GENOME WIDE MAPPING OF CHROMATIN STATES BASED ON HISTONE COMBINATORICS FOR DETERMINATION OF EPIGENETIC EXPRESSION

By

Nighat Noureen
PI131002

A thesis submitted to the
Department of Bioinformatics and Biosciences
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN BIOINFORMATICS

Faculty of Computing
Capital University of Science and Technology
Islamabad
FALL 2015

Dedicated to the Almighty ALLAH, Who is The Creator of this universe and Who bestowed me with the courage and faith to complete this work.

ACKNOWLEDGMENT

All Praises be to Allah Almighty who enabled me to complete this task successfully and my utmost respect to His last Prophet (P.B.U.H.). I would like to express the deepest appreciation to my Supervisor, Associate Professor Dr. Sahar Fazal and Co-Supervisor Dr. Muhammad Abdul Qadir, who continually and convincingly conveyed a spirit of adventure in regard to research. Without the guidance and persistent help of both my mentors this research would not have been possible. I would express my heartiest gratitude towards my family who always stand by me during the pursuit of my educational objective; without their continuous support, it was not possible for me to achieve this goal.

I am also especially thankful to all other members of Center for distributed and Semantic Computing (CDSC) for their motivation and suggestions. Not to forget the anonymous reviewers of our papers, who gave us the direction to move forward, and to improve our research work.

Special thanks to my students, my colleagues who in one way or the other were a source of motivation for me during the completion of this research.

ABSTRACT

Histone proteins wrap DNA around in small globular entities commonly known as nucleosomes. The post translational modifications to the histone tails are referred as histone modifications (HMs). The regulation of DNA in order to access the transcription machinery is epigenetically programmed by specific DNA and chromatin covalent modifications. HMs could either be present or absent at particular genomic loci and the combinatorial patterns of the specific modifications being addressed as 'histone codes', are believed to co-regulate significant biological processes. Regions defined by combinatorial patterns of marks can be referred to as chromatin states. Chromatin states associated with genomic locations correlate with specific functional elements as enhancers, transcription start sites, which can be exclusively inferred from successive combinations of chromatin marks in their contiguous locations. Biologically significant combinatorics of epigenetic modifications and their subsequent functional interplay are still mostly unrevealed. We aimed to use ChIP-Seq data of Histone modifications at different genomic loci to highlight the unbiased genomic grouping and to decode the complex biological network of HMs in association with other chromatin players in defining various chromatin states.

We used different tools and techniques to accomplish our task. Complex biological networks underlying the hidden chromatin states were revealed via merging machine learning and graph theory existing approaches. Histone modification efficient and simple combinatorics was studied at a global and local scale by developing and implementing a clustering and biclustering tool. Results have been compared with the existing approaches. Meanwhile a simple and efficient computational methodology for efficient chromatin states identification for ChromHMM (HMM based chromatin segmentation) has also been developed by utilizing Hidden Markov Models components.

As a result of above study we revealed the role of various factors in maintaining the chromatin state connectivity via focusing chromatin state networks. Our studies highlighted the minimum dominating nodes set and various hubs in chromatin state networks focusing their interaction patterns. Along with we developed and tested a clustering tool ChromClust and a biclustering tool ChromBiSim to highlight histone combinatorics in binarized signal data in an efficient and interactive way. ChromClust operates at global level while ChromBiSim mines local patterns of histone modifications associations.

We conclude that epigenomic landscape is portrayed as interplay of various factors including histone combinations, transcription factors, chromatin modifiers and most importantly the underlying DNA motifs. Each chromatin state has a specific set of these factors which interact with each other to mark that state hence creating the whole chromatin states network.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

HM	Histone Modifications
TF	Trasncription Factors
CM	Chromatin Modifier
TFBS	Transcription Factor Binding Site
MDNS	Minimum Dominatin Nodes Set
DNA	Deoxyribonucleic acid
H	Histone
H2A	Histone 2A
H2B	Histone 2B
H3	Histone 3
H4	Histone 4
MBT	Malignant brain tumor
PHD	Plant homodomain
ADP	Adenosine di-phosphate
HP1	Heterochromatin protein 1
CTCF	CCCTC binding protein
PC2	Polycomb 2
RNA	Ribonucleic acid
siRNA	Small interfering ribonucleic acid
RITS	RNA induced transcriptional silencing
me	methylation
K	Lysine (amino acid)
Swi6	Chromatin associated protein
ES	Embryonic stem
PSSM	Position specific scoring matrix
ChIP	Chromatin immunoprecipitation
SNP	Single nucleotide polymorphism
HMM	Hidden Markov model
BN	Bayesian networks
SPCN	Sparse partial correlation networks
TSS	Transcription start site

BIC	Bayesian information criterion
AIC	Akaike information criterion
CLARA	Clustering large applications
MCL	Markov clustering
EMBL	European molecular biology laboratory
EBI	European biology institute
DB	Database
CLC	Not available but is shared as (Cake-Loving company)
UCSC	UC Santa Cruz Genomics Institute

Chapter 1

INTRODUCTION

The molecular unit of heredity of a living organism is a gene. A modern working definition of a gene is "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions "[1-2]. Now the question arises that are we the mere artifact of the important players of life- GENES? The causal hidden factors contributing towards the composition of the organism phenotype are anticipated to be revealed as a result of the human genome interpretation. This decoding process is just the start of this genomic era which has only shed light upon the surface of the complex mechanisms contributing towards organism's phenotype. For a long time, the genome has been considered as incontrovertible master plan lain down with the inception of our lives.

1.1 Epigenetics

Recent discoveries highlighted the fact that the information and directives for the use of genetic material are the interplay of other genetic and environmental elements. With this view point, the first familial genetic stratum- the DNA- of the cell which is indistinguishable in all tissues of the individual corresponds to "the hardware" and is used as a platform for the execution of all the mechanisms of the body. The cell software on the other hand is represented by the set of connections of the changing environment around us. The software constitutes the second level of organization randomly disseminated across the genome and is continuously read, written and erased as a result of both social and substantial environmental signals [3]. Waddington in 1942 portrayed this theoretical connection between the genes and their instantaneous settings during development and phenotypic determination as epigenetics [4]. Changing patterns in gene expression without affecting the DNA sequential layers is referred as epigenetics. The regulation of DNA in order to access the transcription machinery is epigenetically programmed by specific DNA and chromatin covalent modifications [5]. The causal hidden molecular mechanisms of various epigenetic processes are being highlighted at an outstanding rate; various

techniques and tools being used to unravel the pathways and the key players involved. Many challenges are yet to be met, but we what James Watson said can't be neglected that "the major problem, I think, is chromatin… you can inherit something beyond the DNA sequence. That's where the real excitement of genetics is now" [6].

1.2 Epigenetic marks

Epigenetic marks responsible for phenotypically different tissues having the same genetic information and hereditary features of the organism despite of the varying environmental conditions occur as a result of complex enzymes/chromatin interactivity which maintains and establishes diverse gene expression programs in particular cell types [7].

Epigenetic modifications plunge into two major categories: DNA methylation and histone modifications.

DNA methylation in vertebrates occurs almost exclusively in the context of CpG dinucleotides and in the genome most CpGs are methylated [8-9]. In plants non-CpG methylation has well developed functional role which might also be visualized in mammals [10]. In early mouse embryo and embryonic stem cells this has been viewed at a low rate, which showed a significant decrease in somatic tissues [11-12]. CpA methylation has been highlighted in one of the studies as means for allelic exclusion in sensory neurons [13].

Along with DNA methylation the second major category of epigenetic modifications is histone modifications. They detect important cell states. DNA packed inside the cell highlights one of the important states known as Chromatin. The fundamental unit of chromatin is nucleosome, an octamer of four core histones (H3, H4, H2A, H2B) around which 147 base pairs of DNA are wrapped. Except the N-terminal tails of histones which are unstructured, the core histones are mostly globular. Modification of the histone tails is one of the striking features of histones. Histone tails bear hundreds of modifications of which at least eight are of distinctive nature [14]. Histones have multiple modification sites. Mass spectrometry or antibody specificity has been used to detect over more than 60 different residues on histones where modifications occur. Well this could be a miscalculation of the number of modifications taking place on histones because a true figure could be estimated by looking at the fact that methylation at lysines or arginines may be one of three

different forms: mono-, di-, or trimethyl for lysines and mono- or di- (asymmetric or symmetric) for arginines. The meaningful functionality do exist for this cosmic set of histone modifications, but these modifications could not determined or highlighted one at a time on the same histone. Signaling conditions of the cell decide the appearance or timings of these modifications [14].

Histone modifications are directed by the very important factory- the enzymes, which are being studied intensively over the last few years. Enzymes for various modifications have been identified in some studies like acetylation [15], methylation [16], phosphorylation [17], ubiquitination [18], sumoylation [19], ADP-ribosylation [20], deimination [21-22], and proline isomerization [23].

Modification of each and every nucleosome in one way or the other could be scrutinized comprehensively. This would provide with a static view because histone modifications are dynamic in nature and change rapidly. These modifications have a short duration of few minutes after receiving the stimulus from cell surface. Investigation of bulk histones under one specific set of conditions will highlight only a fraction of the probable modifications [14].

The interaction of different histones in contiguous nucleosomes under the influence of modifications may affect higher-order chromatin structure. Acetylation bears chromatin unfolding potential amongst all modification because it neutralized the basic charge of the lysine.

Protein modifications recruitment and binding include specific domains. Chromo-like domains of the Royal family such as chromo, tudor, MBT and nonrelated PHD domains recognize methylation, whereas acetylation is captured by bromodomains and a domain within 14-3-3 proteins recognize phosphorylation [14].

1.3 Functional Consequences of Histone Modifications

Histone modifications functionality simply can be categorized into two main parts, firstly establishing global chromatin environments and secondly management of DNA based biological responsibilities. Genome is divided into two major portions globally based on histone modification combinatorics; euchromatin and heterochromatin. DNA is accessible for transcription in case of euchromatin while it remains inaccessible in case of heterochromatin. Modifications devise chromatin disentanglement for assisting and facilitating DNA-based and other functions. These

functions range from the local one to the global, meaning from gene transcription and DNA repair to DNA replication or chromosome condensation.

DNA extrication, manipulation, and reverting back to the correct chromatin state, all these biological functionalities require the proper sequential recruitment of the machinery. Facilitation of DNA functions via histone modifications is termed as 'histone code'. The specific set of histone modifications come under this scenario, while the true prediction of this code seems usually unlikely [24].

1.4 Establishing Global Chromatin Environments

Silent heterochromatin and active euchromatin are the two broadly categorized divisions of the genome. These categories are linked with distinctive set of modifications. Enzyme recruitment for chromatin modification by boundary elements segregates the environment in different compartments. CTCF an important transcription factor is an example of boundary element binding protein used for the delivery of modification enzymes [16].

The occurrence of methylation at H3K4 and H3K9 in contiguous euchromatic regions sustain the heterochromatin boundaries shown in studies on fission yeast. Therefore one of the vital roles of chromatin modifications is the maintenance and preservation of chromatin environments into two broad categories. Safeguarding the chromosome ends and their segregation during mitosis is determined by the heterochromatin structure. High levels of methylated sites such as H3K9, H3K27 and H4K20 along with lower levels of acetylation are linked with silent heterochromatin state in mammals [16].

The preservation of inactive X chromosome is dealt with the employment of PC2 to H3K27me3, while on the other hand the maintenance of the pericentric heterochromatin is carried out with the employment of HP1 to H3K9me [16]. In both budding and fission yeast methylation at H3K27 is absent while H3K9 is found in fission yeast and is more likely to the higher organisms [25-27]. The production of small interfering RNAs (siRNAs) from the transcripts originating from centromeric repeats in fission yeast is carried out via the nucleation of heterochromatin instead of its dispersion [16, 27]. Packaging of dicer-mediated siRNAs into RITS complex distributes H3K9 methylation to the heterochromatin development locations.

Spreading and maintenance of heterochromatic state in Swi6 pombe is carried out with the employment of HP1 [16].

Large genomic portion is known as euchromatin, which is the flexible environment for DNA regarding biological inputs. The turning on and off behavior of the genes, untangling the DNA for repair and replication, and other such open choice states are highly reflected by the modification patterns. Genes bear low levels of acetylation, methylation and phosphorylation in transcriptionally inert or inactive state. Actively transcribed euchromatin bears higher acetylation levels along with further enzymatic activities and trimethylation of H3K4, H3K36 and H3K79. One of the recent findings highlighted bivalent domains acquiring both repressive and active modifications, which has shaken the very basic observation that active and silent marks dictate two discrete types of chromatin environments [25]. In mouse ES cells bivalent domains were revealed during the scrutiny of various highly conserved noncoding elements. The coexistence of two contradictory methylation sites H3K27me and H3K4me in the bivalent domains were exposed via ChIP on ChIP technology [25, 26].

Typically H3K4 methylation is implicated in active chromatin while H3K27 methylation is involved in silent chromatin. The low-level expression of developmental transcription factors is associated with enrichment of these conflicting modifications within bivalent domains. In case of differentiation of ES cells, only one of the modifications amongst the repressive or activation were attained. Analysis of these findings highlighted that the bivalent chunk of modifications with in ES cells retain the transcription factors involved in controlling differentiation processes, in poised and low-level expression states. The preservation of pluripotency in ES cells could be implicated through these findings. The selective regulation of modification pathways would be helpful in manipulating the differentiation of stem cells.

1.5 Management of DNA-Based Processes

DNA-bound transcription factors distribute chromatin modifying enzymes for the regulation of gene expression with euchromatin. A cascade of modification events is generated as a result of binding of transcription factor to the promoter region of the specific genes, which then result in expression of repression of those genes. Modifications can be classified into two categories; one associated with the activation while the other with the repression for the purpose of transcription.

Activation involves acetylation, methylation, phosphorylation and ubiquitination while repression occupies methylation, ubiquitination, sumoylation, deimination, and proline isomerization. This fact has been clear from the studies that modification of any kind has the ability to activate or repress under varying environments; e.g. the coding region has a positive effect of methylation at H3K36 while in promoter it produces negative effects; similarly methylation of H3K9 may also have the same effects, like positive in coding and negative in promoter regions [28]. Environmental perspectives matters a lot in case of modifications to work. Histone modifications have vital roles to play in various processes like DNA repair, DNA replication and chromatin condensation.

1.6 Histone Modifications: Truly Epigenetic or not?

Histone modifications deal with various epigenetic processes. Epigenetics could be simplistically defined as those heritable phenotypic changes which don't entail DNA sequence modification. Epigenetics is now being used as the term without "heritable" conveying the meaning of 'passing on information' by the genome which is not encoded on DNA. The term maintenance becomes important because this lies under the shed of non-genetic memory being transmitted from generation to generation. Transmission of various cellular phenotypic phenomenon like X chromosome inactivation, aging, reprogramming, gene silencing, heterochromatin formation, and imprinting occur in the same way. Along with this environment does induce various changes which are transmitted to next generation despite of any original stimulus, well studied in plants [27, 28].

Involvement of histone modifications in epigenetic processes is not the debatable question rather the passing on of the memory of a given state or memory implementation after passing on from a distinctive process are the areas of investigation [27, 28]. There is a need to know whether these processes are under the control of histone modifications or not? If this is so then a mechanism for the broadcast of such modifications onto the chromatin of replicating DNA should exist. Anticipation has been done for H3K9 methylation for the transmission of heterochromatin: employment of HP1 induces H3K9 methylation activity which alters the nucleosomes on the daughter strand this guaranteeing the transmission of H3K9me mark [27, 28]. Lysine methylation was given an epigenetic status because of

monitoring the persistent behavior of H3K4me3 along with the mode of transmission. The mother cell chromatin structure imparting the modification patterns to the daughter cells have not been yet declared as the true structures. So it's also still questionable that does methylation of lysines dictate the memory of chromatin structure? Along with the discovery of demethylases, the existence of histone methylation as a permanent mark becomes unstable [27]. This also makes it questionable whether other types of modifications are epigenetic or not? Does this mean that entire complex chromatin structure of the whole genome is modulated by very few number of histone modifications? Does the correct transmission of assembly of local chromatin structure require other determinants? Association and transmission of histone modifying complexes/enzymes via small RNAs has been studied in fission yeast [28]. Employment of HP1, methylation of H3K9 and heterochromatin formation could also be affected by the deletion of the enzyme dicer which processes small RNAs [29-31].

Working of RNA as a determinant is tempting and some facts do exist which highlighted it so. One of the studies have highlighted the fact that small RNAs in the sperm can be passed on to the offspring intervene an epigenetic phenotype known as paramutation one of the processes very firstly acknowledged in plants [32]. This whole mechanism is extensive rather than limited. Many genomic loci may radiate small RNAs and after passing on to the next generation these RNAs get involved in spreading and passing chromatin-modifying complexes to genes of interest and required loci in order to create the effect of whole chromatin being observed [28, 33].

The extreme accuracy of small RNAs possessing nucleic acid as their guiding machinery adds a tempting feature in their delivery system. Time will decide for the pervasiveness of such assumptive means of chromatin information transfer. This anticipated model for RNA considers it as an ideal and faultless molecule for passing on the memory of a particular chromatin state [33].

This consideration of RNA- intervened processes doesn't neglect the vital role of histone modifications in epigenetic dealings. This simply highlighted that histone modifications may be the executers behind this epigenetic scenario despite of being the carriers of the memory. DNA methylation, post translational histone tail modifications and non-coding RNA-based mechanisms are epigenetically regulated gene expression pathways [34].

1.7 Transcription factors and Transcription factor binding sites

Transcription factors (TFs) are important players of gene expression regulation mechanism. The crucial step in activation and repression of gene is binding of the TFs to their target loci. TFs recognize short and dispersed DNA segments in the genome called as transcription factor binding sites (TFBSs). The shorter length of TFBSs makes their prediction challengeable [35].

The binding of TFs vary with respect to tissue specificity, developmental stages or environmental conditions. This dependency makes the prediction more challenging. TFBSs have been predicted both by experimental as well as computational methods. Genome wide identification of these sites is carried out by ChIP-Seq; one of the well known and powerful methods [36, 37].

The unavailability of ChIP-quality antibodies against each TF obstructs the prediction of all TFs encoded in the genome using ChIP-Seq. A limited number of TFs have been successfully tagged in mammalian genomes via tagging techniques which have been applied to tag every individual TF.

1.7.1 Computational prediction of transcription factor binding sites

Computational prediction of TFBSs is now being carried out [38].Position- specific scoring matrix (PSSM) which reflects the preference of nucleotides at each position has often been used to represent the DNA binding sites of TFs [39]. In order to improve prediction accuracy and to avoid false positives, additional factors as conservation of TFBSs and co-localization are also included. Despite of this the problem still remains challenging.

1.7.2 Chromatin signatures at transcription factor binding sites

Computational prediction of TFBSs doesn't include conditional dependency for their differentiation from one condition to the other. Many studies have reported the association of TFBSs with various chromatin signatures. ChIP-Seq studies have highlighted the contribution of various chromatin modifications for the recruitment of TFs at particular TFBSs. Transcription factor binding events, gene interactions and DNA sequence analysis are the basis of gene regulation studies. Genome wide epigenetic marks specially DNA methylation and histone modifications have revolutionized the study over the past decade. The post-translational modifications of

the histone proteins which form nucleosomes by wrapping about 147 base pairs of DNA are known as histone modifications. Various biological processes such as transcription, dosage compensation, DNA repair, splicing and many more like alteration of chromatin structure or recruitment of key proteins are affected by the histone modifications [14, 40].

1.7.3 Genome-wide mapping of histone modifications via ChIP-Seq

Genome-wide mapping of histone modifications using ChIP-Seq technologies is an alternative approach for TFBSs prediction [41- 42]. Various chromatin signatures are associated with regulatory elements like promoters and enhancers etc, and could be used in their prediction [41, 43-44]. Genome- wide mapping of chromatin modification profiles and TFBSs is carried out by combining chromatin immuno-precipitation (ChIP) with high-throughput sequencing (ChIP-Seq). Huge amounts of data are being generated via ChIP-Sequencing technique and this increases the need of new analysis algorithms for such data [45].

1.7.4 Histone modifications combinatorics mapping onto functional genomic elements

The amalgamation of different histone modifications concedes different functional specificities [46]. For example, H3K4me3 and H3K9ac mark the nucleosomes near the active promoter regions while the inactive promoter regions in *Saccharomyces cerevisiae* generally lack these marks [24, 47-48]. In humans active promoters are collated with H3K4me3 while enhancers are collated with H3K4me1 and are deficient in H3K4me3 [41]. It is therefore clear that H3K4me3 is the histone modification which is liked with promoter regions in the genome and sequencing of these regions will show high enrichment of this modification. On the other hand the enhancer regions have been found deficient of H3K4me3 mark while they have shown high enrichment towards H3K4me1 histone modification. With the detection of plenty of histone modifications, it is plausible that additional patterns of chromatin modifications exist, and may divulge novel functional elements of the genome [49].

1.8 Chromatin combinations and Chromatin States

The dissemination of chromatin domain across the genome is highlighted by histone modifications combinatorics. Regions defined by combinatorial patterns of marks can be referred to as chromatin states. Differences in histone modifications cause

differences in “on-off” transcriptional states and lead to two major chromatin domains or states referred to as euchromatin and heterochromatin [46]. Chromatin states associated with genomic locations correlate with specific functional elements as enhancers, transcription start sites, which can be exclusively inferred from successive combinations of chromatin marks in their contiguous locations [50]. Associations of histone modification combinatorics with the functional DNA elements are being divulged by scrutinizing multiple histone modification maps.

These studies have used various segmentation and visualization methods for dividing the genome into different states and then naming the states based on the combinations of histone modifications but none of them have provided the unbiased criterion to identify the number of states for the data set under consideration. In each study the number of chromatin states has been fixed without providing the computational reason behind it which makes the criterion unbiased.

The unbiased criterion means that there should be some computational way to identify the number of states for the data under consideration. Along with that it is still ambiguous whether these modifications are causes or effects of transcription, these observations evidently reveal an association between various combinations of histone marks and various transcription states [51]. This research will try to segment the genome in an unbiased fashion by validating the existing scoring methods. Along with these validations we will study the link between genetics with epigenetics.

1.9 Problem Statement

Biologically significant combinatorics of histone modifications and their subsequent functional interplay has mostly been unrevealed at present [52]. Regardless of available data about histone modifications and their functions still there is lot more to be explored. Particularly there is a need of understanding the underlying hidden system of multiple histone combinatorics and its relevance to transcription [51-53].

Hypothesis of histone code implies regulation of gene transcription in a collective manner by co-occurring histone modifications. The exact pattern is yet to be untangled and its decoding is highly dependent on deciphering the hidden networks of histone modifications in different genomic regions [52-53].

Therefore there is a need to investigate this hidden combinatorics and networks of histone modifications from different angles. In this PhD research we would critically

work on this aspect. We would be looking at chromatin state networks and the role of histone modification combinatorics in these networks.

We will study the contribution of the underlying DNA motifs in setting the chromatin states and ultimately their effect on chromatin networks. The study of DNA motifs along with the combination of epigenetic marks at particular transcription factor binding sites representing a chromatin state will help in finding out the genetic basis of epigenetic marks. Classification of SNPs and genome wide association studies (GWAS) enrichment in various states will help in the inference of biological and medical data. We will also provide an efficient and unbiased platform for genome segmentation along with providing the computational strategy for chromatin states calculation for existing methods.

1.10 Research Methodology

To evaluate our hypothesis the following strategy will be adopted.

1. Data Retrieval

   ChIP-Seq data of histone modifications and transcription factors binding sites would be retrieved from data repositories [54].

2. Genome Segmentation

   We would segment the genome using Hidden Markov Model (HMM) scripts of the data under study.

3. Genome Mapping and Annotation

   In this phase we shall carry out the mapping and annotation of the segmented data into states (from step 2) to the genome in order to indentify the transcription factor binding sites. This is also a standard process and has been reported in literature [52, 54].

4. TFBSs relation with TFs and Histone Modifications

   Study and analysis of states and peaks of histone modifications and TFs will help in identifying the relationship of TFs, histone modifications and TFBS. It will also help in studying the link between genetics and epigenetics layer.

5. Unbiased genomic segmentation

   We will use various statistical measures to provide a computational stratgey for identifying chromatin states for existing platforms. We will also provide

our own independent platform for identifying patterns in the genome in an unbiased and user friendly way.

6. Linking genetics and epigenetics

   Step 4 would help in proposing the role of DNA sequence motifs in setting chromatin states.

7. Validation

   Validation of role of DNA motifs would be obtained via results comparison with existing methods.

1.11 Summary

We introduced here the basic terminologies regarding epigenetics, histone modifications and their combinations, association of transcription factors with histone modifications, objectives of our research and the major steps to be followed to conduct this research work.

This PhD thesis is organized into 6 chapters; where chapter 1 is the introductory chapter describing the background details of the topic chosen. Chapter 2 describes review of literature mainly focusing the techniques relevant to the field. Chapter 3 highlights the tools (including hardware and softwares) used in study, while chapter 4 highlights the methodologies which have been opted to carry out this PhD research work. Chapter 5 describes the results and their discussions and the last chapter 6 concludes the studies followed by the references and appendices sections.

Chapter 2

LITERATURE REVIEW

Various computational methods have been developed and utilized for the prediction of histone modification combinatorics from ChIP-ChIP/Seq data sets till now. It is complicated to scrutinize and investigate the ChIP-Seq data [55]. Demarcation of the chipped enriched regions from one sample source at a time with optional control sample is the property of many existing analysis tools [56].As enrichments of histone modifications are weak and not very confined and this makes their predictability very tricky and tough. Peak calling approaches have been amplified for broader domains along with systematically signifying these signals beyond read counts [45, 57-58]. Peak calling methods also have been tailored to paired experimental designs in order to associate epigenetic signals to biological functions and processes [59]. Each epigenetic mark contains the information for perceiving the biochemistry and design of its hidden causal genome. As already discussed, histone modifications combinatorics, their variants and TFs mutual functionality is highly informative [60-61]. One of the major limitations of the peak calling methods is their inability to deal with multiple signals at a time and this is a challenging task [62-64].

2.1 Genome Segmentation Approaches

Peak calling algorithms have the drawback of handling single at a time. A challenging task was to handle multiple signals at a time. Therefore analyzing histone modifications data in the form of combinatorics was the hour of need. This has been fulfilled by the introduction of various approaches. These approaches majorly could be classified into two categories: locus-based clustering and genome wide segmentation. Some of the widely used segmentation techniques have been mentioned along with their limitations.

2.1.1 ChromaSig

ChromaSig [49] an unsupervised approach highlighted the recurrence of histone modification ‘motifs’ traversing the human genome. The histone combinatorics has been studied simultaneously with the help of clustering approach. Chromatin signatures are grouped up without the aid of any external annotations or training sets.

Only promoters and enhancers have been focused much ignoring the transcription factors effect. ChromaSig classifies genomic location with globally harmonious chromatin signatures. This method generally assumes that for genome annotation a small set of chromatin states is adequate. ChromaSig is capable of discerning subtle variations in chromatin signatures using histone modifications data only. It was noted that various functional activities related with enhancers like binding of particular transcription factors and co-activators are interrelated with particular histone modifications present at enhancers. This procedure is yet not clear and needs further studies. This mechanism could also be studied at various other genomic loci such as insulators, promoters etc. along with the effect of transcription factors.

2.1.2 Spatial Clustering Approach

Similarly a spatial clustering algorithm [65] utilizing Hidden Markov Model (HMM) was projected for the prediction of combinatorial patterns stretched over adjacent genomic locations. Inspecting DNA regions bearing specific histone modifications is now consistently being achieved via ChIP-Seq [66]. This method is limited to histone combinatorics without studying the effect of TFs and TFBSs.

2.1.3 Chromia

Histone combinatorics was studied at promoters and enhancers using supervised learning approach by Won [43]. Identification of functional regions such as promoters and enhancers solely as well as along with PSSM binding patterns and transcription factor binding sites by amalgamating histone modifications such that a particular pattern represents a signal emanating from a particular HMM state was reported by Won *et al*., [38, 43]. The data set consisting of 8 histone modifications and 13 TFBSs in mES cells were scrutinized which revealed the association of chromatin signatures along with enhancers and promoters.

The TFs didn't show any remarkable association with histone modifications, which is because of the fact that the data set was very limited and didn't include a wide range of TFs and modifications. In order to unravel the hidden mechanisms, the correlation between the TFs and histone marks require further corroboration. The data set utilized only eight marks insufficient to locate any regulatory element in the genome. More marks along with the other regulatory elements will improve the predictions by using an unsupervised approach.

2.1.4 ChromHMM- The Binarization Approach

In a more recent work Ernst and Kellis proposed an alternative HMM algorithm based on the binarization of presence or absence of each histone mark [52]. Epigenetic states such as heterochromatin, various enhancers, splicing regions as well as transcriptional process covering variety of genomic locations have been highlighted using HMMs. HMM states identification has not been supported by any computational view point. The number of states has been fixed to 41 without any sound justification.

2.1.5 CoSBI

Ucar *et al.,* [50] presented a biclustering approach CoSBI for monitoring the histone modifications combinatorics. They identified subsets of chromatin modifications covering diverse genomic regions via a scalable subspace clustering approach using coherent and shifted bicluster identification. This study for the first time introduced the concept of local biclustering in case of histone modifications combinatorics. They presented a complex 3D input formatting of the input data. CoSBI although the first platform but is not user friendly.

2.1.6 Graphical Models

In a more recent study [51] the associations between histone modifications have been highlighted in the form of graphs where nodes represented the modifications and edges formed the associations. Associations among histone modifications specifically in promoter regions by various studies have been confined using Bayesian networks (BNs) [67-70]. These studies intended to ascertain causal links for histone combinatorics considering which modification is required for the occurrence of the other. In existence of hidden mystifying factors occurring recurrently in biological systems, declaration of causality in BNs is contentious [71- 73]. Furthermore, BNs forbid circuits or feedback system, which is not viable in biological systems [51]. The main focal point was the edges that correspond to direct relations.

Associations among histone modifications which have a direct relationship and whose association could not be represented exclusively by mystifying factors; have been enlightened by drawing edges between histone modifications. This work focused sparse partial correlation networks (SPCNs) using graphical Gaussian models. The view point in using PCNs was as histone modifications combinatorics has not been touched via this approach [51]. Focusing interactions among histone modifications is

a remarkable way to pin point their relationships. These interactions are abstract and not physical. These networks could be better explained, defining them as physically existing ones, by including various other existing parameters like chromatin modifiers, enzymes or proteins involved in the process. So it is important to include the role of various transcription factors, transcription factor binding sites along with histone modifications combinations to mark the diverse genomic regions like promoters, enhancers, heterochromatin and insulators etc.

Along with this another parallel study [53] aimed at decoding the complex biological network of HMs in a single region and demonstrating how the HM networks differ in different regulatory regions. According to this work the considerable association between histones and genomic functions is obtained due to variations in the network attributes. This basically indicates unraveling the 'histone code'. This also has been highlighted by demonstrating that different crosstalk mechanisms are used to characterize functionally variant genomic regions as promoters, enhancers and insulators etc. The difference between two networks has been formalized using model-based approach. This study used the proposed graphical model for the selected HMs used in CD4+ T lymphocytes (17, 26) for 3 different types of genomic regions: promoters, enhancers, and insulators [53]. The model is limited for identifying the interactions among HMs only.

One of recent graph based studies [74] highlighted the interactions among histone modifications and chromatin modifiers. Study was limited to the promoter regions only which is the limitation of the study.

Graph theory approaches have been utilized to solve many problems in biological domain. Most of the studies used gene expression profiles and protein-protein interactions for this purpose. Only two to three studies [51, 53, 74] used histone modifications profiles for building graph models and used graph theory approaches to highlight interactions among histone modifications. Therefore much needed to be focused in this area.

2.2 Related Studies

Various other studies [38, 75-84] used HMMs and clustering techniques to segment the genome into different states or clusters. After having a look at various studies we can see that all of them are segmenting the genome into states using HMM or

clustering to segment chromatin into states but in each study number of states vary. Variation in number of chromatin states arose in these studies because the number of histone modifications under consideration was different. The combinatorics patterns could be achieved according to the data set under study. It varies but in total $2^{(\text{number of chromatin marks})}$ combinations could be formed. But in most cases such huge number of patterns do overlap and that's why are combined to achieve one state in order to avoid repetitions.

2.3 Their Limitations and Bottlenecks

Genome segmentation methods based on histone modification combinatorics mostly include clustering based approaches, some focused on HMM based approaches while biclustering have been used in two of the studies. Each study used different data sizes but none of them discussed the effect of data size on defining clusters and chromatin states. Along with what could be the optimal cluster solution or states solution for a specific data under study has also been lacking.

Histone combination profiles have been focused mostly without the effect of other factors involved in transcriptional processes. The role of DNA motifs in setting the chromatin states remained untouched. Keeping these points in our mind we progressed in this direction. We in our experiments tried to address these questions using different approaches including clustering, biclustering and graph theory approach concepts.

2.4 Summary

Genome segmentation methods based on histone combinatorics profiles have been focused in many studies. Along with associations among various histone modifications have also uncovered via graph theory approaches. We discussed these studies and have looked at their limitation in this chapter. We also concluded on how we addressed the some left over questions in the field.

Chapter 3

TOOLS AND TECHNIQUES

In this chapter, details of the tools have been discussed which have been used while conducting out our research work. Tools have been categorized into different sections, including hardware, and softwares. Details of each tool have been given in the respective sections.

3.1 Hardware used with technical specifications

The system used in the research was Dell Inspiron with 4.00GB RAM and Intel(R) Core(TM) i3C CPU. The processor specifications were M380 @ 2.53GHz.

3.2 Software(s), simulation tool(s) used

3.2.1 Windows platform

Windows 7 with 64 bits specification platform has been used for most of the experimental work.

3.2.2 BioLinux 7 platform

Open source BioLinux 7 has been used in parallel with windows for execution of some of the experimental work [85].

3.2.3 Softwares operated via windows platform

3.2.3.1 R

R is a freely available software environment which is used for statistical computing and graphics. It can be operated from many platform including Windows, UNIX and MacOs.

We used 3.2.1 R version for 64 bit windows 7 platform. We have utilized R for writing various scripts to handle matrices and networks data [86].

3.2.3.2 Biolayout express

Biolayout express 3D [87] is a visualization and analysis platform for different networks especially the biological ones. We utilized Biolayout express to visualize chromatin states networks used in our study.

3.2.3.3 ChromHMM

ChromHMM is multivariate Hidden Markov Model (HMM) based freely available tool which is used for characterizing and learning chromatin states. Chromatin states learning is based on integration of multiple ChIP-Seq histone modifications data [88]. We learnt chromatin states at various stages in our study using ChromHMM.

3.2.3.4 HMMSeg

Hidden Markov Model based platform, HMMSeg is used for segmentation of continous genomic data. It is java based tool and can handle multiple datasets at time [84]. We used HMMSeg for segmentation of H1 data.

3.2.3.5 Cluster 3.0

The open source clustering software is used for clustering of gene expression data sets mostly and could be operated from any platform [89]. We used Cluster 3.0 to cluster various emission matrices of HMM used in our study.

3.2.3.6 Visual studio

Visual studio, a product by Microsoft is integrated development environment used for developing applications for different platforms [90]. We used visual studio to develop our applications, ChromClust and ChromBiSim.

3.2.4 Softwares operated via Biolinux platform

3.2.4.1 Homer

Homer is a motif discovery tool used for regulatory element analysis in genomic data and is designed for ChIP-Seq datasets in mind. It could be applied to any of the motif finding problems [91].

3.2.4.2 Bedtools

Bedtools are a set of tools used for the analysis of wide range of genomic data. Bed tools could be operated from the UNIX command lines. We used bed tools to manipulate most of the genomic coordinate bed files [92].

3.3 Summary

We used various linux and windows based tools to achieve various objectives of our research. The hardware utilized for linux and windows platform was Dell Inspiron N5010.

Chapter 4

METHODOLOGIES

This chapter is based on methods to achieve the objectives of the research. It is divided into 6 main sections, each describing the details of approaches utilized.

4.1 A simple computational approach for predicting chromatin states

4.1.1 Dataset

We used ChIP-Seq data of human comprising of nine cell types containing data of nine histone marks; CTCF, H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac [54]. The data used were bed files containing the genomic coordinates and strand orientation of mapped sequence reads. The input control has been used as a separate mark. Chromatin combinatorics has been studied with and without input control in order to check the behavior of input control.

4.1.2 Genome segmentation

The histone marks and CTCF were used for genome segmentation for all cell types. The on and off signals of histone modifications combinations have been utilized to learn the HMM. This multivariate HMM was operated from the platform of ChromHMM software (v1.03) [88]. Firstly the genome was divided into non overlapping intervals containing the presence absence frequency of marks based on the count of tags mapped to the intervals. The non-overlapping intervals (bins) used were of various sizes including 200bp, 400bp, 600bp, 800bp, 1000bp, 5000bp and 50Kbp. The tags mapping followed the Poisson background model [52] with a threshold of 10^{-4}. Data binarization was achieved in two ways; one including the input control to normalize the reads of marks based on Poisson model [52] and second without input control. The two way binarization was achieved to test the role of input control in model learning. A virtual concatenation of all the cell types has been performed so a single model was learnt for all cell types.

After data binarization HMM models were learnt for two streams of data; with input control and without input control. We learnt models with random initialization varying from 2 to 30 states for all bin sizes independently for both data streams. At each state level from 2 to 30, 50 random initializations were used. Therefore 1450 models were learnt for data with input stream and 1450 models for the other.

4.1.3 Bayesian Information Criterion (BIC) score for models

In literature Bayesian information criterion (BIC) [93] and Akaike information criterion (AIC) [94] have been used for model selection based on model parameters. BIC and AIC resolves the over fitting problem by introducing the penalty term for number of parameters in the model. The major difference between approaches is the derivation of this penalty term. The penalty term for AIC, only accounts for the number of free parameters in the HMM, while the BIC penalty term also factors in the amount of training data available. We used BIC for our study in order to consider the data as well.

The BIC scores for all models have been calculated using '(*Eq 4.1*)'. The models with the highest BIC scores at each state level (from 2 to 30) were picked from 50 random initializations to see the behavior of BIC scores for states prediction (for both data streams and all bin sizes).

$$\mathrm{BIC} = -2.\ln \mathrm{L} + \mathrm{k}.\ln(\mathrm{n}) \qquad (Eq\ 4.1)$$

In above equation (*Eq 4.1*) 'n' is the number of data points in the data x, 'k' is the number of free parameters of the model and 'L' is the maximum likelihood of the model. All these values have been obtained from the ChromHMM produced model.

4.1.4 Computational prediction of states

4.1.4.1 HMM parameter evaluation

Hidden Markov Models are defined by five parameters. These parameters include i) the number of states in a model, ii) The number of signals emitted by each state, iii) the initial state probability vector which is the probability of starting at a particular state, iv) the probability of moving from one state to another represented by a probability matrix known as transition matrix. The order of the matrix is reliant on the number of states. So for N states in a model the order of transition matrix will be N*N and v) the emission probability to emit a particular symbol by a state. This is also a

matrix like transition matrix but the order is different. The order of emission matrix depends on the N number of states and M number of signals. So the order is M*N [66]. Among the parameters of HMM emission matrix is important as it contains emission probability of all marks for the defined HMM states. Each state is defined by the values of the emissions in it. We utilized emission matrices of HMM models with the highest BIC score (mentioned in above section) at each state level for evaluating the states number in our study.

