivated Viral Genomes

Overall, only 3.4 % of the high stringency (95 % assembly) contigs from the two libraries showed similarity to known viral sequences. Most of these similarities were to cultivated thermophilic crenarchaeal viruses (Table 6). Similarity to the only non-thermophilic virus, *phage Twort* (Kwan et al. 2005), was limited to the helicase gene, which shares similarity with that of SIFV (see above). The two libraries shared comparable frequencies of sequence similarity to archaeal viruses and bacteriophage. Notable exceptions were *Acidianus rod-shaped virus* and *Sulfolobus islandicus rod-shaped virus* 1 and 2 where the Octopus library demonstrated a higher frequency of homology than the Bear Paw library and the *S. tengchongensis spindle-shaped virus* 1 homology, less common in Octopus than in Bear Paw.

Alignment of the metagenomes to whole genome sequences of six cultivated thermophilic viruses revealed striking conservation of certain sequences (Fig. 5). Almost the entire genome of *Pyrobaculum spherical virus* (PSV) has similarity to sequences in both metagenomic libraries, with median identities of 60 % and 51 % to the Bear Paw and Octopus, respectively. Sequence similarities to the other crenarchaeal viruses and to bacteriophage YS40 were limited to a few specific ORFs, but the degree of similarity was relatively high in those regions. Interestingly, nearly all of the ORFs showing high levels of homology are among the few thermophilic

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 6 Numbers of 95 % contigs with tBLASTx similarities to cultured viral sequences

Virus	References	Accession	Number of tBLASTx similarities	
			Bear paw	Octopus
ARV, <i>Acidianus rod-shaped virus</i>	(Vestergaard et al. 2005)	AJ875026	36	228
SIRV 1 and 2, <i>Sulfolobus islandicus rod-shaped virus 1 and 2</i>	(Blum et al. 2001; Peng et al. 2001)	AJ344259, AJ414696	30	217
PSV, <i>Pyrobaculum spherical virus</i>	(Haring et al. 2004)	AJ635161	44	152
SIFV, <i>S. islandicus filamentous virus</i>	(Arnold et al. 2000)	AF440571	7	46
STSV1, <i>S. tengchongensis spindle-shaped virus 1</i>	(Xiang et al. 2005)	AJ783769	26	22
ATV, <i>Acidianus two-tailed virus</i>	(Prangishvili et al. 2006b)	AJ888457	8	17
YS40, <i>Thermus thermophilus YS40 phage</i>	(Naryshkina et al. 2006)	DQ997624	15	41
TTSV1, <i>Thermoproteus tenax spherical virus 1</i>	(Ahn et al. 2006)	AY722806	6	12
Twort, <i>Staphylococcus phage Twort</i>	(KwanT et al. 2005)	AY954970	4	21

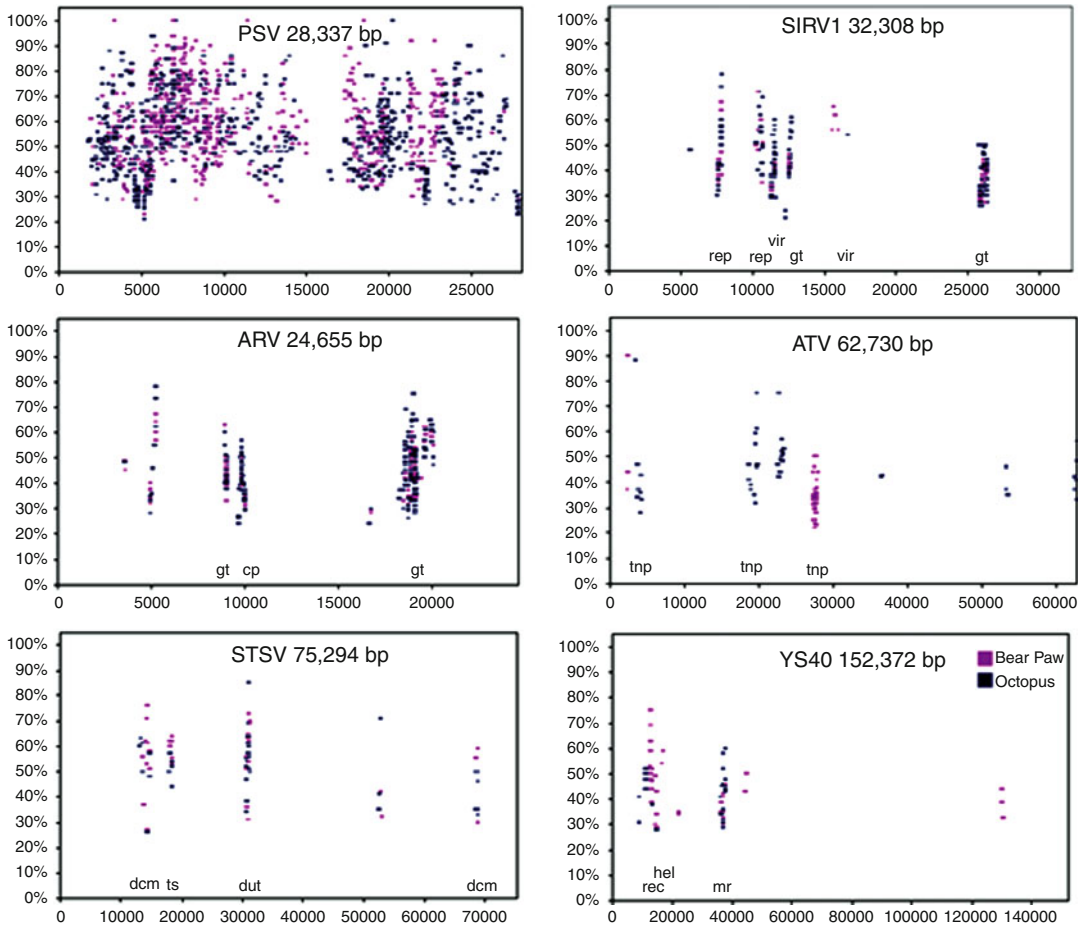
crenarchaeal virus genes for which a function has been assigned or inferred (Fig. 5 and references therein). These regions of high conservation are genes associated with virion components, DNA replication, transposition, recombination, or nucleic acid metabolism.

The degree of alignment to cultivated viruses was surprising. PSV was isolated from Obsidian hot spring (74 °C, pH 5.6), about 30 km away from both Octopus and Bear Paw. The geochemistry of this thermal feature is distinct from the springs in this study (Shock et al. 2005), and life within includes a highly diverse population of *Archaea* and *Bacteria* (Barns et al. 1994; Hugenholtz et al. 1998), most of which have not been detected in Octopus hot spring (Reysenbach et al. 1994; Blank et al. 2002) or elsewhere. In contrast, *Thermoproteus tenax spherical virus*, which is quite similar to PSV in terms of sequence, morphology, and habitat (Ahn et al. 2006), had very limited similarity to the YNP viral metagenomic sequences (not shown). The other viruses showing high similarity to the metagenomic sequences were isolated on different continents and, with the exception of YS40, occurred in highly acidic springs. This observation is more remarkable because the microbial populations of acidic and neutral hot springs are quite distinct (Reysenbach et al. 2002). The one other virus cultivated from Yellowstone,

SSV-RH (Wiedenheft et al. 2004), had no significant tBLASTx similarity to any of the metagenomic samples.

Identification of CRISPR Spacer Cognate Sequences in the Octopus Viral Metagenome

Evidence has been accumulating recently associating CRISPR (clustered regularly interspaced short palindromic repeats) systems with acquired resistance to lateral gene transfer from viruses and episomal elements (reviewed by van der Oost et al. 2009). CRISPRs were first discovered as repetitive sequences found in most bacterial and virtually all archaeal genomes. These systems are functionally analogous, but nonhomologous, with eukaryotic RNA interference and appear to limit the lateral transfer of genes by targeting them for nucleolytic degradation prior to their integration into the genome. The emerging view is that sequences in the repeat region of the CRISPR system correspond to sequences in viral or episomal genes and are transcribed in the host cell as part of a targeting system to neutralize viral infections. However, little direct evidence of conservation between the CRISPR spacer sequences and viral genomes has been found in natural environments. The first



Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Fig. 5 Alignment of Octopus and Bear Paw viral metagenomic library contigs with six cultured virus genomes. Contigs assembled at >95 % identity from the viral metagenomic libraries were compared by tBLASTx to the genomes of PSV, SIRV1, ARV, ATV, STSV, and YS40. Each bar represents the alignment of a unique metagenomic sequence to the indicated location on the cultivated viral genome, shown on the horizontal axis. Percent coding sequence identities are shown in the vertical axis. The threshold for inclusion is $E\text{-value} < 10^{-3}$. Red bars indicate Bear Paw alignments; blue bars indicate

Octopus alignments. Also shown are the known or predicted functions of the conserved coding sequences (*rep* replication related, *vir* virion component, *gt* glycosyl-transferase, *tnp* transposase, *cp* coat protein, *dam* adenine DNA methylase, *ts* thymidylate synthase, *dut* dUTPase, *dcm* cytosine DNA methylase, *hel* helicase, *rec* recombinase, *rnr* ribonucleotide reductase (Arnold et al. 2000; Blum et al. 2001; Peng et al. 2001; Haring 2004; Kessler et al. 2004; Vestergaard et al. 2005; Xiang 2005; Ahn et al. 2006; Naryshkina 2006; Prangishvili 2006b) (Reproduced with permission from (Schoenfeld et al. 2008)

demonstration of a correspondence between CRISPR spacers and viral sequences was in dairy bacteria and their associated phages (Horvath et al. 2008) and, by inference, in acid mine drainages (Andersson and Banfield 2008). In these and other cases, the lack of viral metagenomes limited insight into the coevolution

of these genes in microbial and viral populations. Furthermore, since the CRISPR spacer sequences are generally only 20–50 nucleotides in length, it has been difficult to assign function of the targeted genes by BLASTx or other means.

The Octopus viral metagenome, in conjunction with a microbial metagenome and two



Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 7 Octopus virome sequences showing silent or conservative changes compared to the CRISPR spacer sequences of the *Synechococcus* genome

Sequence	%NAID	Predicated AA sequence	% AASIM/AAID
AGTTTACCCTCAAGTGGGAAGGCGGCTTTGTCCACCATCC		FTLKWEGGFVHH	
.....T.....T.....	95	100/100
.....T.....T.....	95	100/100
.....T..T.....T.....	92	100/100
.....T..T.....T.....	92	100/100
.....T.....T...A.....	92Y..	100/91
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
...C....G..A.....G..G.....C..	86	100/100
.A.....G....G.AC..AA.T....	80Y.N.	100/83
.A.....G....G.AC..AA.T....	80Y.N.	100/83
.A.....G....G.AC..AA.T....	80Y.N.	100/83
.A.....G....G.AC..AA.T....	80Y.N.	100/83
.A.....G....G.AC..AA.T....	80Y.N.	100/83
.C....A..A..A.....T..T.AC..AA...C..	75Y.N.	100/83
.C....A..A..A.....G..T..G.AC..AA...C..	73Y.N.	100/83
.C....A..A..A.....G..T..G.AC..AA...C..	73Y.N.	100/83
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75
.....T.GCGC.....G..T...AC..GA...C..	70	..R...Y.N.	100/75

Synechococcus genomes isolated from the same hot spring within 2 years of one another, provided a unique opportunity to identify the genes targeted by a CRISPR system and observe coevolution of a CRISPR system and its target in host and viral genomes (Heidelberg et al. 2009). The two *Synechococcus* strains contained sequences with the hallmarks of a CRISPR system. Like other such sequences, the CRISPR spacers had no BLASTn or tBLASTx similarity to any sequences in GenBank. When compared to the microbial metagenome, 180 elements had similarity to CRISPR spacer sequences. Of these, four shared similarity with 23 sequences in the Octopus viral metagenome.

Interestingly, two CRISPR spacer sequences shared by the isolates and the microbial and viral metagenomes had similarity to different regions

of one gene in the viral metagenome. The assembly of 23 reads covering this single gene indicates that this was one of the most abundant and conserved element in the entire metagenome, which, by itself, would seem to make it an attractive target for a presumed antiviral system. The data provided by the viral metagenome reveal the target of the spacers was a likely lysozyme gene, the conservation of which may be explained by evolutionary constraints due to the interaction with a host cell wall. Inspection of the *lys* gene assembly revealed the apparent coevolution of the CRISPR system and its viral target (Table 7). Of the 23 viral metagenome reads, only five had detectable nucleic acid identities (NAID).

The sequence of one CRISPR spacer is shown in Line 1. Shown below are sequences from the



virome with similarity to this CRISPR spacer or the same region in reads identified by similarity to a second independent CRISPR spacer or a translation of one of these. Conserved nucleotides are shown as dots; those that diverge from the CRISPR 1 spacer are shown as letters. The percent nucleic acid identities (%NAID) to CRISPR 1 and the percent amino acid similarity and identity (% AASIM and % AAID, respectively) to the predicted translation of CRISPR1 are also shown. (adapted from Heidelberg et al. 2009). The remainder had sequence variations that reduced NAID to as low as 70 %; however, all of these nucleotide variations were silent or conservative with respect to the amino acid sequence, which would likely allow the sequence to evade targeting by the CRISPR system, but not affect the enzymatic function of the gene product. This data suggests a high rate of coevolution or “germ warfare” between the viruses and their hosts in this extreme environment.

Similarities Between the Two Hot Springs' Viral Populations

The two libraries were compared to one another to determine any variation between the viral populations in the two very different thermal environments. Contigs assembled at 95 % from the two libraries were compared to each other by tBLASTx and BLASTn (Table 8). The differences between the two analyses should be the result of noncoding nucleotides. Since gene densities are high in viral genomes and there is very little intergenic sequence, these differences are mainly due to silent codon variations, which should be largely free of selective pressure. Most remarkable is the similarity between the two libraries by either analysis. By tBLASTx, 5,843 of the Octopus contigs (43 %) and 1,593 of the Bear Paw contigs (26 %) shared amino acid coding similarity. By BLASTn, 2,876 (21 %) and 1,339 (21 %) of the respective contigs shared nucleotide similarity. The average percent identities were 74 % and 87 % and the expect values were 1.38E-05 and 3.00E-05, although the

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution, Table 8 Nucleotide and coding similarities between the viral populations of Octopus and Bear Paw hot springs

	tBLASTx	BLASTn
Frequency (number) of Octopus contigs with similarity to Bear Paw contigs	43 % (5,843)	21 % (2,876)
Frequency (number) of Bear Paw contigs with similarity to Octopus contigs	26 % (1,593)	21 % (1,339)
Average length of similarity (nucleotides)	298	175
Average identity	74 %	87 %
Average expect value	1.38E-05	3.00E-05

average length of sequence alignment (298 and 175 bp) was modest in both cases. This level of similarity did not allow extensive assembly of contigs from the two libraries, even at 50 % identity, presumably due to the short lengths of alignment (not shown). Taken together, these data suggest a mosaiclike pattern of overlap of much of the coding content in the two hot springs, although entire viral genomes or even entire genes are not necessarily fully conserved. The fact that the degrees of identity at the nucleotide level and at the translational level were relatively close suggests that this overlap is not due solely to selective pressure on the coding sequence, but must be explained by other mechanisms. This extensive conservation of viral sequences between the two hot springs in this study is surprising, given that microbial populations are highly temperature dependent (Reysenbach et al. 2002) and the surface temperatures of these hot springs differ by 19 °C (74 °C vs. 93 °C).

Conservation and Distribution of Viruses in Thermal Environments

Taken together, the above analyses suggest that (1) viral populations in the water columns are largely independent of microbial populations reported in the pools and (2) viral genomes, particularly certain genes, are more conserved both

regionally and globally than might have been predicted. The regional and global conservation of viral sequences is an intriguing area for further study. There are examples of globally distributed genes among marine viral assemblages (Breitbart and Rohwer 2005; Short and Suttle 2005). Since the oceans are contiguous across the earth, an obvious distribution mechanism exists. Groups of highly similar *Sulfolobus* viruses (Wiedenheft 2004) and *Thermus* phages (Yu 2006) have been isolated from thermal springs on different continents. In these cases, viruses were isolated from environments of similar pH and temperature and were cultivated on the same host under similar laboratory conditions. Gene homologs to these viruses were detected despite the absence of these selective conditions. Conversely, most crenarchaeal virus morphotypes have been detected in enrichments from YNP (Rice et al. 2001; Rachel et al. 2002; Wiedenheft et al. 2004); however, little is known about conservation of genes in these enrichments.

The mechanism and basis of this conservation of viral sequence is open to speculation. It is possible that viruses sharing common genes adapt to the different host populations of the environment. Alternatively, hot springs may be inoculated by airborne viruses from other springs. It is also possible that the viruses acquire genes from mesophilic viruses, although this explanation has no support in this study. Lineages of conserved viral genes may be older than the separation of the continents. Another explanation is proliferation of the viruses deeper in the vent. Thermophilic *Bacteria* and *Archaea*, potential hosts for viruses, have been detected in thermal aquifers several km beneath the earth's surface at abundances similar to those measured in this study (Moser et al. 2005) and many are distributed worldwide. While it is impossible to separate the contribution of the subsurface viruses from any proliferation at the surface in the two pools in this study, samples from thermal springs with no pool at all, collected within seconds of their emergence, have similar or somewhat higher viral abundances to those measured in this report (Breitbart et al. 2004b), suggesting subsurface proliferation is at least a significant

contributor to viral populations at the surface. Subsurface proliferation of viruses would also explain the apparent disconnect between the planktonic viral populations in the pool and the reported sedentary microbial populations, described above. An implication of subsurface proliferation of viruses is that the habitable portion of the subterranean aquifer could be continuous across much of the Yellowstone caldera or even much larger areas. A second implication is that, given the higher pressures in the vents, the temperature limit of life in the subterrestrial aquifers could significantly exceed the temperatures measured at the surface.

Computer Analysis

Availability of computer programs is described in the original publications (Schoenfeld et al. 2008; Heidelberg et al. 2009; Pride and Schoenfeld 2008).

References

- Ahn DG, Kim SI, Rhee JK, Kim KP, Pan JG, et al. TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote *Thermoproteus tenax*. *Virology*. 2006;351:280–90.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
- Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*. 2008;320:1047–50.
- Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*. 2005;6:41.
- Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. The marine viromes of four oceanic regions. *PLoS Biol*. 2006;4:e368.
- Arnold HP, Zillig W, Ziese U, Holz I, Crosby M, et al. A novel lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon *Sulfolobus*. *Virology*. 2000;267:252–66.
- Barns SM, Fundyga RE, Jeffries MW, Pace NR. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci USA*. 1994;91:1609–13.

- Barry ER, Bell SD. DNA replication in the archaea. *Microbiol Mol Biol Rev.* 2006;70:876–87.
- Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, et al. Metagenomic characterization of Chesapeake Bay viroplankton. *Appl Environ Microbiol.* 2007;73:7629–41.
- Blank CE, Cady SL, Pace NR. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol.* 2002;68:5123–35.
- Blum H, Zillig W, Mallok S, Domdey H, Prangishvili D. The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses. *Virology.* 2001;281:6–9.
- Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* 2005;13:278–84.
- Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA.* 2002;99:14250–5.
- Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol.* 2003;185:6220–3.
- Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci.* 2004a;271:565–74.
- Breitbart M, Wegley L, Leeds S, Schoenfeld T, Rohwer F. Phage community dynamics in hot springs. *Appl Environ Microbiol.* 2004b;70:1633–40.
- Brock TD, Brock ML. Measurement of steady-state growth rates of a thermophilic alga directly in nature. *J Bacteriol.* 1968;95:811–5.
- Brown JR, Doolittle WF. Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev.* 1997;61:456–502.
- Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H. Phage as agents of lateral gene transfer. *Curr Opin Microbiol.* 2003;6:417–24.
- Cann AJ, Fandrich SE, Heaphy S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes.* 2005;30:151–6.
- Coenye T, Vandamme P. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett.* 2003;228:45–9.
- Culley AI, Lang AS, Suttle CA. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virology.* 2007;4:69.
- Daubin V, Ochman H. Start-up entities in the origin of new genes. *Curr Opin Genet Dev.* 2004;14:616–9.
- Dionne I, Bell SD. Characterization of an archaeal family 4 uracil DNA glycosylase and its interaction with PCNA and chromatin proteins. *Biochem J.* 2005;387:859–63.
- Feng L, Wang W, Cheng J, Ren Y, Zhao G, et al. Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proc Natl Acad Sci USA.* 2007;104:5602–7.
- Filee J, Forterre P, Sen-Lin T, Laurent J. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol.* 2002;54:763–73.
- Filee J, Forterre P, Laurent J. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res Microbiol.* 2003;154:237–43.
- Forterre P. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie.* 2005;87:793–803.
- Forterre P. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* 2006;117:5–16.
- Fournier RO. Geochemistry and dynamics of the Yellowstone National Park Hydrothermal System. In: Inskeep W, editor. *Geothermal biology and geochemistry in YNP.* Bozeman: Thermal Biology Institute; 2005.
- Geslin C, Le Romancer M, Erauso G, Gaillard M, Perrot G, et al. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, “*Pyrococcus abyssi*”. *J Bacteriol.* 2003;185:3888–94.
- Godiska R, Patterson M, Schoenfeld T and Mead D. Beyond pUC: vectors for cloning unstable DNA. In: Kieleczawa, editor. *DNA sequencing: optimizing the process and analysis.* 2005; Jones and Bartlett Publishers, Sudbury, MA.
- Gold T. The deep, hot biosphere. *Proc Natl Acad Sci USA.* 1992;89:6045–9.
- Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol.* 2005;1:e60.
- Haring M, Peng X, Brugger K, Rachel R, Stetter KO, et al. Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the *Globuloviridae*. *Virology.* 2004;323:233–42.
- Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, et al. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet.* 2006;2:e92.
- Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE.* 2009;4:e4169.
- Hjörleifsdóttir SH, Hreggvidsson GO, Fridjonsson OH, Avarsson A, Kristjánsson JK. Bacteriophage RM 378 of a thermophilic host organism. US Patent; 2002.
- Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, et al. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol.* 2008;190:1401–12.

- Hugenholtz P, Pitulle C, Hershberger KL, Pace NR. Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol.* 1998;180:366–76.
- Jahnke LL, Eder W, Huber R, Hope JM, Hinrichs KU, et al. Signature lipids and stable carbon isotope analyses of Octopus Spring hyperthermophilic communities compared with those of Aquificales representatives. *Appl Environ Microbiol.* 2001;67:5179–89.
- Kahler M, Antrankian G. Cloning and characterization of a family B DNA polymerase from the hyperthermophilic crenarchaeon *Pyrobaculum islandicum*. *J Bacteriol.* 2000;182:655–63.
- Kessler A, Brinkman AB, van der Oost J, Prangishvili D. Transcription of the rod-shaped viruses SIRV1 and SIRV2 of the hyperthermophilic archaeon *Sulfolobus*. *J Bacteriol.* 2004;186:7745–53.
- Kwan T, Liu J, DuBow M, Gros P, Pelletier J. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc Natl Acad Sci USA.* 2005;102:5174–9.
- Kwan T, Liu J, Dubow M, Gros P, Pelletier J. Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. *J Bacteriol.* 2006;188:1184–7.
- Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA.* 2004;101:11013–8.
- Lucchini S, Desiere F, Brussow H. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. *J Virol.* 1999;73:8647–56.
- Lucks JB, Nelson DR, Kudla GR, Plotkin JB. Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol.* 2008;4:e1000001.
- Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998;26:1107–15.
- Martin A, Yeats S, Janekovic D, Reiter WD, Aicher W, et al. SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaeobacterium *Sulfolobus acidocaldarius* isolate B12. *EMBO J.* 1984;3:2165–8.
- McCleskey RB, Ball JW, Nordstrom DK, Holloway JM, Taylor HE. Water-chemistry data for selected hot springs, geysers, and streams in Yellowstone National Park, Wyoming, 2001–2002. U.S. Geological Survey Open-File Report 2004-1316; 2004.
- McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4:63–72.
- Moser DP, Gihring TM, Brockman FJ, Fredrickson JK, Balkwill DL, et al. *Desulfotomaculum* and *Methanobacterium* spp. dominate a 4– to 5-km-deep fault. *Appl Environ Microbiol.* 2005;71:8773–83.
- Naryshkina T, Liu J, Florens L, Swanson SK, Pavlov AR, et al. *Thermus thermophilus* bacteriophage phiYS40 genome and proteomic characterization of virions. *J Mol Biol.* 2006;364:667–77.
- Noble RT, Fuhrman JA. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat Microb Ecol.* 1998;14:113–8.
- Paul JH, Williamson SJ, Long A, Authement RN, John D, et al. Complete genome sequence of phiHSIC, a pseudotemperate marine phage of *Listonella pelagia*. *Appl Environ Microbiol.* 2005;71:3311–20.
- Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, et al. Origins of highly mosaic mycobacteriophage genomes. *Cell.* 2003;113:171–82.
- Peng X, Blum H, She Q, Mallok S, Brugger K, et al. Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology.* 2001;291:226–34.
- Prangishvili D, Garrett RA. Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem Soc Trans.* 2004;32:204–8.
- Prangishvili D, Garrett RA. Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol.* 2005;13:535–42.
- Prangishvili D, Garrett RA, Koonin EV. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res.* 2006a;117:52–67.
- Prangishvili D, Vestergaard G, Haring M, Aramayo R, Basta T, et al. Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. *J Mol Biol.* 2006b;359:1203–16.
- Pride DT, Schoenfeld T. Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics.* 2008;9:420.
- Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics.* 2006;7:8.
- Rachel R, Bettstetter M, Hedlund BP, Haring M, Kessler A, et al. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Arch Virol.* 2002;147:2419–29.
- Reysenbach AL, Wickham GS, Pace NR. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol.* 1994;60:2113–9.
- Reysenbach AL, Gotz D, Yermool D. Microbial diversity of marine and terrestrial thermal springs. In: Staley JT, Reysenbach AL, editors. *Biodiversity of microbial life*. New York: Wiley Liss; 2002.
- Rice G, Stedman K, Snyder J, Wiedenheft B, Willits D, et al. Viruses from extreme thermal environments. *Proc Natl Acad Sci USA.* 2001;98:13341–5.
- Roberts J A, Bell SD, White MF. An archaeal XPF repair endonuclease dependent on a heterotrimeric PCNA. *Mol Microbiol.* 2003;48:361–71.
- Sakaki Y, Oshima T. Isolation and characterization of a bacteriophage infectious to an extreme thermophile, *Thermus thermophilus* HB8. *J Virol.* 1975;15:1449–53.

- Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, et al. Assembly of viral metagenomes from Yellowstone hot springs. *Appl Environ Microbiol.* 2008;74:4164–74.
- Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM. Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. *J Bacteriol.* 2003;185:6434–47.
- Shock EL, Holland M, Meyer-Dombard DR, Amend JP. Geochemical sources of energy for microbial metabolism in hydrothermal ecosystems: Obsidian Pool, Yellowstone National Park. In: Inskeep WP, McDermott TR, editors. *Geothermal biology and geochemistry in YNP*. Bozeman: Thermal Biology Institute; 2005.
- Short CM, Suttle CA. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol.* 2005;71:480–6.
- Snyder JC, Stedman K, Rice G, Wiedenheft B, Spuhler J, et al. Viruses of hyperthermophilic Archaea. *Res Microbiol.* 2003;154:474–82.
- Snyder JC, Spuhler J, Wiedenheft B, Roberto FF, Douglas T, et al. Effects of culturing on the population structure of a hyperthermophilic virus. *Microb Ecol.* 2004;48:561–6.
- Stoner DL, Geary MC, White LJ, Lee RD, Brizzee JA, et al. Mapping microbial biodiversity. *Appl Environ Microbiol.* 2001;67:4324–8.
- Suttle CA. Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol.* 2007;5:801–12.
- Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997;278:631–7.
- van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci.* 2009;34:401–7.
- Vestergaard G, Haring M, Peng X, Rachel R, Garrett RA, et al. A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus *Acidianus*. *Virology.* 2005;336:83–92.
- Villarreal LP, DeFilippis VR. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J Virol.* 2000;74:7079–84.
- Ward DM, Ferris MJ, Nold SC, Bateson MM. A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev.* 1998;62:1353–70.
- Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? *Environ Microbiol.* 2004;6:1–11.
- Wen K, Ortmann AC, Suttle CA. Accurate estimation of viral abundance by epifluorescence microscopy. *Appl Environ Microbiol.* 2004;70:3862–7.
- Wiedenheft B, Stedman K, Roberto F, Willits D, Gleske AK, et al. Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. *J Virol.* 2004;78:1954–61.
- Wommack KE and Colwell RR. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev.* 2000;64:69–114.
- Xiang X, Chen L, Huang X, Luo Y, She Q, et al. *Sulfolobus tengchongensis* spindle-shaped virus STSV1: virus-host interactions and genomic features. *J Virol.* 2005;79:8677–86.
- Yooseph S, Li W, Sutton G. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics.* 2008;9:182.
- Young R. Bacteriophage lysis: mechanism and regulation. *Microbiol Rev.* 1992;56:430–81.
- Yu MX, Slater MR, Ackermann HW. Isolation and characterization of *Thermus* bacteriophages. *Arch Virol.* 2006;151:663–79.

V

Variable Selection to Improve Classification of Metagenomes

Greg Ditzler¹, Yemin Lan², Jean-Luc Bouchot³ and Gail Rosen¹

¹Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA

²School of Biomedical Engineering, Science and Health, Drexel University, Philadelphia, PA, USA

³Department of Mathematics, Drexel University, Philadelphia, PA, USA

Introduction

Metagenomics is the study of DNA extracted from the microbial communities in an environment, in comparison to traditional genomics, which studies the nucleic acids from single organisms (Wooley et al. 2010). In a metagenomic study, a sample is collected directly from the environment, which can be a gram of soil (Rousk et al. 2010; Bowers et al. 2011), milliliter of ocean (Williamson et al. 2008), swab from an object (Caporaso et al. 2011), or a sample of the microbes associated with a host organism, such as humans (Caporaso et al. 2011; Costello et al. 2009). The microbial content of an environmental sample is termed its “microbiome.” There are several questions that are of particular importance when the microbiome is being examined. In particular, who is there, how much of each species is there, and what are they

doing overall? Some of these questions can be addressed using DNA/RNA sequencing followed by homology and taxonomic classification; however, usually hypotheses focus on answering: which organisms and/or their functions (e.g., metabolisms) best differentiate multiple phenotypes in a collection of samples? Consider a collection of gut microbiome samples that were collected from patients with inflammatory bowel disease (IBD) and a control set that do not have IBD. A natural question to ask when examining the differences between the gut microbiomes of the two phenotypes is what organisms or genes can distinguish patients with IBD and healthy controls? Knowing the answers to such question can be useful in developing a better understanding about a disease and aid in developing medicines to target a disease cause.

The question of finding differentiating features, or variables of interest, has been deeply studied in the machine learning community (see Guyon et al. 2006; Saeys et al. 2007), which is commonly referred to as feature selection. Feature selection is the process of finding a subset of features that best differentiate between multiple classes or, in our case, phenotypes in a data set. The process of selecting features is typically achieved by maximizing some objective function (e.g., mutual information) in a greedy fashion. The central motivation for feature selection is to find a smaller subset of features that can be used to differentiate between the multiple phenotypes, which in turn can reduce the computational complexity of the classification algorithm tailored to

Variable Selection to Improve Classification of Metagenomes, Table 1 Functional databases mostly used for creating functional profiles

Large collection of reference sequences	RefSeq	Around 18 million proteins from 18 k organisms, annotations are available for a subset of the database, well-annotated for human sequences
	UniProtKB/Swiss-Prot	Manually curated annotations for 500,000+ sequences, covering 12,930 organisms
Standardized ontologies	Gene Ontology	Well-controlled vocabulary, primarily for eukaryotes
Gene orthologous groups	COG	Gene groups classified into 23 functional categories, inferred from 66 prokaryote and unicellular eukaryote genomes
	KOG	Eukaryote version of COG containing 7 eukaryotic genomes
	eggNOG	Automated annotation of orthologs in 1,133 species
Metabolism	KEGG pathway	400+ manually drawn pathways, based on reactions from multiple species
	BioCyc/Metacyc	2,000+ single-organism, experimentally derived pathways
	SEED	Subsystems that describe metabolic machinery with expert curation
Protein domains and families	Pfam	A large collection of protein families that share the same domain
	FIGfam	Protein families that share domains and pairwise align for their full length sequences, resulting in less sequences per family

do such a task. Furthermore, regression could be used instead of classification in the case of continuous-environmental variables; however, for this entry, we assume that phenotypes take on discrete states, and therefore, classification is the primary focus. Previously, feature selection has been shown useful to reduce the complexity of metagenome classification (Ditzler et al. 2012); however, in this article, its use is expanded to determine relevance of biological features to associated phenotypes, thus aiding researchers in drawing conclusions from metagenomic data.

Feature selection can be applied to a variety of metagenomic data (e.g., 16S rRNA, whole genome shotgun, taxonomic annotations, gene annotations). In addition to selecting species which differentiate microbiomes, many studies wish to map DNA/RNA sequences to functional categories and address enriched/depleted functions between samples. Depending on the type of question being asked and the nature of the data, there are a variety of functional databases to choose from. Table 1 highlights some of the most widely used databases. Large reference sequence databases with a variety of functional descriptions are preferred because they provide detailed annotation of diverse data set. This

raw-labeling of sequences can provide much information; however, it cannot be used to analyze hierarchical functional structure in a data set, such as what high-level functions (e.g., reproduction/cellular transport) are upregulated in my sample. Instead, sequence labeling can answer what genes exist in my sample or which sample is functionally more diverse, because they provide better annotation coverage in the sample than higher-level databases. However, if it is required to annotate with well-defined vocabularies, which is needed to make biological inference and associations, then one wishes to use a standardized ontology database. For example, researchers can use Gene Ontology annotation to examine what functions are enriched in the sample compared to others. In some cases, researchers wish to annotate the function of a gene that appears in multiple organisms rather than just one. In other words, the focus is to accurately assign homologous genes associated with multiple species, which is especially important in metagenomics due to the complex mixture of organisms in a sample. Therefore, orthologous group databases are useful for annotating homologous function of orthologs. For studying a microbiome's metabolism rather than molecular functions, such as asking the questions what

biological processes are enriched/missing from a diseased microbiome or should photosynthesis activity be enhanced in surface soil compared to deeper layer soil samples, several metabolic pathway databases can be used. Finally, protein family databases search for conserved domains and motifs of protein sequences and are important when considering the origin and evolution of proteins. For example, protein motifs that characterize pathogenicity may be used as potential targets for diagnosis and treatment.

Since the diversity of functional databases serves a variety of research questions, it is important to note that many studies would adopt several databases for annotation. Therefore, the optimal feature selection technique may depend on the database choice and the nature of taxonomic or functional data, such as the dimension of feature space, data sparsity, and the possible range of fold change between samples.

This entry is organized as follows: section “[Feature Selection](#)” highlights the components of a general feature selection algorithm and how to design such an algorithm. Section “[A Description of the MetaHit Database](#)” presents the benchmark MetaHit data set, followed by an empirical analysis of feature selection algorithms tested on the MetaHit data set in section “[Data Analysis](#).” Finally, section “[Conclusion](#)” draws concluding remarks for feature selection applied to metagenomic data.

Feature Selection

Feature selection can provide a unique insight about the variables that provide discriminating information about populations, or phenotypes, typically contained in the metadata. This metadata could be as simple as two populations, such as healthy or unhealthy, or significantly more complex by containing many different populations within a data sample. It is natural during the analysis of a biological data set to ask the question: which variables provide the most differentiation between multiple populations? The answer to such questions can be answered using feature selection (Guyon and Elisseeff 2003). There are several

items to consider before applying a feature selection to a (biological) data set. First, how many features should be selected? Most feature selection algorithms assume that the end-user must select this parameter, and the quality of the results will most likely be highly dependent on the value of this parameter. In many situations, cross validation can be used to search for an acceptable value. Second, what is the primary objective for features selection? Is it the goal of the end-user to perform classification, or are they simply looking for the top k features in the data set? The design of the objective function, $J(\cdot)$, for feature selection can be used to emphasize and address these questions.

Let $J(\cdot)$ be a function of the features X_j (for $j = 1; \dots; Q$), the label variables Y , and the current relevant feature set F . Note that the collection of variables (e.g., operational taxonomic units, Pfams, etc.) is denoted by X . The objective function can be designed in a way, such that it reflects the task at hand. For example, if a biologist is interested in the top ranking features that carry the most mutual information between X_j and Y , then the objective function should reflect this goal. In this situation, using a mutual information maximization (MIM) method is sufficient to achieve this goal (Lewis 1992). MIM can be implemented as follows: (a) compute $I(X_j; Y)$ for all j ($I(X_j; Y)$ is the mutual information between X_j and Y), (b) rank the mutual informations in descending order, and (c) select the top k variables with the largest mutual information and place them in F .

However, many times we seek to classify data based on Y , and in such situations designing a more complex objective function is required. For example, it may be more advantageous to select F in such a way that the features contained in F are informative about Y ; however, they are not redundant (i.e., one or more features provide the same amount of information about Y). An example of such an objective function is given by

$$\mathcal{J}(X_j, Y, F) = I(X_j; Y) - \sum_{X_s} I(X_j; X_s)$$

where the first term maximizes the mutual information between the features, X_j , and metadata, Y ,

Variable Selection to Improve Classification of Metagenomes,

Fig. 1 Generic forward feature selection algorithm for a filter-based method

Input: Feature set \mathcal{X} , an objective function \mathcal{J} , k features to select, and initialize an empty set \mathcal{F}

1. Maximize the objective function

$$X^* = \arg \max_{X_j \in \mathcal{F}} \mathcal{J}(X_j, Y, \mathcal{F}) \quad (1)$$

2. Update relevant feature set such that $\mathcal{F} \leftarrow \mathcal{F} \cup X^*$
3. Remove relevant feature from the original set $\mathcal{X} \leftarrow \mathcal{X} \setminus X^*$
4. Repeat until $|\mathcal{F}| = k$

while the second term is penalizing X_j for being redundant with the current relevant feature set in \mathcal{F} . The design of the objective function is quite important to the application to which feature selection is being applied. There are several works that highlight such results on bioinformatics data (Saeys et al. 2007), information theory methods (Brown et al. 2012), and general feature selection techniques (Guyon and Elisseeff 2003).

A simple algorithm for feature selection is the forward selection search, which is shown in Fig. 1. The method begins by initializing the relevant feature set \mathcal{F} to the empty set. Then for k cycles, equation (1) is maximized, and the feature that maximizes the expression is added to the relevant feature set, \mathcal{F} , and removed from the feature set, \mathcal{X} . The forward selection search is used with several feature selection objective functions in the section on “Data Analysis.”

A Description of the MetaHit Database

As mentioned in Introduction, feature selection can allow researchers in metagenomics to interpret the differentiating features in a data set. The interpretation can be insightful and allow the researchers to determine the functional differences between multiple phenotypes. As a case study, let us examine a metagenome data set collected by Qin et al. (2010), which is widely referred to as the MetaHit data set. The data are collected from Illumina-based metagenomic sequencing of 124 fecal samples of 124 European individuals from Spain and Denmark.