The emission matrix of a particular HMM contains the number of rows (vectors) equal to the number of HMM states. We picked emission matrix for each model with highest BIC (each state level) and calculated the correlation amongst the vectors of each emission matrix using the stats package of R [86] by the function shown in the form of '(*Eq 4.2*)'.

$$\text{correlation} = \text{cor}(X, Y) \qquad (Eq\ 4.2)$$

The correlation values of each emission matrix were utilized to calculate the mean emission value. The mean emission values for all state levels (from 2 to 30 for all bin sizes and both data streams separately) were finally compared to see the state limiting point as attaining equilibrium.

4.1.4.2 Combining highly correlated states

Two highly correlated vectors could be clustered into one group. We applied this concept on emission matrices and clustered the highly correlated emission vectors using hierarchical clustering method implemented in Cluster 3.0 [89]. The emission matrix of 30 states model of each bin size (for both data streams) was subjected to clustering. The states number produced by this reduction method has been compared with the mean correlation values. The output from Cluster 3.0 was visualized using Tree View which is a visualizer for Cluster 3.0 output.

4.1.5 Biological annotation of all models

Genomic annotation of states with various genomic elements at each level with highest BIC has been obtained. The percentage of genome overlap for each state and different annotation data was obtained using ChromHMM. The RefSeq Gene, Exon, TSS, CpG islands, repeats and laminB1 annotations were downloaded from the UCSC Genome Browser website. The annotation at each states level has been compared with

the next states level to check the appearance of a new biological state. These annotations at each level have been compared with the emission correlation values in order to deduce the representative model amongst the 30 states.

This methodology has been tested at all bin sizes in order to highlight the importance of bin size in genomic segmentations along with the effect of input control as well. The methodology has been illustrated in detail in the Figure 4.1. This methodology is same for all bin sizes and for data with or without input control.

4.1.6 Comparison of states number with reference model

We compared the identified states number model by our approach (at 200bp bin size with input control) with the reference model [54]. The model was compared in term of states highlighted on the basis of emission parameter. Correlation amongst emission matrix of the reference and our model has been found using an R script1. The segmentations of reference have been compared with the segmentations produced by our model to find the relevance. This has been performed using a shell script2 and an Rscript3.

Rscript1:

```
H1_Data = read.table("D:/ReferenceModel_15.txt")
H2_Data = read.table("D:/TestModel_14.txt")
H1_matrix = as.matrix(H1_Data)
H2_matrix = as.matrix(H2_Data)
correlation = (cor(t(H1_matrix), t(H2_matrix)))
```

Shellscript2:

```
cut -f3 TestModel_14.bed | paste ReferenceModel_15.bed - > Ref_TestCombined.txt
```

Rscript3:

```
library(gplots)
Data = read.table("D:/Ref_TestCombined.txt")
Data_Tabular= table(Data[,c(4,5)])
Data_Tabular_Row_Sum = rowSums(Data_Tabular)
Data_Tabular_Percent_All = Data_Tabular/Data_Tabular_Row_Sum *100
```

The methodology explained is also tested on the real data set used by Ernst *et al*., [52] along with on a simulated data set produced from a randomly generated 3 states model as well. Results were 40 and 3 states model closer to the reference models. This methodology identified works on any ChIP-Seq data but one case (for real data) is presented here due to space constraints.

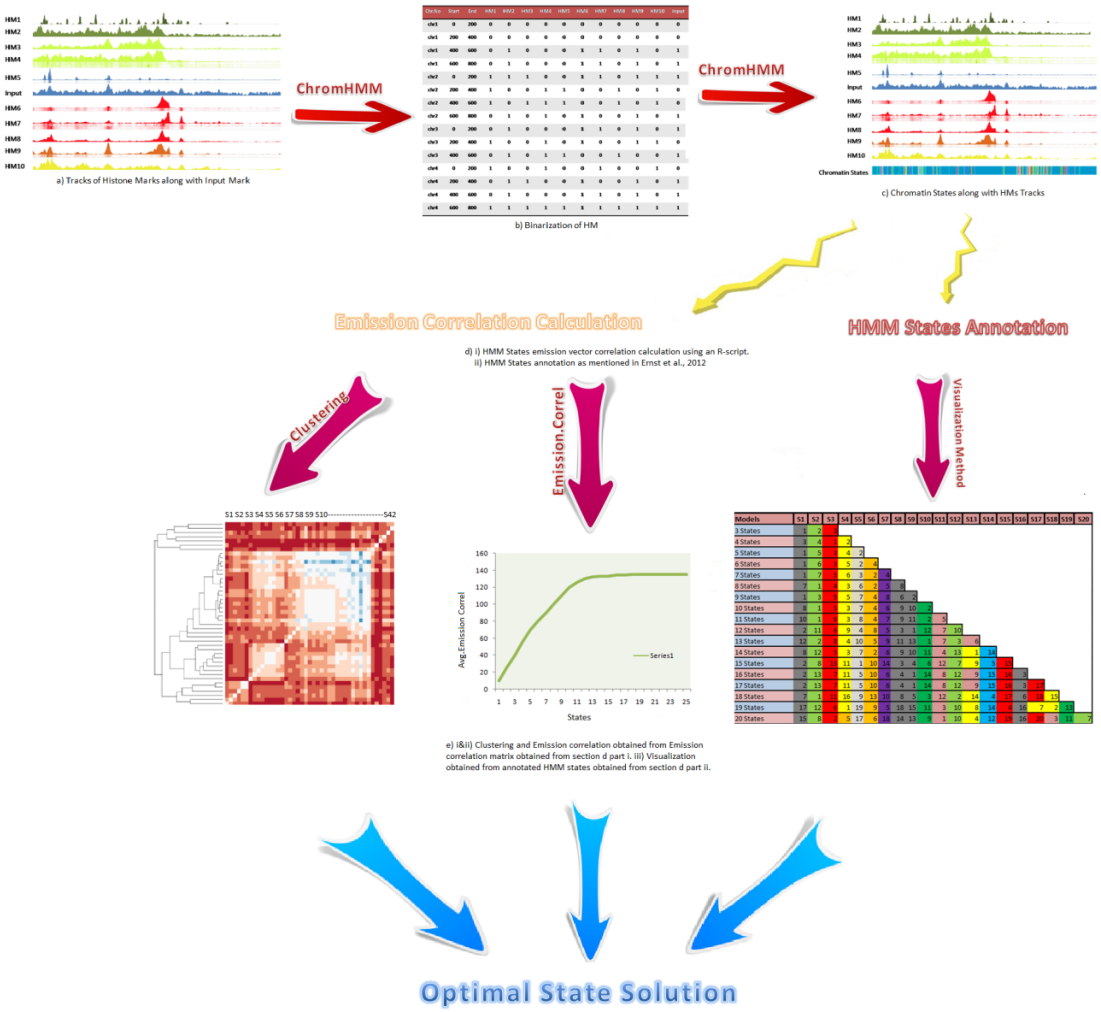


Figure 4.1 Detailed Methodology of unbiased computational method for HMM states prediction. a) Histone modifications signal tracks showing the peaks of histone modifications at the present locations. b) Conversion of histone modification tracks into binary signals using ChromHMM. c) Chromatin states learning using ChromHMM based on presence and absence signals of histone combinations. d) Two different paths followed i) Emission correlation calculation and plotting. ii) HMM States annotation. e) Part i of section d will gives rise to clustering and emission correlation plotting and part ii of d gives rise to visualization method.

4.2 Unbiased segmentation of H1 data

Hidden Markov Models (HMM) provide an unbiased statistical framework for segmentation of multi-dimensional genomic data into states. The average log2Dam-H1/Dam ratios for all H1 variants were binned into 200bp windows and submitted to the HMMSeg algorithm [84]. For a given number of states, the parameters of the HMM (the emission and transition rates) were learnt using maximally 100 iterations of Baum-Welch algorithm [95] and three different random initializations. The model with the best log-likelihood score was chosen. This learning step was repeated for a

variable number of states (2–8) and a segmentation based on 5 states was chosen because models with higher number of states resulted in only a small increase of the log likelihood at the expense of many additional parameters. This choice is supported by a cluster solution using CLARA [96], which is implemented in the ‘‘cluster’’ package of R [86]. Here the optimal number of clusters, defined by a maximal silhouette score, was also found to be 5 [97].

4.3 Semi-supervised mining of histone modifications associations at global level

Inspired by the binarized concept of epigenomic signals from [52] study we propose ChromClust, a **C**hromatin **C**lustering tool which works on binarized data. ChromClust has been tested on various publically available ChIP-Seq data sets [54]. Our tool works extraordinarily with efficient time complexity. Along with clustering our tool maintains the database of clustering records as well. The database stores the clustering features of data sets which could be utilized for clustering some other data. This facility makes it a semi supervised clustering tool where user can also provide some annotated data and can cluster new data based on these annotations. ChromClust efficiently clusters binarized ChIP-Seq signals along with querying facility about the clusters in maintained database. Tool is available freely at **(https://sourceforge.net/projects/chromclust/).**

ChromClust is implemented in C-Sharp via visual studio and can be used on any windows platform with Microsoft .Net Framework 4.5 and above. Windows presentation form (WPF) template for desktop application is used to develop ChromClust. In ChromClust user first creates the database, reads the binarized data, performs clustering and then stores the results in the database along with the added facility of exporting them into a file which could be utilized for further downstream analysis. User can query about the clusters and can visualize clusters in heatmap and chart forms. Overview of the tool is shown in Figure 4.2.

4.3.1 ChromClust Database

The first and foremost thing in using ChromClust is data base creation. User can create the database by providing the name of the file and it is automatically created by using create database facility. Database is maintained in SQLite which is an open source extension of SQL type databases and allows the user to create the database on

the desired location of the hard drive. SQLite is available in many languages as an extension library. We used C-Sharp library to develop ChromClust.

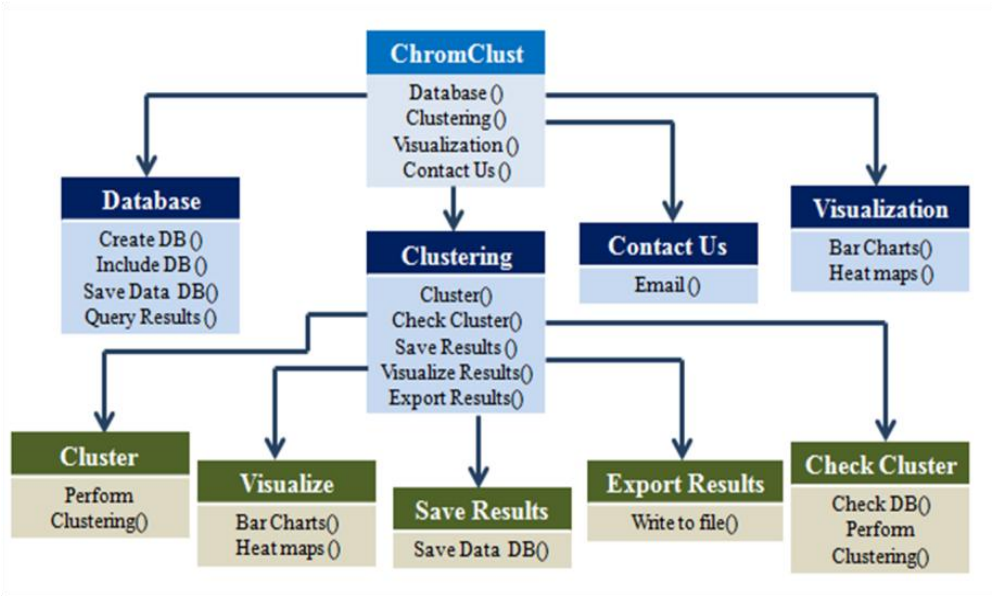


Figure 4.2 Overview of ChromClust tool

4.3.2 Clustering

In order to use the clustering module, it is important to include the database to be used. As mentioned earlier that database is first created. That database is then included for storing results. Clustering modules works on the binarized data matrix of whole genome. Rows represent bins of any fixed size by user while columns represent histone modifications. Clustering is based on the idea presented by Richard Hamming [98]. Overview of algorithm is presented as Scheme 1 and Figure 4.3.

Scheme 1: Overview of algorithm

B = Bin size
S = Similarity measure based on Hamming Distance (Ranges from 60 to 100)
Start
Scan the rows of matrix *M*.
Pick a row and call it cluster head *C*, highlight it as marked *M* and save in database if not there.
Check all rows one by one and find distance *D* of each row with *C*.
If *D* is greater than equal to *S*
Mark the row as *M* and include it in current cluster with *C* as head.
Else
Move ahead
Include cluster in set *Z* and exclude the elements in it for further processing.

Repeat all steps until end.

Data is clustered and results are added to the database. Clusters could be visualized in heatmaps, bar charts and also could be retrieved from the database via querying module. Clusters could further be stored in bed format for further processing.

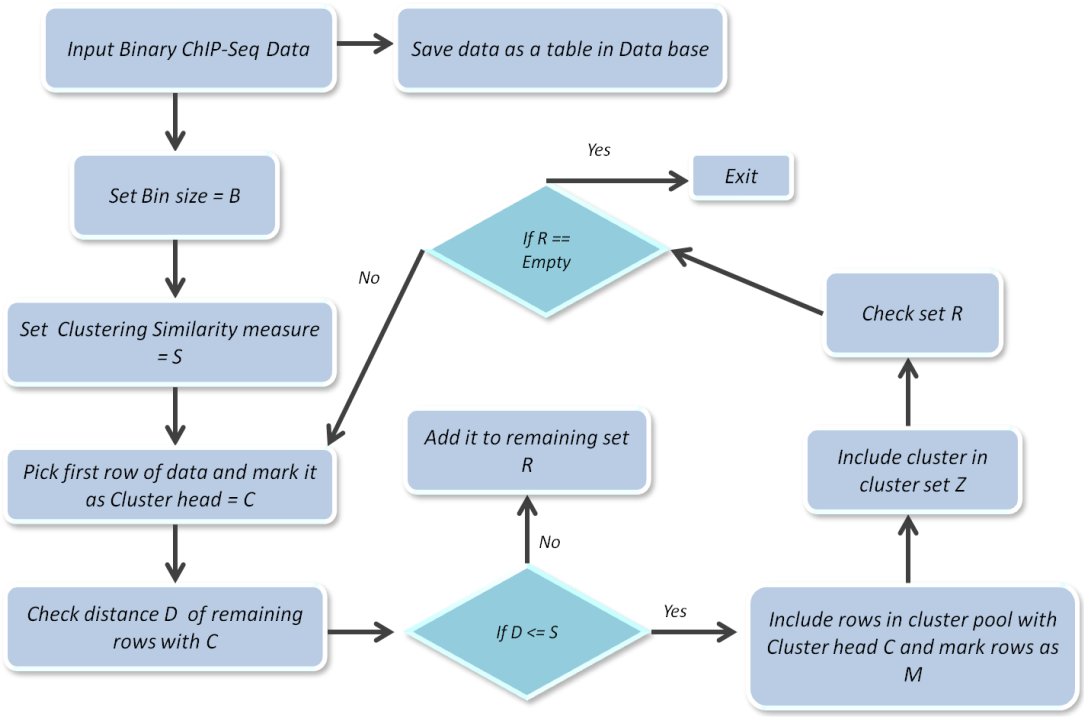


Figure 4.3 ChromClust Algorithm working

4.3.3 Application of ChromClust to Genome wide ChIP-Seq data

One of widely used ENCODE project ChIP-Seq data sets based on nine human cell types [54] have been used in study for results illustration.

4.3.3.1 Data normalization

The initial unprocessed data were bed files containing the genomic coordinates and strand orientation of mapped sequence reads from ChIP-Seq experiments [54]. Data preprocessing involved binarization via ChromHMM [88] platform. The normalization step involved dividing the genome into 5Kbp non-overlapping bins. Each bin marked the presence or absence of the histone modifications by monitoring the total mapping to that bin following Poisson background model. The value of threshold 'T' was set to total count of reads mapped for that mark and was considered as the smallest integer 'T' so that $P(Y>T)<10^{-3}$. Here Y is a random variable which

has a Poisson distribution and the mean is set to the mean of the total count of reads per each bin. In presence of control data the parameter of Poisson distribution is specified by the global average reads multiplied by the local enrichment of control reads [52].

4.3.3.2 Clustering of ChIP-Seq data

Initially the preprocessed binarized data was read and uploaded into the database to make further processing fast and easier. Bin sizes could be specified as per user choice. We used 5Kbp bin size for testing as chromatin signatures are larger than 2Kbp in only promoters and enhancers [41] while the repressed marks are wider than that. Data has been clustered and saved to database. We annotated the clusters with various elements of the genome by using data from UCSC genome browser [99].

4.4 Unsupervised mining of histone modifications associations at local level

Several computational methods have been presented so far for highlighting combinations of histone modifications [43, 49, 65, 84, 88, 100-102]. Multiple studies have shown that combinatorics is being portrayed by only few of the modifications from the set of whole modifications in the data set [50, 103-104].

We present ChromBiSim, a **C**hromatin **Bi**clustering **S**implified toolbox which is based on BiSim algorithm [105]. We proposed BiSim algorithm for biclustering of binarized gene expression data.

We implemented and tested it on binarized ChIP-Seq signals of histone modifications from multiple cell types [54]. We present a novel tool for identification of local histone association patterns from binarized ChIP-Seq signals. The software is written in C-sharp. It's a windows form based desktop application which can run on any windows platform having Microsoft .Net framework 4.5 and above.

The software calculates biclusters (local patterns of histone combinations) by finding the present signals of the histone modifications [105]. It can run on any size of data including whole genome histone modification profiles irrespective of the cell type and organism. Whole genome histone profiles could be analyzed efficiently within 30 minutes on a system with 4.00GB Ram and with 2.53GHz processor. The time reduces to half on a system with greater facilities, while data for whole chromosome takes only few seconds to decode the hidden patterns.

Our tool is one of the most efficient tools to date with respect to time and space complexity. Identified genome wide histone associations can be visualized as heat maps and bar charts. Total biclusters along with amount of genome covered is also calculated. Overview of the tool is shown in the Figure 4.4.

4.4.1 Biclustering algorithm

As mentioned earlier that ChromBiSim is based on BiSim algorithm [105]. We proposed BiSim algorithm for finding biclusters from gene expression data sets.

BiSim algorithm was based on binarized data matrix of gene expression data sets. We improved the algorithm efficiency and proposed that instead of following the divide and conquer recursive approach of BiMax algorithm [106], a simple iterative approach could solve the problem with reduced time complexity.

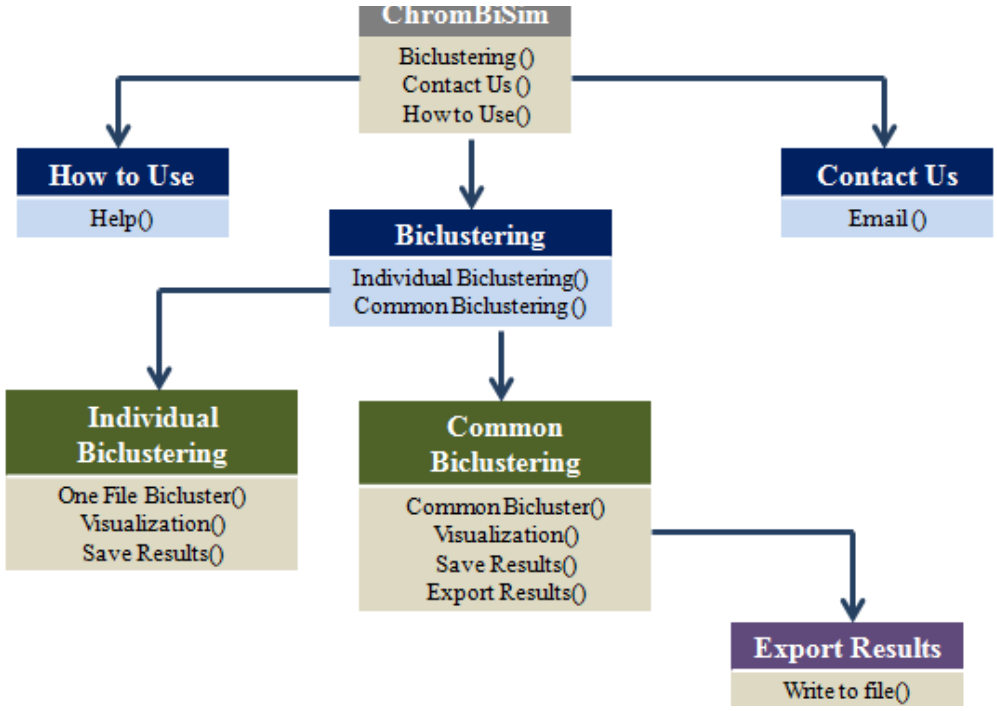


Figure 4.4 Overview of ChromBiSim

We proposed BiSim algorithm [105] for identification of biclusters from gene expression data sets. BiSim algorithm simply scans the binarized data matrix and stores the row wise and column wise on bits along with their indices. It then compares and combine the overlapping indices along both the dimensions. The biclusters constitute of the on bits representing the presence of the signals locally under certain conditions.

We applied and tested BiSim on large scale whole genome ChIP-Seq histone modification profiles. In this case the rows in the data matrix represented the non overlapping bins along the genome while the columns instead of representing the conditions of gene expression matrix, represented the histone modifications reads along the genomic locations. BiSim then simply scans the matrices along the whole genome and finds the local subgroups representing the histone modifications combinations along the whole genome via ChromBiSim platform as shown in the Figure 4.5

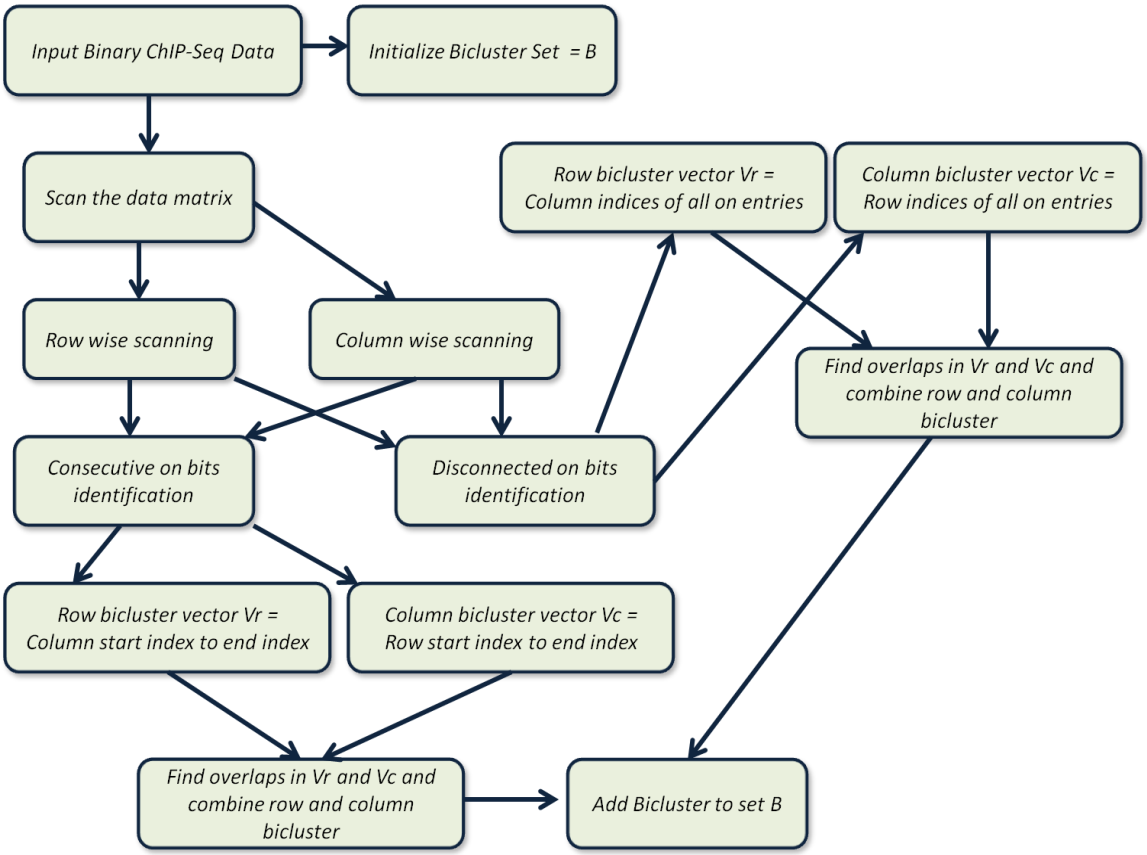


Figure 4.5 BiSim Algorithm

4.4.2 Application of ChromBiSim to Genome wide ChIP-Seq data

ChromBiSim identifies biclusters which could be mapped over various genomic locations. We tested ChromBiSim on various cell types [54]. Our study included 4 cell types of human; H1hesc, Gm12878, Helas3 and K562 comprising of 10 histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, and H4K20me1) and one insulator protein CTCF

along with the input controls of all cell types. We defined a comprehensive landscape of epigenomic profiles based on combinatorics of histone modifications present at certain locations of the genome. Results highlight histone modifications combinations present at certain locations of the genome in all cell types along with the similarities and differences highlighted. Histone modifications data over 5K base pair non overlapping locations was taken and preprocessed as explained in sections 4.3.3.1 and 4.3.3.2. Biclusters were identified and annotated with various genomic elements. ChromBiSim is available at (http://sourceforge.net/projects/chrombisim).

4.5 Date and party hubs in chromatin state networks

4.5.1 Dataset

ChIP-Seq data of HMs over 5 cell types of human including embryonic stem cells (H1hesc), lymphoblastoid (Gm12878), liver carcinoma (HepG2), cervical cancer (Helas3) and myelogenous leukemia (K562) has been retrieved from Encode repository. The bam files of the data were converted to the bed files using bedtools. The peak files of some transcription factors (TFs) for all cell types were also retrieved from Encode repository as shown in the Table 4.1.

Table 4.1 Data set of 5 cell types

S. No	**Gm12878**	**H1hesc**	**Helas**	**HepG2**	**K562**
1	H3K4me3	H3K4me3	H3K4me3	H3K4me3	H3K4me3
2	H3K4me2	H3K4me2	H3K4me2	H3K4me2	H3K4me2
3	H3K4me1	H3K4me1	H3K4me1	H3K4me1	H3K4me1
4	H3K27me3	H3K27me3	H3K27me3	H3K27me3	H3K27me3
5	H3K9me3	H3K9me3	H3K9me3	H3K9me3	H3K9me3
6	H3K36me3	H3K36me3	H3K36me3	H3K36me3	H3K36me3
7	H4K20me1	H4K20me1	H4K20me1	H4K20me1	H4K20me1
8	H3K27ac	H3K27ac	H3K27ac	H3K27ac	H3K27ac
9	H3K9ac	H3K9ac	H3K9ac	H3K9ac	H3K9ac
10	Ctcf	Ctcf	Ctcf	Ctcf	Ctcf
11	Input Control	Input Control	Input Control	Input Control	Input Control
12	Ezh2	Ezh2	Ezh2	Ezh2	Ezh2
13	H2az	H2az	H2az	H2az	H2az
14	P300	P300	P300	P300	P300
15	Pol2	Pol2	Pol2	Pol2	Pol2

4.5.2 Chromatin States

HMs data of all cell types has been used to segment the genome into states using multivariate hidden Markov model (HMM) [88]. Data binarization using Poisson

background model with the threshold of $10^{-4}$and 15 states model learning via virtual concatenation of all cell types has been achieved as discussed [52, 54]. Chromatin states of all cell types have been annotated with RefSeq Genes, RefSeq Exons, RefSeq TSS, RefSeq TES, RefSeq2KbTSS, CpG islands, repeats, laminB1 (obtained from UCSC genome browser) and TF peaks [88]. Fold enrichment of annotations have been normalized with an Rscript4 for equalizing the values as the emission probability values of the histone marks.

Rscript4:

```
library(gplots)
data1 = read.table("D:/EmissionCorrel.txt")
Emissiondatamatrix = as.matrix(data1)
EmissionDataPercent = Emissiondatamatrix*100
png(filename="D:/EmissionCorrels1.png")
heatmap.2(EmissionDataPercent,col =c("darkblue","dodgerblue","cyan"),
trace = "none",dendrogram = "none",keysize = 1.2,density.info =
"none")
dev.off()
data2 = read.table("D:/Annotation.txt")
datamatrix2 = as.matrix(data2)
Check2 = t(t(datamatrix2)/rowSums(t(datamatrix2)))
PercentCols2 = Check2*100
write.table(PercentCol2,"D:/EnrichmentCorrel_Results.txt",sep = "\t",
quote=FALSE)
png(filename="D:/EnrichmentCorrel_Result.png")
heatmap.2(PercentCol2, col =c("darkblue","dodgerblue","cyan"), trace
= "none",dendrogram = "none",keysize = 1.2,density.info = "none")
dev.off()
```

4.5.3 Denovo motif finding

Chromatin states segmentations of all cell types have been exploited for de novo motif discovery using Homer [91]. As chromatin segmentations are obtained via marks combinatorics therefore DNA motifs have been mined for those regions where several HMs combine to emit one state and the state also shows enrichment for certain marks. Homer not only works with peak files but has the ability to utilize the genomic coordinates in bed format. The wrapper *findMotifsGenome.pl* is used for the said purpose which aids HOMER motif discovery algorithm by setting up the data for analysis. Default parameters have been used for *de novo* motif discovery along with checking the enrichment for known motifs. Genome built of hg19 was used for mapping purpose.

4.5.4 Date and part hubs in chromatin state networks

Fold enrichments of TFs, emission values of HMs and enrichment of state motifs have been normalized to one scale of calculation using an Rscript4 mentioned in section

4.5.2. The normalized values have been utilized to construct chromatin state networks for all cell types using BioLayout Express 3D version 3.3 developed by EMBL EBI [87]. TFs, HMs, motifs and states have been used as nodes whereas the interactions between them were considered as edges. A generalized network covering all the cell types has been obtained by overlapping the networks.

Chromatin state networks have been subjected to the calculation of various centrality measures in order to define hubs and non hubs in the networks of 5 cell types. As previously proposed [107] hubs are nodes with maximum interacting partners while non-hubs are the ones with lesser interactions. One of the studies [108] utilized betweenness centrality to define date and party hubs in yeast interaction networks. Many studies utilized closeness centrality as a measure to identify hubs in complex networks and its applications in; extraction of metabolic core of the networks [109], visualization of complex networks [110] and for drug-targets identification [111], [112] along with the identification of cell cycle networks dynamics [113]. We utilized the concept of node degree along with the addition of some more centrality measures in order to define hubs and non hubs. Hubs were further divided into date and party hubs based on the used network centrality measures. We combined various centrality measures including degree, out degree, betweenness, closeness and normalized closeness and then defined a mean centrality measure as *Avg.Centr* which is the arithmetic mean of all centralities used. It adds all centrality measures and divides them by the total of centralities used‘(*Eq 4.3*)’.

$$Avg.Centr = \sum_{i=1}^{n} \ (\deg(i) + outdeg(i) + closeness(i) + norm.closeness(i) + betweeness(i)) \ / \ total \quad (Eq\ 4.3)$$

The centrality measures were calculated using an Rscript5. The degree centrality of a node *i* in a network could be defined as in ‘(*Eq 4.4*)’.

$$deg_j = \sum_{k=1,k\neq j}^{i} A_{ki} \qquad (Eq\ 4.4)$$

where the summation A_{ki} is obtained for the rest of interconnected nodes with the *node j* whose degree is being calculated. Betweenness centrality definitions were used as were discussed in previous studies [114- 115]. Accordingly betweenness centrality

is defined as the number of shortest paths from all nodes to all other nodes that pass through the node. Betweenness centrality of a node is defined as given in equation '*(Eq 4.5)*'.

$$B_i = \sum_{j \neq i \neq k \in N} \sigma jk(i)/\sigma jk \qquad (Eq\ 4.5)$$

Here *σjk* defines the number of shortest paths between nodes j and k whereas σjk(i) indicates number of those paths which pass through node i.

Closeness centrality being the major component of average centrality was computed long with betweenness and node degree centralities. Closeness centralities are one of the key node centrality measures in networks [116-117]. Length of the shortest paths of the nodes in connected graphs is measured as their distance metric. Closeness centrality is referred as the reciprocal of node farness which is the measure of sum of all its distances from the rest of the network nodes [118-119]. The closeness centrality of a node *i* could be defined as in '(*Eq 4.6*)'.

$$Closeness_i = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1} \qquad (Eq\ 4.6)$$

Rscript5:

```
library('tnet')
HMM_Model = read.table("D:/ModelPercent.txt")
HMM_Model_Mat = as.matrix(HMM_Model)
Betweeness= betweenness_w(HMM_Model_Mat, directed=NULL, alpha=1)
Closeness = closeness_w(HMM_Model_Mat, directed=NULL, gconly=TRUE,
precomp.dist=NULL, alpha=1)
Degree = degree_w(HMM_Model_Mat,measure=c("degree","output"),
type="out", alpha=1)
Distancemat = distance_w(HMM_Model_Mat, directed=NULL, gconly=TRUE,
subsample=1, seed=NULL)
W_Richness = weighted_richclub_w(HMM_Model_Mat, rich="k",
reshuffle="weights", NR=1000, nbins=30, seed=NULL, directed=NULL)
```

Along with the degree, betweenness and closeness centralities, out-degree and normalized closeness centralities of nodes were also calculated. Out-degree is defined as the sum of weights on ties originating from a node which indicates the out-strength of that particular node, while the normalized closeness is the closeness which has been normalized as divided by N-1 where N is the total number of nodes. By combining all the centrality measures we calculated the average centrality as in '(*Eq 4.3*)'. The average centrality measure behaves and changes the same way as the closeness centrality and out-degree centrality measures as shown in the results section. But the added advantage of average centrality is that it is comprised of

complete set of degree of a node, its out-degree, its betweenness, closeness and normalized closeness centralities. All these measures are important in defining hubs in one way or the other. Average centrality measure was used to define the hubs and non hubs by defining certain threshold as shown in the equation '(*Eq 4.7*)'.

$$Hubs = Ni > threshold \qquad (Eq\ 4.7)$$

where Ni is the i^{th} node from the node set N. The threshold value was defined by developing consensus as per previous studies [120-124] which in our case was 25% of the highest average centralities of the network. Some studies [120-122] used a cutoff of 5-20% of highly connected nodes as a measure to define hubs, while some used a static cutoff of 5 [124] and 8 [107] as well.

Hubs were divided into date and party hubs by dividing the hubs threshold into 2^{nd} threshold. The nodes above 2^{nd} threshold were marked as party hubs while below were declared as date hubs as shown '(*Eq 4.8*)' and '(*Eq 4.9*).

$$PartyHub = Hubs(i) > 2ndThreshold \qquad (Eq\ 4.8)$$

$$DateHub = Hubs(i) < 2ndThreshold \qquad (Eq\ 4.9)$$

where Hubs is the set of hub nodes and i^{th} node is the node being checked. The hubs greater than 2^{nd} threshold were declared as party hubs (static hubs) while the hubs below 2^{nd} threshold were marked as date hubs (dynamic hubs). There is no mutual agreement upon data and party hubs identification and one of studies [108] utilized betweenness centrality for the purpose. We propose average centrality measure for demarcation of dynamic and static hubs inspired by the concept introduced recently for the identification of date and party hubs [125].The level of 2^{nd} threshold was set as 55% of the highest average centralities.

4.5.5 Regular expressions for chromatin states motifs

Motifs of 15 states from all cell types have been compared with respect to states similarity. The motifs patterns have been compared manually and using multiple sequence alignment and 15 generalized regular expressions have been designed to give generalized patterns for various chromatin states motifs.

4.6 Dominating nodes set in chromatin states networks

4.6.1 Dataset

We obtained ChIP-Seq histone modifications (HM's) data for H1hesc and K562 from Encode repository and chromatin modifiers (CM's) data from gene expression omnibus (GSE32509) for both the cell types. The bam files of the data retrieved were converted to the bed files using bedtools [92]. The bed files of transcription factor (TFs) peak files for both the cell types were also retrieved from Encode repository as shown in the Table 4.2.

Table 4.2 Data set of all marks

TFs-K562		**CFs-K562**	**HMs-K562**	**TFs-H1hesc**		**CFs-H1hesc**	**HMs-H1hesc**
Arid3	Mafk	CBX2_CF	H3K9me3	Atf2	Plu1	CHD1_CF	H3K9me3
Atf1	Max	CBX8_CF	H3K36me3	Atf3	Pol2	CHD7_CF	H3K36me3
Atf3	Mef2a	CHD1_CF	H4K20me1	Bach1	Pou5	HDAC2_CF	H4K20me1
Bach1	Mxi1a	CHD7_CF	H3K79me2	Bcl11	Rad21	HDAC6_CF	H3K79me2
Bcl3	Ncor	ESET_CF	H3K27ac	Brca1	Rbbp5	JARID1A_CF	H3K27ac
Bclaf1	Nelfe	EZH2_CF	H3K9ac	Cebpb	Rfx	JARID1B_CF	H3K9ac
Bdp1	Nfe2	HDAC1_CF	H3K4me3	Chd1a	Rxra	JARID1C_CF	H3K4me3
Bhlhe	Nfya	HDAC2_CF	H3K4me2	Chd2	Sap30	JMJD2A_CF	H3K4me2
Brf1	Nfyb	HDAC6_CF	H3K4me1	Chd7	Sin3	P300_CF	H3K4me1
Brf2	Nr2f2	HP1g_CF	H3K27me3	Cjun	Sirt6	PHF8_CF	H3K27me3
Brg1	Nrf1	JARID1C_CF	Ctcf	Cmyc	Six5	RBBP5_CF	Ctcf
Cbp	Nrsf	KAT3A-CBP_CF		Znf	Sp1	SAP30_CF	
Cbx3	Nsd2	LSD1_CF		Ctbp	Sp2	SIRT6_CF	
Cbx8	P300	MI2_CF		Egr1	Sp4	SUZ12_CF	
Ccnt2	Pcaf	NCOR_CF		Ezh2	Srf	EZH2_CF	
Cebpb	Phf8	NSD2_CF		Fosl1	Suz12		
Cfos	Plu1	P300_CF		Gabp	Taf1		
Chd1a	Pml	PCAF_CF		Gtf2	Taf7		
Chd2	Pol2	PHF8_CF		H2az	Tbp		
Chd4	Pol3	PLU1_CF		Hdac2	Tcf12		
Chd7	Pu1	RBBP5_CF		Hdac6	Tead4		
Cjun	Rad21	REST_CF		Jarid1	Usf1		
Cmyc	Rbbp5	RNAPIIS5P_CF		Jmjd3	Usf2		
Znf1	Rfx5	RNF2_CF		Jund	Yy1		
Ezh2	Rnf2	SAP30_CF		Mafk			
H2az	Rpc155	SIRT6_CF		Max			
Hdac1	Sap30	SUZ12_CF		Mxi2			
Hdac2	Setdb1			Nanog			

Hdac6	Sin3			Nrf2			
Jund	Sirt6			Nrsf			
Kap1	Six5			P300			
Lsd1	Smc3			Phf8			
Maff	Sp1						
Tblr1	Sp2						
Tbp	Srf						
Tead4	Stat1						
Tf3	Stat2						
Thap1	Stat5						
Tr4	Suz12						
Trim	Taf1						
Ubtf	Taf7						
Usf1	Tal1						
Usf2	Zbtb						
Yy1							

4.6.2 Chromatin states learning

The histone marks along with the input control and CTCF have been utilized to partition the genome into various segments based on combinatorics as discussed in section 4.5.2. A 15 states model has been learnt by creating a virtual concatenation of both cell types [54].