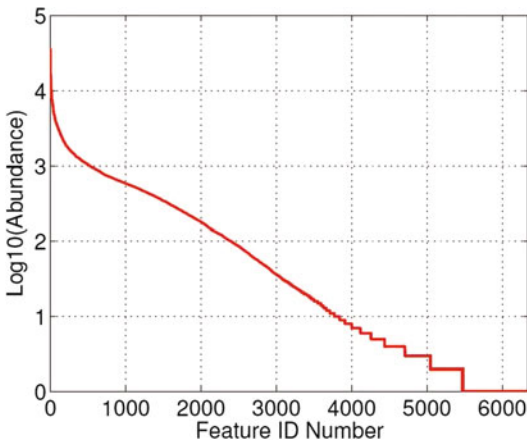
The MetaHit data set represents one of the most comprehensive studies of the human gut microbiome. Among the 124 individuals in the database, 25 are from patients who have inflammatory bowel disease (IBD), and 42 patients are also obese. It is interesting to note that only three of the individuals who have IBD are also obese. Let us consider two different labeling schemes for the data: IBD and obesity, both of which are binary prediction problems. The sequences from each individual are functionally annotated using the Pfam database (Finn et al. 2010), in a recent study that utilized the MetaHit data set for feature selection on patient age (Lan et al. 2013). There are a total of 6,343 unique functional features detected in the data set, and Fig. 2 shows the \log_{10} of the total abundance for each of the 6,343 functional features over the 124 observations in the data set.

One way to (loosely) access the separability of the IBD and no IBD patients (or obese and not obese) in the data is to examine the principal coordinate analysis (PCoA) plots of the patients’ Pfam data (Gower 1967). Figure 3 shows the PCoA scatter plots of the two sample labeling schemes using PCoA implemented with the Euclidean distance. From these plots we observe that there is a significant amount of overlap between the classes for both labeling schemes.

Data Analysis

In this section, the classification accuracy and area under the receiver operating characteristic (auROC)

curve for the MetaHit data set are examined when feature selection is applied. The accuracy is measured using the standard 1–0 loss, and the auROC is interpreted as the probability of ranking a target data instance higher than a randomly selected nontarget data instance (Fawcett 2006). The IBD/obese class label is identified as the target for the calculation of the auROC.

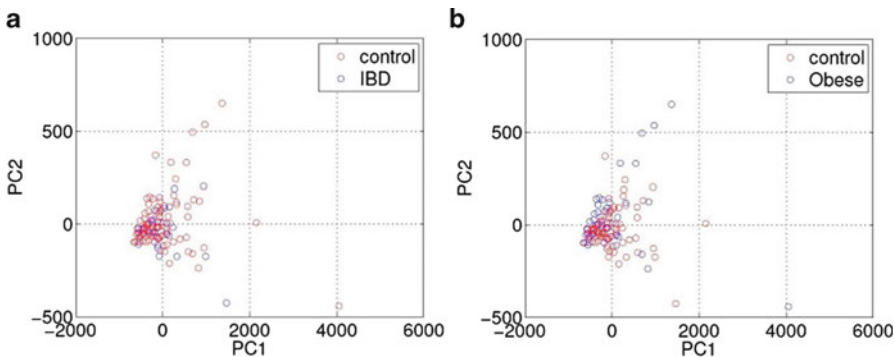


Variable Selection to Improve Classification of Metagenomes, Fig. 2 Logarithm of the total abundance of each feature detected by the Pfam database for Qin et al. (2010)’s human gut microbiome data set. The x-axis represents rank of each feature corresponding with the number of detections sorted in descending order. From the plot, it is obvious that there are few Pfams with a large abundance and many Pfams with a very low abundance count. For example, there are 2,572 Pfams with 10 or fewer occurrences across the 124 observations

The joint-mutual information feature selection algorithm (JMI) is implemented with a forward selection search, and the naïve Bayes classifier is implemented with a multinomial model. The FEAST feature selection toolbox implements the JMI algorithm (Brown et al. 2012). All statistics are presented as averages from tenfold cross validation using stratified sampling. Stratified sampling assures that instances from each class will be in each cross-validation data set. Note that completely random cross-validation data set partitions do not guarantee this property.

The auROC and loss for the multinomial naïve Bayes classifier are measured using the two labeling schemes described in section “A Description of the MetaHit Database” (i.e., IBD and obese). Table 2 contains the classification assessments from the different labeling schemes as well as a variation in the number of features that are selected via JMI. From Table 2, it is clear that feature selection can have a significant outcome in the classification results. This is best shown in Fig. 4 which shows the number of features selected by the MIM algorithm versus the loss (Fig. 4b) and the auROC (Fig. 4a). Note that these results are generated using the mutual information maximization approach; however, similar results/trends are observed for other feature selection methods.

Figure 5a presents a visualization of the MetaHit data set before and after MIM feature



Variable Selection to Improve Classification of Metagenomes, Fig. 3 (a) IBD (b) Obese. Multi-dimensional scaling of the MetaHit data set with the IBD

and obese labeling of the samples. There appears to be a significant amount of overlap between the controls and targets for both prediction problems

selection is applied. The features are sorted from high to low in terms of overall abundance, and the patients are represented such that samples 1–99 do not have IBD and samples 99–124 have IBD. Clearly, this shows a large amount of sparsity that is inherent in the data, which would also be evident if taxonomic abundances were used over Pfams. Figure 5b shows that most of the features being selected by MIM are relatively abundant features; however, simply because a feature is abundant does not imply that the feature is relevant. This can be observed near the 44th feature in Fig. 5b. Note that the features in Fig. 5b are ordered by the time they were selected by the forward search.

The top Pfams that maximize the mutual information for the MetaHit data set are shown in

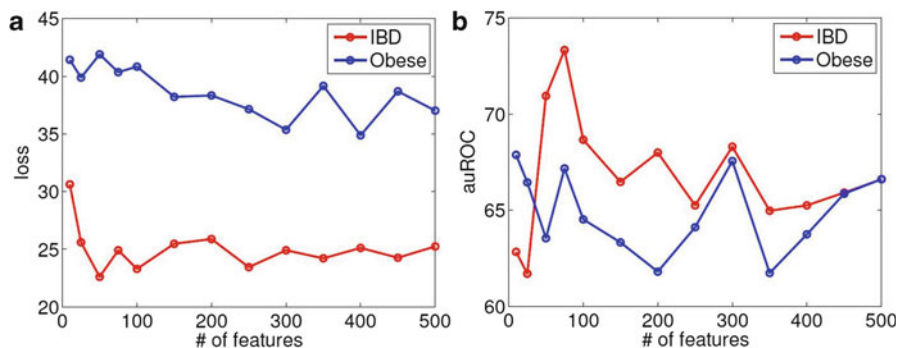
Variable Selection to Improve Classification of Metagenomes, Table 2 Area under the ROC (auROC) curves and classification error for a naïve Bayes classifier tested using tenfold cross validation

	auROC (IBD)	Error (IBD)	auROC (obese)	Error (obese)
10	0.706	0.233	0.640	0.395
15	0.624	0.290	0.672	0.352
25	0.616	0.292	0.660	0.403
50	0.750	0.223	0.649	0.422
100	0.660	0.249	0.659	0.397
200	0.654	0.257	0.643	0.389
500	0.635	0.277	0.641	0.378
All	0.665	0.238	0.622	0.240

Table 3. It is known in IBD patients that the expression of ABC transporter protein (PF00005, the first feature MIM selected for classifying IBD versus no IBD samples) is decreased which limits the protection against various luminal threats (Deuring et al. 2011). The feature selection for IBD also identified glycosyltransferase (PF00535), whose alternation is hypothesized to result in recruitment of bacteria to the gut mucosa and increased inflammation (Campbell et al. 2001). And the genotype of acetyltransferase (PF00583) plays an important role in the pathogenesis of IBD, which is useful in the diagnostics and treatment of IBD (Baranska et al. 2011). It is not surprising that ABC transporter (PF00005) is also selected for obesity, which is known to mediate fatty acid transport that is associated with obesity and insulin-resistant states (Ashrafi 2007) and ATPases (PF02518) that catalyze dephosphorylation reactions to release energy.

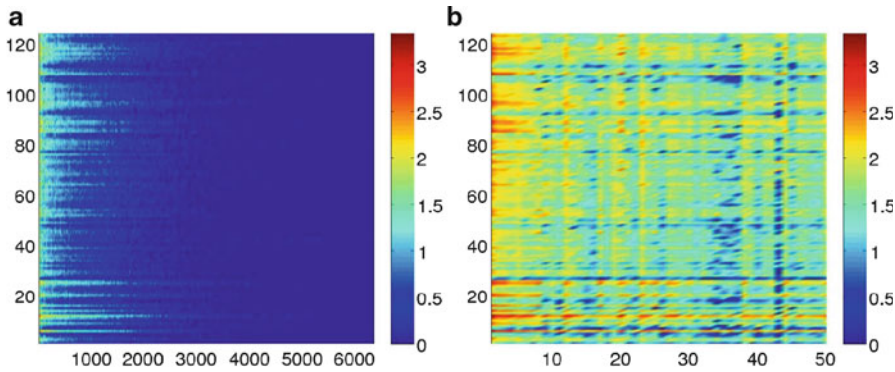
Conclusion

This entry has presented a broad overview about how feature selection algorithms can be used to facilitate and interpret data in the field of metagenomics. Recall that metagenomic abundance data can be of very large dimension (e.g., MetaHit), and feature selection reduces the



Variable Selection to Improve Classification of Metagenomes, Fig. 4 (a) Loss of naïve Bayes. (b) auROC of naïve Bayes. The effect of the number of features selected by the MIM algorithm versus the loss (left) and the auROC (right). The number of features being

selected has a larger effect on the auROC (i.e., detection of target population examples) than the accuracy of the system. Similar results are observed with JMI and other feature selection methods



Variable Selection to Improve Classification of Metagenomes, Fig. 5 (a) No feature selection. (b) Feature selection. Visualization of the abundance matrix (on a log10 scale) (a) Before and (b) after MIM feature selection. The *x-axis* represents a feature and *y-axis* represents

samples. Samples 1 through 99 do not have IBD, and samples 99 through 124 have IBD. (b) contains the top 50 features relevant to the 124 data sets. Differences between the two classes cannot be visualized; however, classification auROCs are 10–15 % above chance

Variable Selection to Improve Classification of Metagenomes, Table 3 List of the “top” Pfams as selected by the MIM feature selection algorithm. Note that redundancy terms are not accounted for in the objective of MIM. Hence, the features below are the ones that provide the largest amounts of mutual information. The ID in parentheses is the Pfam accession number

	IBD features	Obese features
Feature 1	ABC transporter (PF00005)	ABC transporter (PF00005)
Feature 2	Phage integrase family 2 (PF00589)	MatE (PF01554)
Feature 3	Glycosyltransferase family 2 (PF00535)	TonB-dependent receptor (F00593)
Feature 4	Acetyltransferase (GNAT) family (PF00583)	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase (PF02518)
Feature 5	Helix-turn-helix (PF01381)	Response regulator receiver domain (PF00072)

dimensionality of the space to allow for a quick interoperation of the data. Furthermore, feature selection is also useful for classification because it allows us to remove potentially irrelevant features from the data set, which allows the classifier to focus on learning from the relevant information rather than attempt to decipher what is or is not relevant.

References

Ashrafi K. Obesity and the regulation of fat metabolism. *WormBook*. 2007;9:1–20. Review. PMID:18050496.

Baranska M, Trzcinski R, Dziki A, Rychlik-Sych M, Dudarewicz M, Skretkiewicz J. The role of n-acetyltransferase 2 polymorphism in the etiopathogenesis of inflammatory bowel disease. *Dig Dis Sci*. 2011;56(7):2073–80. doi: 10.1007/s10620-010-1527-4. Epub 2011 Feb 15. PMID:21321790.

Bowers RM, McLetchie S, Knight R, Fierer N. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *ISME J*. 2011;5:601–12.

Brown G, Pocock A, Zhao M-J, Luj'an M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J Mach Learn Res*. 2012;13:27–66.

Campbell BJ, Yu LG, Rhodes JM. Altered glycosylation in inflammatory bowel disease: a possible role in cancer development. *Glycoconj J*. 2001;18(11–12):851–8. Review.

Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. Moving pictures of the human microbiome. *Genome Biol*. 2011;12(5).

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science*. 2009;326:1694–7.

Deuring JJ, Peppelenbosch MP, Kuipers EJ, van der Woude CJ, de Haar C. Impeded protein folding and

- function in active inflammatory bowel disease. *Biochem Soc Trans.* 2011;39:1107–11.
- Ditzler G, Polikar R, Rosen G. Information theoretic feature selection for high dimensional metagenomic data. In: *International Workshop on Genomic Signal Processing and Statistics*, 2012.
- Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett.* 2006;27:861–74.
- Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Guneseakaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy S, Bateman A. The pfam protein families database. *Nucleic Acids Res.* 2010;38:D211–222.
- Gower J. Multivariate analysis and multidimensional geometry. *J R Stat Soc.* 1967;17(1):13–28.
- Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003; 3:1157–82.
- Guyon I, Gunn S, Nikravesh M, Zadeh LA. *Feature extraction: foundations and applications*. Berlin: Springer; 2006.
- Lan Y, Kriete A, Rosen GL. Selecting age-related functional characteristics in the human gut microbiome. *Microbiome.* 2013;1(1):2. doi: 10.1186/2049-2618-1-2.
- Lewis DD. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*. p. 212–217.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Jian M, Zhou Y, Li Y, Zhang X, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:59–65.
- Rousk J, Baath E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* 2010;4:1340–51.
- Saeys Y, Inza I, Larra naga P. A review of feature selection techniques in bioinformatics. *Oxf Bioinforma.* 2007;23(19):2507–17.
- Williamson S, Rusch D, Yooseph S, Halpern A, Heidelberg K, Glass J, Andrews-Pfannkoch C, Fadrosh D, Miller C, Sutton G, Frazier M, Venter JC. The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS Biol.* 2008;3(1).
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol.* 2010;6(2):1–13.

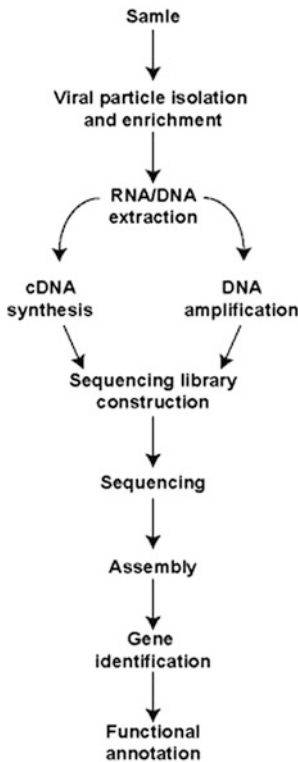
Viral Metagenome Annotation Pipeline

Hernan Lorenzi

Informatics, J. Craig Venter Institute, Rockville, MD, USA

Introduction

Viruses are the most abundant and diverse organisms on Earth, yet only a small fraction of the viral genome sequence space has been decoded. Based on analyses of environmental viral communities, it is estimated that only 1 % of the existent viral diversity has been explored. During more than a century, cultivation of viruses has remained the gold standard for virus discovery and characterization. One major limitation of this approach is that for most viral species, their hosts (predominantly microbes) are either unknown or cannot be grown in culture. Viral metagenomics (VM) circumvents this limitation by sequencing viral genetic material isolated directly from the environment. A typical viral metagenomics workflow is depicted in Fig. 1. Viral metagenomic methods have evolved significantly since their beginnings. Initially, they involved viral particle purification and enrichment from environmental samples, sharing of isolated nucleic acids followed by an optional cDNA synthesis step in the case of RNA viruses, cloning into shotgun libraries, and direct sequencing of the total DNA content by Sanger. This low-throughput approach has been used in the past for the characterization of viral communities from many different environments (Steward and Preston 2011; Bench et al. 2007). During the last decade, the advent of high-throughput sequencing technologies and the development of novel viral particle purification methods are revolutionizing the field of VM facilitating the rapid expansion of viral genome data and boosting the number of associated metagenomics publications in PubMed (Fig. 2). Among the many sequencing



Viral Metagenome Annotation Pipeline, Fig. 1 Schematic representation of a viral metagenomic workflow

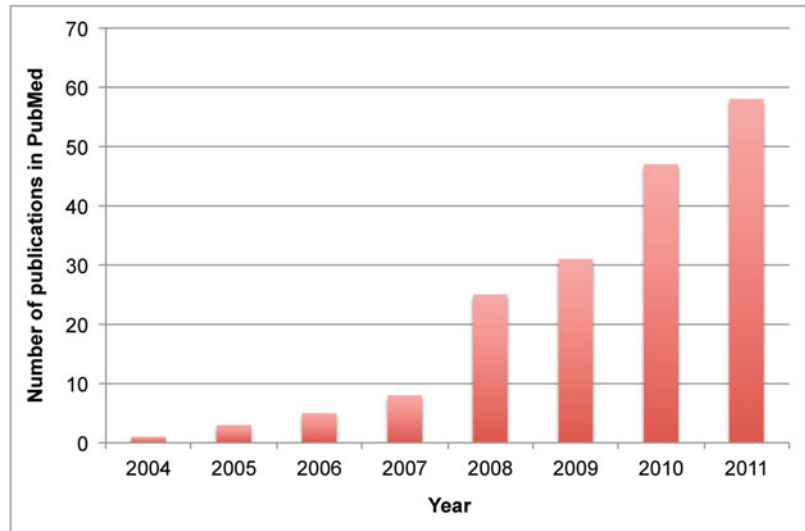
platforms currently available, Illumina and 454 FLX/titanium pyrosequencing have been the most frequently used for the characterization of VM samples. One Illumina sequencing lane generates approximately 125–150 million reads of up to 150 bp in length while one full-plate run of 454 titanium produces ~1 million reads of about 350–450 bp. A VM project usually involves two or more Illumina/454 runs, and therefore, the volume of sequence produced by these studies is in the order of several gigabases. This huge volume of sequencing data makes downstream annotation and analyses methods very challenging and computationally expensive. Therefore, it is critical to preprocess sequencing data whenever possible in order to reduce the amount of sequence to be annotated. Pre-annotation processing methods include

elimination of redundant and low-quality reads, assemblage, and viral gene identification (Fig. 1). How the sequencing data will be processed and annotated will depend on the goals of the project and characteristics of the viral community but basically can be divided into two major annotation strategies, read-based annotation (RBA) and gene-based annotation (GBA). The former approach is more straightforward and involves the identification of similarities to protein sequences or domains directly on the reads to classify them into phylogenetic or functional groups. Gene-based annotation, on the other hand, requires an optional assemblage of sequencing reads into contigs and/or scaffolds, gene identification, and functional prediction of predicted proteins. In this entry, we describe current tools, databases, and methodologies that have been developed in the past few years for RBA and GBA of viral metagenomic datasets and discuss their advantages and drawbacks.

Read-Based Annotation

Direct annotation of sequencing reads is frequently used when the goal of the VM study is to investigate and compare the type of species or gene functions that are present in one or more viral communities. In general, read-based annotation assumes that each read encodes for a single gene. Before proceeding with any annotation, it is important to preprocess sequencing reads to eliminate regions with low-quality base calls and duplicated reads. This is particularly important when working with next generation sequencing (NGS) data, since pyrosequencing and Illumina technologies have a higher error rate compared with Sanger. In particular, 454 pyrosequencing is prone to the generation of artifactual indels in regions containing homopolymers (Kunin et al. 2010; Gilles et al. 2011) while Illumina reads have a higher substitution error rate than 454 dealing better with homopolymeric regions (Minoche et al. 2011). Also, NGS platforms have a tendency to produce a significant number of

Viral Metagenome Annotation Pipeline, Fig. 2 Number of articles in PubMed about viral metagenomics during the period 2004–2011



duplicated reads, in particular when sequencing libraries are constructed from very limited quantities of starting RNA/DNA material (<10 ng). There are several programs that can be used to remove exact or near exact duplicated reads or trimming low-quality bases and vector sequences without requiring a large computer infrastructure. Some examples are BIGpre (Zhang et al. 2011), Bolger et al. 2014 (<http://www.usadellab.org/cms/index.php?page=trimmomatic>), PyroCleaner (Jerome et al. 2011), CD-HIT (Huang et al. 2010), NGS QC Toolkit (Patel and Jain 2012), LUCY (Chou and Holmes 2001), and SeqClean (Chen et al. 2007). For example, Trimmomatic is a java-based program that can run in Linux, Windows, and Mac OSX operating systems and has several different options for trimming low-quality bases and adaptor sequences from Illumina reads. BIGpre is compatible with both 454 and Illumina platforms and detects and removes redundant reads after taking sequencing errors into account and trimming low-quality reads from raw data as well. BIGpre and NGS QC Toolkit also output a number of quality stats about NGS reads that are useful to assess the presence of sequence bias and the correlation between forward and reverse reads among other tools.

Once raw sequencing reads have been processed, it is possible to proceed with the

annotation stage. Because viruses have a fast evolutionary rate, any comparison at the nucleotide level is not sensitive enough to detect similarities between reads from a studied metagenome and nucleotide databases of characterized viral genes or genomes. In consequence, all searches should be done using translated sequences. The simplest annotation approach is to compare the six-frame translations of each read against a collection of well-annotated protein databases using TBLASTN or equivalent algorithms to identify the types of viral species or functions encoded by the viral metagenome. The main advantage of RBA is that it does not involve previous gene identification or assembly of reads, processes that require some level of user expertise. Another benefit is that translation-based similarity searches are independent of gene structure and therefore may prove to be more sensitive than GBA at the time of studying viral communities whose genomes are enriched in intron-containing genes. However, RBA has several disadvantages. First, sequence similarity searches using TBLASTN or equivalent programs are computationally demanding and time consuming. Second, many databases of conserved protein domains or motifs cannot be queried using nucleotide sequences or on the fly translations. Third, when reads code for more than one gene, the molecular functions associated with the most divergent

genes on the read are usually masked by the gene with the best (lowest) e-values and hence are difficult to detect. Fourth, further characterization and phylogenetic analysis of protein families are complicated by the fact that it is difficult to generate multiple sequence alignments from evolutionary-related genes that start at different positions on their respective reads. Lastly, the higher indels rate of NGS reads, in particular in 454-derived sequences, creates artifactual translation frameshifts that can lead to an overestimation of gene family diversity and complicates the interpretation of results from protein database searches.

Gene-Based Annotation

A more thorough and efficient way to annotate sequencing datasets from viral communities is to identify protein-coding genes before carrying out any comparison against protein databases. This approach reduces considerably the amount of sequencing data to be queried and hence computing time, expands the spectrum of databases that can be searched, and simplifies the interpretation of results and further evolutionary studies.

Although GBA may involve different bioinformatics tools, databases, and cutoffs, it is usually composed of the following consecutive steps: (i) sequence assembly; (ii) protein-coding gene identification; (iii) similarity searches of predicted proteins against generic or specialized databases of characterized proteins, conserved protein domains or motifs; and (iv) functional assignments of predicted proteins following a series of predefined rules. Below, we will discuss each of these steps in more detail.

Assembling Viral Metagenomes

Metagenomic sequence assembly is a fundamental way to improve metagenomic annotation. For example, the sensitivity of both phylogenetic assignment methods based on nucleotide composition and metagenomic ab initio gene finders increases with sequence length (McHardy et al. 2007; Li 2009; Yok and Rosen 2010). Single-genome assemblers usually do not

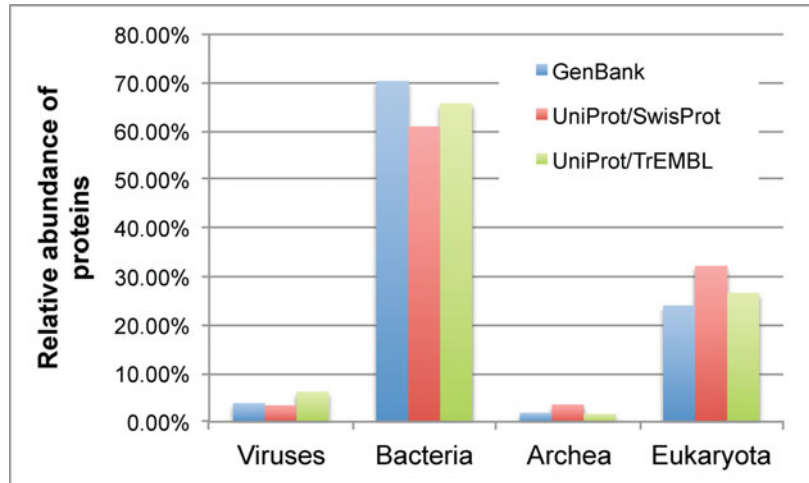
perform well on metagenomic datasets because they are not designed to handle a mixture of reads derived from different strains and species with distinct relative abundances. In this context, sequences of highly abundant species are likely misidentified as repeats in a single genome, resulting in a number of small fragmented scaffolds. There are a number of programs and websites specifically designed for generating de novo contigs and scaffolds of overlapping metagenomic NGS reads. The CAMERA website (Sun et al. 2011) offers a meta-assembly procedure for 454 reads which consist of running a number of single-genome assemblers with carefully optimized parameters on the metagenomic dataset, then it collects all the resulting contigs and assigns quality scores by consensus analysis, and finally, it uses an adaptation of *phrap* (<http://www.phrap.org>) to reassemble the contigs based on computed quality scores. There are also a number of metagenome-specific de novo assemblers, such as MetaVelvet (Namiki et al. 2012), Meta-IDBA (Peng et al. 2011), IDBA-UD (Peng et al. 2012), and Genovo (Laserson et al. 2011). These programs deal better with a mixed population of species with different abundances compared to single-genome de novo assemblers (Namiki et al. 2012) and seem to reduce the number of chimeric contigs. Also, depending on the species diversity of the metagenome, some of these programs may perform differently (Namiki et al. 2012), and therefore, it is better to try a variety of assembly programs before starting to work on the annotation of a particular dataset.

Ab Initio Gene Identification

Gene features in viral genomes are strongly dictated by the genetic characteristics of their host. Thus, bacterial viruses, or bacteriophages, are mostly composed of single-exon genes while eukaryotic-infecting viruses may contain genes with more than one exon. In spite of this property, the majority of genes encoded by viral genomes do not have introns. Therefore, there are a number of gene finders that are suitable for the ab initio identification of viral genes on either NGS reads or assembled sequences, although

Viral Metagenome Annotation Pipeline,

Fig. 3 Relative number of viral, bacterial, archaeal, and eukaryotic proteins in GenBank, UniProt/Swiss-Prot, and UniProt/TrEMBL. Numbers are relative to the total number of protein in each database



none of them have been specifically developed for viral metagenomic samples. Two of the most widely used gene finders are MetaGeneAnnotator (Noguchi et al. 2008) and FragGeneScan (Rho et al. 2010). MetaGeneAnnotator integrates statistical models from prophage, bacterial and archaeal genes, and ribosomal-binding sites, and it also uses a self-training model from input sequences for making predictions. FragGeneScan incorporates sequencing error models and codon usage information in a hidden Markov model (HMM) that improves the prediction of protein-coding genes in NGS reads and assemblies. FragGeneScan is able to compensate for artifactual frameshifts in pyrosequencing reads caused by the higher frequency of indels at homopolymeric regions. An alternative strategy to the identification of genes with gene finders is using naïve six-frame translations (NSFT) that identify each possible ORF of at least 80 nt of length. In this case, 5' and 3' ends of reads can be considered as start and stop codons, respectively, to also incorporate partial genes truncated by read ends. In those cases where there are two or more overlapping ORFs, it is possible to analyze all of them or select candidate genes based on their properties: length, dn/ds ratio, similarity at the protein level, etc. An alternative to this approach is to combine the results of NSFT with gene predictions from FragGeneScan or MetaGeneAnnotator and pick the longest predicted gene per region.