4.6.3 Segment overlap Enrichments

Annotation data including RefSeq Genes, RefSeq Exons, RefSeq TSS, RefSeq TES, RefSeq2KbTSS, CpG islands, repeats and laminB1 have been retrieved from UCSC genome browser, while the bed files of CMs and TFs peak files have also been used for the overlap enrichment. Fold enrichment of all marks for the chromatin states have been calculated using the overlap enrichment facility of ChromHMM.

4.6.4 Correlation of Factors

HMM learning provided the emission matrix and fold enrichment resulted in enrichment matrix. The emission matrix contained the probability of a mark, while the enrichment matrix contained the fold enrichment. In order to find correlation between all the factors, the scale of measurement for all of them was equalized. The fold enrichments were converted into the probability values by the following formula for both the cell types ‘(*Eq 4.10*)’;

$$Probability\ of\ a\ Factor = Ci \div \sum_{i=1}^{n} Ci \qquad (Eq\ 4.10)$$

Where *"Ci"* corresponds to a column entry, it means a particular column entry divided by the summation of all entries of that column produced the probability of enrichment for a factor at that location. The formula was executed using Rscript4.

The emission matrix and the probability enrichment matrix were utilized to find the correlation between the marks for each state. Equation '(*Eq 4.2*)' is utilized for this purpose. The equation is utilized to find correlations of all the marks with the chromatin states and with each other for 15 states for both cell types. The total number of correlations for one state in this way was equal to 'm' factorial where 'm' is the total number of entries in a vector.

4.6.5 Complete Networks and Network complexes

Inter-factors correlation and correlation of factors with states were utilized to uncover the local network of each chromatin state for both cell types. Local networks were obtained for each chromatin state independently for both cell types. In order highlight the complete network of all chromatin states for each cell type, we ignored the interactions amongst the factors in each state (mainly highlighted in local networks). Complete network of each cell type was obtained via chromatin states correlations with all factors and connections within the states. BioLayout Express 3D version 3.3 developed by EMBL EBI [87] had been utilized to develop the interaction networks. The factors acted as nodes and the interactions as the edges between them.

Complete networks were subjected to Markov Clustering Algorithm (MCL) [126] in order to mine the closely related complexes in the network (via BioLayout platform). The entities having strong correlation were clustered to obtain complexes within the network. The inflation parameter was set to 2.2, the pre-inflation parameter was set to 3.0, and scheme parameter to manage computing resources in MCL was set to 6.0.

4.6.6 Minimum Dominating Nodes set in Chromatin States Networks

Dominating nodes are the set of nodes which fulfill certain properties of minimum dominating set for a given network [127] and act as network controllers. A minimum dominating set is defined as a subset of nodes from which rest of the nodes can be reached by one step [117, 127].

Following the principle of centrality lethality rule [114] and the idea presented in [115], we highlighted the MDNS in the chromatin state networks of both the cell types used in our study.

We calculated the degree, closeness and betweenness centralities of all nodes in complete networks of both cell types using an Rscript5 utilizing the Tnet R package [86]. All node centralities were then averaged to take one combined centrality measure for each Node using equation '(*Eq 4.3*)'. The averaged centralities of all nodes were compared to find the MDNS in the chromatin state Networks. The MDNS of both cell types were compared.

In order to highlight the importance of MDNS in chromatin networks we subjected the networks vulnerable to attack against the determined MDNS. We sorted the node members of the network according to their combined centralities. Starting from the higher combined centrality nodes successively we deleted the MDNS from the network until the network started to disrupt and the importance of MDNS became clear.

4.6.7 Cross validation and Reduced Chromatin States Networks

Chromatin states learning has been obtained by keeping one mark out of the set, means one fold cross validation has been performed in order to check the fold enrichment behavior of the CMs and TFs and its effect on the interaction networks. In this way 11 models of 15 states have been obtained. Comparison of cross validated models with the complete learned reference model had been performed using an R-script3 for both the cell types. This step was performed to identify the dependence of CMs and TFs enrichment on the presence or absence of a certain histone mark.

Along with the one fold cross validation, we learnt an HMM model bearing 10 states in order to check the effect of enrichment and the change in the interaction networks in both the cell types with respect to the MDNS.

4.7 Summary

Histone modifications combinations from ChIP-Seq profiles are retrieved via various computational approaches including HMM and clustering algorithms. We based our studies on ChIP-Seq histone modifications profiles and identified the patterns using various approaches including HMM, clustering and biclustering. We proposed a simple computational approach to segment the genome in an unbiased way using any

HMM platform. We considered ChromHMM as a case study in our case. Along with we applied clustering and the use of silhouette score to identify patterns in the genomic data of H1 profiles and used this as the basis for HMM states identification. We proposed efficient clustering and biclustering platforms for identifying histone modifications associations at both global and local scales using binarized data signals as input. Our approaches efficiently identify histone modifications combinatorics at both levels and would be useful for further downstream analysis.

Along with the identification of histone modifications combinations we studied the relationship between various factors of chromatin states networks. We studied the role of DNA motifs in setting the chromatin states and the role of different kinds of states in setting chromatin networks using graph centrality measures.

Chapter 5

RESULTS AND ANALYSIS

In this chapter, we explain step by step the results obtained as a result of methodologies followed in Chapter 4 described above. It is mainly comprised of 6 main sections and the summary concluding all the findings.

5.1 A simple computational approach for finding chromatin states

The analysis of the proposed computational approach for predicting chromatin states has been explained step by step.

5.1.1 Comparison of BIC scores

Chromatin states have been learnt for the defined non overlapping segments of different sizes. For each bin size various models from 2 to 30 states have been learnt. This learning at each level was subjected to 50 random initializations of model learning. At each state level (from 2 to 30) we picked model with the highest BIC score for data with both streams. The highest BIC Score and the log likelihood of the models (from 2 to 30 for 200bp bin size) have been plotted in the Figures 5.1a and 5.1b respectively. Both the plots show the continuous increase in the likelihood and the BIC score. This indicated the fact that at a fine resolution of 200bp, due to data dependency, BIC and log likelihood don't converge. So at this resolution BIC and log likelihood scales could not be used to obtain the converging points for states number. We tested the BIC for all bin sizes. We found the convergence at 50Kbp binning as shown in the Figure 5.1c. This convergence showed the highest BIC Score at 11 states model after which the BIC score started to decrease.

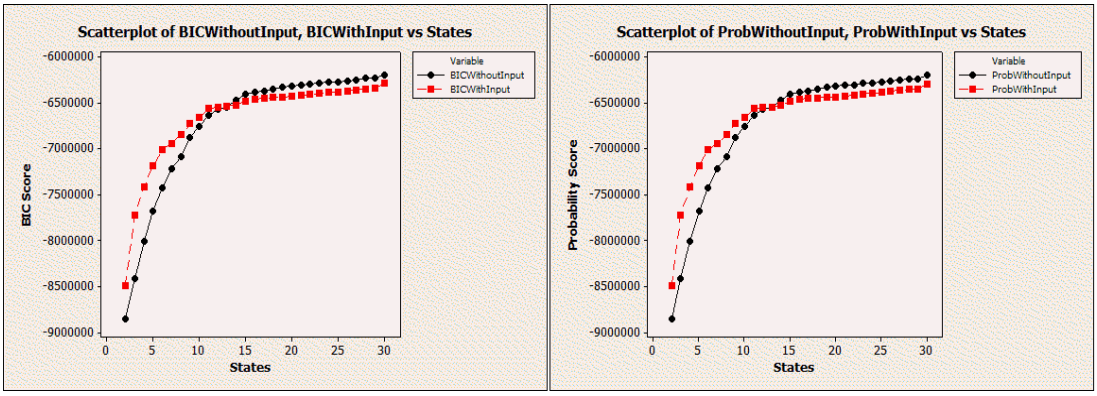


Figure 5.1 a) BIC Score of HMM Models for cases using input control versus without Input control. b) Probability score of HMM Models for cases using input control versus without input control.

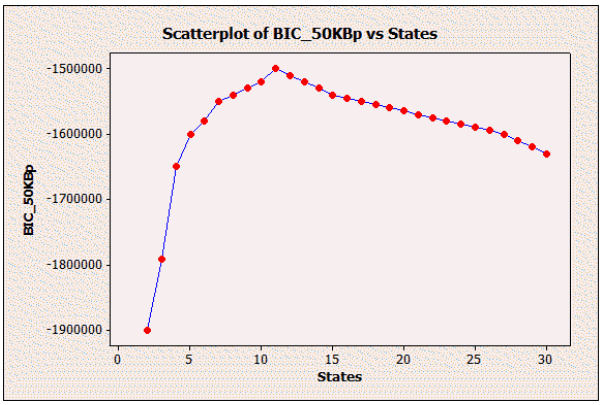


Figure 5.1c BIC Score of HMMs over 50Kbp non overlapping regions (binning) across various chromatin states

5.1.2 Computational states number identification

5.1.2.1 Emission means correlation

The vectors of each emission matrix (at each level) when subjected to correlation produced the values ranging from -1 to 1 where the values from 0 to -1 defined no correlation and the values between 0 to 1 defined positive correlation. The mean correlation of emission matrices at each states level (2 to 30) has been plotted in the Figure 5.2 for 200bp bin size to monitor the change. As shown in the plot the mean correlation of data with input control peaked at 14 states model and for data without input control at 16 states model and then both the lines smoothed out to show no further increase.

The mean correlations for other bin sizes have been shown in the Figure 5.3a and 5.3b for the data with input control. The plot 5.3a showed that 400bp binning gave rise to

13 states model, 600bp bin size produced 12 states model while 800, 1000, 5000 and 50Kbp binning produced 11 states models. The Figures 5.3c and 5.3d showed 15 states model for 400 bin size, 14 states model for 600 and 13 states models for 800, 1000, 5000 and 50Kbp bin sizes for data without input control. Plots 5.3b and 5.3d show combined plots of all the bin sizes for both cases respectively.

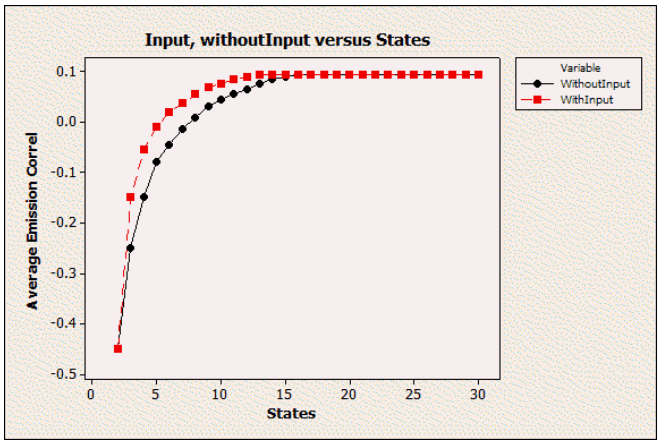


Figure 5.2 Mean correlation plot of Emission matrices for HMMs ranging from 2 to 30 states at 200bp resolution

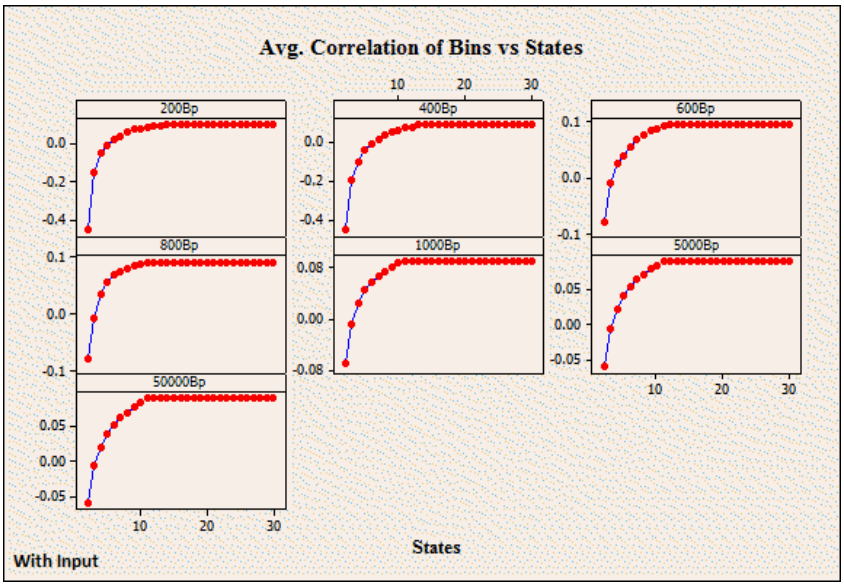


Figure 5.3a Cumulative Average Emission correlation plot of HMMs ranging from 2 to 30 states for all Bin sizes using input control.

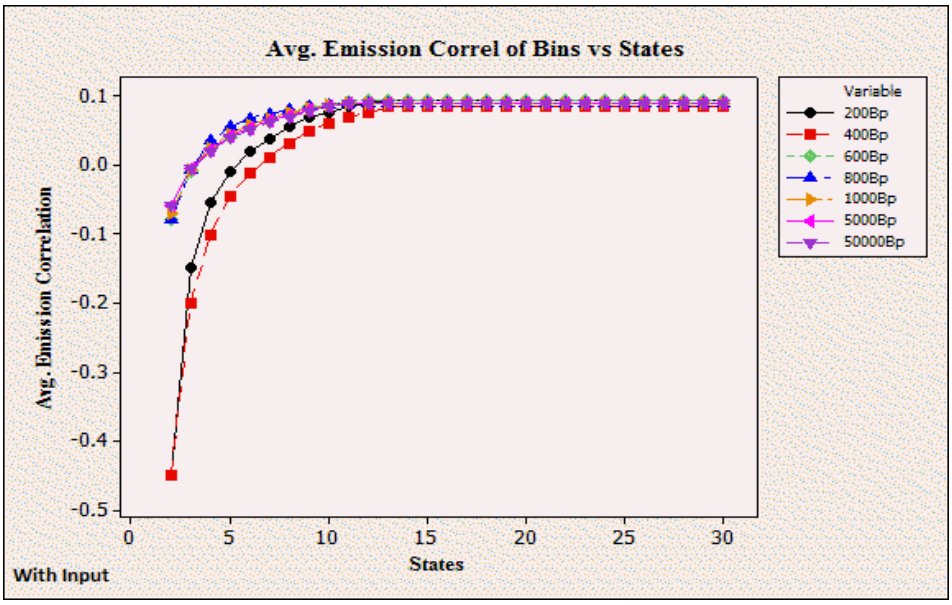


Figure 5.3.b. Comparison plot of Cumulative Average Emission correlation plot for all Bin sizes using input control.

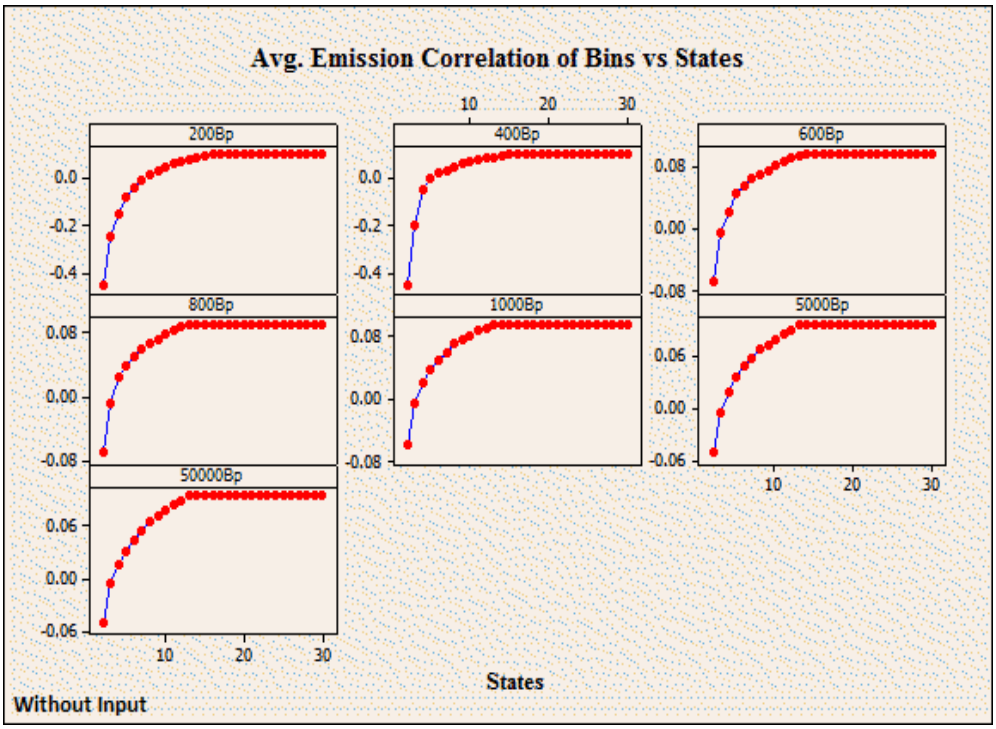


Figure 5.3c Cumulative Average Emission correlation plot of HMMs ranging from 2 to 30 states for all Bin sizes without input control.

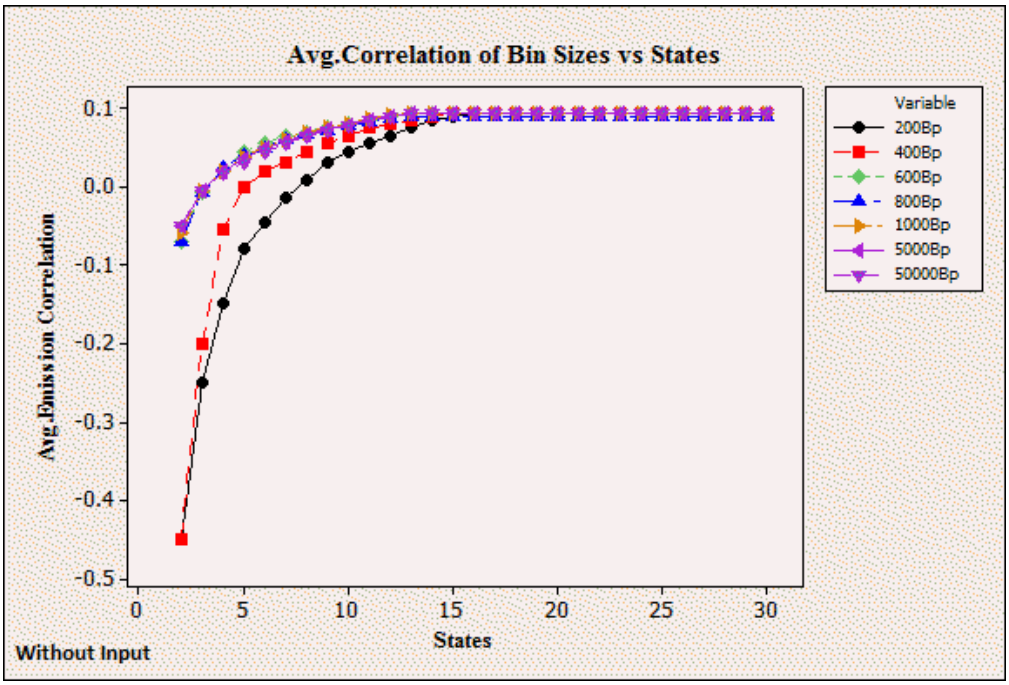


Figure 5.3d) Comparison plot of Cumulative Average Emission correlation plot for all Bin sizes without input control.

5.1.2.2 Clustering emission vectors

We used hierarchical clustering to cluster up the closely related and least distant vectors of the 30 states model for all the bin sizes. The highly correlated emission vectors (probability equivalent to 0.85 or greater) of each matrix were grouped up. Clustering results for 200bp bin size highlighted the fact that least distant vectors merged to give rise to 14 clusters for the data with input control and 16 clusters for the data without input control as shown in the Figures 5.4a and 5.4b. The correlation values of clusters have been shown in the Table 5.1. The emission matrices for 14 and 16 states models have been represented in the form of circular graphs obtained from Circos [128] as shown in the Figures 5.5a and 5.5b respectively.

The results of clustering of other bin sizes with and without input control have been shown in the Figures 5.6a and 5.6b respectively. The clustering results were similar to the convergence points of mean correlations for the respective bins, as represented in above section. The numbers of clusters were the same as the number of highly peaked mean correlation values.

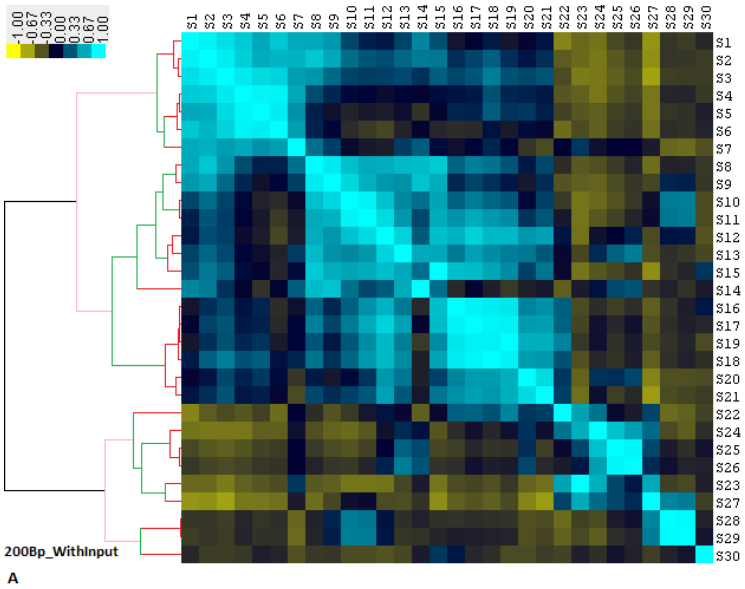


Figure 5.4a Hierarchical clustering of 30 states HMM for the data with input control for 200bp bin size producing 14 clusters.

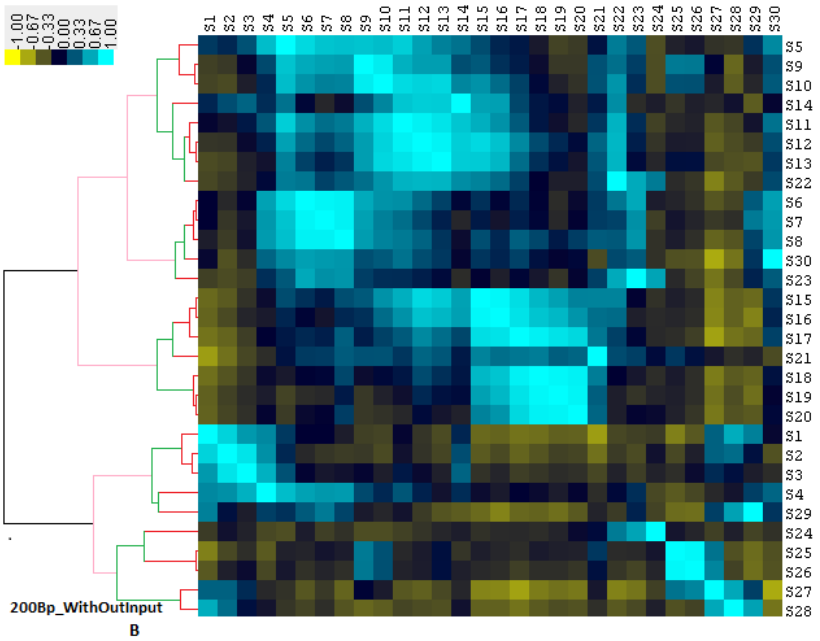


Figure 5.4b Hierarchical clustering of 30 states HMM for the data without input control for 200bp bin size producing 16 clusters.

Table 5.1 Clustering probabilities of 30 states HMM for 200bp Bin size data.

State Clusters- Data with input control	Correlation probability	State Clusters- Data without input control	Correlation probability
{S1,{S2,S3}}	0.92	{{S9,S10},S5}	0.87
{{S4,S5},S6}	0.90	{S11, {{S12,S13}}	0.82
S7, {{S4,S5},S6}	0.85	{S22,{S11,{S12,S13}}}	0.85
{S8,S9}	0.90	[S14{S22,{S11,{S12,S13}}}]	0.85
{S10,S11}	0.93	{S8,{S6,S7}}	0.85
{S15, {S12,S13}}	0.90	[S30,{S8,{S6,S7}}]	0.87
S14[{S15,{S12,S13}},{S8,S9}{S10, S11}]	0.90	(S23[S30,{S8,{S6,S7}}])	0.86
[{{S16,S17},S19},S18]	0.92	{S17, {S15,S16}}	0.89
{S20,S21}	0.91	[S21, {S17, {S15,S16}}]	0.85
{S24,{S25,S26}}	0.90	{S18, {S19, S20}}	0.89
S23,S27	0.90	{S1,{S2,S3}}	0.86
S22[{(S25,S26),S24}, {(S23)(S27)}]	0.91	S4	0.85
{S28,S29}	0.90	S29	0.87
S30,{S28,S29}	0.85	{S25,S26}	0.89
		{ S24, {S25,S26}}	0.85
		S27,S28	0.85

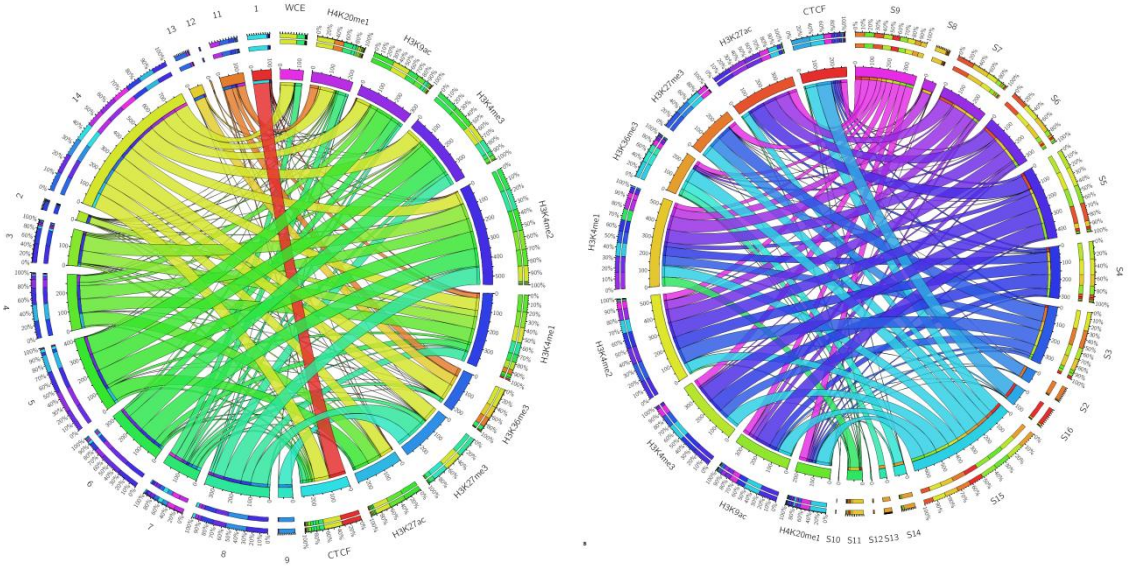


Figure 5.5a Emission matrix of 14 states HMM with input control for 200bp bin size. **b.** Emission matrix of 16 states HMM for data without input control for 200bp bin size.

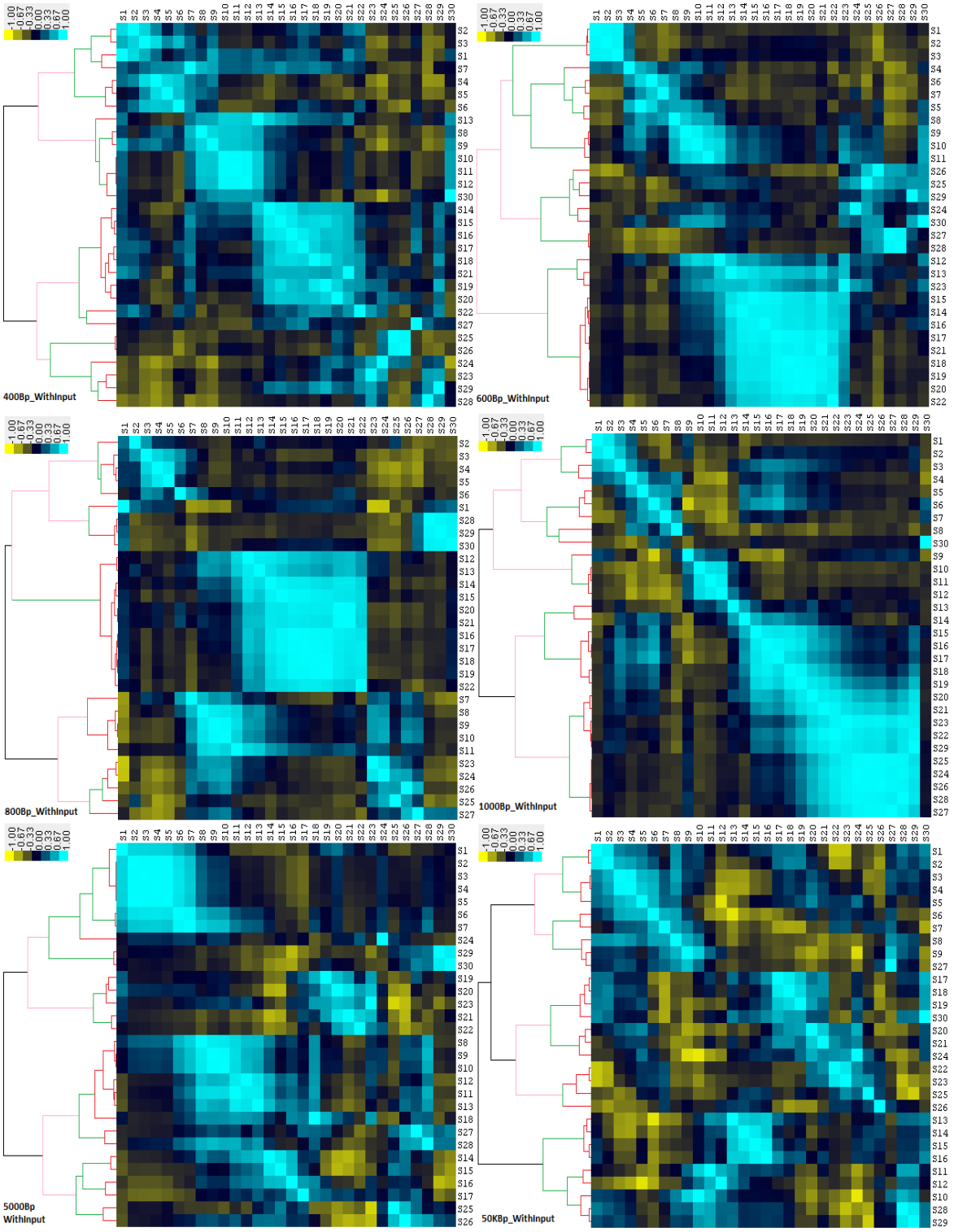


Figure 5.6a Hierarchical clustering of Emission correlation of 30 states model for the data with input control **a)** for 400bp bin size **b)** for 600bp bin size **c)** for 800bp bin size **d)** for 1000bp bin size **e)** for 5000bp bin size **f)** for 50,000bp bin size.

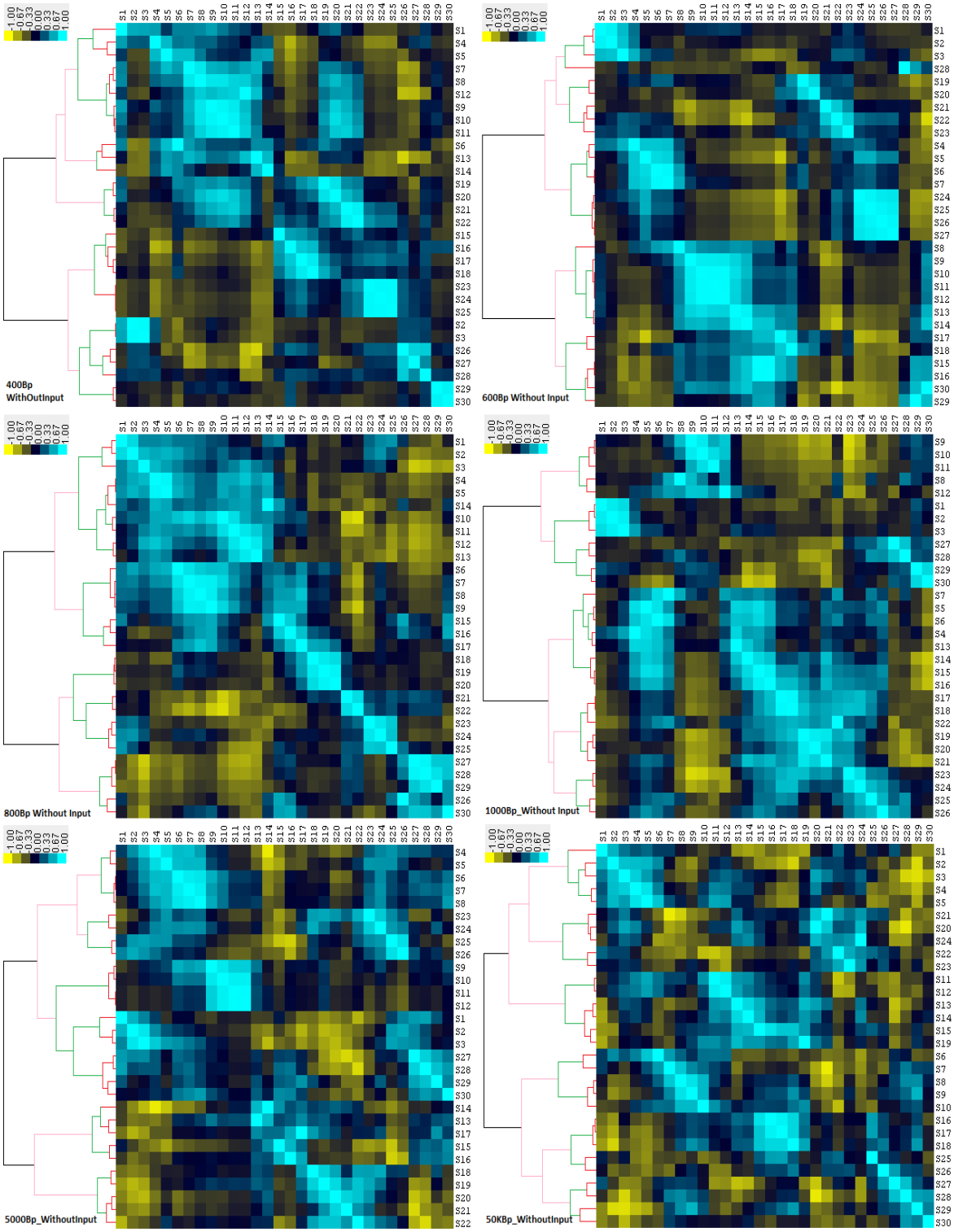


Figure 5.6b Hierarchical clustering of Emission correlation of 30 states model for the data without input control **a)** for 400bp bin size **b)** for 600bp bin size **c)** for 800bp bin size **d)** for 1000bp bin size **e)** for 5000bp bin size **f)** for 50,000bp bin size.

5.1.3 Biologically annotated models

We performed functional annotation of the states with respect to genomic elements at each states level as proposed earlier [52, 54]. In the annotations achieved; the active promoter regions were enriched with H3K4me2/3 marks along with RefSeq TSS, enhancer states were enriched with H3K4me1 and H3K27ac, repressed states contained high frequency of H3K27me3, H3K36me3 mark showed up towards the transcribed states with RefSeq genes and RefSeq exons enrichment, heterochromatin had very low signals for all marks ,the repeat regions contained either the high frequency of all marks or the presence of H3K9me3 only while the CTCF marked the insulator region. We developed a visualization strategy to identify the states number amongst 2 to 30 states models for the data with and without input control for all bin sizes. At each level we marked the states with respect to the annotations. We observed the appearance of the new functionally annotated state at each level.

This annotation was observed in all cases whereas we have shown the results of 200bp bin size only because there was an existing reference [54] for it and the other annotation followed the same procedure. This visualization scheme produced newly functional states till 14 states model (for 200bp data with input control), after that annotations started repeating. There was no addition of new functional annotations among the states. Data without input showed newly emerging states till 16 states model. This scheme worked both for data with input and without input as shown in the Figures 5.7a and 5.7b respectively for 200bp bin sizes. The 14 states model states were as: state1 an insulator, states 2, 3 the weak enhancers, state 4 was the strong enhancer, the active promoter regions belonged to state 5, state 6 was comprised of weak promoter, states 7 and 14 represented the repeats regions, poised promoter has been represented by state 8, state 9 shows the repressed region, the heterochromatin comprising of the maximum part of the genome was showed by state 10 and the strongly transcribed regions of the genome have mapped to states 11 and 13 while the weak ones to state 12.

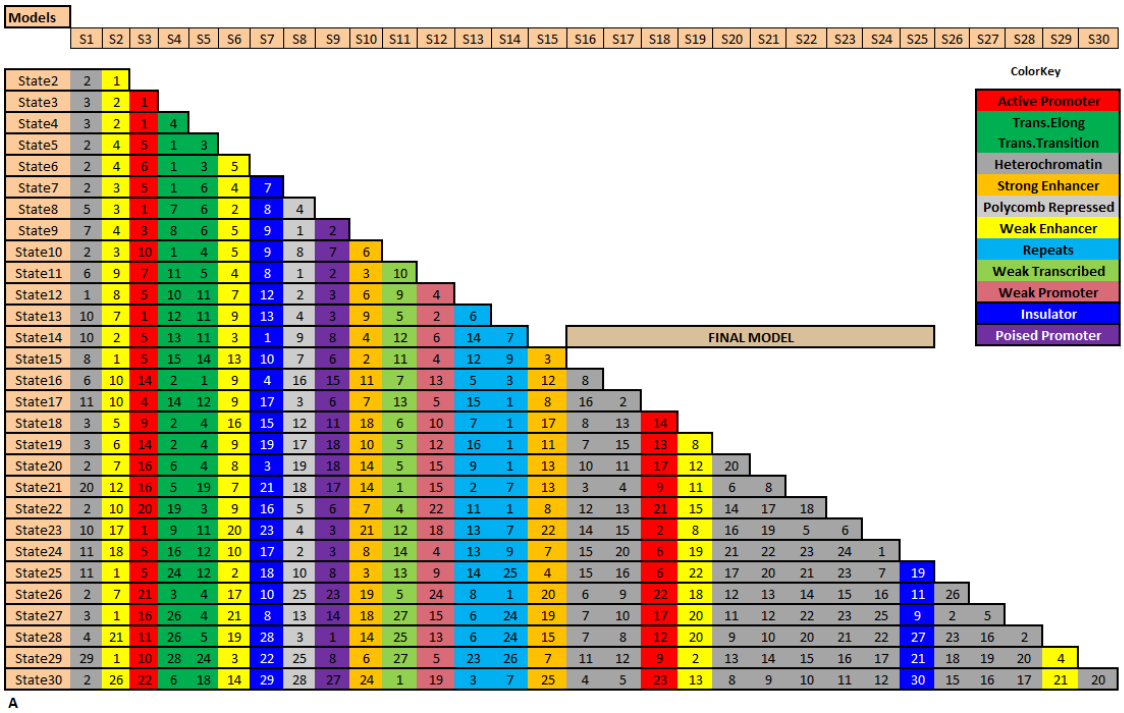


Figure 5.7a HMM states (ranging from 2 to 30 states) annotated visualization of data using input control over 200bp bin size. Similar annotated states at each states level are shown in same columns in order to observe the appearance of new annotated state. 14 states model is highlighted as the final model in visualization scheme.