Functional Annotation of Predicted Genes

Functional predictions of protein sequences are usually done in two consecutive steps: (A) similarity searches against very well-curated protein databases and (B) functional assignments based on database hits. A fundamental problem in functional annotation of viral genes is how to assign functional roles to their encoding proteins when viral sequences are highly divergent from those already present in well-annotated protein databases. To make the situation even more complicated, proteins of viral origin represent a tiny fraction of the proteins deposited in public repositories (Fig. 3). In consequence, in a typical VM project, only a very small proportion of viral peptides give significant hits ($e\text{-value} \leq 1 \times 10^{-5}$) against protein databases. Therefore, protein database searches have to be complemented with other bioinformatics tools to increase the number of functionally predicted viral proteins. In this section we describe a strategy for functional annotation of viral metagenomic datasets as implemented in the Viral MetaGenomic Annotation Pipeline (VMGAP) at the J. Craig Venter Institute (Lorenzi et al. 2011). This pipeline makes use of databases of conserved protein domains, mobile genetic elements, and environmental peptides to improve the sensitivity and quality of the annotation. The first step in the VMGAP is to perform several similarity searches

between the VM peptides to be annotated and the following databases:

(i) *BLASTP searches against public nonredundant protein databases*

Several generic nonredundant protein databases can be used for functional assignment of viral proteins: GenBank NR, UniProtKB (UniProt Consortium 2012), and UniRef (Suzek et al. 2007). UniProtKB consists of two databases, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. Protein records in UniProtKB/Swiss-Prot are annotated and reviewed by a curator, while entries in UniProtKB/TrEMBL are automatically annotated and classified. UniRef is a group of nonredundant protein databases derived from clustering UniProtKB entries at different percentages of identity. Thus, UniRef100 combines identical complete and fragmented protein sequences from any organism into a single UniRef entry. UniRef90 and UniRef50 are built by clustering UniRef100 sequences at the 90 % or 50 % sequence identity levels. One of the main advantages of using a clustered protein database such as UniRef90 and UniRef50 is that they significantly reduce the time required for similarity searches and improve detection of distant relationships, since all closely related proteins are collapsed in a single representative sequence (Suzek et al. 2007).

(ii) *BLASTP searches against the ACLAME database*

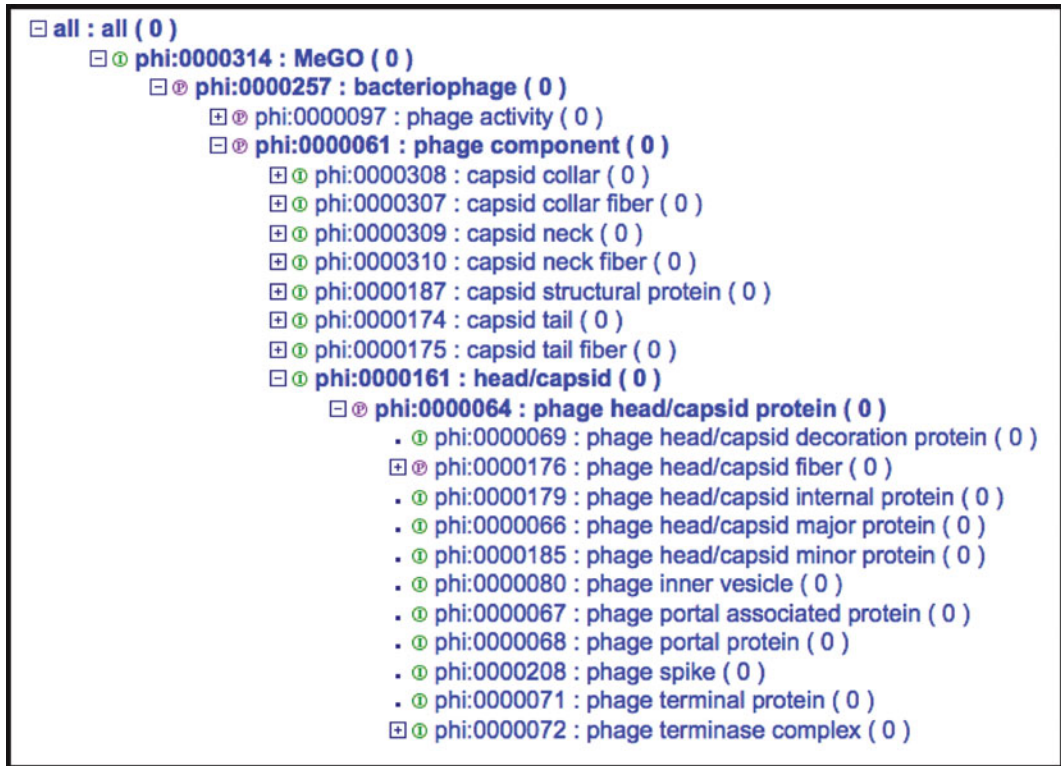
The ACLAME database is a repository of mobile genetic elements such as bacteriophages, plasmids, and transposons (Leplae et al. 2010). Entries in ACLAME are organized into families based on their sequence similarity and function. Those families with more than three members are manually annotated with functional assignments using gene ontology terms from GO (Shoop et al. 2004) and MeGO (Toussaint et al. 2007). MeGO is an ontology developed by ACLAME to describe biological functions, processes, and components specific to mobile genetic elements that are not present in the GO database (for an example see Fig. 4).

(iii) *BLASTP searches against GenBank environmental nonredundant database*

An intriguing aspect of VM is the fact that the majority of viral predicted proteins do not share similarity with any known sequence. This collection of unknown proteins, which are usually discarded as “junk” sequences, may represent a formidable source for the discovery of new viral species. One way to exploit these protein sequences is to compare them against the proteins from other metagenomic datasets to gain some insight about the viral entities that are shared between them. GenBank environmental nonredundant database (env_nr) is a collection of all the protein sequences derived from metagenomic projects deposited in GenBank.

(iv) *HMM searches against PFAM database*

PFAM is a database of hidden Markov models of conserved protein domains (Punta et al. 2012). Because these domains are usually associated with a particular molecular function or protein family and evolve at a lower pace compared to other protein regions, they are excellent tools for identifying functional domains in highly divergent protein sequences as the ones from viruses. PFAM HMM searches can be run with the HMMER2 suite of programs (Eddy 2011) in two different modes, global or local, allowing for total or partial alignments of the HMMs to the queried protein sequences, respectively. If gene predictions are done on reads, it is expected a high proportion of partial (truncated) proteins. In that case, local HMM searches are a more sensitive approach. HMM searches using global alignments are more specific than locals and perform better on complete proteins. However, even in assembled VM datasets the proportion of truncated genes is very high, since assemblies tend to be very partitioned. Recently, PFAM released a new generation of HMM models compatible with a new development of the HMMER package, the HMMER3. These HMMs can only be run in local mode but have similar specificity and



Viral Metagenome Annotation Pipeline, Fig. 4 Example of MeGO terms as they appear in ACLAME using AmiGO

sensitivity to those from the two PFAM HMMER2 models (local and global) combined. HMMER3 uses a faster algorithm and hence is a better choice for performing HMM searches on VM protein datasets.

(v) *RPS-BLAST against NCBI-CDD database*

The NCBI Conserved Domain Database (CDD) (Marchler-Bauer et al. 2011) is a compendium of position-specific scoring matrices (PSSMs) representing conserved protein domains, protein families, and super-families gathered from SMART (Letunic et al. 2012), COG database (Tatusov et al. 2003), NCBI-curated protein domains (Sayers et al. 2012), and PFAM. In spite of having some overlap with PFAM HMMs, PSSMs derived from PFAM domains do not behave exactly the same as their HMMs counterparts, and therefore, they complement each other. CDD-PSSMs are usually associated with a molecular function or represent

signatures of protein families and therefore provide useful functional information. Since PSSMs are BLAST scoring matrices specific of conserved protein domains or motifs, their use gives better sensitivity than regular BLASTP at the time of detecting these domains on more divergent proteins.

(vi) *Additional bioinformatic tools for functional annotation*

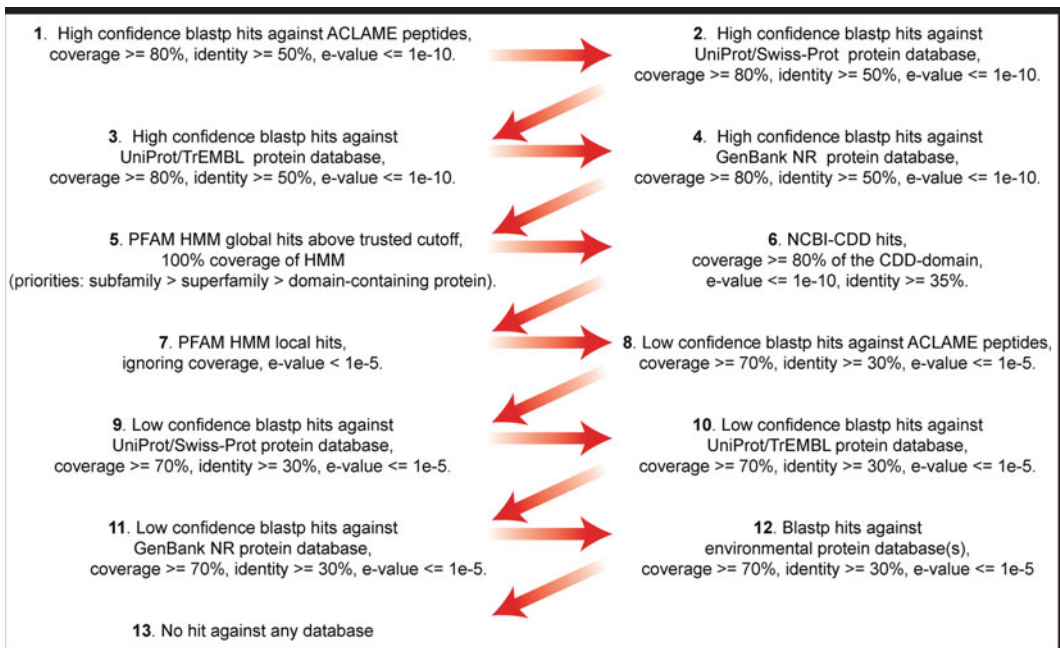
Because a significant proportion of the proteins encoded by viral metagenomes are unknown, it is useful to take advantage of in silico protein-signal prediction tools that could provide hints about their putative function. An important first step toward understanding the biological role of unknown viral proteins is determining their subcellular localization while infecting their host. A set of popular protein localization prediction programs has been developed for the identification of protein signals that

dictate the subcellular destination of peptides: SignalP (Petersen et al. 2011), ChloroP (Emanuelsson et al. 1999), TargetP (Emanuelsson et al. 2000), and Krogh et al. 2001 (<http://www.cbs.dtu.dk/services/TMHMM/>). None of these programs are specifically designed for viral genes. However, once in the host, viruses can use prokaryotic/eukaryotic signals to target their own proteins to defined subcellular locations. SignalP 4.0 uses a neural network-based method to predict signal peptides from gram positive, gram negative, and eukaryotic peptides, and it has been recently improved to distinguish between signal peptides and transmembrane domains located near the N-terminus of proteins. ChloroP also uses a neural network approach to predict chloroplast transit peptides, and therefore, it might be useful for the functional annotation of viruses that infect plants. TargetP is a program that predicts the subcellular location of eukaryotic proteins. The location assignment is based on the presence of any of the following N-terminal signals:

chloroplast transit peptide, mitochondrial targeting peptide, or signal peptide. TMHMM is a program that predicts transmembrane domains based on HMM searches. Each of these programs outputs a *p-value* that can be used to select highly significant predictions.