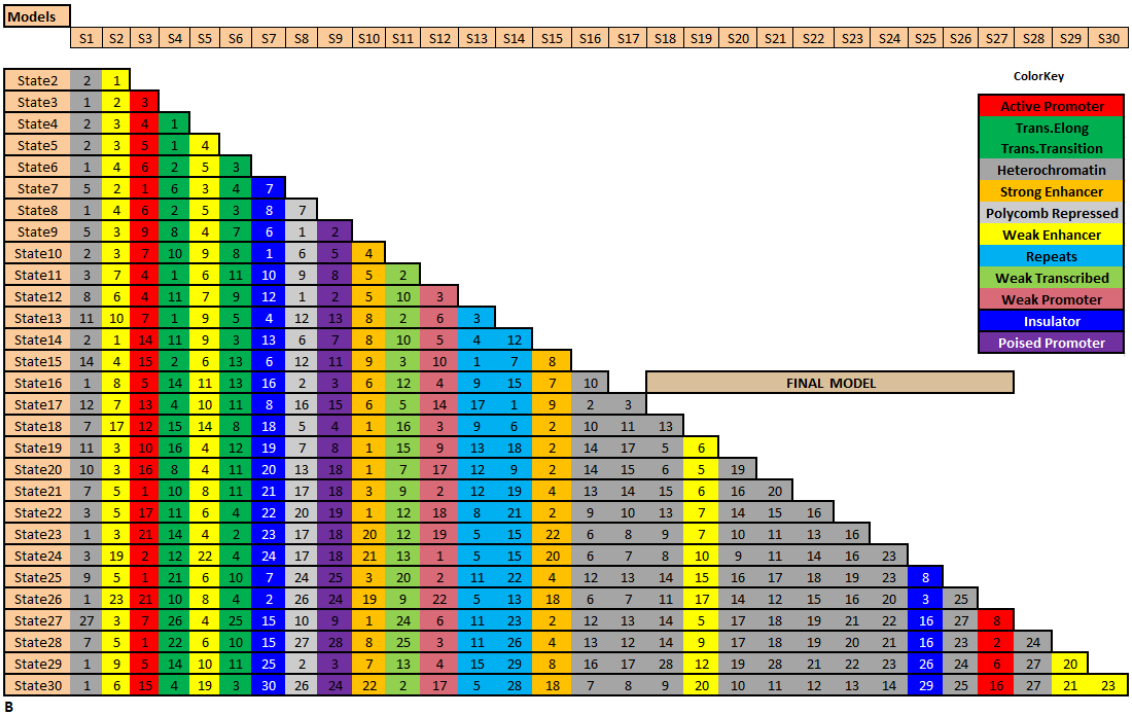


Figure 5.7b HMM states (ranging from 2 to 30 states) annotated visualization of data without input control over 200bp bin size. 16 states model is highlighted as the final model in visualization scheme.

5.1.4 Comparison with the reference model

5.1.4.1 Emission based comparison

Emission means correlation along with clustering and annotation produced 14 states model as the optimal model for 200bp in our study. We therefore compared the biologically annotated state of the reference model [54] with the 14 states model of 200bp from our study on the basis of emission vectors. The states are highly correlated having correlation greater than 90% in each case as shown in the Figure5.8a where R represents the state for the reference model and E represents our model which we called unbiased model. The color coding in Figure 5.8a represents states annotation for both models. Our model contained one strong enhancer while the reference contained 2 as shown in the Figures 5.8b and 5.8c respectively. Remaining states were exactly similar. The strong enhancer state in our model showed more than 90% correlation with both the states which means the emission vectors are highly correlated with the reference model. The emission matrices of both the models shown in the Figure 5.8b are highly correlated.

A

States	ComparisonStates	Correlation
Insulator	R8VersusE1	0.999
WeakEnhancer	R7VersusE2	0.998
WeakEnhancer	R6VersusE3	0.997
StrongEnhancer	R4VersusE4	0.912
StrongEnhancer	R5VersusE4	0.959
ActivePromoter	R1VersusE5	0.993
WeakPromoter	R2VersusE6	0.984
PoisedPromoter	R3VersusE8	0.998
Repeats	R15VersusE14	0.978
Repeats	R14VersusE7	0.930
Transcribed	R9VersusE11	0.963
Transcribed	R10VersusE13	0.996
WeakTranscribed	R11VersusE12	0.980
PolycombRepressed	R12VersusE9	0.999
Heterochromatin	R13VersusE10	1.000

B

Unbiased States	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	89.6	0.5	0.3	1.4	6.4	1.4	0.2	0.8	1.8	1.0
2	1.4	1.0	0.3	0.9	35.4	3.4	0.3	3.6	2.8	1.0
3	11.0	1.8	0.5	7.3	70.8	86.3	5.8	5.7	6.2	1.0
4	7.1	0.2	1.4	3.1	78.2	75.7	17.7	75.7	38.6	1.0
5	14.0	0.2	0.3	4.2	14.7	98.2	99.1	79.9	98.3	2.0
6	12.2	4.7	0.3	10.8	35.7	100.0	99.6	5.9	46.3	1.0
7	22.0	25.0	21.0	43.0	7.0	7.0	20.0	6.0	10.0	37.0
8	12.3	94.7	0.1	13.5	56.3	98.2	61.3	0.7	14.3	1.0
9	2.3	73.6	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	5.2	0.0	43.0	44.3	22.4	5.0	0.2	3.9	3.5	1.0
12	0.0	0.0	3.7	3.5	0.0	0.0	0.0	0.0	0.0	0.0
13	1.7	0.0	57.0	9.0	0.5	0.1	0.1	0.8	1.3	1.0
14	83.0	84.0	90.0	80.0	70.0	70.0	89.0	70.0	80.0	78.0

C

Reference States	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me	H3K27	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

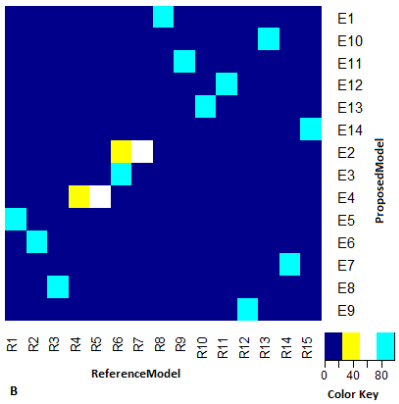
Figure 5.8 **a)** Emissions comparison of the Reference and the unbiased model. Correlation between similar states for the Reference and the unbiased model has been shown across the annotated states. **b)** Emission matrices of the Unbiased and **c)** the Reference Model are shown where each matrix represents the enrichment of a factor in the respective state.

5.1.5 State segmentation based comparison

The genome segments of our 14 states model has been mapped onto the 15 states reference model [54]. The comparison has been shown in the Figure 5.9 where part a represents the comparison in the percentage and part b represents the pictorial heat map representation. As highlighted in the Figures 5.9a and 5.9b that each of the biologically annotated state of our model mapped exactly on the similar state of the reference model e.g. insulator of the unbiased model mapped onto the insulator of the reference with more than 90% efficiency. Same is the case for others, except for one strong enhancer state of the unbiased model which mapped up to 25% on one of the enhancer of the reference and up to 71% onto the other enhancer state of the reference. The models are highly correlated both with respect to the segmentation and emission vectors.

	Reference	Ac.Prom	WeakProm	PoisedProm	StrongEnh	StrongEnh	WeakEnh	WeakEnh	Insulator	TranTransition	TransElong	WeakTrans	PolyRep	Heterochrom	Repeats	Repeats
UnbiasedModel	States	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
Insulator	E1	0.0	0.0	0.1	0.0	0.1	1.3	1.7	94.3	0.2	0.2	0.6	0.3	0.9	0.3	0.2
WeakEnh	E2	0.0	0.0	0.2	0.0	1.3	26.5	69.6	0.3	0.6	0.0	0.1	1.0	0.4	0.1	0.0
WeakEnh	E3	0.0	0.2	0.3	0.3	0.9	92.3	1.6	0.6	3.4	0.0	0.1	0.0	0.2	0.0	0.0
StrongEnh	E4	0.7	0.1	0.0	25.3	71.2	0.9	1.0	0.0	0.4	0.0	0.1	0.0	0.0	0.0	0.1
Act.Prom	E5	98.3	1.3	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
WeakProm	E6	1.8	97.1	0.2	0.6	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Repeats	E7	0.5	0.2	4.6	0.1	0.3	0.7	2.0	1.1	1.5	0.0	0.7	1.0	0.1	96.7	2.7
PoisedProm	E8	0.0	0.9	97.9	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.5
PolyRep	E9	0.0	0.0	0.2	0.0	0.0	0.1	0.8	0.3	0.0	0.0	0.5	93.7	2.2	1.9	0.3
Heterochrom	E10	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.0	0.9	0.5	97.9	0.1	0.0
TransTransition	E11	0.0	0.0	0.0	0.0	0.4	1.9	1.0	1.5	92.3	0.9	0.7	0.3	0.1	0.4	0.4
WeakTrans	E12	0.0	0.0	0.0	0.0	0.0	0.1	0.7	0.0	1.0	0.6	95.3	0.2	1.9	0.1	0.0
TransElong	E13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	6.8	90.8	2.3	0.0	0.0	0.0	0.0
Repeats	E14	1.7	0.3	0.8	0.9	1.5	1.4	0.2	0.1	0.3	0.1	0.2	0.0	0.0	1.7	90.8

A



B

Figure 5.9 a) HMM states segmentation comparison of the Reference and the unbiased model. **b)** Heat map of the comparison in part a shows the overlap between various genomic regions for both the models

5.2 H1 subtypes segmentation

In order to identify distinct combinations of H1 subtypes that are recurrent throughout the genome, we performed segmentation using an unbiased hidden Markov model (HMM) approach.

We found that a 5 states model optimally described our DamID data [97]. In this model, state 3 is the most abundant one, covering 50% of all probed locations, and corresponds to an average steady-state level of all 5 H1 subtypes binding (Figure 5.10a).

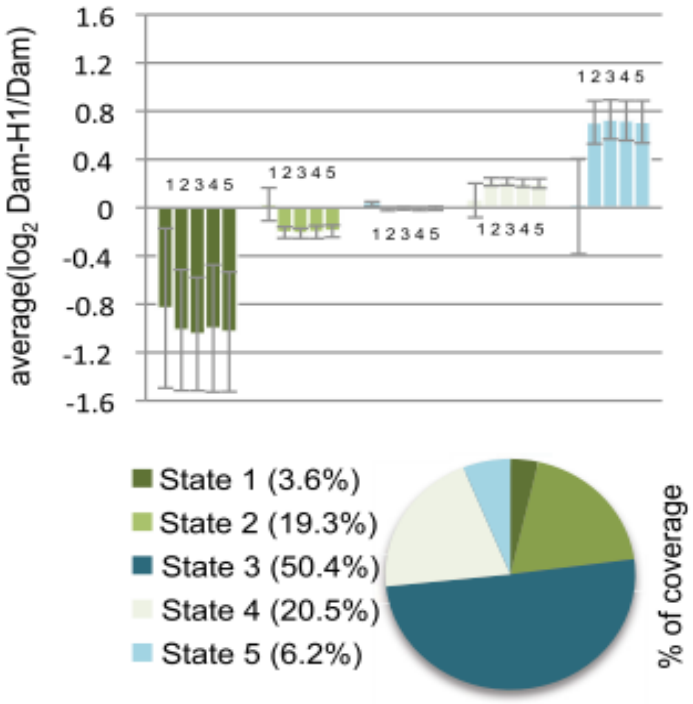


Figure 5.10a H1 binding can be described by 5 principal states, each consisting of a unique combination of H1 subtypes. The bar graph shows the average DamID values (log Dam-H1/Dam) of H1 subtypes H1.1–H1.5 (1, 2, 3, 4, and 5) within each of the 5 HMM states. Error bars correspond to the SD. The pie chart shows the percentage of genome coverage of each of the 5HMM states [97].

State 1 corresponds to regions strongly depleted of all H1 subtypes in the DamID data sets, and shows 3.6% genome coverage. States 2, 4, and 5 show a similar level of H1.1 binding and different levels of H1.2–H1.5 enrichment. State 2, which cover 19.3% of the genome, and state 4, with 20.5% coverage, show a slight depletion and enrichment, respectively, of H1.2–H1.5. State 5, covering 6.2% of probed locations,

corresponds to high enrichments of H1.2– H1.5 and highly variable H1.1 binding. The fact that none of the identified states distinguished between the subtypes H1.2–H1.5 supports the similarity of their distribution. We further corroborated this observation by independently calculating the correlation coefficients among our data sets for all somatic H1 subtypes. Indeed, the DamID results suggest the existence of two distinct groups: one consisting of H1.1 and one comprised of H1.2–H1.5 (Figure 5.10b) [97].

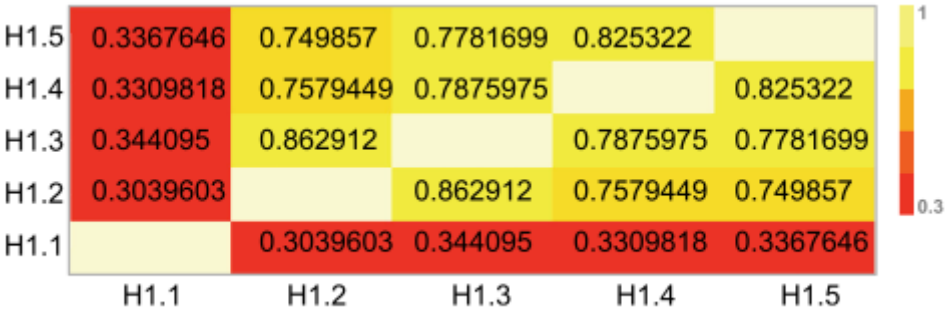


Figure 5.10b Correlation of H1 subtypes genomic distribution. The heat map shows the Pearson correlation coefficients calculated between H1 DamID binding values (log2Dam-H1/Dam) across all probes on the array [97].

5.3 ChromClust case study

5.3.1 Input data

Starting with ChIP-Seq data based on 10 histone marks (H3K4me3, H3K4me2, H3K4me1, H4K20me1, H3K9ac, H3K27me3, H3K9me3, H3K9ac, H3K36me3 and H3K27ac), one insulator protein CTCF and input control [54], we used our approach to identify signal enriched regions containing these modifications in nine cell types. We used binarized approach for clustering purpose based on present and absent signals of histone modifications. Our approach recovered possible combinations mapping various genomic locations including promoters, enhancers, transcribed regions, heterochromatin, repressed and repeat regions. This method is the first of its kind amongst the clustering algorithms which is not restricted to already defined promoters and enhancers [49]. We divided the genome into 5K base pair non overlapping regions. The algorithm followed semi- supervised approach by first creating the database (DB). Data of first cell type is processed for clustering and results have been stored in the DB. This was achieved in an unsupervised way

irrespective of any a priori knowledge base. Remaining eight cell types were clustered one by one based on previous clustering knowledge, in a supervised fashion. This could be achieved without using prior knowledge as per user choice. The added facility of DB is used to handle this very specialty of ChromClust.

5.3.2 Identified Clusters for various genomic regions

The clustering approach allows the user to select the similarity basis for clustering. Similarity basis is the way of achieving relaxed or strict grouping of chromatin combinations. It can vary from 60 percent to 100 percent. As the similarity measure approaches to 100 percent it tights up the clustering criteria and only exact clusters are observed. In this way number of clusters increases. We tested various similarity measures ranging from 70 to 100 percent match criteria. In case of a similarity measure e.g. 95%, the hamming distance between two non-overlapping regions is measured; if it is greater than equal to 95 percent then both the regions are part of one cluster and so on.

In case of 70 percent we obtained 11 clusters, 31 clusters resulted in case of 80% match, 90 percent match criteria retrieved 68 clusters while 100 percent resulted into 364 clusters. This cleared the picture that there were 364 unique regions in the genome with respect to binarized data in all nine cell types.

5.3.2.1 Annotations of clusters

We annotated the clusters with various regions obtained from UCSC genome browser. As an example case we describe here results for 70 percent match criteria where we got 11 clusters.

Cluster 1 spanning over 44% region of the 5Kbp non overlapping segments of the genome marked the heterochromatin region. Cluster 1 showed the low frequency or absent signals of all histone marks. A high enrichment of lamin B1 has been seen in these regions. Cluster 2 covering over almost 5% region showed high enrichment of all the marks and acts as a repeat region. Cluster 3 enriched in RefSeq genes, RefSeq exons, RefSeq TES and mark H3K36me3 showed the transcribed regions. Cluster 4 enriched with lamin B1 and H3K9me3 marks the repeat area. Enhancers span over clusters 5 and 9 (strong and weak respectively) with high enrichment in H3K4me1. Clusters 6, 7 and 8 represent the class of promoters (strong, weak and poised respectively). These regions show enrichment with CpG islands, RefSeq TSS (in case

of Cluster 6) and RefSeq TSS2Kb (in case of Cluster 6), H3K4me3, H4K20me1, H3K9ac, H3K27me3 (in case of poised one). Repressed regions have been captured in clusters 10 and 11 with enrichment in lamin B1 and H3K27me3.The enrichments and genome percents have been shown in Figure 5.11. The enrichments of clusters showed correlation with previous studies [42, 54, 61, 129-130].

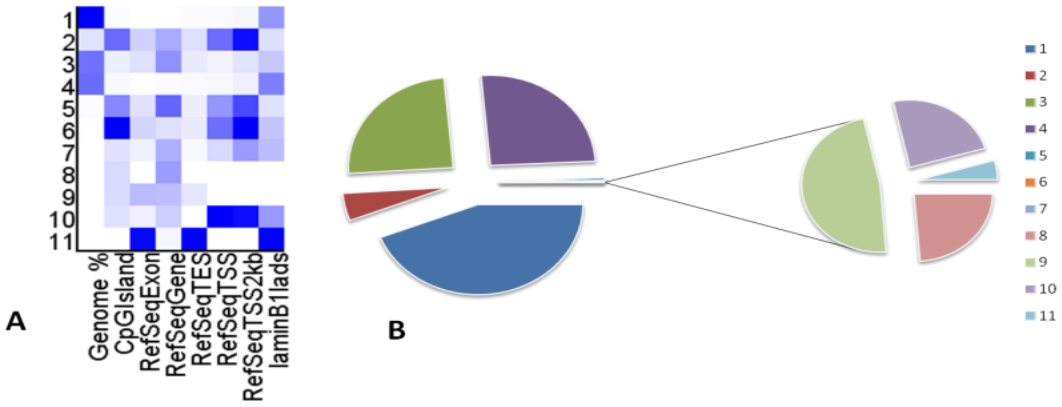


Figure 5.11 a Clusters annotation. **b.** Genome wise percentage of each cluster.

5.3.3 Comparison of ChromClust with other tools

The results of our tool for the data set used showed that there were 364 unique regions (in 5KBp non overlapping regions) in the genome based on histone modifications combinatorics. However using tools like ChromHMM [88], ChromaSig [49], spatial clustering [65] and HMMSeg [84], it's hard to identify the unique regions in the whole genomic structure. Our tool has this added facility of peeping deep into the genomic regions and studying the regions with respect to any similarity measure (starting from 60 percent and going up to 100 percent). Size of non overlapping regions again is a user based choice in our tool. ChromClust outperforms the others in terms of identifying similarity based regions and it easily splits the genome into unique regions as well thereby providing us the deep view of the genome under study as shown in the Figure 5.12. Therefore we conclude that splitting the whole genome into unique regions and studying the genomic structure with respect to presence of all possible histone combinations is the uniqueness of ChromClust.

Tools	Unique Regions
ChromClust	364
ChromHMM	Uniqueness not defined, Only provides user defined states
HMMSeg	Uniqueness not defined, Only provides user defined states
Spatial Clustering	Uniqueness not defined, Only provides user defined states/clusters
ChromaSig	Only works on promoter and enhancer regions

Figure 5.12 Comparison of ChromClust with existing tools for 5Kbp non over lapping regions.

5.4 ChromBiSim case study

Histone modification data over 5K base pair non overlapping locations was taken. Raw data of 11 marks for each of the four cell types were preprocessed. The preprocessing step normalized the marks with respect to the input controls and converted them into binary format.

The binarized data of each cell type was used for bicluster identification. In total 803 biclusters were retrieved for all cell types where 192 biclusters were mined for Gm12878 cell type, 190 for H1hesc, 218 for Helas3 and 203 for K562 (Figure 5.13a). Each bicluster contains information of the total number of non-overlapping regions which it covers along with the histone modifications present in it (Figure 5.13b). Along with the total number of bins, it also contains information of their chromosome wise location as well. This information of biclusters was exported to the output files which were used for further downstream analysis.

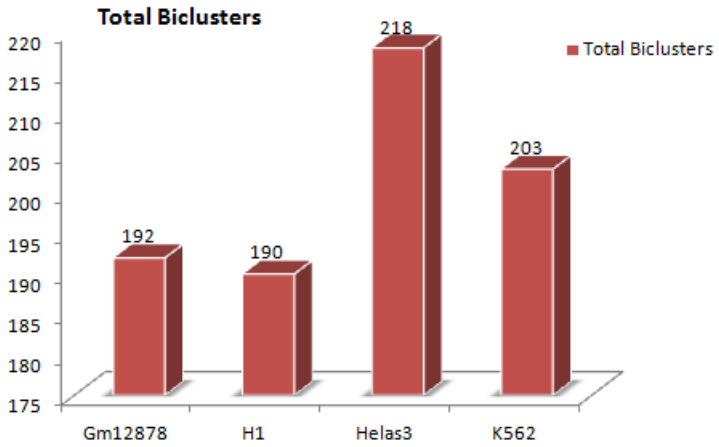


Figure 5.13a Bar chart representation of total number of biclusters in all cell types.

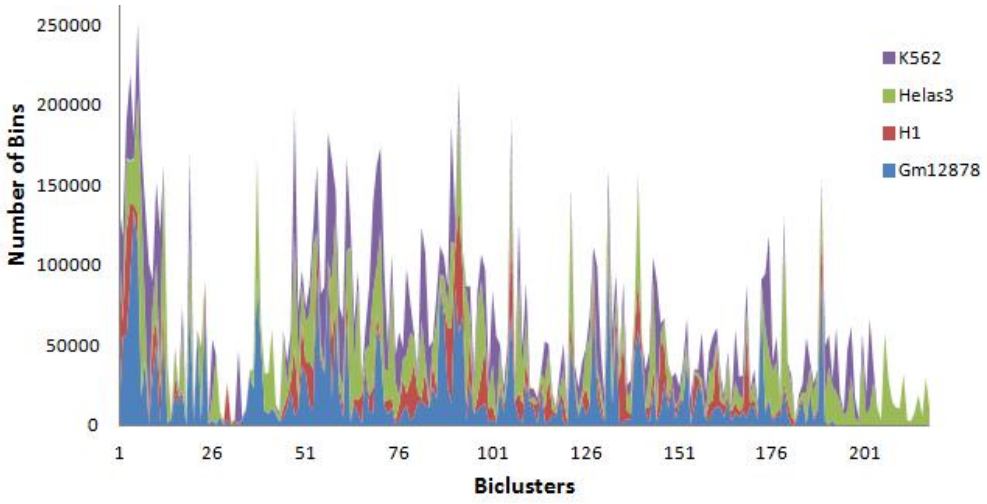


Figure 5.13b Bar chart representation of total number of non overlapping regions (bins) per biclusters in all cell types.

5.4.1 Comparison of biclusters across cell types

We used the output files generated by ChromBiSim for comparison purpose. We compared the biclusters of each cell type manually as well as overlapped the biclusters of all cell types using interactive tool Venny [131]. Overlapping results from both methods were obtained. In total 109 biclusters out of 803 were present in all cell types, whereas 25 were Gm12878 specific, 14 were K562 specific, 37 were Helas3 specific and 30 were H1hesc specific. H1hesc and Gm12878 contained 4 biclusters which were not found in Helas3 and K562, whereas Helas3 and K562 contained 17 biclusters which were not present in Gm12878 and H1hesc. It means the normal cell types contains 4 such combinations of histone modifications which were not found in cancerous cell types, while cancerous cell types contains 17 such combinations which were not found in normal cell types. This analysis shows that ChromBiSim could easily highlight the differences between normal and diseased histone modification cell types combinatorics. Interesting combinations were also observed like 12 biclusters were found which were present in H1 and K562, 6 biclusters were found in Helas3 and H1hesc only, 4 biclusters in Helas3 and Gm12878 while 8 biclusters were found in K562 and Gm12878 only. Among these combinations we observed such combinations as well which were missing only in one cell type, like 11 biclusters were such which were present in H1hesc, Helas3 and K562 while missing in Gm12878, similarly 10 biclusters were found which were

missing in K562 only, 8 biclusters were not present in Helas3 only and 24 were such which were not present in H1hesc only (Figure 5.14a). The analysis highlighted the cell type specific combinations along with combinations present in two cell types and missing in others, or three cell types and missing in one of them. This analysis and tool would be helpful for any other studies based on comparisons of epigenomic profiles as well.

Along with similarities and differences among the biclusters we annotated the biclusters with the various genomic regions obtained from UCSC genome browser. Annotation results marked the biclusters for various regions like insulators, repeats, polycomb repressed regions, enhancers, transcribed regions and poised and active promoters. ChromBiSim mined almost all possible combinations of histone modifications. Amongst them some of the annotated biclusters are highlighted (Figure 5.14b).

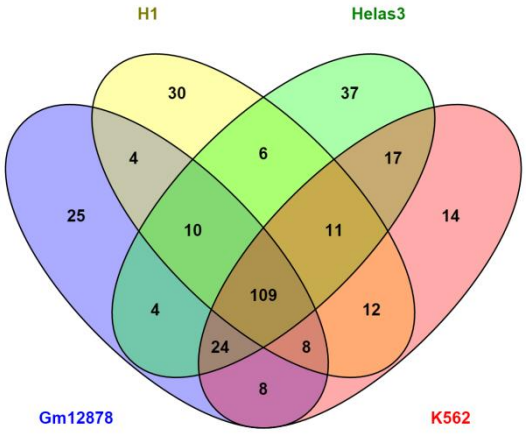


Figure 5.14a Venn diagram representing similarities and differences of biclusters in all cell types.

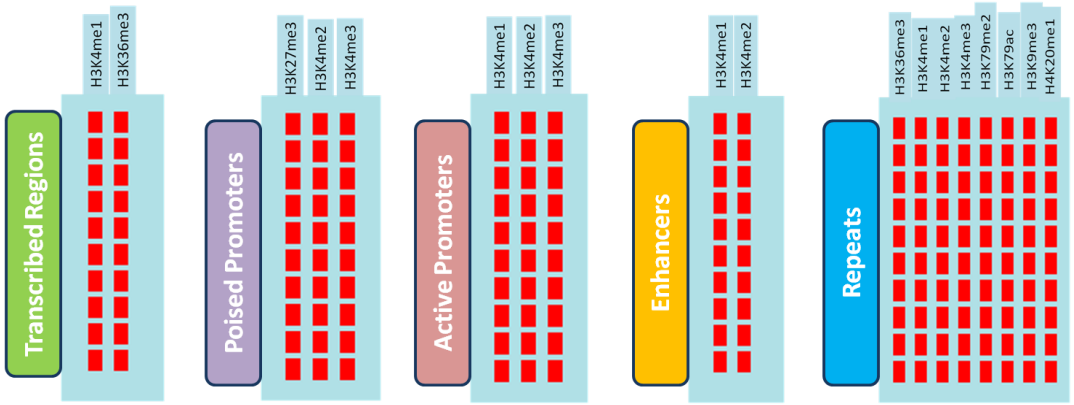


Figure 5.14b Some annotated biclusters with various genomic locations from all (4) cell types

5.4.2 Comparison with other tools

ChromBiSim outperforms other methods as it works on presence of marks and decodes the associations within seconds from the whole genome data set. The 1Kbp non overlapping data of four cell types took only few minutes to mine the patterns. If the binning (non overlapping regions) is narrowed down like 200bp which we also tested on various cell types, it took over maximum 30 minutes to identify whole genome patterns depending on the machine being used.

ChromBiSim is the first of its kind to mine biclusters from genome wide data based on binarized signals. ChromHMM [88] mined the combinatorics from the binarized data but it contained both the present and absent mark. Therefore ChromBiSim like ChromHMM [88] works on binarized data but unlikely it works on local signals and portrays epigenomic landscape from the present marks signals neglecting the absent ones.

It outperforms the two biclustering algorithms [50, 104] for histone modifications in several aspects as being user friendly, not dependent on peak calling methods to identify peaks for bicluster analysis, producing graphical results along with exporting the output in files for further downstream analysis.

5.5 Date and party hubs in chromatin state networks

5.5.1 Chromatin states learning

ChIP-Seq data of 5 cell types (H1hesc, K562, HepG2, Helas and Gm12878) comprising of 10 HMs along with input control and an insulator protein (CTCF) have been retrieved from public repositories. Genome segmentation was achieved using multivariate hidden Markov model (HMM) [88]. A virtual concatenation of cell types was created to learn 15 states HMM based on previous studies [54]. States of each cell type were individually annotated with RefSeq Genes, RefSeq Exons, RefSeq TSS, RefSeq TES, RefSeq2KbTSS, CpG islands, laminB1, P300, Ezh2, H2az, Pol2 and Ctcf binding site. The 15 annotated states were; State 1 being the repeat region was defined by the presence of H3K9me3 mark, state 2 marked as heterochromatin showed the low frequency of all marks. Both the states showed enrichment with LaminB1. The transcribed states numbered 3 to 5 showed high enrichment of H3K36me3, H3K79me2 marks along with strong annotation with Refseq exons, Refseq gene and Refseq TES. Enhancer regions spreading over states 6,7,11 and 12

form two major groups; states 6 and 7 closely related to transcribed regions lie immediately after them and it could be linked to the role of elongation factors in posing enhancers in ES cells in one of the studies [132]. This state behavior was shown in all cell types. Enrichment of H3K79me2 and H4K20me1 was seen in states 6 and 7. States 11 and 12 were immediately followed by promoter states (9 and 10). Enrichment of P300 and H3K4me1 was observed in states 11 and 12. States 9 and 10 marked as promoters were found highly enriched with H3K4me3, H3K4me2, H3K9ac, H3K27ac, Refseq TSS and Refseq TSS2Kb. State 8 marked as repeat region showed enrichment in all marks used. Enrichment of H3K27me3 was observed in states 13 and 14 which behave as polycomb repressed regions [133-134]. The last state 15 being highly enriched in CTCF was marked as the insulator state. The enrichment mapping shows high correlation with previous studies [54]. States distribution along with sequence of steps followed in study has been shown in Figure 5.15 while state enrichments of all cell types have been shown in Figure 5.16.

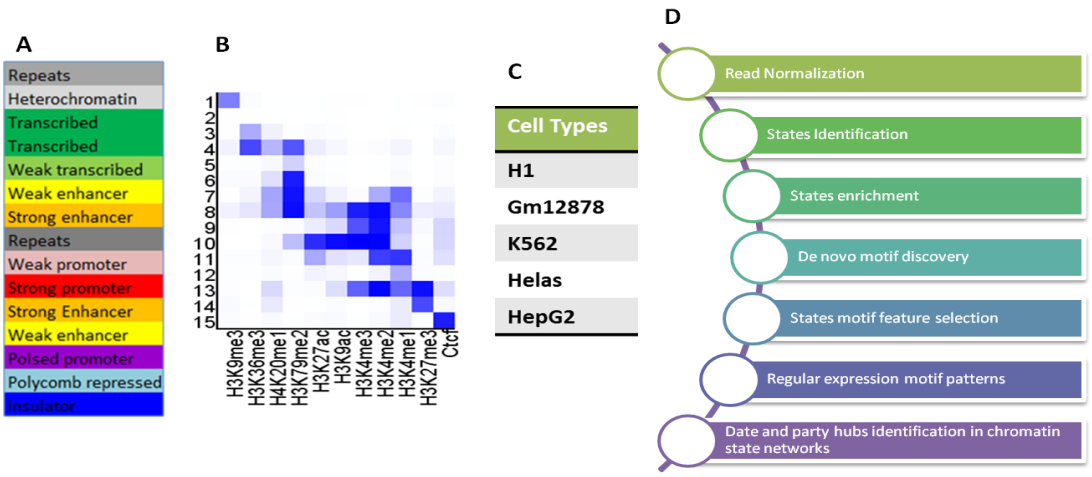


Figure 5.15 A. Sequentially annotated chromatin states key. **B**. Emission matrix representing histone marks patterns in each state. **C**. Cell types used in study. **D**. Sequence of steps followed in study.

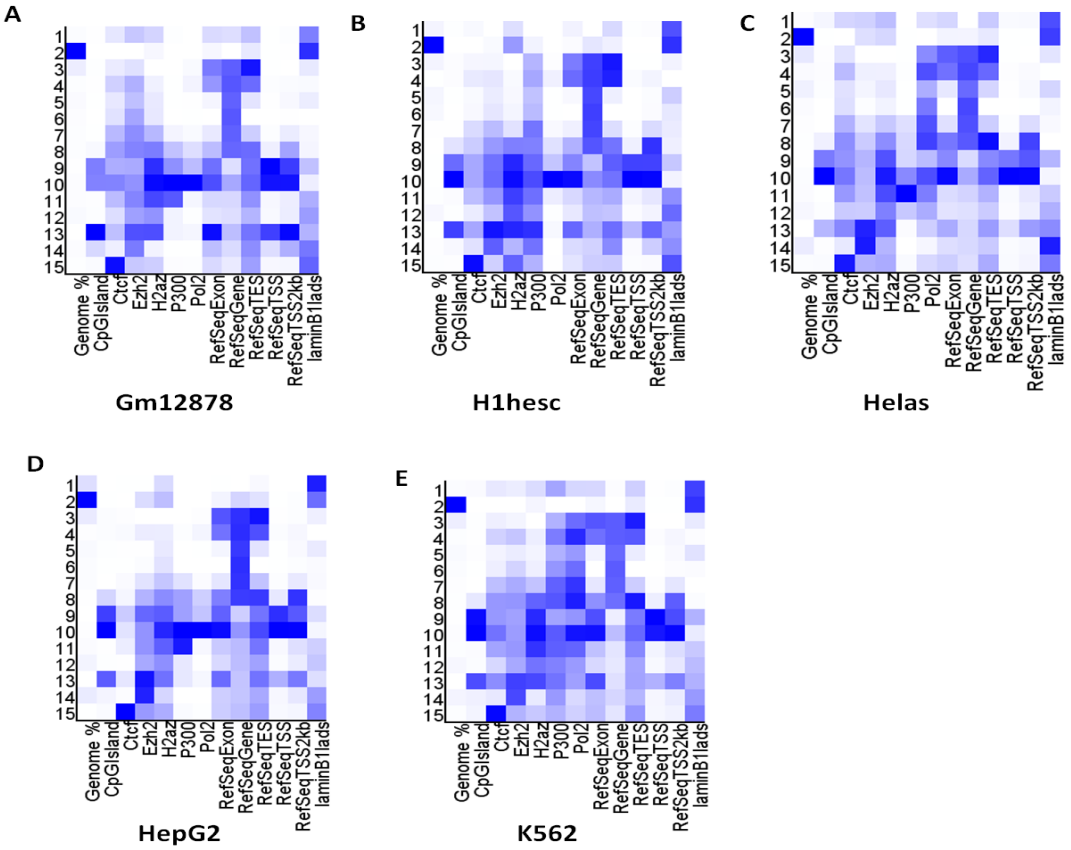


Figure 5.16 State annotations of chromatin states in all cell types (Gm12878, H1hesc, Helas, HepG2 and K562)

5.5.2 Denovo motif learning

DNA motifs as representatives of chromatin states of 5 cell types have been obtained via Homer software [91]. The chromatin state segmentations of each cell type were subjected to de novo motif discovery using default parameters used for histone marks. As each chromatin state is represented by combination of histone modifications therefore we tried to pick the motifs for each state and compared their behavior across each cell type. In total 1845 motifs for all cell types were obtained with significant p-value i-e <= $1e\text{-}^{10}$ [91]. Of the 1845 motifs 57 were found in state1, 128 in state2, 134 in state3, 129 in state4, 86 in state5, 94 in state6, 151 in state7, 140 in state8, 138 in state9, 144 in state10, 131 in state11, 139 in state12, 83 in state13, 171 in state14 and 120 in state15. Percentage distribution of each state with respect to motifs has been shown in the Figure 5.17a.

In each chromatin state, two types of motifs have been found, one cell type specific and the other which were present overlapping in different cell types and are marked as state specific ones. Cell type and state specific motif percent counts have been

represented in Figure 5.17b along with detailed combination of motifs in cell types in Appendix-A**.** It has been demonstrated that the shared motifs contribute almost 20% towards the total motifs count.

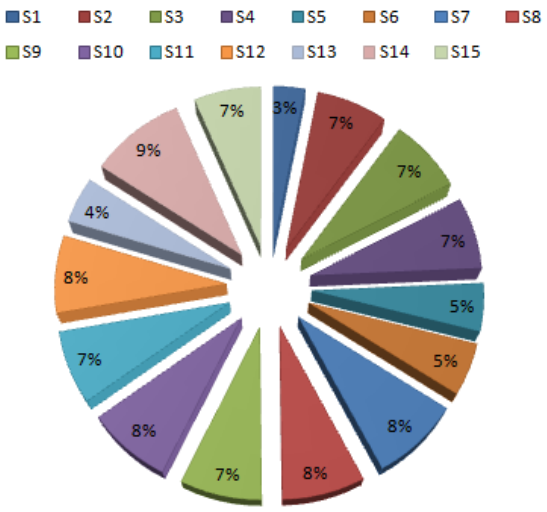


Figure 5.17a Pie chart representing the percentage distribution of DNA motifs in each chromatin state where S1 to S15 represents the 15 states of the HMM.

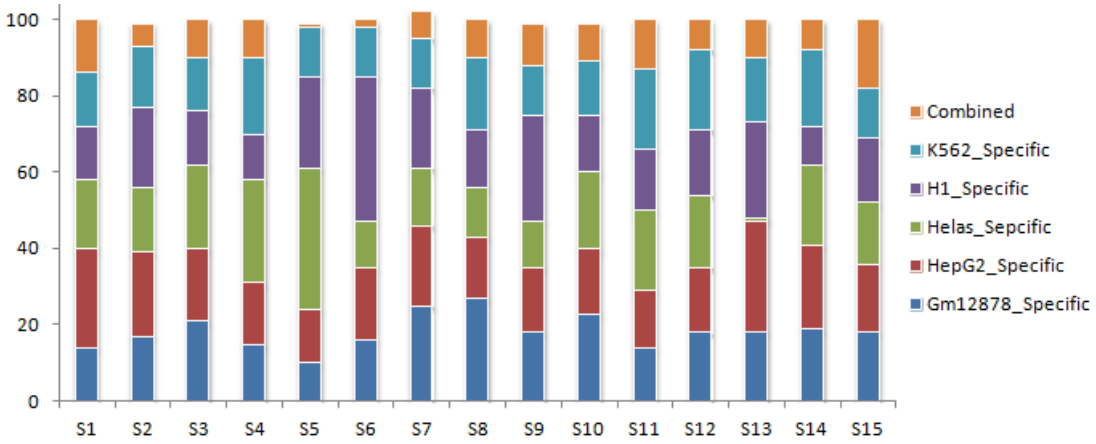


Figure 5.17b Bar chart representation of various types of DNA motifs in 4 cell types.

Among the total motifs the highest count goes to enhancer states of all cell types presenting their ability for more cell type specific regulators [130]. The overlapping motifs for enhancer states were 30% in total of the 27.9% while the cell type specific were 70% of the 27.9%. Promoter states follow the enhancers and

contain the highest number of sharing motifs by multiple cell types [135], while the transcribed regions have lesser shared motifs than enhancers and promoters. Transcribed states were more inclined towards the unique motifs rather than shared ones [135]. Heterochromatic states, repeat regions, repressed states and insulator have lesser number of motifs than three significant states with respect to transcription process (promoter, enhancer and transcribed) of chromatin. The details of motif counts are highlighted in Figure 5.17c. Motifs for each state in all cell types have been highlighted along with their p-values, ranks and best match for known motifs in supplementary information. Top scoring motifs of all states (of all cell types) have been shown in Table 5.2 along with their motif sequences and p-values.

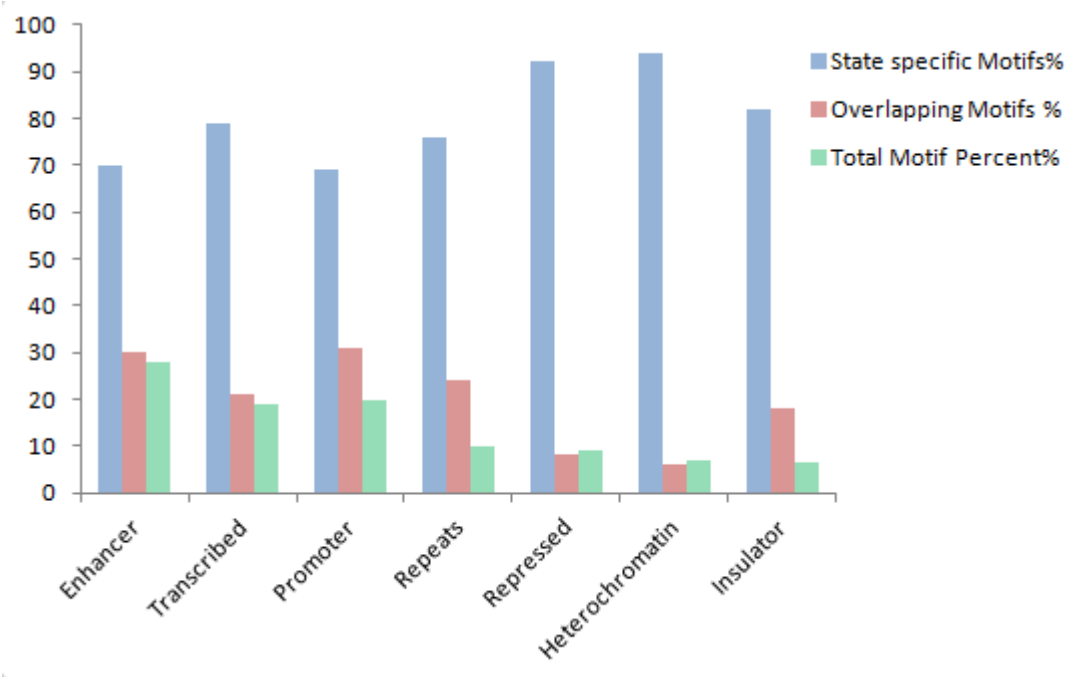


Figure 5.17c Bar chart representing chromatin state specific motif percent counts.

As mentioned earlier in chromatin states, the active enhancer states contained the highest number of motifs which is because of the presence of the active enhancer mark H3K27ac [137] at those sites. This is not surprising as chromatin mark across enhancers show dynamic nature across cell types [130]. The transcribed states have shown mostly the unique motifs because of the presence of transcriptional activity mark H3K36me3 and hence it makes the motif based regulation mostly unique [135]. Active promoter states sharing most of the motifs in all cell types show such behavior because of the combined effect of the H3K4me3 and H3K27ac at those sites [135].

Table 5.2 Top scoring DNA motifs of chromatin states in 4 cell types

States	Cell types	Rank	Motif	P-Value
1	Gm12878	1	CAYACTGGAGAG	1e-76
2	Gm12878	1	CTATCCCCAGGC	1e-30
3	Gm12878	1	ATGTAGCAAGTC	1e-26
4	Gm12878	1	CAGGGGATGCCG	1e-28
5	Gm12878	1	TATGGAAGSGGT	1e-16
6	Gm12878	1	ACTGCCAGCTAC	1e-21
7	Gm12878	1	CAACCTACCAAG	1e-36
8	Gm12878	1	GTTAAAACTGGT	1e-48
9	Gm12878	1	ACTTTCACTTTC	1e-126
10	Gm12878	1	ASTTTCABTTTC	1e-127
11	Gm12878	1	GRAASTGAAA	1e-242
12	Gm12878	1	TTTCACTTCC	1e-182
13	Gm12878	1	HAAATCGV	1e-45
14	Gm12878	1	TGTTCCCAGCTT	1e-42
15	Gm12878	1	GCCCCCTAGTGG	1e-12635
1	HepG2	1	CATACTGGAGAG	1e-25
2	HepG2	1	CCAGTTCRACTC	1e-29
3	HepG2	1	CACACGGCGCAC	1e-25
4	HepG2	1	GACGCCACTWGA	1e-28
5	HepG2	1	TTDATAGGTCCT	1e-20
6	HepG2	1	CAAGTGGCAGTC	1e-30
7	HepG2	1	AGAGCCAGTGTG	1e-36
8	HepG2	1	CGYWWWWW	1e-51
9	HepG2	1	GCCMYCTGGTGG	1e-146
10	HepG2	1	GNCGGAAT	1e-97
11	HepG2	1	GDWCAAAGTTCA	1e-400
12	HepG2	1	TGRACTTTGRMC	1e-249
13	HepG2	1	STTTCRVTTT	1e-53
14	HepG2	1	CAGTTACAATTG	1e-48
15	HepG2	1	GCCCCCTAGTGG	1e-12483
1	K562	1	ACACTGGAGAGA	1e-58
2	K562	1	CATGGCTGTTGG	1e-33
3	K562	1	CCAGGTGCACCT	1e-24
4	K562	1	CCACCAGGGGGC	1e-27
5	K562	1	CCTGCCGACAGC	1e-22
6	K562	1	GAAGACRTCCCT	1e-26
7	K562	1	GGTAAGCCACCT	1e-33
8	K562	1	CTACATTGGATC	1e-31
9	K562	1	GCCCCGCCCCCH	1e-156
10	K562	1	SCGRWATT	1e-145
11	K562	1	CSTTATCT	1e-294
12	K562	1	BCCTTATCWN	1e-158
13	K562	1	CGCTYTSC	1e-20
14	K562	1	ATGSTGAACCTW	1e-43
15	K562	1	GCCCCCTAGTGG	1e-10910
1	H1hesc	1	AGAGAAACCHTA	1e-35
2	H1hesc	1	CTGTCCAGATGT	1e-33
3	H1hesc	1	GACCATGGGTCA	1e-33

4	H1hesc	1	CAKCATCA	1e-28
5	H1hesc	1	ATGAGGCCCTGC	1e-24
6	H1hesc	1	CTTGCTCTTGGG	1e-42
7	H1hesc	1	GTCTTTGTCAAA	1e-54
8	H1hesc	1	CGNWNWNN	1e-107
9	H1hesc	1	CGNWWWTD	1e-217
10	H1hesc	1	SCGNWWWW	1e-181
11	H1hesc	1	ATGCAAATGA	1e-272
12	H1hesc	1	TTGTYATGCAAA	1e-272
13	H1hesc	1	CGNWTTTW	1e-62
14	H1hesc	1	GTATACTGTAAG	1e-23
15	H1hesc	1	CCACTAGGGGGC	1e-14580
1	Helas	1	GAGAAACCCTAT	1e-47
2	Helas	1	CGTTTTCATCAG	1e-38
3	Helas	1	CTCCCTGAGTGC	1e-35
4	Helas	1	TGGTGATGGTAC	1e-47
5	Helas	1	CAVRTTGACG	1e-27
6	Helas	1	CAGACTTAGCAG	1e-24
7	Helas	1	TAAAAAGTCGCA	1e-25
8	Helas	1	TAGGCAAGACTT	1e-33
9	Helas	1	CCACTAGATGGC	1e-186
10	Helas	1	GSKSTGAGTCAG	1e-176
11	Helas	1	NNATGASTCATN	1e-883
12	Helas	1	DATGASTCATHW	1e-394
13	Helas	1	TASTTTCAYTTT	1e-36
14	Helas	1	CTGATCCATCTA	1e-55
15	Helas	1	CCACYAGRKGGC	1e-8301

In order to systematically identify motifs involved in setting histone modifications resulting in particular chromatin states, we identified the ones which are also similar to the known motifs.

Amongst the known motifs Mafk_2, ZBTB33, RUNX, GATA3, CTCF(Zf), ETS1, Atf3, CEPBP and HOXA2 matched with known families of TFs. Mafk_2 is well known as transcription regulator. It may act as an oncogene or as a tumor suppressor depending on cell type specificity [136, 138-140]. Mafk_2 type matching DNA motif [TTG (T|G) ATTTTT (T|C) T] was observed in chromatin states 10 (strong promoter state) of HepG2 and K562 cell types. As is mentioned earlier that it act as a tumor suppressor and is cell type specific, it could be seen that this motif was observed in promoter states of two cancerous cell types, hence acting as cell type specific strong promoter motif. ZBTB33 is a transcriptional regulator which promotes histone deacetylation by recruitment of N-CoR repressor complex and leads to the development of repressive chromatin structures in gene promoters of the targets [141]. ZBTB33 type matching motif [CTCTCGCG (T|C) CAC] has been found in chromatin

state 10 (strong promoter state) of HepG2 and H1hesc cell types. ZBTB33 being transcriptional regulator factor has been found in strong promoter state of two cell types.

RUNX is also a class of TFs and plays its role in osteoblastic differentiation and skeletal morphogenesis [142]. RUNX type motif [CCCCGTGTGGTG] has been observed in strong enhancer states of Gm12878 and K562 cell types. GATA3 is one of the TF activator binding to the enhancers of T-cells. T-helper differentiation process is mediated by GATA3 [143]. GATA3 motif [CGAGATAA] has been observed in chromatin state 11 and 7 (strong enhancer states) of HepG2, K562 and Helas cell types. It is noted that this motif was observed in all cancer cell types. So it is a cell type specific enhancer state motif.

CTCF a TF with 11-zinc finger is included in various processes of gene regulation covering hormone-responsive gene silencing, chromatin insulation and promoter activation and repression. CTCF targets oncogenes along with tumor suppressor genes [144]. Due to its ability of activation or repression and targeting oncogene [57], CTCF type motif [GCC (A|C) (C|T) CT (A|G|T) GTGG] has been observed in enhancer state (state 10) of H1hesc, K562, Helas and HepG2 cell types. Gm12878 was the cell type which was lacking this motif in enhancer state. Hence all cancerous cell types and the embryonic stem cell type contained this motif which is in return marking the enhancer state due to histone combinatorics at these genomic locations.

ETS1 belongs to the class of TFs involved in Gata4 regulation via enhancer sites [145] and the DNA motif of ETS1 [(A|G) CTTCCT (G|C) (T|C) (T|A|G|C)] has been found in strong enhancer state (state10) of HepG2, K562 and Gm12878 cell types. Atf3 acts as transcriptional repressor and stabilizes the binding of inhibitory cofactors at promoter regions [146]. DNA motif [(A|C|G) TGA (C|G) TCA (T|C|G) (C|T|A)] of Atf3 has been found in state 10 of H1hesc and Helas cell types hence making it cell type specific enhancer motif. CEPBP is an enhancer binding and activator protein involved in TF dimerization [147] and its motif is found in Helas cell type in state 11 (enhancer state). HOXA2 proteins are involved in spatial and temporal regulation of embryonic development [148] and its motif [AGATGGATGGCG] is found in polycomb repressed state marked by H3K27me3 in K562 and Helas cell types. It is clear from above discussion that DNA motifs are important in setting histone

modification combinations and in return marking the chromatin states. This could also been seen in the cell types used in one of the earlier studies [135].

5.5.3 Chromatin state networks

Chromatin states obtained by histone marks combinations, enrichment of states with various elements of genome and enrichment of states with DNA motifs have been used to make chromatin state networks. Histone modifications, TFs, CMs, and DNA motifs have been used as nodes in these networks. The interaction between each of these factors was used as the connecting edge between the nodes. The probability of interaction was used as the weight of the edge in the network. The networks of the five cell types contained in average more than 300 nodes, 600 edges, average connectivity of around 3.6 and maximum connectivity of around 60. The exact detail of each cell type is mentioned in the Table 5.3.

Table 5.3 Network information of all cell types

Cell Types	Total Nodes	Total Edges	Avg. Connectivity	Maximum Connectivity
Gm12878	367	648	3.53	65
H1hesc	337	621	3.69	53
K562	337	610	3.62	58
Helas	360	643	3.6	60
HepG2	364	655	3.6	60

All networks were subjected to Markov clustering (MCL) algorithm in order to find clusters in the network. Each network was divided into 15 clusters, each cluster representing the chromatin state as shown in the Figure 5.18 where cluster1 to cluster 15 comprised of states 4, 14, 10, 11, 12, 3, 15, 5, 8, 7, 9, 2, 1, 6 and 13 respectively . The figure represents a generic network of all cell types. Cell type specific networks have been shown in Figures 5.19a, b, c, d and e.

Each network has been visualized as a planar grid in order to have a closer look at the paths available through the network. The edges connecting the nodes arranged themselves in a plane to give a planar representation as shown in the Figure 5.20. A closer look at the figure marks differences at some points of connectivity pointing the

differences in networks of cell types hence proving the point of their being independent cell types.

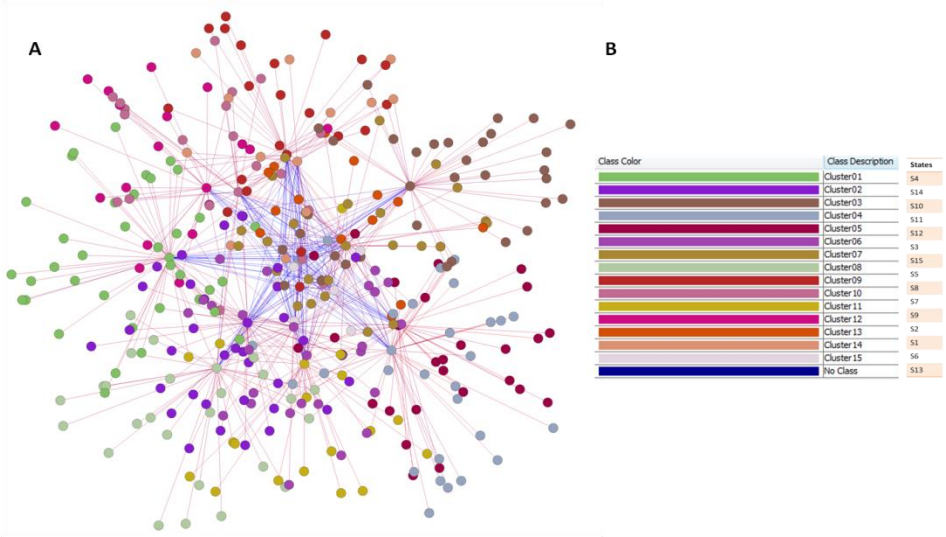


Figure 5.18 A. Generic chromatin states network of all cell types. **B.** Network clusters key with respect to states

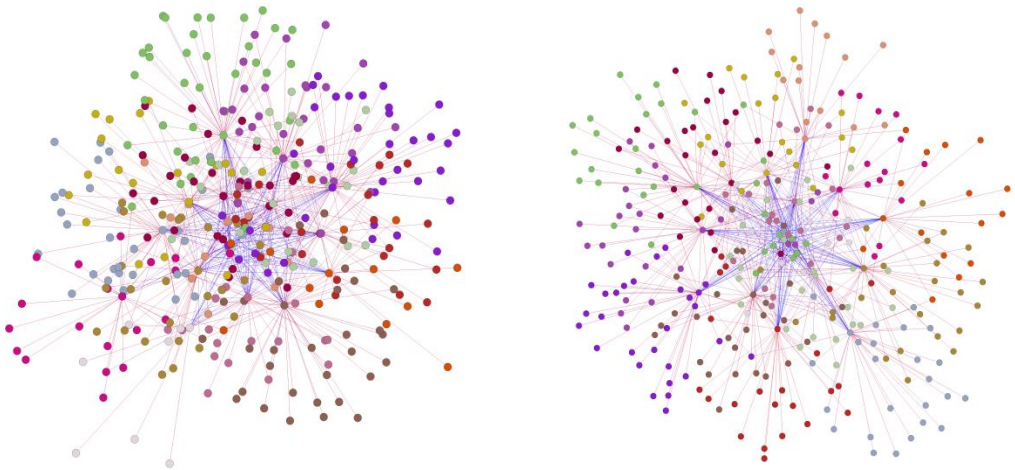


Figure 5.19 a. Gm12878 chromatin states network. **b.** H1hesc chromatin states network

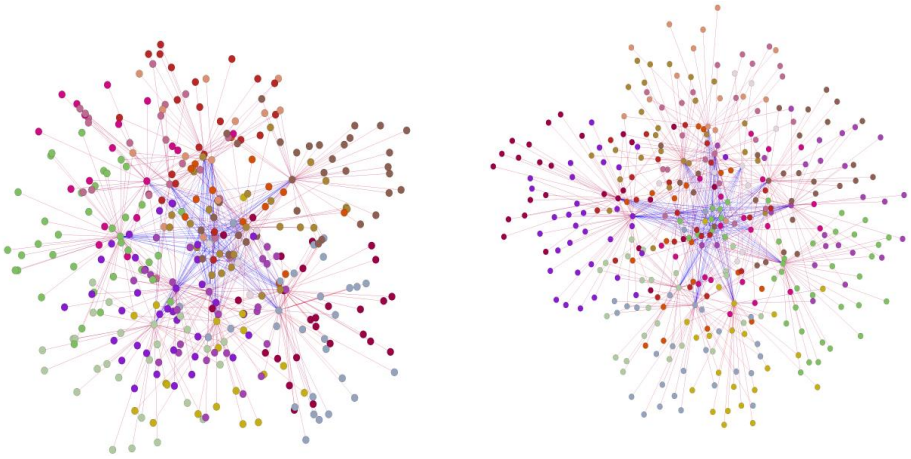


Figure 5.19 c Helas chromatin states network. **d**. HepG2 chromatin states network

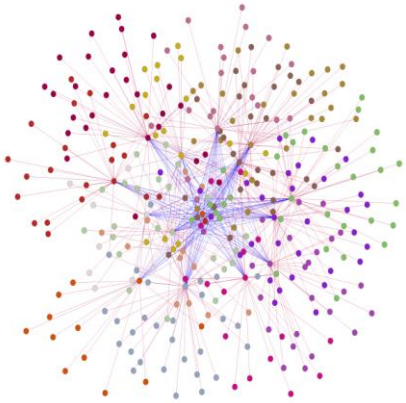


Figure 5.19 e. K562 chromatin states network

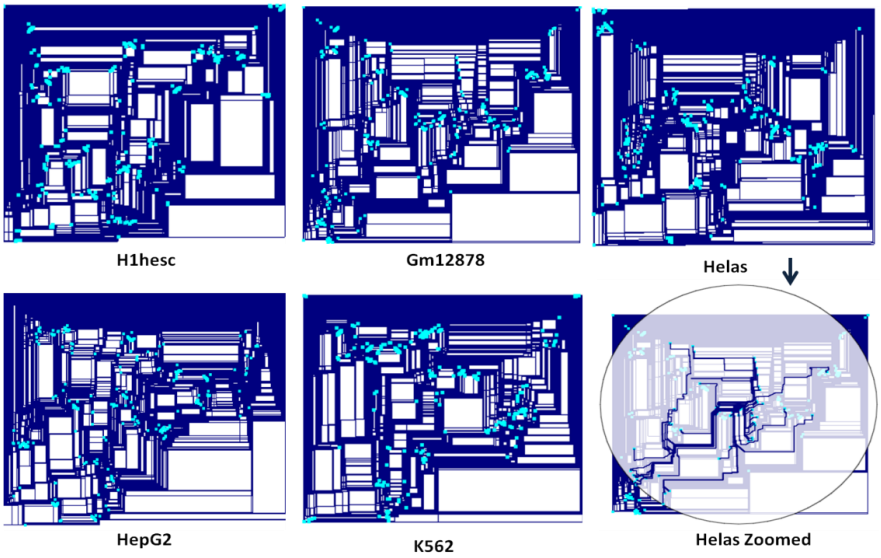


Figure 5.20 Planar grids of chromatin state networks of all cell types.

5.5.4 Date and party hubs in networks

Chromatin state networks were subjected to the calculation of various centrality measures in order to identify the hubs playing an important role in network connectivity. All centrality measures including closeness, degree, and betweenness were used to calculate a combined score represented as average centrality score as shown in the Table 5.4 and Figure 5.21a.

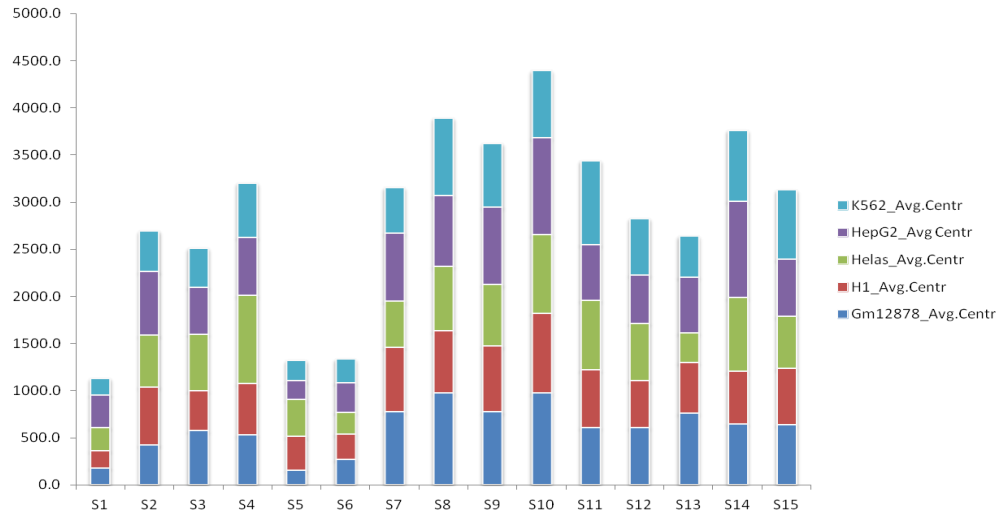


Figure 5.21a Average centralities of chromatin state networks of all cell types.

Individual centrality measures of all cell types have been illustrated in Appendix-B (Tables S5.1, S5.2, S5.3, S5.4 and S5.5). As clear from the average centrality scores state 10 (strong promoter) is the most significant hub among the networks of all cell types. Repeats regions (state 8) marked significantly by all HMs is the 2nd most significant hub followed by state 14 the Polycomb repressed state. State 1 (repeats marked by H3K9me3) along with states 5 and 6 (the transcribed regions) bear the lowest of the average centrality scores, lesser than the 25% of the most significant hub. Keeping the criterion of hubs identification; states 1, 5 and 6 were marked as non- hubs because they were lesser than the threshold values as was shown earlier in various studies [120-124]. Remaining states were above the threshold of hub criterion therefore they were also the part of hub entity set. Hence the hub entity set contained the heterochromatin, promoter states, strong enhancer states, polycomb repressed regions, poised promoters, strong transcribed states and insulator regions.

.

Table 5.4 Average centrality measures of chromatin states.

Nodes	Gm12878	H1	Helas	HepG2	K562
S1	179.4	182.5	243.8	348.6	172.6
S2	422.3	616.0	546.0	675.6	431.8
S3	571.5	424.5	602.4	496.6	415.2
S4	527.9	543.0	936.6	616.9	575.0
S5	156.8	354.6	392.4	198.7	215.6
S6	267.1	269.8	233.0	311.0	250.2
S7	775.6	684.8	484.0	725.6	480.9
S8	973.2	657.3	688.2	748.4	822.4
S9	771.2	698.5	651.8	824.1	673.3
S10	970.9	849.7	829.4	1028.5	714.4
S11	606.3	610.1	738.2	593.6	890.5
S12	605.6	501.4	607.0	511.2	597.8
S13	756.3	539.6	313.6	592.9	434.8
S14	647.2	554.3	781.4	1026.0	751.7
S15	637.9	593.7	556.4	602.8	735.5

Hub entity set was further classified into date and party hubs on the basis of 2nd threshold which was defined as 55% of the most significant hub. Hubs below the threshold were marked as date (dynamic) hubs while the ones above the threshold were declared as party (static) hubs. States 2, 3 and 13 were declared as date hubs in all cell types whereas the remaining ones were marked as party hubs as shown in Figures 5.21b and 5.21c. Date and party hubs are marked in Figure 5.21b in a bar chart representation while Figure 5.21c highlights connectivity of the respective hubs in the network. Among the states heterochromatic regions, one of the transcribed regions and poised promoter were the ones which behaved as the dynamic players in the chromatin and have been declared as the date hubs, while the enhancers, set of promoters, repeats, insulator and some of the transcribed regions were shown to be the static players and remain consistent within cell types in order to declare their role as party hubs. Though the values of party hubs varied amongst the cell types which showed the variation of connections within cell types but the overall effect of centrality scores remained the same for the demarcation of date and party hubs.

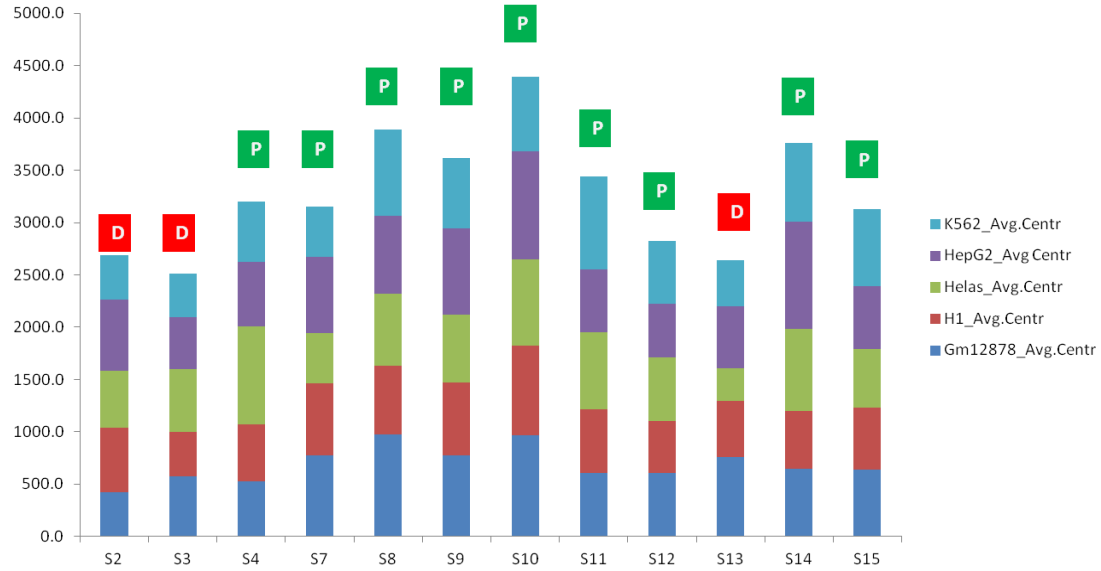


Figure 5.21b Bar chart representation of date and party hubs in chromatin state networks

Hence based on the idea introduced by [125] our introduced measure of average centrality worked in the same way as their combined centrality measure. Our average centrality measure has shown significant similar patterns like the closeness centrality as was introduced in several studies for hubs identification [109-113]. We therefore conclude that average centrality measure is a significant measure to define hubs along with their further classification of date and party hubs.

Along with network centralities role of DNA motifs was also observed in setting the date and party hubs. In case of date hubs, combinations of motifs in multiple cell types decreased while in case of party hubs it was opposite. Average of motif combinations was below 10 in case of date hubs while it was above 14 in case of party hubs as shown in the Tables 5.5 and 5.6. Party hubs contain motifs which are found in two, three, four and even all cell types. State 4, 8, 9, 10, 11, 12 and 15 contains combined motifs in 2, 3,4 and all cell types, while states 7 and 14 contains motifs in combinations in 2 cell types. Date hubs on the other hand contained combinations of 2 with the exception of one motif which was present in 3 cell types in state 3. Total percentage of motifs was around 70% in case of party hubs while around 18% and 12% in date and non-hubs respectively as shown in Figure 5.21d.

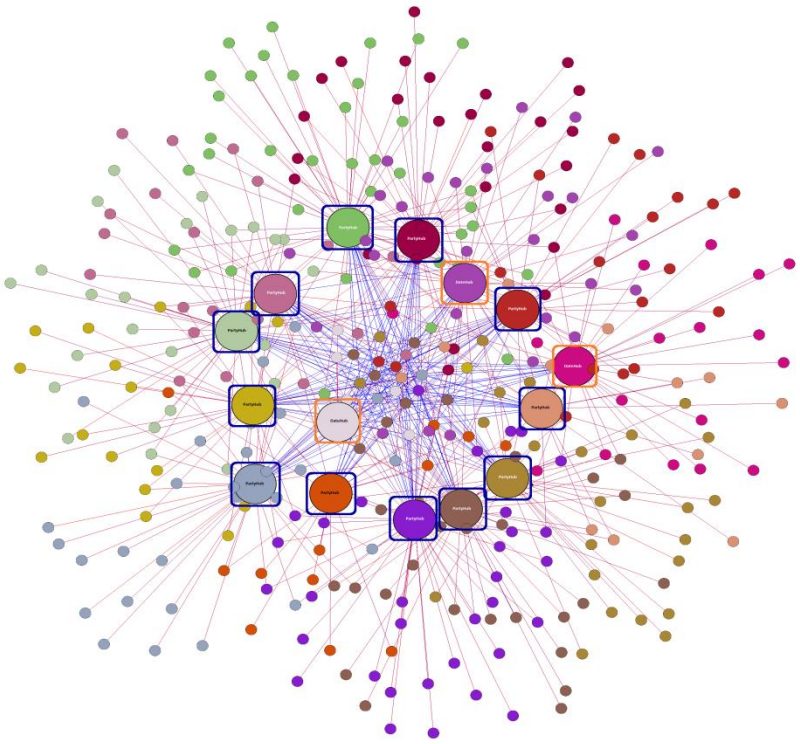


Figure 5.21c Network connectivity of dynamic and static hubs.

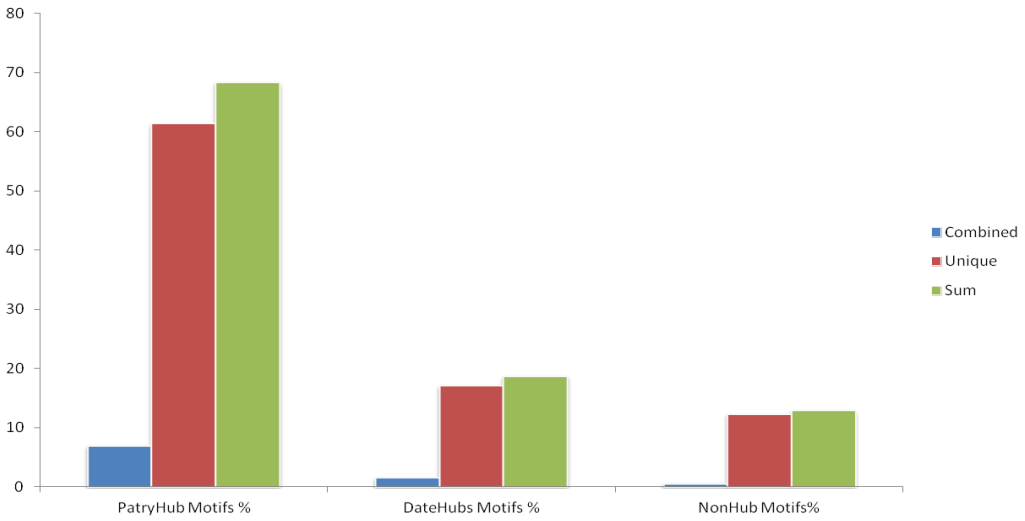


Figure 5.21d Percent count of date and party hubs in all networks

Table 5.5 Party hubs chromatin states with overlapping motifs in 4 cell types.

State 4	State 7	State 8	State 9	State 10	State 11	State 12	State 14	State 15	Avg
Gm12878_HepG2_Helas	Gm12878_H1	Gm12878_K562_H1	Gm12878_K562_Helas	Gm12878_H1	Gm12878_H1	Gm12878_K562	Gm12878_HepG2	Gm12878_Helas	
Gm12878_Helas	Gm12878_K562	Gm12878_K562_H1	All	Gm12878_Helas	Gm12878_H1	Gm12878_K562	Gm12878_HepG2	Gm12878_Helas	
Gm12878_Helas	Gm12878_K562	Gm12878_K562	Gm12878_Helas	Gm12878_Helas	Gm12878_H1	Gm12878_K562	Gm12878_Helas	Gm12878_Helas	
Gm12878_Helas	Gm12878_K562	Gm12878_Helas	Gm12878_H1	Gm12878_HepG2	Gm12878_HepG2	Gm12878_HepG2	Gm12878_Helas	Gm12878_K562	
Gm12878_Helas	Gm12878_Helas	Gm12878_HepG2	Gm12878_K562	Gm12878_K562	Gm12878_HepG2	Gm12878_HepG2	HepG2_K562	Gm12878_K562	
Gm12878_HepG2	HepG2_H1	HepG2_H1	AllExcept_Helas	HepG2_K562_Helas	Gm12878_HepG2_K562	Gm12878_H1	HepG2_K562	Gm12878_K562	
Gm12878_H1	HepG2_H1	HepG2_Helas	Gm12878_HepG2	HepG2_K562	Gm12878_K562_H1	Gm12878_Helas	HepG2_K562	Gm12878_HepG2	
HepG2_H1_Helas	HepG2_Helas	HepG2_Helas	HepG2_K562	HepG2_K562	Gm12878_Helas	Gm12878_Helas	HepG2_Helas	Gm12878_HepG2	
HepG2_H1_Helas	HepG2_K562	HepG2_Helas	HepG2_K562	HepG2_K562	Gm12878_Helas	All_Except Gm12878	K562_Helas	HepG2_H1	
HepG2_K562	K562_H1	HepG2_Helas	HepG2_H1	HepG2_H1	Gm12878_K562	HepG2_Helas	K562_Helas	HepG2_H1	
K562_H1_Helas	K562_H1	HepG2_K562	HepG2_H1	HepG2_H1	HepG2_K562_Helas	H1_Helas	K562_Helas	HepG2_Helas	
K562_Helas		HepG2_K562	HepG2_Helas	HepG2_H1	AllExcept_Gm12878		K562_Helas	K562_H1	
		K562_Helas	HepG2_Helas	K562_H1	HepG2_K562		H1_Helas	K562_H1	
			HepG2_K562_Helas	K562_H1	HepG2_K562			K562_H1	
			H1_Helas	K562_H1	K562_Helas			K562_H1	
					K562_Helas			Gm12878_HepG2_K562_H1	
					H1_Helas			Gm12878_K562_H1_Helas	
								HepG2_K562_H1_Helas	
								HepG2_H1_Helas	
								All	
								All	
12	11	13	15	15	17	11	13	21	14

Table 5.6 Date hubs chromatin state with overlapping motifs in 4 cell types.

State 2	State 3	State 13	Avg.
Gm12878_K562	Gm12878_Helas	Gm12878_H1	
Gm12878_K562	Gm12878_K562_H1	Gm12878_H1	
Gm12878_Helas	Gm12878_K562	Gm12878_H1	
Gm12878_Helas	Gm12878_K562	Gm12878_HepG2	
Gm12878_Helas	Gm12878_HepG2	Gm12878_HepG2	
K562_Helas	Gm12878_HepG2	HepG2_K562	
K562_H1	HepG2_Helas	HepG2_H1	
H1_Helas	HepG2_Helas	K562_Helas	
	K562_H1		
	K562_H1		
	H1_Helas		
	H1_Helas		
	H1_Helas		
8	**13**	**8**	**9.67**

5.5.5 Regular Expressions of state motifs

Sharing motifs of all states in multiple cell types were utilized to form the consensus pattern. The consensus pattern of DNA motifs represents state specific motifs, which has been obtained via multiple sequence alignment through CLC sequence viewer 7.5 (CLC Seq-Viewer).

The consensus sequences for all states along with the regular expressions have been shown in the Table 5.7. Gaps in alignment have been represented by alphabetical letter ‘Z’ while the letter ‘N’ represents any of the ‘A’, ‘T’, ‘G’ and ‘C’ letters of the DNA sequence.

The regular expression patterns and consensus sequences have been generated keeping in view the data of 5 human cell types. Any cell type bearing the same sharing motifs could follow the same regular expression patterns in their chromatin state motifs.

Table 5.7 Regular expressions for state specific motifs.

States	Consensus Sequence	Regular expression
S1	NGTANTTTNCAC	(A|G|T|C)GTA(A|G|T|C)TTT(T|A|C|G)CAC
S2	GTGCCTGC----	GTGCCTGCZZZZ
S3	CTGT-GTCATATG	CTGTZGTCATATG
S4	GNNCAGC---NG	G(A|T|G|C)(A|T|C|G)CAGCZZZ(A|T|G|C)G
S5	NNNAGNNNCNGT	(A|T|G|C)(A|T|G|C)(A|T|G|C)AG(A|T|G|C)(A|T|G|C)(A|T|G|C)C(A|T|G|C)GT
S6	CAAGCNGNAGNG	CAAGC(A|T|C|G)G(A|T|C|G)AG(A|T|C|G)G
S7	--NANNTAGGTT	ZZ(A|T|C|G)A(A|T|C|G)(A|T|C|G)TAGGTT
S8	AAAC-ACTANATG	AAACZACTA(A|T|C|G)ATG
S9	CGAGTGGGTNTCG	CGAGTGGGT(A|T|C|G)TCG
S10	G--TCGNNGTAT	GZZTCG(A|T|C|G)(A|T|C|G)GTAT
S11	ACCTGACCC—N	ACCTGACCCZZ(A|T|C|G)
S12	NANGG-CTAG-G	(A|T|C|G)A(A|T|C|G)GGZCTAGZG
S13	CCCCTCTTGTGG	CCCCTCTTGTGG
S14	NTNCNGNGCACC	(A|T|C|G)T(A|T|C|G)C(A|T|C|G)G(A|T|C|G)GCACC
S15	GCANCCC------G	GCA(A|T|C|G)CCCZZZZZZG

5.6 Minimum Dominating nodes set in chromatin states networks

5.6.1 Chromatin States of Human cell types

We used 15 states HMM models of both cell types from section 5.5.1. To unravel the chromatin states associations based on epigenetic marks we analyzed the correlations between them. The correlation highlighted positive as well as negative correlations as shown in the Figure 5.22. Positive correlations indicate the association between the states while negative correlation indicated that based on epigenetic features the states has no relevance. In the light of this state1 and state2 showed a positive correlation where state1 being the repeat region defined by the presence of H3K9me3 mark while state2 was predicted as the heterochromatin state with low level of all marks. Both the states showed strong enrichment of LaminB1. The transcription initiation and elongation states (3-5) bearing high enrichment of H3K36me3 and H3K79me2 marks showed positive correlation amongst them.