Functional Assignments to VM Proteins Based on Annotation Rules

The second stage of functional annotation is the processing of the functional information produced from database searches to generate a file containing a summary of the functional characteristics (product names, GO/MeGO terms, EC numbers, etc.) for each viral peptide. Each of the evidences generated by the analyses described above is more or less informative or accurate depending on the origin of the VM, the queried databases, and the programs used. Therefore, the best approach is to apply a series of hierarchical rules to prioritize the use of a certain piece of evidence over another based on how trustful and useful that evidence is. Figure 5 shows a potential hierarchical scheme similar to the one used as



Viral Metagenome Annotation Pipeline, Fig. 5 Hierarchical scheme for functional annotation of viral proteins

part of the VMGAP at the JCVI. Under this scheme, hits against ACLAME database are the highest ranked supporting evidence for functional assignments. Hence, any viral protein that hits an ACLAME entry with an e-value $\leq 1 \times 10^{-10}$, with at least 50 % identity spanning 80 % of the length of the shortest sequence, will automatically inherit the functional annotation associated with that particular ACLAME peptide. The second, third, and fourth tiers of evidence correspond respectively to highly significant BLASTP hits against UniProt/Swiss-Prot (US), UniProt/TrEMBL (UT), and GenBank NR (GB). US has a higher hierarchy than UT and GB because entries in US are manually curated. BLASTP hits against UT have a higher priority than GB hits because UT annotation is usually more comprehensive compared with GB. Ranked fifth and sixth are hits against almost complete PFAM HMMs and CDD-PSSMs, respectively. PFAM hits are more reliable than CDD hits because they can be selected based on their e-value but also using pre-calibrated domain-specific bit score cutoffs named trusted cutoff. Any protein that hits a PFAM HMM with a bit score above its trusted cutoff is considered to contain that particular domain with a very high level of confidence. CDD domains, on the other hand, are being selected just based on the e-value of the RPS-BLAST hit and coverage of the CDD domain, and hence, hits are less reliable compared with tier five. Local hits against PFAM HMM domains with e-values $\leq 1 \times 10^{-5}$ are ranked seventh below CDD hits. Because these hits span just a portion of the HMM model, they are solely selected by their e-value and not by their bit score. Tiers eight to 11 look for less reliable hits against ACLAME, US, UT, and GB databases in that order using more permissive cutoffs (e-value $\leq 1 \times 10^{-5}$; coverage ≥ 70 %, identity ≥ 30 %) compared to tiers 1–4 (e-value $\leq 1 \times 10^{-10}$; coverage ≥ 80 %, identity ≥ 50 %). Ranked 12th are BLASTP hits against environmental protein databases, such as GenBank env_nr, with e-values of at least 1×10^{-5} . Entries in environmental protein databases are likely to lack any functional annotation. However, associated metadata such as geographic

location, body site, type of disease, etc., may still provide some clues about the biology of the viruses present in the VM sample. Finally, if the viral protein does not contain a database hit that falls within any of the first 12 tiers, then it is considered an unknown protein.

Note that the rules described above can be further improved by, for example, the incorporation of results from subcellular localization predictions (TargetP, SignalP, ChloroP, and TMHMM) between tiers 12 and 13 or any other functional analysis.

Applying the rules described above, it is possible to assign product names, EC numbers, and GO/MeGO terms to predicted proteins from the VM sample. For example, if a viral predicted protein has a hit against a peptide from ACLAME database above the cutoffs from tier 1, then it can inherit the product name as well as the GO or MeGO terms associated with that particular ACLAME entry. UniProt entries, in particular from US, are also a very good source of product names, EC numbers, and GO terms. However, these assignments should be done from high confidence hits only.

Web Resources for Functional Annotation of VM Datasets

Currently, there are a number of publicly available bioinformatics tools that can be used for the structural (gene identification) and functional annotation of viral metagenomes. MG-RAST (Glass et al. 2010; Meyer et al. 2008) is a popular web resource able to perform structural and functional annotations on both NGS reads and assembled metagenomic data. One main advantage is that all computes are run by the MG-RAST server, and therefore, the user is not required to have a big computer infrastructure. It also handles Illumina and 454 reads and provides several read preprocessing tools such as elimination of duplicated or contaminated reads and deletion of low-quality sequences and short reads. Structural annotation is carried out either on reads or assemblies using FragGeneScan while functional annotation is being done by

similarity searches against a protein nonredundant database that compiles the following public protein databases: GenBank NR, KEGG (Tanabe and Kanehisa 2012), IMG (Markowitz et al. 2012), InterPro (Hunter et al. 2012), PATRIC (Gillespie et al. 2011), Dwivedi et al. 2012 (<http://www.phantome.org/>), RefSeq (Pruitt et al. 2012), SEED (DeJongh et al. 2007), UniProt/Swiss-Prot, UniProt/TrEMBL, COG (Tatusov et al. 2003), GO, KO (Mao et al. 2005), and eggNOG (Powell et al. 2012). Among these databases is Phantome, a protein database of complete phage genomes manually curated by experts using a subsystem approach (Overbeek et al. 2005). Another nice feature of MG-RAST is that it allows the comparison among the annotated VM samples provided by the user and the more than 10,000 metagenomic datasets that are publicly available at the MG-RAST server.

Another useful web resource is CAMERA (Sun et al. 2011), which allows the construction of customized workflows for the analysis of external metagenomic data. Among the many bioinformatic tools available are an assembly pipeline for 454 reads, protein clustering with CD-HIT, clustering of duplicated 454 reads, gene predictions based on different gene finders, and a general pipeline for annotation of metagenomic datasets.

References

- Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE. Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol.* 2007;73(23):7629–41. 2168038.
- Bolger AM1, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014; doi:10.1093/bioinformatics/btu170
- Chen YA, Lin CC, Wang CD, Wu HB, Hwang PI. An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics.* 2007;8:416. 2194723.
- Chou HH, Holmes MH. DNA sequence quality trimming and vector removal. *Bioinformatics.* 2001;17(12): 1093–104.
- DeJongh M, Formsa K, Boillot P, Gould J, Rycenga M, Best A. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinforma.* 2007;8:139. 1868769.
- Dwivedi B, Schmieder R, Goldsmith DB, Edwards RA, Breitbart M. PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. *BMC Bioinformatics.* 2012;13:37.
- Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7(10):e1002195. 3197634.
- Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 1999;8(5):978–84. 2144330.
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol.* 2000;300(4):1005–16.
- Gilles A, Meglec E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. *BMC Genomics.* 2011;12:245. 3116506.
- Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun.* 2011;79(11):4286–98. 3257917.
- Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc.* 2010; 2010(1):pdb prot5368.
- Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* 2010;26(5):680–2. 2828112.
- Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 2012;40(Database issue):D306–12. 3245097.
- Jerome M, Noirot C, Klopp C. Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Res Notes.* 2011;4:149. 3117718.
- Krogh A1, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001;305(3):567–80.
- Kunin V, Engelbrektsen A, Ochman H, Hugenholtz P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol.* 2010;12(1):118–23.
- Laserson J, Jojic V, Koller D. Genovo: de novo assembly for metagenomes. *J Comput Biol.* 2011;18(3):429–43.
- Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a classification of mobile genetic elements, update 2010. *Nucleic Acids Res.* 2010;38(Database issue): D57–61. 2808911.
- Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 2012;40(Database issue):D302–5. 3245027.

- Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinforma.* 2009;10:359. 2774329.
- Lorenzi HA, Hoover J, Inman J, Safford T, Murphy S, Kagan L, Williamson SJ. The Viral MetaGenome Annotation Pipeline (VMGAP): an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand Genomic Sci.* 2011;4(3):418–29. 3156399.
- Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics.* 2005;21(19):3787–93.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 2011;39(Database issue):D225–9. 3013737.
- Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* 2012;40(Database issue):D123–9. 3245048.
- McHardy AC, Martin HG, Tsirigos A, Hugenholz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods.* 2007;4(1):63–72.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 2008;9:386. 2563014.
- Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 2011;12(11):R112. 3334598.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012;40(20):e155.
- Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 2008;15(6):387–96. 2608843.
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, De Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33(17):5691–702. 1251668.
- Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One.* 2012;7(2):e30619. 3270013.
- Peng Y, Leung HC, Yiu SM, Chin FY. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics.* 2011;27(13):i94–101. 3117360.
- Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420–8.
- Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods.* 2011;8(10):785–6.
- Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* 2012;40(Database issue):D284–9. 3245133.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012;40(Database issue):D130–5. 3245008.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. The Pfam protein families database. *Nucleic Acids Res.* 2012;40(Database issue):D290–301. 3245129.
- Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010;38(20):e191. 2978382.
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvermin V, Church DM, Dicuccio M, Federhen S, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2012;40(Database issue):D13–25. 3245031.
- Shoop E, Casaes P, Onsongo G, Lesnett L, Petursdottir EO, Donkor EK, Tkach D, Cosimini M. Data exploration tools for the gene ontology database. *Bioinformatics.* 2004;20(18):3442–54.
- Steward GF, Preston CM. Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. *Virology.* 2011;8:287. 3128862.
- Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* 2011;39(Database issue):D546–51. 3013694.
- Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics.* 2007;23(10):1282–8.
- Tanabe M, Kanehisa M. Using the KEGG database resource. *Curr Protoc Bioinform.* 2012. Chapter 1: Unit1 12.
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 2003;4:41. 222959.
- The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2012;40(Database issue):D71–5. 3245120.

- Toussaint A, Lima-Mendez G, Leplae R. PhiGO, a phage ontology associated with the ACLAME database. *Res Microbiol.* 2007;158(7):567–71.
- Yok N, Rosen G. Benchmarking of gene prediction programs for metagenomic data. *Conf Proc IEEE Eng Med Biol Soc.* 2010;2010:6190–3.
- Zhang T, Luo Y, Liu K, Pan L, Zhang B, Yu J, Hu S. BIGpre: a quality assessment package for next-generation sequencing data. *Genomics Proteomics Bioinforma.* 2011;9(6):238–44.

Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach

Jian Yang

MOH Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS&PUMC), Beijing, People's Republic of China

Synonyms

Metagenomic detection of viral agents in clinical samples

Definition

Viral pathogens in clinical samples here refer to the human viruses isolated (or been discovered) in clinical samples that associate with known human diseases. Viruses from environmental samples or nonhuman biological samples as well as the large amount of commensal viruses in human virome are not discussed in this entry.

Introduction

Viral diseases continue to threaten public health and medicine in the twenty-first century by causing significant disease burden globally. Accurate and rapid identification of the viral agents is the key step towards better control and prevention of

the associate diseases. Traditional techniques for virus discovery such as cultivation-, morphology-, serology-, and immunology-based methods have contributed significantly to the identification of most important viral pathogens during the last century. In addition, modern molecular methods such as PCR and microarray also play more and more important roles in clinical virology practices in the past decade. The newly emerged metagenomic-based method is a particularly powerful approach for virus identification since genetic materials can be analyzed directly from clinical samples, bypassing the need for culturing, cloning, or a priori knowledge of what viruses may be present. The recent advent of next-generation sequencing technologies (NGS), which have dramatically improved the speed and cost-effectiveness of sequencing, fueled the clinical application of metagenomic method for viral diagnosis. Herein, we summarized the most recent studies that have successfully identified viral pathogens from clinical samples by using the NGS-based metagenomic approach.

Viral Pathogens in Diseased Human Tissues

The astonishing power of NGS-based metagenomic approach for clinical diagnosis was first illustrated by two remarkable studies reported in 2008. Merkel cell carcinoma (MCC) is a rare but aggressive human skin cancer that typically affects elderly and immune-suppressed individuals. By high-throughput metagenomic sequencing of the cDNA library of tumor tissues and digital subtraction of human transcriptome, Feng et al. identified a novel polyomavirus that may be a contributing factor in the pathogenesis of MCC (Feng et al. 2008). Another study used a similar strategy to discover a new arenavirus that likely associated with a cluster of fatal transplant-associated diseases, after many traditional and molecular assays including culture, PCR, and oligonucleotide microarray had failed to identify any potential infectious agents (Palacios et al. 2008). The success of

NGS-based metagenomic approach in clinical diagnosis provided a new route for the identification of pathogens from clinical samples and was believed to be the herald of a breakthrough in the field of pathogen discovery (MacConaill and Meyerson 2008). Indeed, the metagenomic approaches were further applied to screen post-mortem tissues for potential viral agents by the same group from Columbia University, and they successively identified a new hemorrhagic fever-associated arenavirus named Lujo virus from Southern Africa and an astrovirus as a causative agent for encephalitis in a patient with agammaglobulinemia (Briese et al. 2009; Quan et al. 2010).

Viral Pathogens in Fecal Samples

Due to the relatively feasible accessibility, stool specimens are the most intensively investigated clinical samples by using metagenomic approaches to date (Table 1). Diarrhea is one of the major infectious causes of death worldwide, but about 40 % of the diarrhea cases are of unknown etiology. Metagenomic approaches were recently used by different groups to screen stool samples from diarrhea patients, and many known eukaryotic viruses as well as several new viral species/genus were discovered, including a novel gyrovirus species GyV3 and a potential new parvovirus genus (Nakamura et al. 2009; Phan et al. 2012a, b). The same group from Blood Systems Research Institute also analyzed fecal samples from 35 South Asian children with nonpolio acute flaccid paralysis and identified a large number of known enteric viruses as well as several new viral species (Victoria et al. 2009). But numerous viruses were also detectable in the samples from six healthy contacts of the patients. In addition, two groups dedicated to the unknown etiology of gastrointestinal illness with the metagenomic approach and revealed a new astrovirus VA1 and a novel picornavirus named klassevirus, respectively (Finkbeiner et al. 2009; Greninger et al. 2009). But further studies are still required to fully characterize these newly identified potential viral pathogens.

Viral Pathogens in Respiratory Specimens

The respiratory tract is one of the most heavily exposed organs in human body to microorganisms. Therefore, the new NGS-based metagenomic approaches were extensively used by different studies to identify viral agents from patients with respiratory infections (Table 1). However, the quantities of samples from respiratory tract, either swabs or aspirates, are much lower than those of fecal samples mentioned above. Detection of potential viral agents from the minute respiratory samples using the metagenomic approach is therefore particularly challenging and tricky. All of the aforementioned studies targeting human tissues or stools employed the Roche/454 platform for metagenomic sequencing as it produced longer reads (but lower overall throughput) than other NGS platforms. Nevertheless, three of the five published studies working on respiratory specimens tried the Illumina platform instead. Actually, the ultrahigh throughput of Illumina platform can largely compensate the disadvantage in reads length as compared to the Roche/454 platform (Yang et al. 2011). In addition, a simulation study showed that the Illumina technology was more sensitive than the Roche/454 technology in detection viruses from biological samples (Cheval et al. 2011). Indeed, using only 36 bp reads, our group identified seven known respiratory viral agents from 16 clinical samples, including a case of coinfection that would have been misdiagnosed by conventional PCR assays (Yang et al. 2011). Moreover, when utilizing the paired-end sequencing strategy, the novel enterovirus 109 was readily identified from a case of acute respiratory illness in a Nicaraguan child (Yozwiak et al. 2010), whereas 90 % of the viral genome of H1N1 influenza A virus can even be assembled *de novo* (Greninger et al. 2010).