S1-S2	S1-S3	S1-S4	S1-S5	S1-S6	S1-S7	S1-S8	S1-S9	S1-S10	S1-S11	S1-S12	S1-S13	S1-S14	S1-S15
0.61	-0.11	-0.14	-0.12	-0.11	-0.26	-0.34	-0.23	-0.31	-0.24	-0.15	-0.28	-0.11	-0.09
S2-S3	S2-S4	S2-S5	S2-S6	S2-S7	S2-S8	S2-S9	S2-S10	S2-S11	S2-S12	S2-S13	S2-S14	S2-S15	
-0.01	-0.16	-0.25	-0.25	-0.33	-0.51	-0.48	-0.09	-0.27	-0.05	-0.46	-0.06	-0.30	
S3-S4	S3-S5	S3-S6	S3-S7	S3-S8	S3-S9	S3-S10	S3-S11	S3-S12	S3-S13	S3-S14	S3-S15		
0.70	-0.08	-0.09	-0.19	-0.32	-0.28	-0.34	-0.27	-0.17	-0.35	-0.16	-0.11		
S4-S5	S4-S6	S4-S7	S4-S8	S4-S9	S4-S10	S4-S11	S4-S12	S4-S13	S4-S14	S4-S15			
0.64	0.64	0.45	0.09	-0.42	-0.40	-0.38	-0.24	-0.51	-0.23	-0.20			
S5-S6	S5-S7	S5-S8	S5-S9	S5-S10	S5-S11	S5-S12	S5-S13	S5-S14	S5-S15				
1.00	0.83	0.46	-0.27	-0.17	-0.23	-0.12	-0.33	-0.16	-0.12				
S6-S7	S6-S8	S6-S9	S6-S10	S6-S11	S6-S12	S6-S13	S6-S14	S6-S15					
0.81	0.48	-0.25	-0.16	-0.24	-0.15	-0.31	-0.14	-0.14					
S7-S8	S7-S9	S7-S10	S7-S11	S7-S12	S7-S13	S7-S14	S7-S15						
0.59	-0.03	-0.10	0.30	0.37	-0.05	-0.20	-0.18						
S8-S9	S8-S10	S8-S11	S8-S12	S8-S13	S8-S14	S8-S15							
0.68	0.58	0.37	0.08	0.35	-0.34	-0.30							
S9-S10	S9-S11	S9-S12	S9-S13	S9-S14	S9-S15								
0.70	0.59	0.10	0.67	-0.22	-0.04								
S10-S11	S10-S12	S10-S13	S10-S14	S10-S15									
0.37	-0.05	0.26	-0.36	-0.19									
S11-S12	S11-S13	S11-S14	S11-S15										
0.76	0.51	-0.18	-0.03										
S12-S13	S12-S14	S12-S15											
0.27	-0.03	-0.05											
S13-S14	S13-S15												
0.54	-0.15												
S14-S15													
-0.10													

Figure 5.22 Chromatin states correlation based on emission probabilities for 15 states HMM (both cell types).

Interesting results have been revealed in case of enhancer states where states (6-7, 11-12) have been demarcated into two clusters. First cluster (states 6 and 7) showed positive correlation with the transcribed states which could be related to the recently predicted role of elongation factors in posing enhancers in ES cells [132]. This behavior remained the same in case of K562 cells as well. The second cluster (states 11-12) comprising of the most active enhancers showed positive correlations with the promoters cluster (states 9-10 and 13). The active promoter state was enriched with mediator complexes and RNA PolII whereas state 11 (the very strong enhancer) showed high enrichment in H3K4me1, H3K27ac, P300_TF and P300_CM along with various other TFs which is consistent with its role as active enhancer. State 13 and 14 showed positive correlation with relevance to enrichment in H3K27me3 mark involved in inactivation along with high enrichment in polycomb repressive complexes (CBX2 , CBX8, EZH2, SUZ12 REST and SETDB1) [133-134]. The insulator state marked by the presence of CTCF had not shown positive correlation with any other state. All the correlations of this state were negative. It showed enrichment in some of the TFs. The enrichments have been highlighted in Figures

5.23a and 5.23b for H1hesc and K562 respectively. Both cell types are highly correlated in their enrichment patterns.

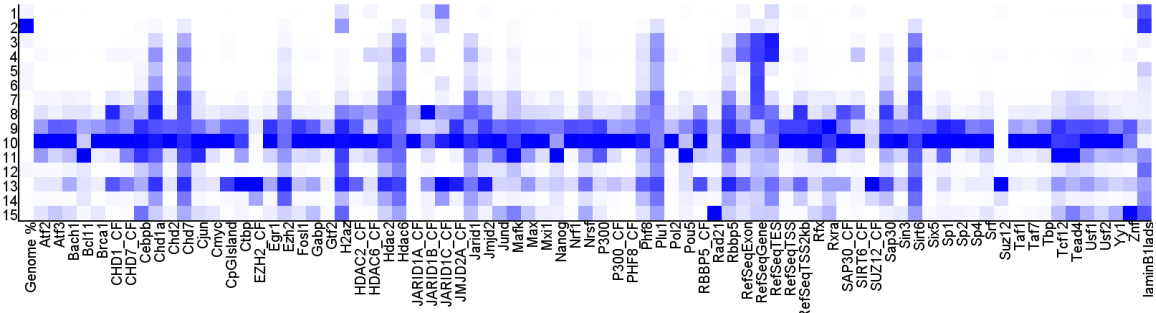


Figure 5.23 a. TFs and CMs marks enrichments of 15 states HMM for H1hesc cell type. Each state shows signal enrichment with respect to its specificity.

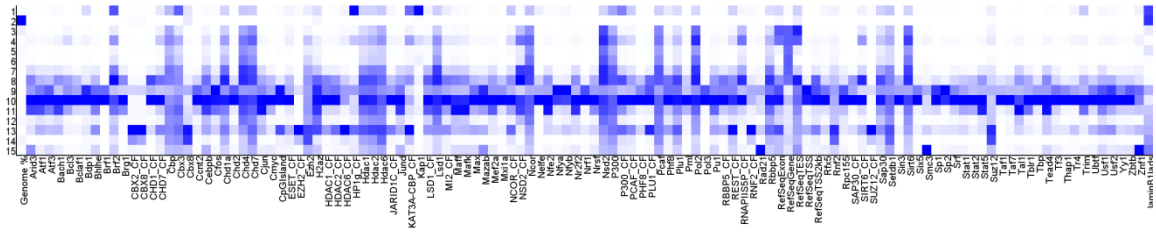


Figure 5.23 b. TFs and CMs marks enrichments of 15 states model for K562 cell type. Each state shows signal enrichment with respect to its specificity.

5.6.2 Chromatin States Networks

The genomic regions and epigenetic marks associated with each state have been utilized in network creation. The chromatin states complete (global) network for both the cell types (Figures. 5.24a and 5.24b) highlights the direct association of the states with the factors (HMs, TFs and CMs) and within states correlations to show network connectivity. The complete network in case of H1hesc comprised of total 104 nodes connected with 534 edges having average network connectivity of 10.26 and maximum connectivity of 88. K562 network contained 148 and 1050 edges with an average network connectivity of 14 and maximum network connectivity of 137. Both complete networks after subjection to Markov clustering (MCL) formed certain complexes/clusters in the network based on edge correlation association. In case of both the cell types 6 clusters have been seen; where cluster 1 contained states 8, 9 and 10, states 3,4,5,6 and 7 formed cluster2, cluster 3 contained states 13 and 14, while cluster 4 contained states 11 and 12, cluster 5 revolved around states 1 and 2, and the

last was cluster 6 having state 15 along with its associated features/factors as discussed in the above section. The clusters formed in the global networks showed strong association with the correlations mentioned in the above section which supports the positive correlation among the states. Despite of network connectivity amongst states, the global networks represent bipartite property where set one contained the states and set 2 contained the factors.

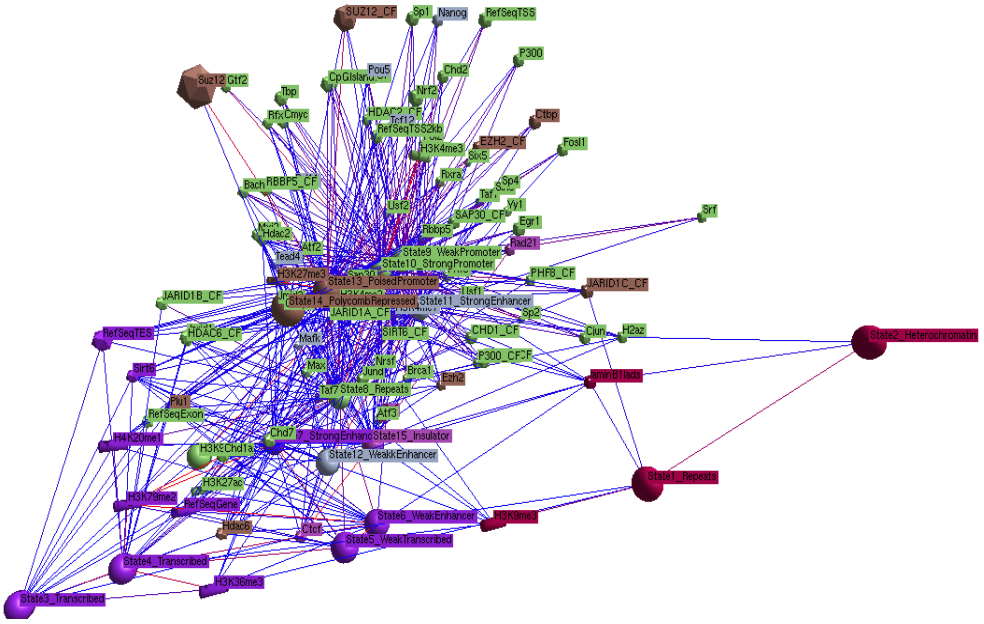


Figure 5.24a Complete chromatin states network for H1hesc cell type where nodes represent states, HMs, TFs and CMs.

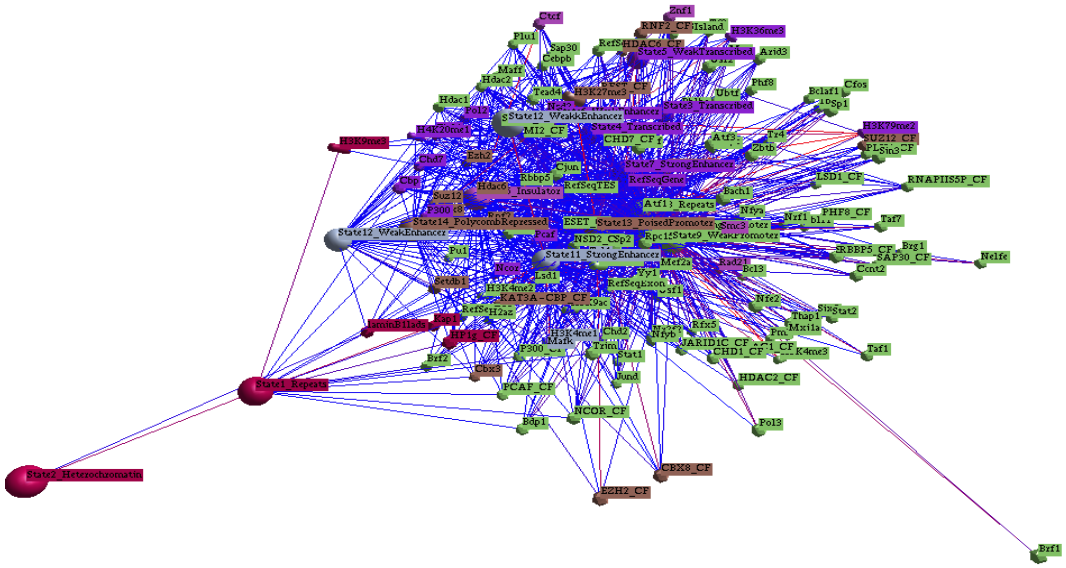


Figure 5.24b Complete chromatin states network for K562 cell type.

In order to study the detailed interactions at individual states level each complete network has been further zoomed in to obtain the local networks. The local networks of each state were obtained by the correlation of factors with the states along with inter correlation among the factors. This networking produced K complete (K=15) graphs, where each node was connected to every other node in the network, where K is the number of complete graphs as we had 15 states so it resulted in 15 complete graphs. So in total 30 complete graphs (local networks), 15 each for two global networks have been formed as shown in the Figure 5.25a (H1hesc) and Figure 5.25b (K562). Around 15% of interactions were common in all local networks while others were state specific which was due to the inclination of certain factors towards the associated states. State to factors direct links in local networks were around 30% while remaining were the indirect paths created by the inter factor interactions. The marks/factors associated with the states have been explained in detail in section above, these marks were seen in the networks both global and local.

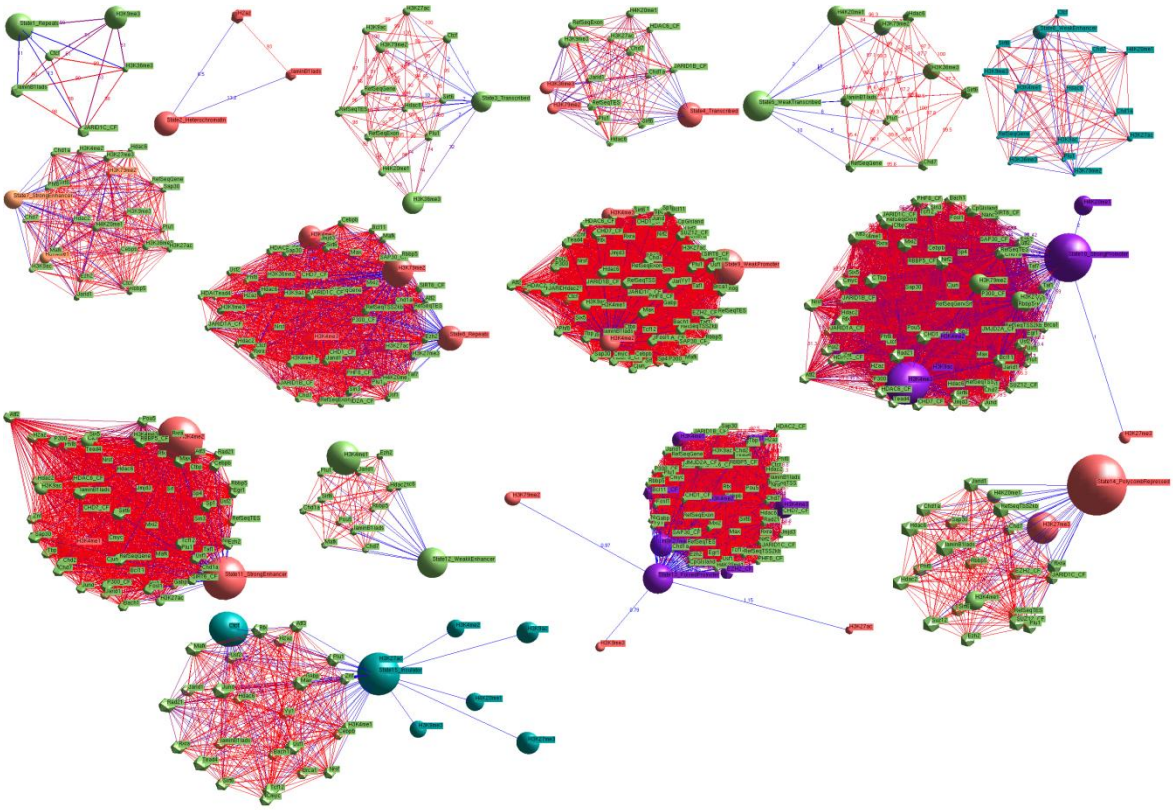


Figure 5.25a Local chromatin states network for H1hesc cell type. Each network represents a single state network hence in total 15 local networks.

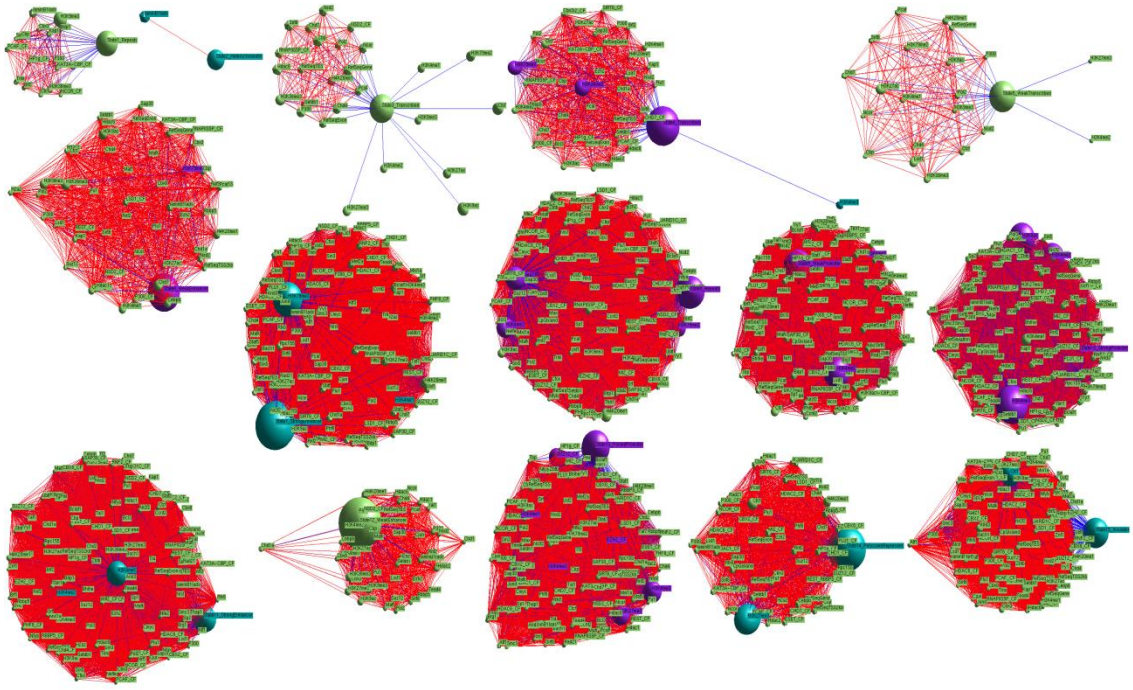


Figure 5.25b Local chromatin states network for K562 cell type. Each network is a local network of a chromatin state ranging from state 1 to state 15 of HMM.

5.6.3 Minimum Dominating nodes set (MDNS)

The complete chromatin state networks discussed in above section have been subjected to the identification of set of minimum dominating nodes which could play an important role in guaranteeing the connectivity of underlying set of chromatin states mapping to the functionally genomic elements.

H1hesc complete network when subjected to the combined centralities as mentioned above produced the results as shown in the Table 5.8a and Figure 5.26a. It is clear that state 10 which is the strong promoter produced the result with highest combined centrality where the weak promoter state is the second highest in ranking. The states after the promoter states are the strong enhancer closely correlated with the promoters' states and the poised promoter. The other correlated enhancers and transcribed states lie after that. The state with the lowest centrality measure was the heterochromatin while the one before that was the repeats state.

Chromatin states network of K562 after subjection to centrality measures produced similar patterns to that of H1hesc as shown in the Table 5.8b and Figure 5.26b. The node with the largest combined centrality measure was the strong promoter and then the weak promoter exactly as in case of H1hesc. The lowest centrality node was the

heterochromatin as in the H1hesc. The internal sequence of descending order in both the cell types vary at some nodes as the input set contained cell type specific factors and the factors of H1hesc varied from that of K562. The TFs and CMs of H1hesc were comprised of 56 and 15 elements respectively whereas K562 contained 27 CMs and 87 TFs. The larger set of K562 contained some of the specific factors more oriented towards the polycomb states including CBX2, CBX8, REST and SETDB1 which were not the part of H1hesc set; this resulted in adding the centralities of that state which changed the internal sequence of states. Our test on similar marks highlighted the fact that the sequence of states for centralities remained the same for both the cell types.

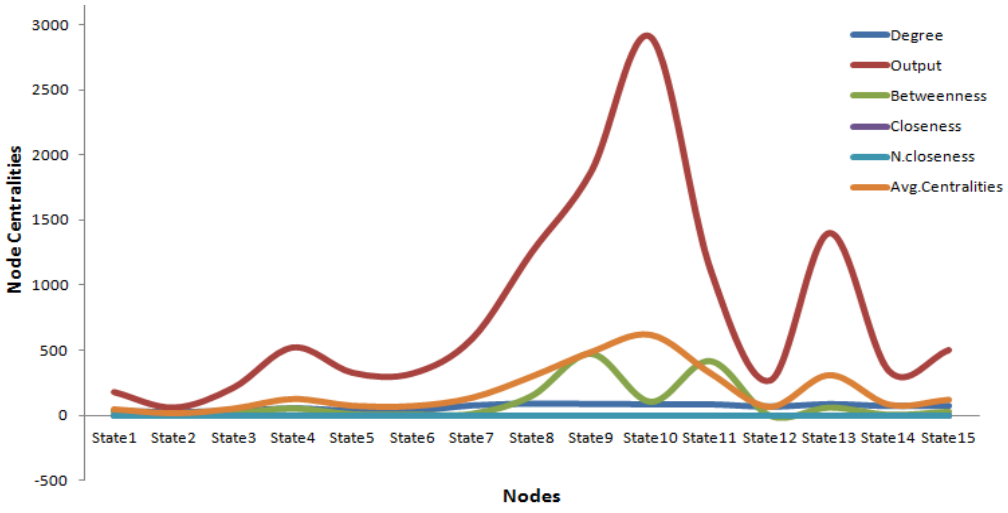


Figure 5.26a Combined centralities for H1hesc cell type chromatin states network

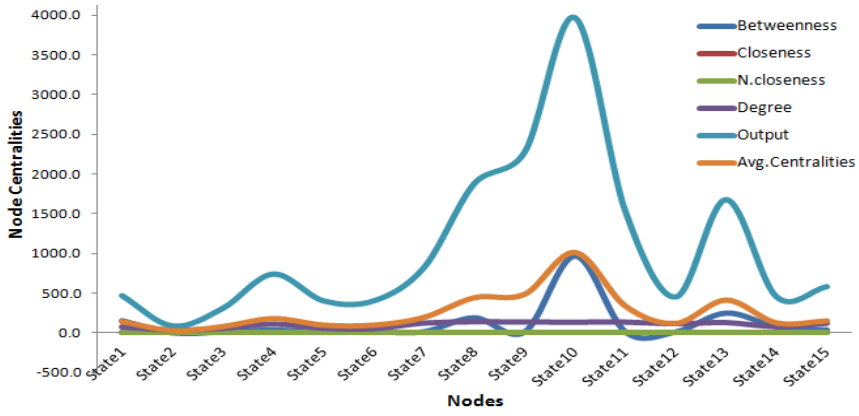


Figure 5.26b Combined centralities for K562 cell type network.

Table 5.8a Node Centralities of H1hesc cell type.

Node	Degree	Output	Betweenness	Closeness	N.closeness	Avg.Centralities
State1	32	178	29	0.01	0.00	47.80
State2	23	59	4	0.01	0.00	17.20
State3	33	211	18	0.02	0.00	52.40
State4	53	520	57	0.03	0.00	126.01
State5	34	326	11	0.02	0.00	74.20
State6	33	320	8	0.03	0.00	72.21
State7	77	585	14	0.04	0.00	135.21
State8	93	1247	151	0.05	0.00	298.21
State9	89	1866	477	0.10	0.01	486.42
State10	87	2901	107	0.13	0.01	619.03
State11	86	1122	420	0.06	0.00	325.61
State12	67	266	0	0.03	0.00	66.61
State13	88	1398	63	0.06	0.00	309.81
State14	72	338	7	0.03	0.00	83.41
State15	74	500	31	0.03	0.00	121.01

Table 5.8b Node Centralities of K562 cell type.

Node	Betweenness	Closeness	N.closeness	Degree	Output	Avg.Centralities
State1	157.0	0.1	0.0	75.0	466.0	139.6
State2	5.0	0.0	0.0	37.0	83.0	25.0
State3	16.0	0.0	0.0	58.0	305.0	75.8
State4	40.0	0.0	0.0	109.0	737.0	177.2
State5	14.0	0.0	0.0	59.0	401.0	94.8
State6	17.0	0.0	0.0	54.0	400.0	94.2
State7	20.0	0.1	0.0	121.0	820.0	192.2
State8	193.5	0.1	0.0	136.0	1878.0	441.5
State9	17.0	0.1	0.0	133.0	2274.0	484.8
State10	963.0	0.2	0.0	131.0	3964.0	1011.6
State11	17.0	0.1	0.0	132.0	1527.0	335.2
State12	16.0	0.0	0.0	113.0	449.0	115.6
State13	252.0	0.1	0.0	126.0	1674.0	410.4
State14	81.0	0.0	0.0	78.0	452.0	122.2
State15	41.0	0.0	0.0	118.0	579.0	147.6

In order to further highlight the importance of MDNS we sorted the nodes with combined centralities and then deleted the top centrality nodes including promoters, enhancers and transcribed states. The left over states and all the factors showed a disrupted and disconnected network where many connections were missing and the closeness centrality equalized to infinity as shown in the Figures 5.27a and 5.27b for H1hesc while Figures 5.27c and 5.27d for K562 respectively. The 5.27a and 5.27c illustrates the disrupted network of H1hesc and K562 while 5.27b and 5.27d shows the node centralities of disrupted networks.

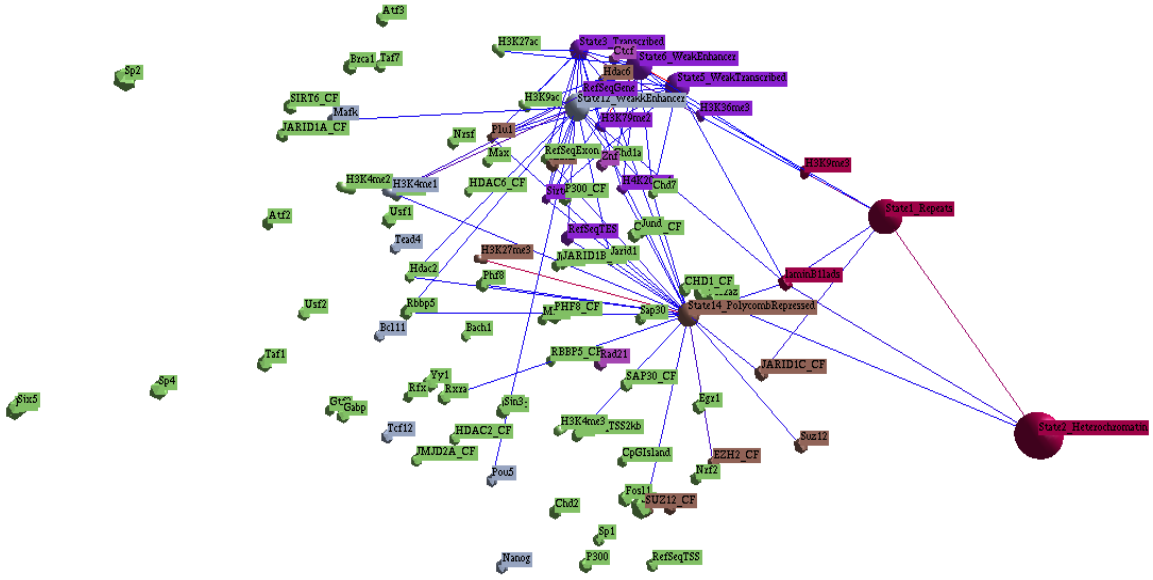


Figure 5.27a Disrupted chromatin states network for H1hesc cell type.

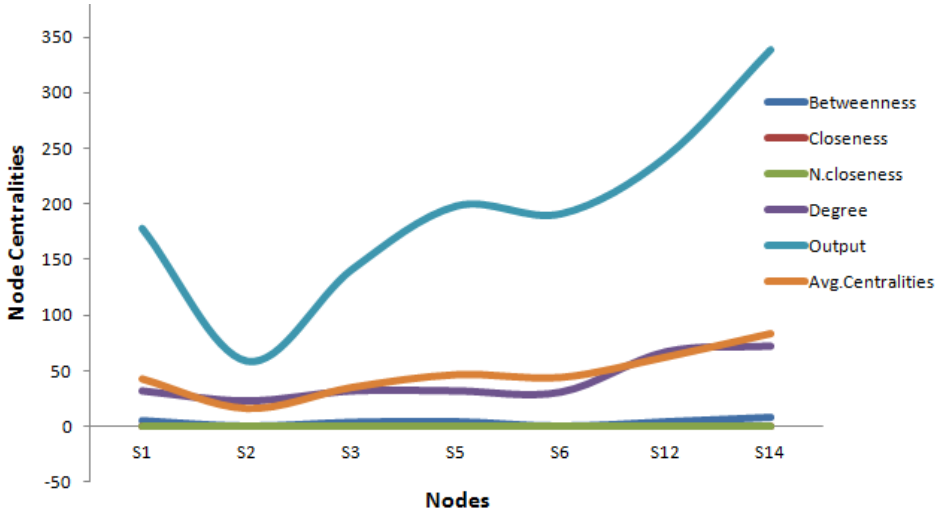


Figure 5.27b Disrupted chromatin states network centralities for H1hesc cell type.

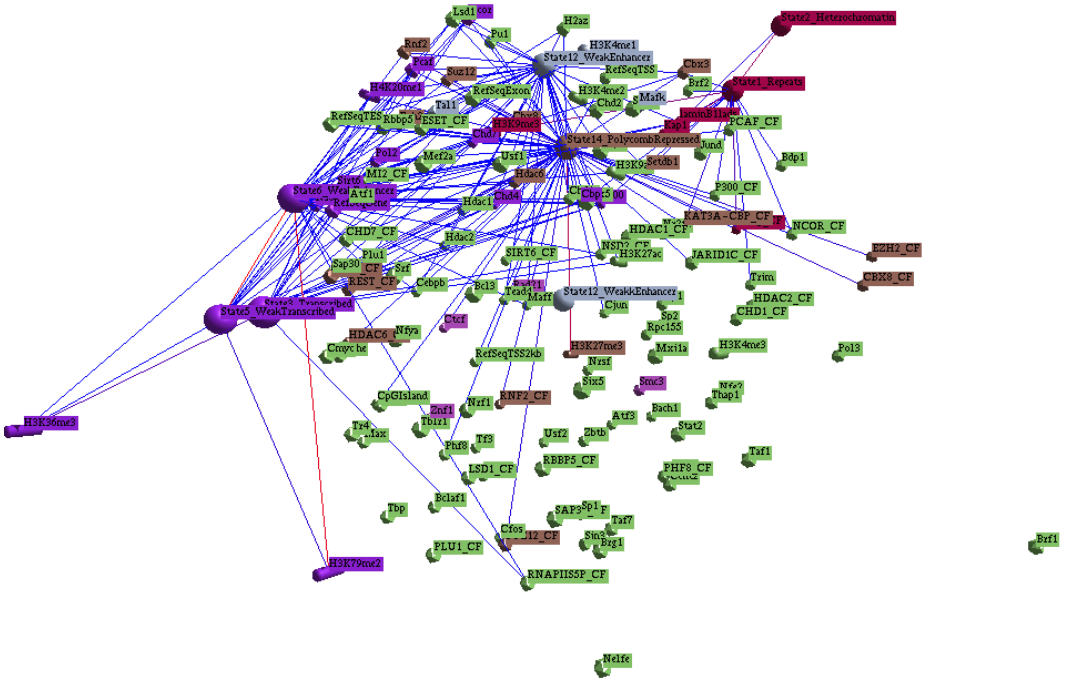


Figure 5.27c Disrupted chromatin states network for K562 cell type.

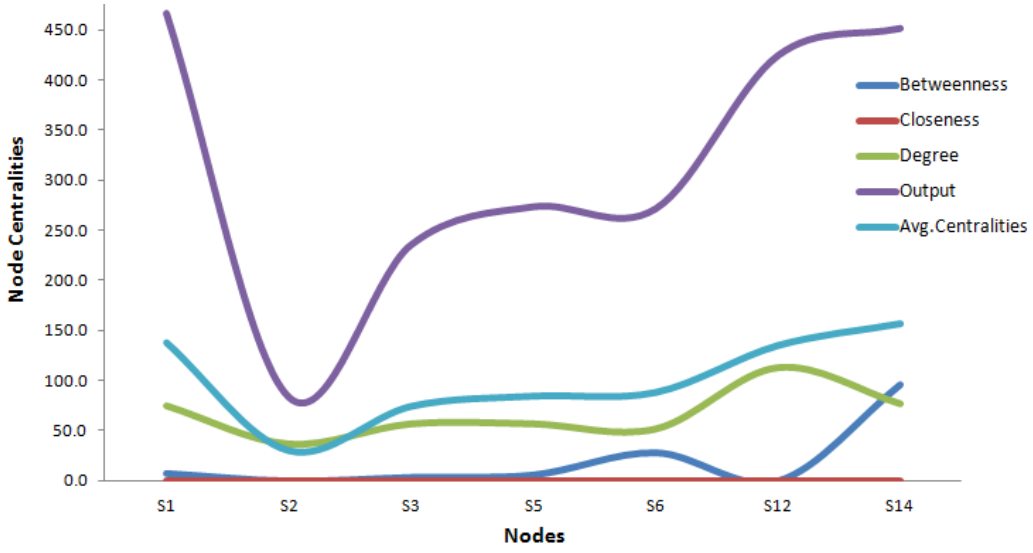


Figure 5.27d Disrupted chromatin states network centralities for K562 cell type.

5.6.4 Cross validation and reduced chromatin states networks

The leave one out mark cross validation models have been compared with the reference model of all marks for both cell types. The enrichment and emissions patterns of both the cell types showed the difference in case of mapping to the states as has been shown in Figures 5.28a and 5.28b for H1hesc and 5.28c for K562 respectively. Each mark which is left out e.g. CTCF for the learning purpose, the enrichment level varied in that state with respect to the reference model. Such

variance has been observed in case of all the marks. This shows the importance of the presence of a certain mark contributing towards the integrity of the state network.

The combined centralities test for a reduced 10 states model for both cell types have shown the similar behavior as of the 15 states model as shown in the Figures 5.29a (H1hesc) and 5.29b (K562) respectively. State 3 the active promoter showed the highest centrality values preceded by the enhancer which is state 4. States distributions and enrichments have been shown in the Figures 5.30a (both cell types), 5.30b (both cell types), 5.30c (H1hesc) and 5.30d (K562). This test also highlighted the promoter, enhancers followed by the transcribed states as the MDNS of the chromatin states network. So any network of chromatin states bearing these states would mark promoters, enhancers and transcribed states as the MDNS.

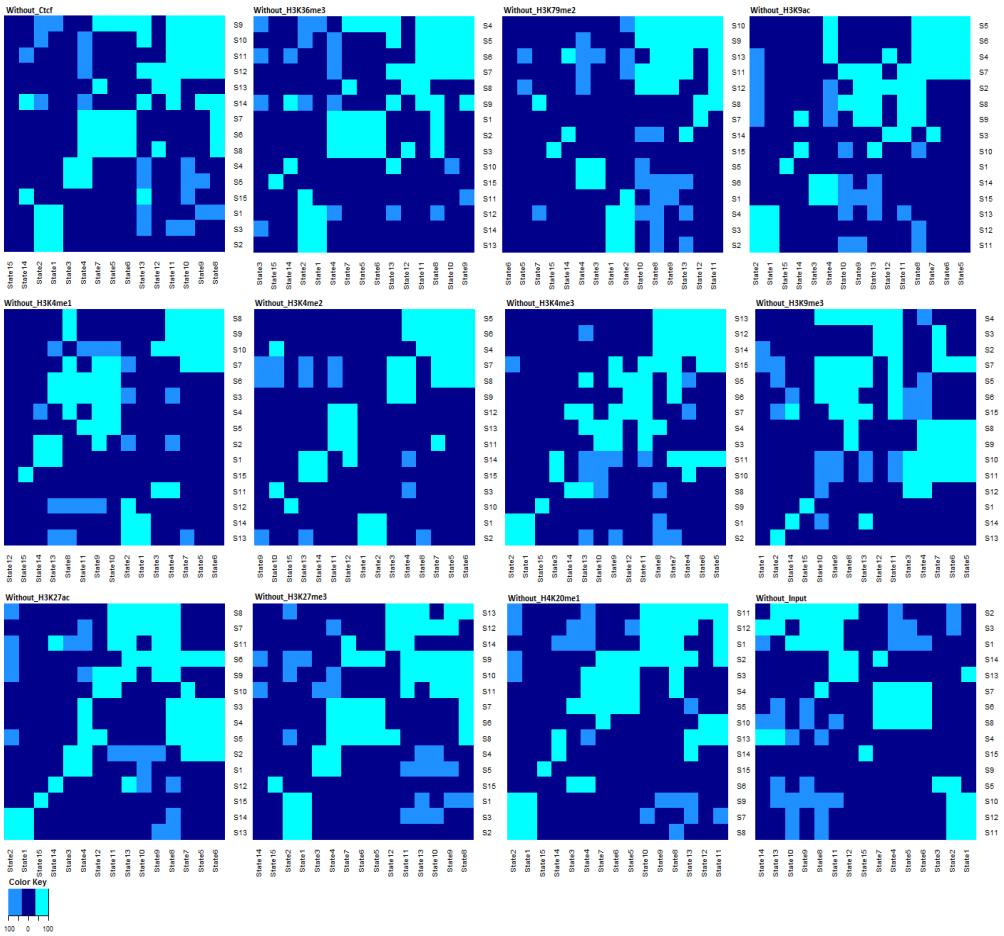


Figure 5.28a Emission correlations of cross correlation models with the reference 15 states models.

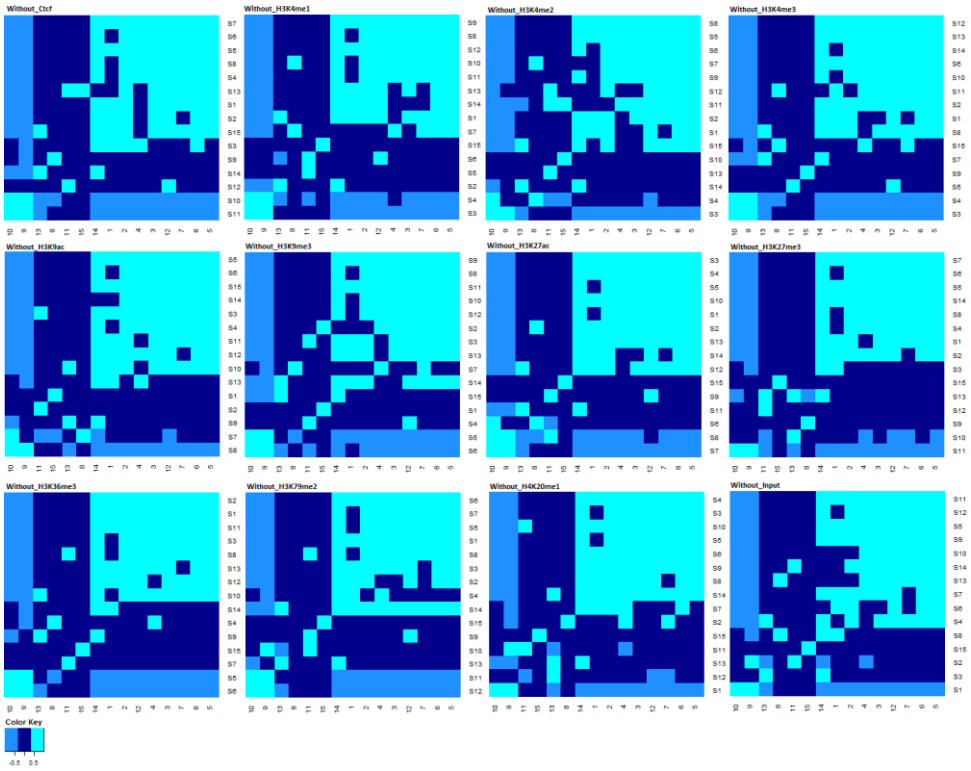


Figure 5.28b Enrichment correlations of cross correlation models with the reference 15 states models for H1hesc

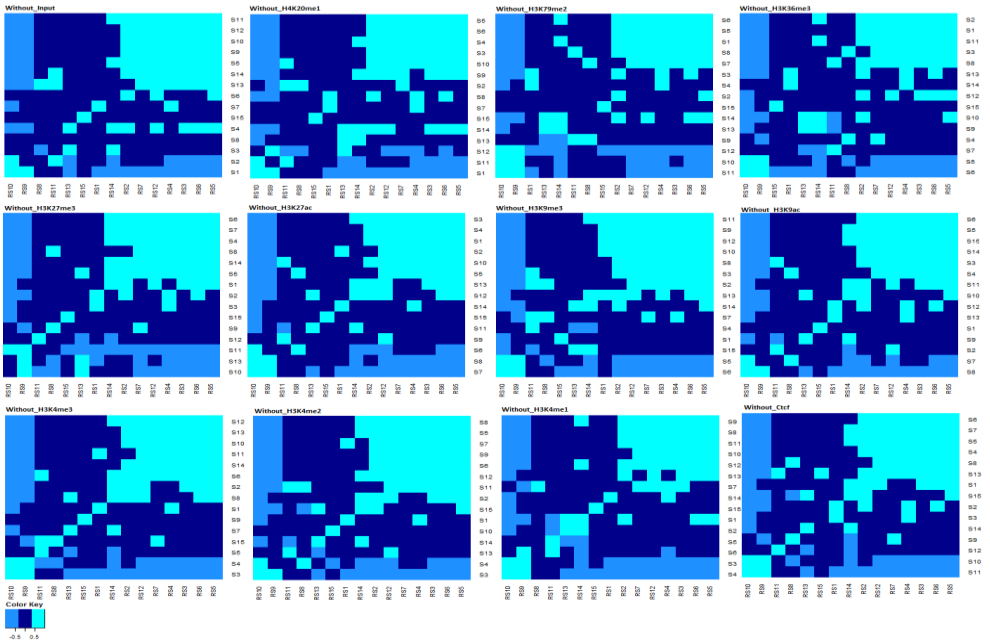


Figure 5.28c Enrichment correlations of cross correlation models with the reference 15 states models for K562

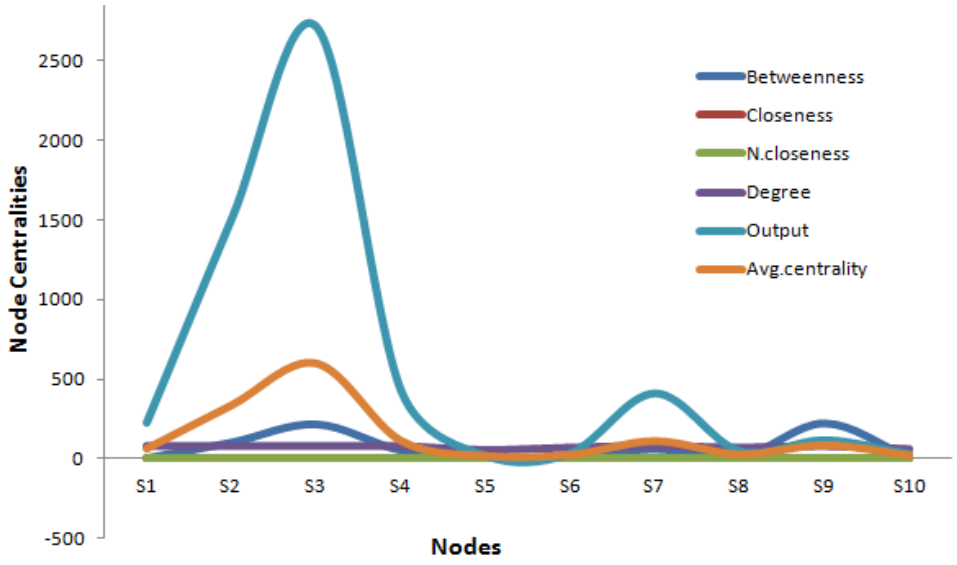


Figure 5.29a Combined centralities for reduced H1hesc cell type network model

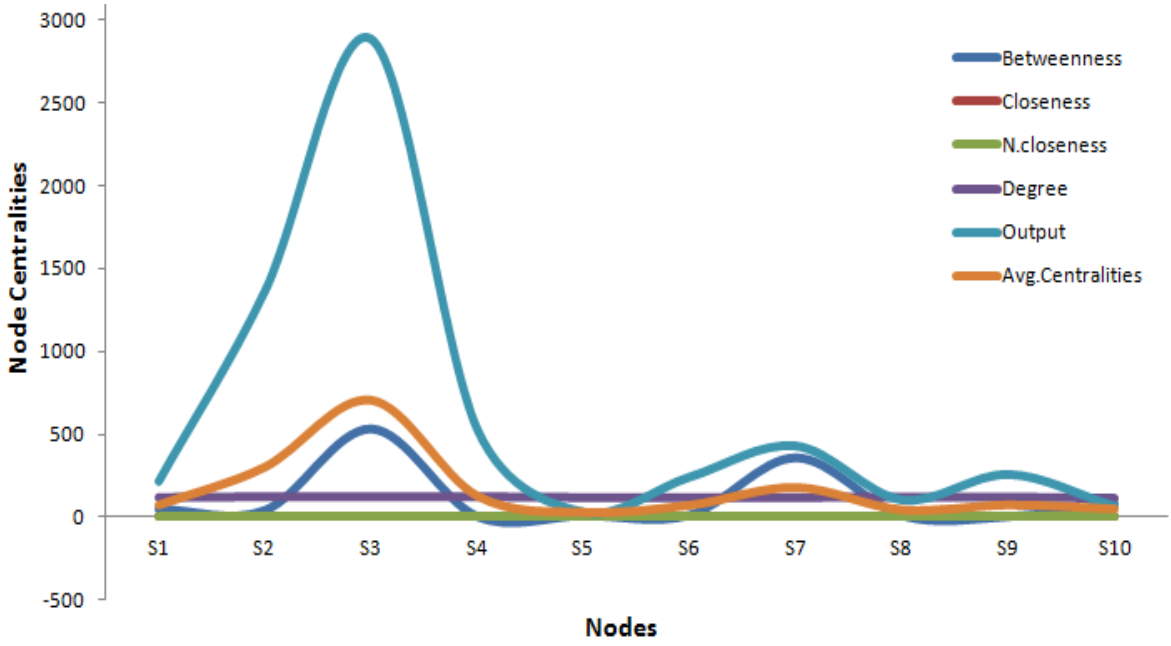


Figure 5.29b Combined centralities for reduced K562 cell type network model

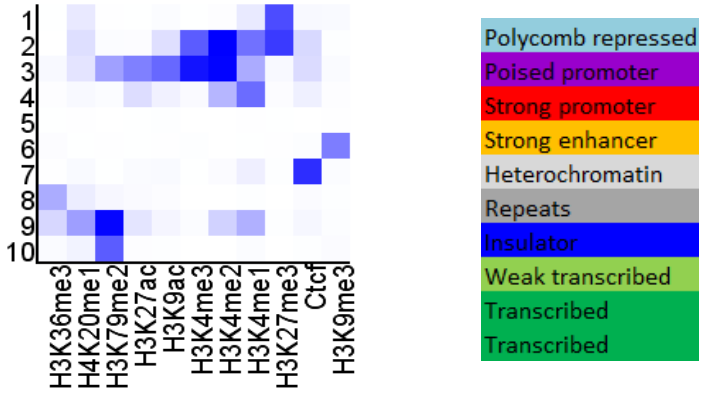


Figure 5.30 a Emission probabilities of reduced 10 states models. **b** Sequence wise genomic elements distribution of chromatin states.

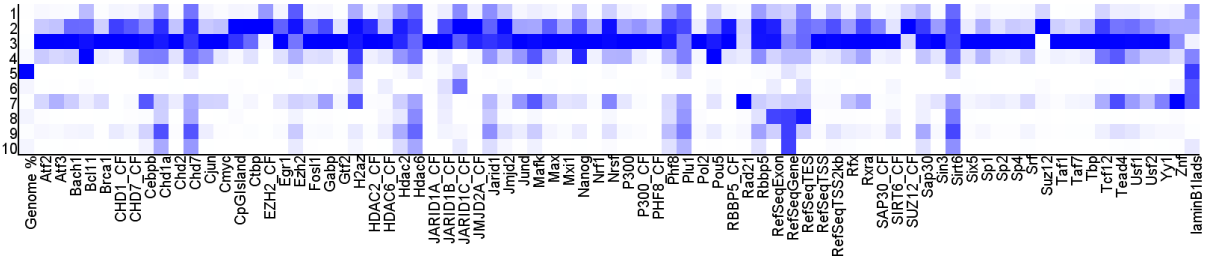


Figure 5.30 c. States enrichments of H1hesc for 10 states model.

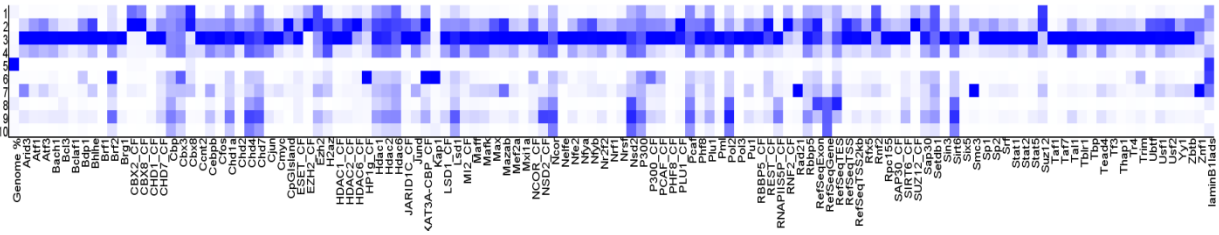


Figure 5.30 d . States enrichments of K562 for 10 states model

5.7 Discussion of the findings

Each of the results obtained mentioned in the above sections i.e. 5.1 to 5.6 have been compared with the previous studies in the field.

5.7.1 Computational identification of chromatin states

We proposed a computational scheme to find out the underlying chromatin states in given genomic ChIP-Seq histone modifications profiles data. We tested various

measures for different non-overlapping bin sizes of genomic data. One of studies [82] provided BIC as a criterion at 1000bp resolution for states number identification but our results have shown (section 5.1) that there is high data dependency at such scale so BIC cannot be the true scale of measurement. The BIC score converged at 11 states model for 50Kbp bin size which is very wide. It is not reasonable to study the behavior of sharp and narrow marks like H3K4me3 which has a range of 200bp to 400bp. This shows that BIC is not a valid criterion to get optimal states in case of genomic data at fine resolution due to data dependency. The BIC scores for other bin sizes have not been shown because of the same behavior as of 200bp. This had also been shown by Ernst *et al*., [52] with a different set of data where BIC didn't converge due to data dependency.

We used emission matrices parameters for chromatin states identification purpose. We tested emission matrices for different bin sizes to test the effect of non-overlapping bin sizes in chromatin states identification. The earlier findings of narrow and broad peaked marks like H3K4me3 and H3K9me3 showed the range of these marks from 200bp to 600bp respectively [149]. This supports the use of 200bp as the proper resolution scale to peep deep into the chromatin states which has been supported by many earlier studies [52, 54]. We identified correlation among vectors of emission matrices and calculated the mean correlation values. The mean correlation starting from negative value moved towards positive and adds up to the positivity till a certain level, after that it attained equilibrium. When the probability of accepting the next move is zero or closer, then equilibrium is reached as is defined in case of simulated annealing to find the next move [150]. We also found no change in mean correlation at the next state level after attaining the equilibrium so that was the peak value achieved in our case.

Emission matrices correlation and clustering of emission matrices vectors in our study both showed 100% correlation in their results. Clustering of highly correlated vectors combines to produce distinct classes [151]. We applied the same concept and clustered up vectors with correlations of 0.85 and above. The reduced model represents the states number where the emissions vectors are distinct with low correlations. This highlighted the fact that underlying chromatin states in genomic data could be identified using this approach. Along with emission matrices, the simple

biological annotation scheme could also be utilized for chromatin states identification. Our results showed strong correlation with the reference model [54] in case of emission matrices technique as well as the visualization scheme.

The use of input in data normalization is important in order to get true signals at various locations for different marks [152], which is also clear in this case as well. On the whole we can say that data normalization using input is essential in learning HMM states for a particular dataset under study. As the data in case of having input control is normalized therefore the cumulative average correlation has been converged earlier than the data without input control.

So we conclude that HMM parameter based evaluation and annotation based visualization is a good way to identify underlying hidden states for the ChIP-Seq data under study even for a lay man who could not identify how many significant states the data under study contains. The process will help users to find states by any of the ways we adopted, the easiest being the clustering of emission matrix and visualization of annotations.

5.7.2 Grouping of H1 data

Our analysis suggests that, based on their DamID binding values, H1 subtypes can be divided into two groups: (1) H1.1 and (2) H1.2–H1.5. We found that the differences between the enrichment of H1.1 and that of H1.2–H1.5 seem to be more pronounced at regulatory regions marked by activating histone modifications such as promoters, enhancers, and CpG islands. For example, we observed in our DamID profiling that although H1.1 is not depleted at inactive and poised HCPs, H1.2–H1.5 HCPs are often associated with developmentally regulated genes [153], which need to be activated in response to external stimuli and therefore do not require a stable repressive chromatin organization. Moreover, although CG richness can affect H1 subtype binding to chromatin similarly, H1.1 is the only subtype that displays a rather positive correlation with DNA methylation independently of its genomic location. Based on this, we propose that H1.1 is somehow special among H1 subtypes [97]. The chromatin structure promoted by H1.1 binding might support a level of compaction that facilitates rapid conversion into either an active or repressed state. Previous data also agree with this special role for H1.1. For example, it was shown that H1.1 had the least ability to condense nucleosomes in vitro [154], and had the

lowest affinity to chromatin both in fluorescence recovery after photobleaching experiments in vivo and biochemical studies in vitro [155-156]. The dynamic binding of H1.1 might at least in part explain why chromatin-containing H1.1 does not necessarily impede gene expression. In line with this, Alami *et al*. [157] showed that H1.1 has a unique role in activating a reporter transgene in mice. In view of its distinct distribution and its potential association with more active and open chromatin, it is not surprising that H1.1 expression is restricted to certain cells, thereby conferring tissue specificity [158]. Interestingly, estimation of the degree of nucleotide substitutions during evolution also pointed toward a functional differentiation of H1.1 from the other H1 subtypes, further strengthening the uniqueness of H1.1 among the somatic H1 subtypes from an evolutionary perspective [159].

5.7.3 Unbiased genomic segmentation via ChromClust

Several computational methods have been presented so far for highlighting combinations of histone modifications. One of the supervised approaches marked histone modification signals at known promoter and enhancer regions [43]. Among the unsupervised approaches, the well-known ChromaSig [49] mined clusters mapping to previously defined promoter and enhancer regions hence missing the larger chunk of the genome. Among clustering algorithms spatial clustering introduced by [65] combined clustering with hidden Markov models (HMM) to find histone combinations. Ambiguity in the concept transpired, by the concept of sharing of functional elements by consecutive genomic regions [65].

Various HMM approaches modeling chromatin mark signal levels via multivariate normal's have been used earlier [43, 49, 65, 84, 88, 101-103]. A binarized multivariate HMM signaling the presence or absence of histone modifications is in wide use these days [52]. The HMM approaches pose a restriction to the user in defining the number of states. Each approach then segments the genome based on user restricted number of states.

Our tool ChromClust based on semi-supervised learning approach splits the genome based on histone modification combinatorics in a very specialized way. It works on the idea of binarized signals inspired by one of the previous approaches [52] but has the added facility of not restricting the user to specify the clusters. Existing tools don't

provide a proper user friendly environment. ChromClust is a standalone desktop application which allows the user to conduct semi-supervised clustering in an interactive way. It executes efficiently on whole genome data and produces results in less than 30 minutes. Amongst the existing tools so far in this domain, our tool outperforms others in terms of cluster identification, storing results in DB, allows supervised learning along with un-supervised learning abilities as well. Comparison of our tool with others is shown in Table 5.9.

Table 5.9 Comparison of ChromClust with other tools

Tools	Language / Platform	Facilities	Approach	Time Consumption	GUI
ChromClust	C-Sharp	1. Cluster identification 2. Data base management 3. Results storing and querying posing with respect to each cluster facility	Hamming distance based method	Around 20 mins for whole genome	Yes
ChromHMM	Java	1. Chromatin states identification 2. Annotation of states	Hidden Markov Model	4 to 5 hours	No
ChromaSig	-	Cluster identification	Euclidean distance based method	2 hours	No
HMMSeg	Java	States identification	Hidden Markov Model	4 to 5 hours	No

5.7.4 Unbiased genomic segmentation via ChromBiSim

ChromBiSim outperforms other methods as it works on presence of marks and decodes the associations within seconds from the whole genome data set. The 5Kbp non overlapping data of K562 took only few seconds to mine the patterns. If the binning (non-overlapping regions) is narrowed down like 200bp which we also tested on various cell types, it took over maximum 30 minutes to identify whole genome patterns.

ChromBiSim like ChromHMM [88] works on binarized data but unlikely it works on local signals and portrays epigenomic landscape from the present marks signals neglecting the absent signals. It outperforms the two biclustering algorithms [50, 104] for histone modifications in several aspects as being user friendly, not dependent on peak calling methods to identify peaks for bicluster analysis.

5.7.5 Date and party hubs of chromatin networks

In order to study the details of chromatin states networks, we utilized the combinations of histone modifications in chromatin states learning and then highlighted the mapping of states to functional regions of the genome. We predicted the DNA motifs in setting the combinations of histone modifications ultimately giving rise to a chromatin state. State specific DNA motifs along with cell type specificities have been mined in this experimentation. Combined information of histone modifications correlation, annotations of functional genomics regions and DNA motifs prediction have been visualized as chromatin state networks for 5 chosen cell types.

Chromatin state networks were subjected to various network centrality measures in order to identify the hubs and non-hubs in the networks for the first time. In previous studies [108, 160-165] these concepts have been used for gene expression profiling studies along with PPI networks. We for the first time targeted chromatin state networks. Identified hubs were further divided into date and party hubs which according to literature [165] are the dynamic and static nodes in the network. Heterochromatin, one of the transcribed regions and poised promoter states were identified as date hubs which very clearly are dynamic regions of the genome and separate the boundaries of repressed and active states, low level marks and high level marks [52]. We found that enhancer and promoter regions histone mark combinations show stable behavior across networks of various cell types marking them as party hubs. This was also shown by the virtual concatenation and chromatin state learning of various cell types by Ernst *et al*., [88]. We highlighted the role of DNA motifs in setting the date and party hubs in chromatin state networks as well. DNA motif combinations involved in setting chromatin states, hubs and non-hubs were utilized to find the consensus patterns covering 5 cell types. We covered 5 cell types in our study as they have been widely in literature. This study could further be replicated over multiple cell types covering different species to define consensus patterns of state specific DNA motifs along with studying the variations and similarities in patterns of date and party hubs.

5.7.6 Minimum dominating nodes sets in chromatin networks

Epigenetically defined chromatin states have been utilized to classify the genome for any cell type under consideration [52, 54]. ChromHMM [88] based on multivariate hidden Markov model (HMM) have become a de facto standard for classification of chromatin with respect to genomic elements based on the presence absence of marks combinatorics. State enrichments with various CMs and TFs in our study showed strong relevance with the previous studies [54, 80,166]. The interactions among different histone modifications in our state networks and their mapping to functionally active genomic elements as mentioned in the results section in detail strongly correlate with the previous studies[38,41-44, 51-53,61,74,133,134,137, 167-171]. The interactions of HMs shown in the study [15] have been clearly seen in our networks while their study focused only on the TSS regions of various genes for HMs only. The positive and negative correlation among various marks shown in our study along with the stability of the networks across different cell types is also in strong concordance with the previous study [51]. The crosstalk mechanism of HMs presented for promoters, insulators and enhancers using Bayesian approach [53] could easily be seen in our states network for both the cell types. Well this study like the previous [51] focused HM combinatorics only. One of the recent studies [74] utilized the same cell types we have used in our study. It [74] focused on HMs along with CMs in order to show the chromatin signaling networks around TSS regions using linear regression approach. The interactions in their network have also been mined in our study along with the additional interactions of heterochromatin, repeat regions, enhancers and transcribed states based on TFs, HMs and CMs. Another work by the same group [168] presented the chromatin state networks in mouse ES cells using the same set of factors we used in our study. The interactions pinpointed in mouse cells have been recovered in ES networks of human cells in our study. Mouse networks show strong correlation with human networks. The [168] study focused on one cell type while we compared two cell types and also revealed the network stability across them. Despite of above all we excavated the minimum dominating set in chromatin state networks.

In order to highlight the minimum dominating sets in networks of different areas several studies have been presented so far [115, 172-179]. Biologically central nodes (genes/proteins) bear some topological centrality as compared to the rest of nodes in the network and centrality measures are discerning in revealing such nodes [127,

180]. High-degree and high-betweenness nodes (proteins) have been considered to be the controllers [115] based on centrality-lethality rule [114] in such networks. These two measures have been utilized in highlighting the minimum dominating set in these networks [115].

A chromatin states network defining the classification of the genome contains several important nodes which could be considered as the MDNS as in case of protein-protein interaction networks [115]. According to the centrality- lethality rule [114] the deletion of highly connected nodes (gene, protein) affects the network architecture three times more than a node with smaller connectivity and node centralities play an important role in identifying the driver nodes in biological networks (proteins). As genes and proteins are important part of the chromatin and cannot be considered without it therefore above rule implies the same for the chromatin as well. Our study on the identification of MDNS cleared the fact that driver nodes are important part of any chromatin network like PPI or other networks [115].

This study presents the promoter and enhancer regions followed by transcribed regions as MDNS in chromatin networks. So based on our results and the facts of previous approaches mostly presented for histone combinatorics and transcription factors [38, 41-44, 54, 167] we conclude that promoter and enhancers regions followed by the transcribed states act as the controllers of the chromatin states networks which could be visualized in case of any cell type. Therefore deletion of such nodes would lead to a chromatin without drivers like a vehicle without the controllers. As the controllers would be missing, the vehicle would move anywhere in its vicinity destroying not only the encountered things but also itself. The same scenario could be visualized if controllers of the chromatin are missing. Removing promoters, enhancers or transcribed regions would affect the chromatin badly hence leading genome vehicle to enter the territory of disease networks. Therefore the presence of controllers is important for the signaling activities across various elements of the chromatin and disruption at these nodes would be a great threat to the chromatin integrity.

This study would help in future to uncover various disease networks including various cancers and diseases caused by disruption in histone modifications networks. This would also provide a great breakthrough to inter relate genomes of various species and their networks, a study leading to disease evolutionary networks. Furthermore the

targeted therapies could be easily focused when disease networks and their underlying mysteries would be resolved, a step leading towards evolution in disease genetics and epigenetics.

5.8 Limitations

Two major objectives of the study; one the unbiased genomic segmentation and secondly the role of DNA motifs in chromatin state networks have been achieved with valid results. Due to shortage of time we only tested the working of methodologies on limited data sets. Our methods could be tested on large scale data sets in order to see the behavior across multiple species.

5.9 Recommendations

We recommend the users to test the cross species comparisons when using our methodology. The methods could be easily replicated across any platform and could easily be applied on any genomes irrespective of the sizes.

5.10 Summary

The unbiased genome segmentation based on histone combinatorics and the mystery of interacting factors in those segments is a challenging task. We in this study tried to find out a computational approach to aid the HMM segmentation to segment the genome in an unbiased way. Along with we provided 2 very simple, efficient and unbiased genome segmentation platforms, one is a global one based on clustering of binarized data and the second is local one based on a biclustering approach. In order to highlight the interactions and associations among various partners in setting chromatin states we used graph theory measures. We highlighted the role of DNA motifs in setting the chromatin states and have also seen the difference in their role with respect to cell types.

Chapter 6

CONCLUSION

Building of chromatin is settled because of histone units and modifications in these units regulate several processes including DNA-templated processes, such as DNA repair, DNA replication and mitosis. Therefore focusing histone modifications with perspective of biomarker discovery and therapeutic purpose would be a great breakthrough in the field. One of the great contributions in the field would be to resolve the issue on the bigger canvas covering a blend of interactions among individual modifications on histones, their combinatorics along with their readers, writers and erasers. Epigenetically defined chromatin states have been utilized to classify the genome for any cell type under consideration. The interactions among various factors setting these states is important to be decoded and the true meaning of these interactions needs to be pointed out. We in our study tried to highlight one of the points in this area.

We tried to provide an unbiased approach for any HMM platform to find the hidden number of states in a genome under study using specific data sizes. we conclude based on our observations that HMM parameter based evaluation and annotation based visualization is a good way to identify underlying hidden states for the ChIP-Seq data under study even for a lay man who could not identify how many significant states the data under study contains. The process will help users to find states by any of the ways we adopted, the easiest being the clustering of emission matrix and visualization of annotations.

We provided a ChromClust platform for identification of histone modifications combinatorics in an unbiased way. We present an efficient tool for clustering and classification of genome based on histone modifications combinations. Our tool is the first generic tool to successfully identify patterns enriched in various genomic regions in any species. It is robust with an easy to handle GUI (graphical user interface). It can easily handle data of any amount ranging from a single chromosome to whole genome. ChromClust represents the global clustering model. In order to highlight the patterns of histone modification combinatorics at local scale we present ChromBiSim tool. ChromBiSim is a freely available and an easy to handle tool. ChromBiSim

extracts local patterns and gives exact histone combinatorics present at various genomic locations excluding those modifications which are not present at those locations. Hence true associations of epigenomic modifications are extracted by our tool by clearly epitomizing the histone code which is foremost aspiration of epigenomics.

We also highlighted the role of DNA motifs in setting chromatin states by studying the interactions among TFs, HMs and DNA motifs. Our study indicated that whole chromatin network could be visualized as connections among various hubs, some act as dynamic ones while others as static one, each being the important players. We also highlighted that each chromatin state is defined by the set of specific motifs which could be stated as state specific motifs while some are the cell type specific ones as well. Conclusively we converge at the point that chromatin is a broad canvas portrayed by the interactive blend of histone modifications, transcription factors, chromatin modifiers along with underlying DNA motifs. We tried to highlight this blend in our own way but still lots needed to be revealed.

REFERENCES

1. H. Pearson. “Genetics: what is a gene?”. *Nature,* vol. 441.no.7092, pp. 398–401, May 2006.
2. E.Pennisi. “DNA Study Forces Rethink of What It Means to Be a Gene.” *Science,* vol. 316, no.5831, pp. 1556–1557, Jun.2007.
3. M. C. Vinci, G. Polvani and M. Pesce. “Epigenetic Programming and Risk: The Birthplace of Cardiovascular Disease?” *Stem Cell Rev.,* vol.9,no.3,pp.241-253,Jun. 2013.
4. C.H. Waddington. “The epigenotype.” *Int. J. Epidemiol*., vol.41, no.1, pp.10-13, Feb.2012.
5. P.A. Callinan and A.P.Feinberg. “The emerging science of epigenomics.” *Hum.Mol.Genet.,* vol.15,pp R95-R101, Apr.2006.
6. J.D. Watson. “Celebrating the genetic jubilee: a conversation with James D. Watson. Interviewed by John Rennie”. *Sci. Am*., vol. 288,no.4, pp. 66-69, Apr. 2003.
7. J.B. Gurdon and D.A.Melton. “Nuclear reprogramming in cells.” *Science,* vol. 322, no. 5909, pp. 1811-1815 ,Dec. 2008.
8. A. Bird. “DNA methylation patterns and epigenetic memory.” *Genes Dev.*, vol.16, no.1,pp. 6–21,Jan.2002.
9. M.G. Goll and T.H. Beston. “Eukaryotic cytosine methyltransferases.” *Annu. Rev. Biochem*., vol.74, pp. 481–514,2005.
10. S.W. Chan, I.R. Henderson and S.E. Jacobsen. “Gardening the genome: DNA methylation in Arabidopsis thaliana.” *Nat. Rev. Genet.*, vol. 6, pp.351–360, 2005.
11. T.R. Haines, D.I. Rodenhiser and P.J. Ainsworth. “Allelespecific non-CpG methylation of the Nf1 gene during early mouse development.” *Dev. Biol.,* vol.240, no.2, pp. 585–598, 2001.
12. B.H. Ramsahoye, D. Biniszkiewicz, F .Lyko, V .Clark, A.P. Bird and R. Jaenisch. “Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.” *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 5237–5242, 2000.
13. S .Lomvardas, G. Barnea, D.J. Pisapia, M . Mendelsohn, J . Kirkland and R. Axel. “Interchromosomal interactions and olfactory receptor choice.” *Cell,* vol.126, pp. 403–413,2006.
14. T. Kouzarides. “Chromatin Modifications and Their Function.” *Cell,* vol.128, pp.693–705,2007.
15. D.E. Sterner and S.L.Berger. “Acetylation of histones and transcription-related factors.” *Microbiol. Mol. Biol. Rev.*, vol. 64, pp. 435–459, 2000.
16. Y. Zhang and D. Reinberg. “Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails.” *Genes Dev.,* vol.15,pp. 2343–2360, 2006.
17. S.J. Nowak and V.G. Corces. “Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation.” *Trends Genet.,* vol. 20, pp. 214–220, 2004.
18. A. Shilatifard. “Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression.” *Annu. Rev. Biochem.,* vol. 75, pp. 243–269, 2006.

19. D .Nathan, K. Ingvarsdottir, D.E. Sterner, G.R. Bylebyl, M. Dokmanovic, J.A. Dorsey, K.A. Whelan, M. Krsmanovic, W.S. Lane, P.B. Meluh, et al. "Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive acting histone modifications." *Genes Dev.*, vol. 20, pp.966–976, 2006.
20. P.O. Hassa, S.S. Haenni, M. Elser, and M.O. Hottiger. "Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going?" *Microbiol. Mol. Biol. Rev.*, vol. 70, pp. 789–829, 2006.
21. G.L. Cuthbert, S. Daujat, A.W. Snowden, H. Erdjument-Bromage, T. Hagiwara, M. Yamada, R. Schneider, P.D. Gregory, P . Tempst, A.J. Bannister and T. Kouzarides. "Histone deimination antagonizes arginine methylation." *Cell*, vol. 118, pp. 545–553,2004.
22. L. Wang, J.L. Brown, R. Cao, Y. Zhang, J.A. Kassis, R.S. Jones. "Hierarchical Recruitment of Polycomb Group Silencing Complexes." *Molecular Cell*, vol.14,no. 5, pp. 637–646, Jun.2004.
23. C.J. Nelson, H. Santos-Rosa and T. Kouzarides. "Proline isomerization of histone H3 regulates lysine methylation and gene expression." *Cell*, vol. 126, pp. 905–916, 2006.
24. C.L. Liu, T. Kaplan, M. Kim, S. Buratowski, S.L. Schreiber, et al. "Single nucleosome mapping of histone modifications in S. cerevisiae." *PLoS Biol*., vol. 3, no. 10, pp. e328, 2005.
25. E. Bernstein and C.D. Allis. "RNA meets chromatin." *Genes Dev.*, vol. 19, no.14, pp. 1635–1655, 2005.
26. V.Azuara, P.Perry, S.Sauer, M.Spivakov, H.F.Jørgensen, R.M.John, Gouti Mina, M.Casanova, G.Warnes, M.Merkenschlager and A.G. Fisher. "Chromatin signatures of pluripotent cell lines." *Nature Cell Biology*, vol. 8, no.5, pp. 532 – 538, Mar. 2006.
27. C.R. Vakoc, S.A. Mandat, B.A. Olenchock and G.A. Blobel. "Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin." *Mol. Cell*, vol. 19, pp. 381–391, 2005.
28. A. Verdel, S. Jia, S. Gerber, T. Sugiyama, S. Gygi, S.I. Grewal and D. Moazed. "RNAi-mediated targeting of heterochromatin by the RITS complex." *Science*,vol.303, no.5658,pp.672-6, Jan. 2004..
29. T. Fukagawa, M. Nogami, M. Yoshikawa, M. Ikeno, T. Okazaki, Y. Takami, T. Nakayama and M. Oshimura. "Dicer is essential for formation of the heterochromatin structure in vertebrate cells." *Nat Cell Biol.*, vol. 6, no. 8, pp. 784-91, Aug. 2004.
30. C. Kanellopoulou, S.A. Muljo, A.L. Kung, S. Ganesan, R. Drapkin, T. Jenuwein, D.M. Livingston and K. Rajewsky. "Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing." *Genes and Dev.*, vol. 19, pp. 489–501, 2005.
31. M.Zaratiegui, D.V.Irvine, R.A.Martienssen. "Noncoding RNAs and gene silencing." *Cell*, vol.128, pp.763–776,2007.
32. M. Rassoulzadegan, V. Grandjean, P. Gounon, S. Vincent, I. Gillot and F. Cuzin. "RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse." *Nature*, vol. 441, pp. 469-474, May 2006.

33. M. Bühler, A. Verdel and D. Moazed. “Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing.” *Cell*, vol. 125, no.5, pp.873-86, Jun. 2006.
34. A.D. Goldberg, C.D. Allis and E. Bernstein. “Epigenetics: a landscape takes shape.” *Cell*, vol. 128, no.4, pp.635-8, Feb.2007.
35. M.Levine and R.Tjian. “Transcription regulation and animal diversity.” *Nature*, vol.424, pp.147-151, 2003.
36. V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, P.O. Brown. “Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.” *Nature*, vol. 409, no. 6819, pp.533-8, Jan. 2001.
37. T.H. Kim , L.O. Barrera , M .Zheng, C. Qu, M.A. Singer, T.A. Richmond, Y. Wu, R.D. Green and B. Ren. “A high-resolution map of active promoters in the human genome.” *Nature* vol.436, pp. 876-880, 2005.
38. K.J. Won, B. Ren and W. Wang. “Genome-wide prediction of transcription factor binding sites using an integrated model.” *Genome Biology*, vol.11, no.1, pp. R7. 2010.
39. G.D.Stormo. “DNA binding sites: representation and discovery.” *Bioinformatics*, vol.16, pp.16-23, 2000.
40. T. Suganuma and J.L. Workman. “Signals and Combinatorial Functions of Histone Modifications.” *Annual Review Biochemistry*, vol. 80, pp. 473–499, 2011.
41. N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, et al. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.” *Nat. Genet.*, vol.39, pp.311–318, 2007.
42. A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G.Wei, I. Chepelev, K. Zhao. “High-resolution profiling of histone methylations in the human genome.” *Cell*, vol.129, pp. 823-837, 2007.
43. K.J. Won, I. Chepelev, B. Ren and W. Wang. “Prediction of regulatory elements in mammalian genomes using chromatin signatures.” *BMC Bioinformatics*, vol. 9, pp. 547, 2008.
44. X. Wang, Z. Xuan, X. Zhao, Y. Li and M.Q. Zhang. “High-resolution human core- promoter prediction with CoreBoost_HM.” *Genome Res*. vol. 19, pp.266-275, 2009.
45. C. Zang, D.E. Schones, C. Zeng, K. Cui, K. Zhao and W. Peng. “A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.” *Bioinformatics*, vol. 25, no.15, pp.1952–1958, 2009.
46. T. Jenuwein and C.D. Allis. “Translating the histone code.” *Science*, vol. 293, pp. 1074–1080, 2001.
47. C.B. Millar and M. Grunstein. “Genome-wide patterns of histone modifications in yeast.” *Nat. Rev. Mol. Cell Biol.*, vol. 7, pp. 657–666, 2006.
48. D.K. Pokholok, C.T. Harbison, S. Levine, M. Cole, N.M. Hannett, et al. “Genome-wide map of nucleosome acetylation and methylation in yeast.” *Cell*, vol. 122, pp.517–527, 2005.
49. G. Hon, B. Ren and W. Wang. “ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.” *PLoS Comput. Biol.*, vol.4, pp. e1000201, 2008.
50. D. Ucar, Q. Hu and K. Tan. “Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering.” *Nucleic Acids Research*, pp. 1–13, 2011.

51. J. Lasserre, H.R. Chung and M. Vingron. “Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks.” *Plos Comput. Biol.*, vol. 9, no.9, pp.e1003168,2013.
52. J. Ernst and M. Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome.” *Nat. Biotechnol*., vol. 28, pp. 817–825, 2010.
53. R. Mitra, P. Müller, S. Liang, Y. Xu and Y. Ji. “Towards Breaking the Histone Code - Bayesian Graphical Models for Histone Modifications.” *Circulation: Cardiovascular Genetics*, vol. 6, no.4, pp.419-426, 2013.
54. J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, et al. “Mapping and analysis of chromatin state dynamics in nine human cell types.” *Nat*., vol. 473, pp. 43–49, 2011.
55. S.G. Landt, G.K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B.E. Bernstein, et al. “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” *Genome Res*., vol. 22, no.9, pp.1813–1831, 2012.
56. E.G. Wilbanks and M.T. Facciotti. “Evaluation of algorithm performance in ChIP-Seq peak detection.” *PLoS ONE*, vol. 5, no.7, pp.e11471,2010.
57. Q. Song and A. Smith. “Identifying dispersed epigenomic domains from ChIP-Seq data.” *Bioinformatics*, vol.27,no.6, pp.870, 2011.
58. S.A. Hoang, X. Xu and S. Bekiranov. “Quantification of histone modification ChIP-seq enrichment for datamining and machine learning applications.” *BMC Res. Notes*, vol.4, pp.288, 2011.
59. H. Xu, C.L. Wei, F. Lin and W.K. Sung. “An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data.” *Bioinformatics*, vol. 24, no.20, pp. 2344–2349, 2008.
60. B.D. Strahl and C.D. Allis. “The language of covalent histone modifications.” *Nature*, vol. 403, no. 6765, pp.41–45, 2000.
61. Z. Wang, C. Zang, J.A. Rosenfeld, D.E. Schones, A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, W. Peng, M.Q. Zhang and K. Zhao. “Combinatorial patterns of histone acetylations and methylations in the human genome.” *Nature Genet.*, vol.40, no.7, pp.897–903, 2008.
62. T.Ye, A.R. Krebs, M.A.Choukrallah, C.Keime, F.Plewniak, I.Davidson and L.Tora. “seqMINER: an integrated ChIP-seq data interpretation platform.” *Nucleic Acids Res.*, vol.39, pp.e35, 2011.
63. H. Kim, J. Kim, H. Selby, D. Gao, T. Tong, T.L. Phang and A.C. Tan. “A short survey of computational analysis methods in analysing ChIP-seq data.” *Human Genomics*, vol.5, no.2, pp.117-123, 2011.
64. F.A. Santoni. “EMdeCODE: a novel algorithm capable of reading words of epigenetic code to predict enhancers and retroviral integration sites and to identify H3R2me1 as a distinctive mark of coding versus non-coding genes.” *Nucleic Acids Res*., vol. 41, no. 3, pp. e48, Feb. 2013.
65. R. Jaschek and A. Tanay. “Spatial clustering of multivariate genomic and epigenomic information.” *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*, Springer-Verlag, 2009, pp. 170–183.
66. T.M. Przytycka and J .Zheng. “Hidden Markov Models.” *In: eLS. John Wiley & Sons, Ltd: Chichester*, 2011, doi: 10.1002/9780470015902.a0005267.pub2
67. G. Parmigiani, S. Garrett, R. Anbazhagan and E. Gabrielson. “A statistical framework for expression-based molecular classification in cancer.” *Journal*

of Royal Statistical Society. Series B (Statistical Methodology), vol. 64, pp.717–736, 2002.
68. M .Yuan and Y. Lin. "Model selection and estimation in the gaussian graphical model." *Biometrika*, vol. 94, pp.19–35, 2007.
69. J .Friedman, T. Hastie, R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, vol. 9, pp.432–441, 2008.
70. J. Scott and C. Carvalho. "Feature-inclusion stochastic search for gaussian graphical models." *Journal of Computational and Graphical Statistics*, vol.17, pp.790–808, 2008.
71. J. Besag. "Spatial interaction and the statistical analysis of lattice systems." *Journal of Royal Statistical Society. Series B (Methodological),* vol. 36, pp. 192–236, 1974.
72. C. Caragea and S. Kaiser. "Autologistic models with interpretable parameters." *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 14, pp. 281–300, 2009.
73. R. Bartfai, W.A.M. Hoeijmakers, A.M. Salcedo-Amaya, A.H. Smits, E. Janssen-Megens, A. Kaan, et al. "H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3." *PLoS Pathog.* vol. 6, pp. e1001223, 2010.
74. J. Perner, J. Lasserre, S. Kinkley, M. Vingron, H. Chung. "Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling." *Nucleic Acids Research*, vol. 42.no. 22, pp.13689-13695, 2014.
75. F. Roudier, I. Ahmed, C. Bérard, A. Sarazin, T. Mary-Huard, S. Cortijo, D. Bouyer, E. Caillieux, E. Duvernois-Berthet, L. Al-Shikhley, et al. "Integrative epigenomic mapping defines four main chromatin states in Arabidopsis." *EMBO Journal*, vol. 30, pp. 1928–1938, 2011.
76. T. Liu, A. Rechtsteiner, T.A. Egelhofer, A. Vielle, I. Latorre, M.S. Cheung, S. Ercan, K. Ikegami, M. Jensen, P. Kolasinska-Zwierz, et al. "Broad chromosomal domains of histone modification patterns in C.elegans." *Genome Research*, vol.21, pp. 227–236, 2011.
77. M.B. Gerstein, Z.J. Lu, E.L. Van Nostrand, C. Cheng, B.I. Arshinoff, T. Liu, K.Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, et al. "Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project." *Science*, vol. 330, pp. 1775–1787, 2010.
78. S. Roy, J. Ernst, P.V. Kharchenko, P. Kheradpour, N. Negre, M.L. Eaton, J.M. Landolin, C.A. Bristow, L. Ma, M.F. Lin, S. Washietl, B.I. Arshinoff, et al. "Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE." *Science*, vol. 330, pp.1787–1797, 2011.
79. N.C. Riddle, A. Minoda, P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, M.Y. Tolstorukov, A.A. Gorchakov, J.D. Jaffe, C. Kennedy, D. Linder-Basso, S.E. Peach, et al. "Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin." *Genome Research*, vol. 21, pp.147–163,2011.
80. P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, A. Minoda, N.C. Riddle, J. Ernst, P.J. Sabo, et al. "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster." *Nature*, vol.471, pp. 480–486, 2010.
81. G.J. Filion, G.J.V. Bemmel, U.Braunschweig, W. Talhout, J. Kind, L.D. Ward, W. Brugman, I.J. de Castro, R.M. Kerkhoven, H.J. Bussemaker, B. van

Steensel. “Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells.” *Cell*, vol. 143, pp. 212–224, 2010.
82. R. Bonneville and V.X. Jin. “A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor alpha target genes.” *Bioinformatics*, vol. 29, no.1, pp.22-28, 2013.
83. J.L. Larson and G.C. Yuan. “Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model.” *BMC Bioinformatics*, vol.11, pp.557, 2010.
84. N. Day, A. Hemmaplardh, R.E. Thurman, J.A. Stamatoyannopoulos and W.S. Noble. “Unsupervised segmentation of continuous genomic data.” *Bioinformatics*, vol. 23, pp. 1424–1426, 2007.
85. http://nebc.nerc.ac.uk/nebc_website_frozen/nebc.nerc.ac.uk//tools/bio-linux-7/
86. https://www.r-project.org/
87. D.W. Wright, T. Angus, A.J. Enright and T.C. Freeman. “Visualisation of BioPAX Networks using BioLayout Express3D.” *F1000 Research*, vol. 3, pp. 246, 2014.
88. J. Ernst and M. Kellis. “ChromHMM: automating chromatin-state discovery and characterization.” *Nature Methods*, vol. 9, pp.215-216, 2012.
89. M.J.L. de Hoon, S. Imoto, J. Nolan, S. Miyano. “Open Source Clustering Software.” *Bioinformatics*, vol. 20, no.9, pp. 1453-1454, 2004.
90. https://www.visualstudio.com/
91. S. Heinz, C. Benner, N. Spann, E. Bertolino, et al. “Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Mol. Cell*, vol. 38, no.4, pp.576-589, May 2010.
92. A.R. Quinlan and I.M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features.” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.
93. G. Schwarz. “Estimating the dimension of a model.” *Annuals of Statistics*, vol. 6, pp. 461-464, 1978.
94. H. Akaike. “Information theory and an extension of the maximum likelihood principle.” *In: 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR. Budapest: Akadémiai Kiadó*, 1973, pp. 267-281.
95. A.P. Dempster, N. M. Laird and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp.1-38, 1977.
96. L. Kaufman and P.J. Rousseeuw. “Finding Groups in Data: An Introduction to Cluster Analysis.” *Wiley, New York*, 1990.
97. A. Izzo, K. Kamieniarz-Gdula, F. Ramirez, N. Noureen, J. Kind, T. Manke, B. van Steensel, and R. Schneider. “The Genomic Landscape of the Somatic Linker Histone Subtypes H1.1 to H1.5 in Human Cells.” *Cell Reports*, vol.3, pp. 2142–2154, Jun. 2013.
98. R.W. Hamming. “Error detecting and error correcting codes” *Bell System Technical Journal*, vol. 29, no.2, pp. 147–160, 1950.
99. Ucsc.edu
100. L. Jia, G.Landan, M. Pomerantz, R. Jaschek, P. Herman, D.Reich, C.Yan, et al. “Functional enhancers at the gene-poor 8q24 cancer-linked locus.” *PLoS Genet.*, vol.5, pp. e1000597, 2009.

101. R.E. Thurman, N.Day, W.S.Noble and J.A.Stamatoyannopoulos. “Identification of higher-order functional domains in the human ENCODE regions.” *Genome Res.*, vol.17, pp.917, 2007.
102. B. Schuettengruber, M.Ganapathi, B.Leblanc, M. Portoso, R.Jaschek, B.Tolhuis, M.van Lohuizen, A.Tanay, G.Cavalli. “Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos.” *PLoS Biol.*, vol.7, pp. e13, 2009.
103. D.E. Schones and K. Zhao. “Genome-wide approaches to studying chromatin modifications.” *Nat. Rev. Genet.*, vol. 9, pp.179–191, 2008.
104. L. Teng and K. Tan. “Finding combinatorial histone code by semi- supervised biclustering.” *BMC Genomics*, vol. 13, pp.301, 2012.
105. N. Noureen and M.A. Qadir. “BiSim: A Simple and Efficient Biclustering Algorithm.” *International Conference of Soft Computing and Pattern Recognition SOCPAR 09*, 2009, pp. 1 – 6.
106. A.Prelić, S.Bleuler, P.Zimmermann, A.Wille, P.Bühlmann, W.Gruissem, L.Hennig, L.Thiele and E.Zitzler."A systematic comparison and evaluation of biclustering methods for gene expression data." *Bioinformatics*, vol.22(9), pp.1122-9, 2006.
107. D. Ekman, S. Light, A.K. Björklund and A. Elofsson. “What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?” *Genome Biology*, vol. 7, pp. R45, 2006.
108. S. Agarwal, C.M. Deane, M.A. Porter and N.S. Jones. “Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks.” *PLoS Computational Biology*, vol. 6, no.6,pp. e1000817, 2010.
109. Md.Silva,HMH.Ma and A.P.ZAP.Zeng. "Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks." Proc IEEE, 2008,vol. 96(8),pp.1411–1420.
110. A.J.Ignacio, L.DallAsta, A.Barrat and A.Vespignani. "k-core decomposition : a tool for the visualization of large scale networks." WorldWide Web Internet And Web Info Syst 2005. abs/cs/050. [http://arxiv.org/abs/cs/0504107]
111. B.Vogelstein, D.Lane and A.J.Levine. "Surfing the p53 network." Nature, vol.408(6810),pp.307–310,2000.
112. H.Jeong, A.L.Barab and Z.N.Oltvai."Prediction of protein essentiality based on genomic data." Complexus, vol.1, pp.19–28, 2003.
113. Y.Katzir, Y.Elhanati, I.Averbukh and E.Braun. "Dynamics of the cell-cycle network under genome-rewiring perturbations." Phys. Biol., vol. 10(6),pp.066001,2013.
114. H. Jeong, S.P. Mason, A.L. Barabási and Z.N. Oltvai. “Lethality and centrality in protein networks.” *Nature*, vol.411, no.6833,pp.41-42, 2001.
115. X.F. Zhang, L.O. Yang, Y. Zhu, M.Y. Wu and D.Q. Dai. “Determining minimum set of driver nodes in protein-protein interaction networks.” *BMC Bioinformatics*, vol.16, pp.146, 2015.
116. T. Opsahl, F. Agneessens and J. Skvoretz. “Node centrality in weighted networks: Generalizing degree and shortest paths.” *Social Networks*, vol.32, pp. 245-25, 2010.
117. L.C. Freeman. “A set of measures of centrality based on betweenness.” *Sociometry*, vol. 40, no.1, pp.35-41, 1977.
118. A. Bavelas. “Communication patterns in task-oriented groups.” *Journal of the Acoustical Society of America*, vol. 22, pp.725-730, 1950.

119. G. Sabidussi. “The centrality index of a graph.” *Psychometrika*, vol.31, pp. 581-603, 1966.
120. N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher,B.J. Breitkreutz, L.D. Hurst, et al. “Stratus not altocumulus: a new view of the yeast protein interaction network,” PLoS Biol.,vol. 4, pp. e317,2006.
121. N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher,B.J. Breitkreutz, L.D. Hurst, et al. “Still stratus not altocumulus: further evidence against the date/party hub distinction,” PLoS Biol., vol.5, pp. e154,2007.
122. T.Reguly, A.Breitkreutz, L.Boucher, B.J. Breitkreutz, G.C.Hon,C.L.Myers, A.Parsons, H.Friesen, R.Oughtred, A.Tong, C.Stark, Y.Ho, D.Botstein, B.Andrews, C.Boone, O.G.Troyanskya,T.Ideker, K.Dolinski, N.N.Batada, and M.Tyers."Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae." J. Biol.,vol.5 (4),pp. 11,2006.
123. G.Jin, S.Zhang, X.S.Zhang,L.Chen."Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast." PLoS One,vol.2 (11), pp.e1207, 2007.
124. J. D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F.Berriz, L.V. Zhang, et al. “ Evidence for dynamically organised modularity in the yeast protein-protein interaction network,” Nature, vol.430(6995), pp. 88-93,2001.
125. M.Bhattacharyya and S.Chakrabarti."Identification of important interacting proteins (IIPs) in Plasmodium falciparum using large-scale interaction network analysis and in-silico knock-out studies".Malaria Journal,vol.14(70), 2015.
126. A.J. Enright, S. Van Dongen and C.A. Ouzounis. “An efficient algorithm for large-scale detection of protein families.” *Nucleic Acids Research*, vol.30, no.7, pp.1575-1584, 2002.
127. S. Wuchty. “Controllability in protein interaction networks.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no.19, pp.7156-7160, 2014.
128. M.I. Krzywinski, J.E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones and M.A. Marra. “Circos: an Information Aesthetic for Comparative Genomics.” *Genome Res*., vol.19, pp.1639-1645, 2009.
129. L . Zhang, E.E. Eugeni, M.R. Parthun and M.A. Freitas. “Identification of novel histone post-translational modifi- cations by peptide mass fingerprinting.” *Chromosoma*, vol.112, pp.77–86, 2003.
130. N.D. Heintzman, G.C. Hon, R.D. Hawkins, P. Kheradpour, A. Stark, et al. “Histone modifications at human enhancers reflect global cell-type-specific gene expression” *Nature*, vol. 459, pp.108–112, 2009.
131. J.C.Oliveros."Venny. An interactive tool for comparing lists with Venn's diagrams."2007-2015. http://bioinfogp.cnb.csic.es/tools/venny/index.html
132. C. Lin, A.S. Garruss, Z. Luo, F. Guo and A. Shilatifard. “The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation.” *Cell*, vol. 152, pp.144-156, 2013.
133. J.A. Simon and R.E. Kingston. “Mechanisms of polycomb gene silencing: knowns and unknowns.” *Nature Reviews Molecular Cell Biology*, vol. 10, no.10, pp.697-708, 2009.
134. R. Margueron and D. Reinberg. “The Polycomb complex PRC2 and its mark in life.” *Nature*, vol. 469, no.7330, pp.343-349, 2011.
135. J.W. Whitaker, Z. Chen and W. Wang. “Predicting the human epigenome from DNA motifs.” *Nature Methods*, vol.12, pp.265–272, 2015.

136. S. Dhakshinamoorthy and A.K. Jaiswal. “c-Maf negatively regulates ARE-mediated detoxifying enzyme genes expression and anti-oxidant induction.” *Oncogene*, vol.21, pp. 5301-5312, 2002.
137. M.P. Creyghton, A.W. Cheng, G.G. Welstead, T. Kooistra, B.W. Carey, E.J. Steine, et al. “Histone H3K27ac separates active from poised enhancers and predicts developmental state.” *Proceedings of the National Academy of Sciences of the United States of America*, vol.107, pp. 21931-21936, 2010.
138. E.M.Hurt, A. Wiestner, A. Rosenwald, A.L. Shaffer, E. Campo, T. Grogan, et al. “Overexpression of c-maf is a frequent oncogenic event in multiple myeloma that promotes proliferation and pathological interactions with bone marrow stroma.” *Cancer Cell*, vol.5, pp.191-199, 2004.
139. C. Pouponnot, K. Sii-Felice, I. Hmitou, N. Rocques, L. Lecoin, S. Druillennec, M.P. Felder-Schmittbuhl and A. Eychene. “Cell context reveals a dual role for Maf in oncogenesis.” *Oncogene*, vol. 25, pp.1299-1310, 2006.
140. A. Eychene, N. Rocques and C. Pouponnot. “A new MAFia in cancer.” *Nature Reviews Cancer,* vol.8, pp.683-693, 2008.
141. H.G. Yoon, D.W. Chan, A.B. Reynolds, J. Qin, J. Wong. “N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso.” *Molecular Cell*, vol.12, pp.723-734, 2003.
142. N. Pelletier, N. Champagne, S. Stifani and X.J. Yang. “MOZ and MORF histone acetyltransferases interact with the Runt-domain transcription factor Runx2.” *Oncogene*, vol.21, pp.2729-2740, 2002.
143. T. Sasaki, A. Onodera, H. Hosokawa, Y. Watanabe, S. Horiuchi, J. Yamashita, H. Tanaka, Y. Ogawa, Y. Suzuki and T. Nakayama. “Genome-wide gene expression profiling revealed a critical role for GATA3 in the maintenance of the Th2 cell identity.” *PLoS ONE*, vol. 8, pp.E66468-E66468, 2013.
144. G.N. Filippova, C.F. Qi, J.E.Ulmer, J.M. Moore, M.D. Ward, Y.J. Hu, et al. “Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity.” *Cancer Research*, vol. 62, no.1, pp. 48-52, 2002.
145. W. Schachterle, A. Rojas, S.M. Xu and B.L. Black. “ETS-dependent regulation of a distal Gata4 cardiac enhancer.” *Developmental Biol.*, vol. 361, no.2, pp.439-49, 2012.
146. B.P.C. Chen, G. Liang, J.Whelan and T.J. Hai. “ATF3 and ATF3 delta Zip. Transcriptional repression versus activation by alternatively spliced isoforms.” *The Journal of Biological Chemistry*, vol. 269, pp.15819-15826, 1994.
147. T. Hai and T. Curran. “Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity.” *Proceedings of the National Academy of Sciences of the United States of America*, vol.88, pp. 3720–3724, 1991.
148. F. Alasti, A. Sadeghi, M.H. Sanati, M. Farhadi, E. Stollar, T. Somers, G.V. Camp. “A Mutation in HOXA2 Is Responsible for Autosomal-Recessive Microtia in an Iranian Family.” *The American Journal of Human Genetics*, vol.83, no.3, pp. 424, 2008.
149. J. Feng, T. Liu, B. Qin, Y. Zhang and X.S. Liu. “Identifying ChIP-Seq enrichment using MACS.” *Nature Protocols*, vol. 7, no.9, 2012.
150. M. Malek, M. Guruswamy, H. Owens and M. Pandya. “A Hybrid algorithm technique." *TR-89-06.Austin Texas*, 1989,pp. 78712-1188.

151. C. Bohm, K. Kailing, P. Kroger and A. Zimek. “Computing Clusters of Correlation Connected Objects.” *In: Proceedings of ACM International Conf. Management of Data (SIGMOD). Paris, France*, 2004, pp.455-466.
152. K. Liang and S. Keles. “Normalization of ChIP-Seq data with control.” *BMC Bioinformatics*, vol.13, pp. 199, 2012.
153. J.Zhu, F.He, S. Hu and J. Yu. “On the nature of human housekeeping genes.” *Trends Genet.*, vol.24, pp.481–484, 2008.
154. J.R.Khadake, and M.R. Rao. “DNA- and chromatin-condensing properties of rat testes H1a and H1t compared to those of rat liver H1bdec; H1t is a poor condenser of chromatin.” *Biochemistry*, vol.34, pp.15792–15801, 1995.
155. M.Orrego, I.Ponte, A.Roque, N.Buschati, X.Mora, and P. Suau. “Differential affinity of mammalian histone H1 somatic subtypes for DNA and chromatin.” *BMC Biol.*,vol.5, pp.22, 2007.
156. J.P.Th’ng, R.Sung, M.Ye and M.J. Hendzel . “H1 family histones in the nucleus. Control of binding and localization by the C-terminal domain.” *J. Biol. Chem.*, vol.280, pp. 27809–27814, 2005.
157. R.Alami, Y.Fan, S.Pack, T.M.Sonbuchner, A.Besse, Q.Lin, J.M.Greally, A.I. Skoultchi and E.E. Bouhassira. “Mammalian linker-histone subtypes differentially affect gene expression in vivo.” *Proc. Natl. Acad. Sci. USA*, vol.100, pp.5920–5925, 2003.
158. B. Pin˜a, P. Martı´nez and P. Suau. “Changes in H1 complement in differentiating rat-brain cortical neurons.” *Eur. J. Biochem.*, vol.164, pp.71–76, 1987.
159. I. Ponte, J.M.Vidal-Taboada and P. Suau. “Evolution of the vertebrate H1 histone class: evidence for the functional differentiation of the subtypes.” *Mol. Biol. Evol.*, vol. 15, pp. 702–708, 1998.
160. H.B. Fraser. “Modularity and evolutionary constraint on proteins.” *Nature Genetics*, vol.37, no.4, pp.351–352, 2005.
161. J.D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, et al. “Evidence for dynamically organized modularity in the yeast protein-protein interaction network.” *Nature*, vol.430, pp.88–93, 2004.
162. X. He and J. Zhang. “Why do hubs tend to be essential in protein networks?” *PLoS Genetics*, vol. 2, no.6, pp.e88, 2006.
163. E. Yoneki, P. Hui and J. Crowcroft. “Distinct Types of Hubs in Human Dynamic Networks.” *Proceedings of the 1st Workshop on Social Network Systems, ACM New York, NY, USA*, 2008, pp.7-12.
164. E. Fadhal, J. Gamieldien and E.C. Mwambene. “Protein interaction networks as metric spaces: a novel perspective on distribution of hubs.” *BMC Systems Biology*, vol.8, no.6, 2014.
165. M. Mirzarezaee, B.N. Araabi and M. Sadeghi. “Features analysis for identification of date and party hubs in protein interaction network of Saccharomyces Cerevisiae.” *BMC Systems Biology*, vol. 4, no. 172, 2010.
166. J.Göke, M. Jung, S. Behrens, L . Chavez, S. O'Keeffe, B. Timmermann, et al. “Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development.” *PLoS Computational Biol.*, vol. 7, no.e1002304,2011.
167. J. Wang. “Computational study of associations between histone modification and protein-DNA binding in yeast genome by integrating diverse information.” *BMC Genomics*, vol.12, no.172, 2011.

168. E.C.de Santa Pau, J. Perner, D. Juan, S. Marsili, D. Ochoa, H.R. Chung, et al. "Functional analysis and co-evolutionary model of chromatin and DNA methylation networks in embryonic stem cells." *bioRxiv.*, 2014.doi: http://dx.doi.org/10.1101/008821.
169. G.C. Hon, R.D. Hawkins and B. Ren. "Predictive chromatin signatures in the mammalian genome." *Human Molecular Genetics,* vol.18, pp. R195-R201, 2009.
170. R.C. McLeay, C.J. Leat and T.L. Bailey. "Tissue-specific prediction of directly regulated genes." *Bioinformatics*, vol. 27, pp. 2354-2360, 2011.
171. Z. Zhang and M.Q. Zhang. "Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes." *BMC Bioinformatics*, vol.12, no.155, 2011.
172. Y.Y. Liu, J.J. Slotine and A.L. Barabási. "Controllability of complex networks." *Nature*, vol.473, no.7346, pp.167-173, 2011.
173. M. Egerstedt. "Complex networks: Degrees of control." *Nature*, vol. 473, no.7346, pp.158-159, 2011.
174. F.J. Müller and A. Schuppert. "Few inputs can reprogram biological networks." *Nature*, vol.478, no.7369, pp.4, 2011.
175. Y. Tang, H. Gao, W. Zou and J.Kurths. "Identifying controlling nodes in neuronal networks in different scales." *PLoS ONE,* vol. 7, no.7, pp. 41375, 2012.
176. Y.Y. Liu, J.J. Slotine and A.L. Barabási. "Observability of complex systems." *Proceedings of the National Academy of Sciences of the United States of America,* vol.110, no. (7), pp.2460-5, 2013.
177. J. Gao, Y.Y. Liu, R.M. D'Souza and A.L. Barabási. "Target control of complex networks." *Nature Communications,* vol.5, no.5415, pp.1-7, 2014.
178. J.C. Nacher and T. Akutsu. "Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control." *New Journal of Physics*, vol.14, no.7, pp. 073005, 2012.
179. S.T. Hedetniemi and R.C. Laskar. "Bibliography on domination in graphs and some basic definitions of domination parameters." *Discrete Mathematics,* vol.86, no.1, pp. 257-77, 1990.
180. T. Milenkovi´c, V. Memiševi´c, A. Bonato and N. Pržulj. "Dominating biological networks." *PLoS ONE*, vol.6, no.8, pp.23016, 2011.

APPENDICES

Appendix – A

Motifs combinations in chromatin states

States	Cell_Type	Motifs
1	**AllExcept_K562**	YML081W(MacIsaac)/Yeast
	Gm12878_HepG2_K562	MA0020.1_Dof2/Jaspar
	Gm12878_HepG2	PB0013.1_Eomes_1/Jaspar
	HepG2_Helas	dl-A/dmmpmm(Bergman)/fly
	HepG2_Helas	MA0130.1_ZNF354C/Jaspar
	HepG2_K562	Su(H)/dmmpmm(Papatsenko)/fly
	H1_Helas	MA0085.1_Su(H)/Jaspar
	H1_Helas	PB0196.1_Zbtb7b_2/Jaspar
	H1_Specific	MA0260.1_che-1/Jaspar
	Helas_Specific	PH0017.1_Cux1_2/Jaspar
	K562_Specific	vnd/dmmpmm(Noyes_hd)/fly
	HepG2_Specific	PB0170.1_Sox17_2/Jaspar
	Gm12878_Specific	RPN4/RPN4_H2O2Lo/[](Harbison)/Yeast
2	**Gm12878_K562**	ASH1/Literature(Harbison)/Yeast
	Gm12878_K562	MA0319.1_HSF1/Jaspar
	Gm12878_Helas	MA0267.1_ACE2/Jaspar
	Gm12878_Helas	CRZ1(MacIsaac)/Yeast
	Gm12878_Helas	TATA-Box(TBP)/Promoter/Homer
	K562_Helas	toy/dmmpmm(Pollard)/fly
	K562_H1	SOK2/SOK2_BUT14/4-SUT1(Harbison)/Yeast
	H1_Helas	brk/dmmpmm(Down)/fly
	H1_Specific	PB0193.1_Tcfe2a_2/Jaspar
	Helas_Specific	Chop(bZIP)/MEF-Chop-ChIP-Seq(GSE35681)/Homer
	K562_Specific	PB0150.1_Mybl1_2/Jaspar
	HepG2_Specific	YML081W(MacIsaac)/Yeast
	Gm12878_Specific	MA0524.1_TFAP2C/Jaspar
3	**Gm12878_Helas**	MA0364.1_REI1/Jaspar
	Gm12878_K562_H1	MA0488.1_JUN/Jaspar
	Gm12878_K562	CHA4(MacIsaac)/Yeast
	Gm12878_K562	MA0349.1_OPI1/Jaspar
	Gm12878_HepG2	MA0379.1_SIG1/Jaspar
	Gm12878_HepG2	SeqBias: GCW-triplet

	HepG2_Helas	MA0298.1_FZF1/Jaspar
	HepG2_Helas	MA0121.1_ARR10/Jaspar
	K562_H1	tin/dmmpmm(Bigfoot)/fly
	K562_H1	Unknown6/Drosophila-Promoters/Homer
	H1_Helas	PB0196.1_Zbtb7b_2/Jaspar
	H1_Helas	PCF/Arabidopsis-Promoters/Homer
	H1_Helas	Unknown4/Arabidopsis-Promoters/Homer
	H1_Specific	Hr46/dmmpmm(Bergman)/fly
	Helas_Specific	MA0503.1_Nkx2-5_(var.2)/Jaspar
	K562_Specific	sna/dmmpmm(Bergman)/fly
	HepG2_Specific	PHO2/PHO2_H2O2Hi/[](Harbison)/Yeast
	Gm12878_Specific	CEBP:AP1(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer
4	**Gm12878_HepG2_Helas**	grh/dmmpmm(Papatsenko)/fly
	Gm12878_Helas	PB0091.1_Zbtb3_1/Jaspar
	Gm12878_Helas	Unknown4/Arabidopsis-Promoters/Homer
	Gm12878_Helas	MA0089.1_NFE2L1::MafG/Jaspar
	Gm12878_Helas	MA0553.1_SMZ/Jaspar
	Gm12878_HepG2	ACE2/ACE2_YPD/2-SWI5(Harbison)/Yeast
	Gm12878_H1	PB0032.1_IRC900814_1/Jaspar
	HepG2_H1_Helas	MA0193.1_Lag1/Jaspar
	HepG2_H1_Helas	MA0124.1_NKX3-1/Jaspar
	HepG2_K562	MA0095.2_YY1/Jaspar
	K562_H1_Helas	Tag/dmmpmm(Papatsenko)/fly
	K562_Helas	pho/dmmpmm(Bergman)/fly
	H1_Helas	XBP1/Literature(Harbison)/Yeast
	H1_Specific	Aef1/dmmpmm(Pollard)/fly
	Helas_Specific	Unknown2/Drosophila-Promoters/Homer
	K562_Specific	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer
	HepG2_Specific	MA0264.1_ceh-22/Jaspar
	HepG2_Specific	CEBP:AP1(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer
	Gm12878_Specific	MA0364.1_REI1/Jaspar
5	**HepG2_k562**	PB0130.1_Gm397_2/Jaspar
	H1_Specific	SKN7(MacIsaac)/Yeast
	Helas_Specific	MF0002.1_bZIP_CREB/G-box-like_subclass/Jaspar
	K562_Specific	Tbx20(T-box)/Heart-Tbx20-ChIP-Seq(GSE29636)/Homer
	HepG2_Specific	ttk/dmmpmm(Bigfoot)/fly
	Gm12878_Specific	MA0305.1_GCR2/Jaspar
6	**K562_Helas**	PB0199.1_Zfp161_2/Jaspar
	H1_Helas	MA0354.1_PDR8/Jaspar
	H1_Specific	PB0191.1_Tcfap2c_2/Jaspar
	Helas_Specific	MafK(bZIP)/C2C12-MafK-ChIP-

		Seq(GSE36030)/Homer
	K562_Specific	shn-ZFP2/dmmpmm(Bergman)/fly
	HepG2_Specific	Tbox:Smad(T-box,MAD)/ESCd5-Smad2_3-ChIP-Seq(GSE29422)/Homer
	Gm12878_Specific	MA0100.2_Myb/Jaspar
7	**Gm12878_H1**	MA0193.1_Lag1/Jaspar
	Gm12878_K562	HAP2/Literature(Harbison)/Yeast
	Gm12878_K562	MA0130.1_ZNF354C/Jaspar
	Gm12878_K562	MA0403.1_TBF1/Jaspar
	Gm12878_Helas	MA0325.1_LYS14/Jaspar
	HepG2_H1	MA0211.1_bap/Jaspar
	HepG2_H1	PHD1(MacIsaac)/Yeast
	HepG2_Helas	MET31(MacIsaac)/Yeast
	HepG2_K562	GATA3(Zf)/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer
	K562_H1	Srebp2(bHLH)/HepG2-Srebp2-ChIP-Seq(GSE31477)/Homer
	K562_H1	tll/dmmpmm(Bergman)/fly
	H1_Specific	PB0166.1_Sox12_2/Jaspar
	Helas_Specific	gt/dmmpmm(SeSiMCMC)/fly
	K562_Specific	MA0285.1_CRZ1/Jaspar
	HepG2_Specific	brk/dmmpmm(Down)/fly
	Gm12878_Specific	PB0179.1_Sp100_2/Jaspar
8	**Gm12878_K562_H1**	ovo/dmmpmm(Bigfoot)/fly
	Gm12878_K562_H1	MA0161.1_NFIC/Jaspar
	Gm12878_K562	br-Z2/dmmpmm(SeSiMCMC)/fly
	Gm12878_Helas	MA0126.1_ovo/Jaspar
	Gm12878_HepG2	RPN4/RPN4_H2O2Lo/[](Harbison)/Yeast
	HepG2_H1	MA0291.1_DAL82/Jaspar
	HepG2_Helas	SD0002.1_at_AC_acceptor/Jaspar
	HepG2_Helas	MA0524.1_TFAP2C/Jaspar
	HepG2_Helas	ACE2/ACE2_YPD/2-SWI5(Harbison)/Yeast
	HepG2_Helas	PB0002.1_Arid5a_1/Jaspar
	HepG2_K562	TCFL2(HMG)/K562-TCF7L2-ChIP-Seq(GSE29196)/Homer
	HepG2_K562	MA0114.2_HNF4A/Jaspar
	K562_Helas	PB0046.1_Mybl1_1/Jaspar
	H1_Specific	MA0375.1_RSC30/Jaspar
	Helas_Specific	Meis1(Homeobox)/MastCells-Meis1-ChIP-Seq(GSE48085)/Homer
	K562_Specific	AARE(HLH)/mES-cMyc-ChIP-Seq/Homer
	HepG2_Specific	RIM101/Literature(Harbison)/Yeast
	Gm12878_Specific	CG11617/dmmpmm(Noyes_hd)/fly
9	**Gm12878_K562_Helas**	IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer
	All	Sp1(Zf)/Promoter/Homer

	Gm12878_Helas	CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski et al.)/Homer
	Gm12878_H1	Deaf1/dmmpmm(Pollard)/fly
	Gm12878_K562	NFY(CCAAT)/Promoter/Homer
	AllExcept_Helas	GFY(?)/Promoter/Homer
	Gm12878_HepG2	NRF(NRF)/Promoter/Homer
	HepG2_K562	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer
	HepG2_K562	MA0437.1_YPR196W/Jaspar
	HepG2_H1	HAP2/Literature(Harbison)/Yeast
	HepG2_H1	MA0542.1_ELT-3/Jaspar
	HepG2_Helas	kni/dmmpmm(Down)/fly
	HepG2_Helas	GFX(?)/Promoter/Homer
	HepG2_K562_Helas	PB0056.1_Rfxdc2_1/Jaspar
	H1_Helas	IME1(MacIsaac)/Yeast
	H1_Specific	Unknown3/Arabidopsis-Promoters/Homer
	Helas_Specific	PB0179.1_Sp100_2/Jaspar
	K562_Specific	PB0185.1_Tcf1_2/Jaspar
	HepG2_Specific	MA0508.1_PRDM1/Jaspar
	Gm12878_Specific	PB0052.1_Plagl1_1/Jaspar
10	**Gm12878_H1**	STB5(MacIsaac)/Yeast
	Gm12878_Helas	PDR1/PDR1_YPD/[](Harbison)/Yeast
	Gm12878_Helas	RLM1(MacIsaac)/Yeast
	Gm12878_HepG2	GCN4/GCN4_SM/121-GCN4(Harbison)/Yeast
	Gm12878_K562	MA0360.1_RDR1/Jaspar
	HepG2_K562_Helas	MA0325.1_LYS14/Jaspar
	HepG2_K562	MA0393.1_STE12/Jaspar
	HepG2_K562	MA0289.1_DAL80/Jaspar
	HepG2_K562	PB0146.1_Mafk_2/Jaspar
	HepG2_H1	MA0527.1_ZBTB33/Jaspar
	HepG2_H1	MA0565.1_FUS3/Jaspar
	HepG2_H1	dl/dmmpmm(Down)/fly
	K562_H1	YY1(Zf)/Promoter/Homer
	K562_H1	ZNF711(Zf)/SHSY5Y-ZNF711-ChIP-Seq(GSE20673)/Homer
	K562_H1	MA0420.1_YBR239C/Jaspar
	H1_Specific	MA0275.1_ASG1/Jaspar
	Helas_Specific	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer
	K562_Specific	Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer
	HepG2_Specific	MA0512.1_Rxra/Jaspar
	Gm12878_Specific	IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer
11	**Gm12878_H1**	MA0507.1_POU2F2/Jaspar

	Gm12878_H1	MA0079.3_SP1/Jaspar
	Gm12878_H1	MET31(MacIsaac)/Yeast
	Gm12878_HepG2	PHA-4(Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer
	Gm12878_HepG2	Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer
	Gm12878_HepG2_K562	Ets1-distal(ETS)/CD4+-PolII-ChIP-Seq(Barski et al.)/Homer
	Gm12878_K562_H1	MA0460.1_ttk/Jaspar
	Gm12878_Helas	ARR1/Literature(Harbison)/Yeast
	Gm12878_Helas	PB0128.1_Gcm1_2/Jaspar
	Gm12878_K562	RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/Homer
	HepG2_K562_Helas	GATA3(Zf)/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer
	AllExcept_Gm12878	CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski et al.)/Homer
	HepG2_K562	CRZ1(MacIsaac)/Yeast
	HepG2_K562	MAC1/Literature(Harbison)/Yeast
	K562_Helas	EKLF(Zf)/Erythrocyte-Klf1-ChIP-Seq(GSE20478)/Homer
	K562_Helas	RUNX(Runt)/HPC7-Runx1-ChIP-Seq(GSE22178)/Homer
	H1_Helas	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer
	H1_Specific	MA0514.1_Sox3/Jaspar
	Helas_Specific	MF0006.1_bZIP_cEBP-like_subclass/Jaspar
	K562_Specific	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer
	HepG2_Specific	MA0114.2_HNF4A/Jaspar
	Gm12878_Specific	IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer
12	**Gm12878_K562**	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq(GSE29180)/Homer
	Gm12878_K562	Mef2a(MADS)/HL1-Mef2a.biotin-ChIP-Seq(GSE21529)/Homer
	Gm12878_K562	PB0097.1_Zfp281_1/Jaspar
	Gm12878_HepG2	Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer
	Gm12878_HepG2	PB0200.1_Zfp187_2/Jaspar
	Gm12878_H1	MA0160.1_NR4A2/Jaspar
	Gm12878_Helas	P(MYB)/Zea mays/AthaMap
	Gm12878_Helas	PB0044.1_Mtf1_1/Jaspar
	All_ExceptGm12878	MET31(MacIsaac)/Yeast
	HepG2_Helas	MA0382.1_SKO1/Jaspar
	H1_Helas	bZIP911(2)(bZIP)/Antirrhinum majus/AthaMap
	H1_Specific	MA0142.1_Pou5f1::Sox2/Jaspar
	Helas_Specific	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer

	K562_Specific	Gata4(Zf)/Heart-Gata4-ChIP-Seq(GSE35151)/Homer
	HepG2_Specific	MA0114.2_HNF4A/Jaspar
	Gm12878_Specific	PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIPSeq(GSE21512)/Homer
13	**Gm12878_H1**	MA0291.1_DAL82/Jaspar
	Gm12878_H1	MA0330.1_MBP1::SWI6/Jaspar
	Gm12878_H1	SeqBias: CA-repeat
	Gm12878_HepG2	MA0020.1_Dof2/Jaspar
	Gm12878_HepG2	MA0266.1_ABF2/Jaspar
	HepG2_K562	CST6(MacIsaac)/Yeast
	HepG2_H1	ROX1(MacIsaac)/Yeast
	K562_Helas	IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer
	H1_Specific	kni/dmmpmm(Down)/fly
	Helas_Specific	MA0362.1_RDS2/Jaspar
	K562_Specific	prd/dmmpmm(Down)/fly
	HepG2_Specific	IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer
	Gm12878_Specific	Unknown3/Arabidopsis-Promoters/