Viral Pathogens in Blood Samples

Viral hemorrhagic fever (VHF) is a severe illness characterized by high fever and bleeding, which

Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach, Table 1 Selected clinical viral diagnosis reports using a metagenomic approach based on next-generation sequencing technologies

Sample types	Related diseases	Viral pathogens detected	Sequencing platform	Reference
Diseased human tissues	Tumor tissues	Merkel cell carcinoma (a type of human skin cancer)	Roche/454	Feng et al. 2008
	Postmortem tissues	Fatal transplant-associated diseases	Roche/454	Palacios et al. 2008
	Postmortem tissues and sera	Hemorrhagic fever	Roche/454	Briese et al. 2009
	Biopsy and postmortem tissues	Encephalitis	Roche/454	Quan et al. 2010
Fecal samples	Stools	Diarrhea	Roche/454	Nakamura et al. 2009
	Stools	Nonpolio acute flaccid paralysis	Sanger, Roche/454	Victoria et al. 2009
	Stools	Pediatric gastroenteritis	Roche/454	Greninger et al. 2009
	Stools	Acute gastroenteritis	Sanger, Roche/454	Finkbeiner et al. 2009
	Stools	Diarrhea	Roche/454	Phan et al. 2012a
	Stools	Pediatric acute diarrhea	Roche/454	Phan et al. 2012b
Respiratory specimens	Nasopharyngeal aspirates	Influenza	Roche/454	Nakamura et al. 2009
	Nasopharyngeal swabs	Acute pediatric respiratory illness	Roche/454	Yozwiak et al. 2010
	Nasopharyngeal swabs	Influenza	Roche/454	Greninger et al. 2010
	Nasopharyngeal aspirates	Acute lower respiratory tract infections	Roche/454	Yang et al. 2011
	Nasopharyngeal aspirates	Acute lower respiratory tract infections	Roche/454	Lysholm et al. 2012
Blood samples	Blood	Hemorrhagic fever	Roche/454	Towner et al. 2008
	Sera	Fever, thrombocytopenia, and leukopenia syndrome	Roche/454	Xu et al. 2011
	Sera	Hemorrhagic fever	Roche/454	McMullan et al. 2012
	Sera	Dengue-like disease	Roche/454	Yozwiak et al. 2012

may be caused by a number of viruses. Recently, a group from the Centers for Disease Control and Prevention dedicated to screen viral agents in blood samples from VHF patients in Uganda

using the Roche/454-based metagenomic approach. They successfully identified a new Ebola virus likely responsible for a large hemorrhagic fever outbreak in western Uganda

(Towner et al. 2008). In another study on the suspected hemorrhagic fever endemic in northern Uganda, using the same strategy, they not only recognized yellow fever virus but also generated 98 % of the virus genome sequence, which facilitated the follow-up phylogenetic analyses (McMullan et al. 2012). The Illumina platforms are also employed for the detection of viral pathogens in blood samples by using a metagenomic approach (Table 1). During a tick-transmitted-like outbreak of fever, thrombocytopenia, and leukopenia syndrome in China, most patients are tested negative for the former-suspected human granulocytic anaplasmosis. Hence, a metagenomic approach based on paired-end Illumina sequencing was applied to screen potential viral agents from the sera of patients, and a novel bunyavirus was successfully identified (Xu et al. 2011). In addition, the novel virus was confirmed to present in 78 % of the acute-phase serum samples by further RT-PCR testing.

Summary

Since the first introduce in 2008, we have witnessed the emergence and extensive applications of the NGS-based metagenomic approach as a powerful tool in diagnostic virology. The intrinsic properties of metagenomics provide the method prominent advantages in speed and sensitivity for parallel screening of known viral pathogens as well as detection of new unexpected viral agents in clinical samples. With the continuous development and improvement of high-throughput sequencing technologies, the metagenomic approach will probably become an essential diagnostic method in clinical routines.

However, in current stage, several issues should be kept in mind for the application of the metagenomic approach in viral diagnostic practices. First, the selection of different NGS platforms will be critical to both preceding sample nucleotides preparation and further sequence data analyses. Though the majority of published applications used Roche/454 platform, the Illumina technology is increasingly employed in most

recent studies as the higher throughput do offers greater sensitivity as compared with the former. Second, differ from traditional methods the metagenomic approach rely heavily on subsequent bioinformatics data analyses, which can be very tricky particularly in case of detection potential novel viruses. Lacking of standard protocols for metagenomic data analysis has hampered the further extensive applications of metagenomic approach in the future. Third, results from metagenomic approach only indicate the presence of given viruses in the clinic samples screened, and it cannot directly deduce that the viruses identified are responsible for the human diseases investigated. Hence, the biological and medical interpretations of metagenomic results may require further evidences from epidemiology, morphology, immunology, etc.

Cross-References

- ▶ [Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing](#)
- ▶ [Viral MetaGenome Annotation Pipeline](#)

References

- Briese T, Paweska JT, McMullan LK, et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog.* 2009;5:e1000455.
- Cheval J, Sauvage V, Frangeul L, et al. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *J Clin Microbiol.* 2011;49:3268–75.
- Feng H, Shuda M, Chang Y, et al. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science.* 2008;319:1096–100.
- Finkbeiner SR, Li Y, Ruone S, et al. Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *J Virol.* 2009;83:10836–9.
- Greninger AL, Chen EC, Sittler T, et al. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One.* 2010;5:e13381.
- Greninger AL, Runckel C, Chiu CY, et al. The complete genome of klassevirus – a novel picornavirus in pediatric stool. *Virology.* 2009;6:82.

- Lysholm F, Wetterbom A, Lindau C, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One*. 2012;7:e30875.
- MacConaill L, Meyerson M. Adding pathogens by genomic subtraction. *Nat Genet*. 2008;40:380–2.
- McMullan LK, Frace M, Sammons SA, et al. Using next generation sequencing to identify yellow fever virus in Uganda. *Virology*. 2012;422:1–5.
- Nakamura S, Yang CS, Sakon N, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One*. 2009;4:e4219.
- Palacios G, Druce J, Du L, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med*. 2008;358:991–8.
- Phan TG, Li L, O’Ryan MG, et al. A third gyrovirus species in human faeces. *J Gen Virol*. 2012a;93:1356–61.
- Phan TG, Vo NP, Bonkoungou IJ, et al. Acute diarrhea in West African children: diverse enteric viruses and a novel parvovirus genus. *J Virol*. 2012b;86:11024–30.
- Quan PL, Wagner TA, Briese T, et al. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerg Infect Dis*. 2010;16:918–25.
- Towner JS, Sealy TK, Khristova ML, et al. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog*. 2008;4:e1000212.
- Victoria JG, Kapoor A, Li L, et al. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol*. 2009;83:4642–51.
- Xu B, Liu L, Huang X, et al. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog*. 2011;7:e1002369.
- Yang J, Yang F, Ren L, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol*. 2011;49:3463–9.
- Yozwiak NL, Skewes-Cox P, Gordon A, et al. Human enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua. *J Virol*. 2010;84:9047–58.
- Yozwiak NL, Skewes-Cox P, Stenglein MD, et al. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis*. 2012;6:e1485.

List of Entries

- A 123 of Metagenomics
A De Novo Metagenomic Assembly Program for Shotgun DNA Reads
Ab Initio Gene Identification in Metagenomic Sequences
AbundanceBin, Metagenomic Sequencing
Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads
All-Species Living Tree Project
antiSMASH
Approaches in Metagenome Research: Progress and Challenges
Arbuscular Mycorrhizal Fungi Assemblages in Chernozems
Bacterial Diversity in Tree Canopies of the Atlantic Forest
Bacteriocin Mining in Metagenomes
Binning Sequences Using Very Sparse Labels Within a Metagenome
Biological Treasure Metagenome
Carbohydrate-Active Enzymes Database, Metagenomic Expert Resource
Challenge of Metagenome Assembly and Possible Standards
CLUSEAN, Overview
Computational Approaches for Metagenomic Datasets
Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes
Culture Collections in the Study of Microbial Diversity, Importance
Culturing
Customizable Web Server for Fast Metagenomic Sequence Analysis
DACTAL
Diversity and Distribution of Marine Microbial Eukaryotes
DNA Methylation Analysis by Pyrosequencing
Environmental Shaping of Codon Usage and Functional Adaptation Across Microbial Communities
Evaluating Putative Chimeric Sequences from PCR-Amplified Products
Extended Local Similarity Analysis (eLSA) of Biological Data
Extraction Methods, Variability Encountered in Extradiol Dioxygenases Retrieved from the Metagenome
Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences
Fosmid System
FragGeneScan: Predicting Genes in Short and Error-Prone Reads
FR-HIT Overview
Functional Metagenomics of Bacterial-Cell Crosstalk
Functional Metagenomics of Human Intestinal Microbiome β -Glucuronidase Activity
Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing
Genome Atlases, Potential Applications in Study of Metagenomes
Genome Portal, Joint Genome Institute
Genome-Based Studies of Marine Microorganisms
GeoChip-Based Metagenomic Technologies for Analyzing Microbial Community Functional Structure and Activities
GHOSTM
Horizontal Gene Transfer and Bacterial Diversity

- Host-Virus Interaction: From Metagenomics to Single-Cell Genomics
- Human Gut Microbial Genes by Metagenomic Sequencing
- Human Oral Microbiome Database (HOMD) Insights into Environmental Microbial Denitrification from Integrated Metagenomic, Cultivation, and Genomic Analyses
- Integrated Database Resource for Marine Ecological Genomics
- Integrans as Repositories of Genetic Novelty
- IPRStats, Overview
- I-rDNA and C16S: Identification and Classification of Ribosomal RNA Gene Fragments
- KEGG and GenomeNet, New Developments, Metagenomic Analysis
- Krona: Interactive Metagenomic Visualization in a Web Browser
- Lateral Gene Transfer and Microbial Diversity
- Lessons Learned from Simulated Metagenomic Datasets
- MEMOSys: Platform for Genome-Scale Metabolic Models
- MetaBin
- MetaBioME
- MEtaGenome ANalyzer (MEGAN): Metagenomic Expert Resource
- Metagenome of Acidic Hot Spring Microbial Planktonic Community: Structural and Functional Insights
- Metagenomes: 23S Sequences
- Metagenomic Analysis of Bile Salt Hydrolases in the Human Gut Microbiome
- Metagenomic by RAPD Profiling
- Metagenomic Potential for Understanding Horizontal Gene Transfer
- Metagenomic Research: Methods and Ecological Applications
- Metagenomics Potential for Bioremediation
- Metagenomics, Metadata, and Meta-analysis
- MetaRank: Ranking Microbial Taxonomic Units or Functional Groups for Comparative Analysis of Metagenomes
- METAREP, Overview
- MetaTISA: Metagenomic Gene Start Prediction with
- Metaxa, Overview
- Microbial Diversity, Bar-Coding Approaches
- Microbial Ecology in the Age of Metagenomics: An Introduction
- Microbial Ecosystems, Protection of
- Mining Metagenomic Datasets for Antibiotic Resistance Genes
- Mining Metagenomic Datasets for Cellulases
- Mock Community Analysis
- Molecular Ecological Network of Microbial Communities
- Monitoring Lactic Acid Bacterial Diversity During Shochu Fermentation
- MRL and SuperFine+MRL
- New Computational Methodologies to Understand Microbial Diversity
- New Method for Comparative Functional Genomics and Metagenomics Using KEGG MODULE
- Next-Generation Sequencing for Metagenomic Data: Assembling and Binning
- NGS QC Toolkit: A Platform for Quality Control of Next-Generation Sequencing Data
- Novel Alkalistable and Thermostable Xylanase-Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome
- Novel Approaches to Pathogen Discovery in Metagenomes
- Nucleotide Composition Analysis: Use in Metagenome Analysis
- Open Resource Metagenomics
- Phylogenetics, Overview
- PhyloPythia(S)
- Plasmid Capture from Metagenomes
- Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies
- Proteomics and Metaproteomics
- RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data
- SATe-Enabled Phylogenetic Placement
- Serial Analysis of V1 Ribosomal Sequence Tags
- SILVA Databases
- Simultaneous Quantification of Multiple Bacteria
- STAMP: Statistical Analysis of Metagenomic Profiles
- Subtractive Hybridization Magnetic Bead Capture: Molecular Technique for Recovery of Full-Length ORFs from Metagenomes

Taxa Counting Using Specific Peptides of Aminoacyl tRNA Synthetases

Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3

The Vaginal Microbiome in Health and Disease

tRNA Gene Database Curated Manually by Experts

Use of Bacterial Artificial Chromosomes in Metagenomics Studies, Overview

Use of Viral Metagenomes from Yellowstone Hot Springs to Study Phylogenetic Relationships and Evolution

Variable Selection to Improve Classification of Metagenomes

Viral Metagenome Annotation Pipeline

Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach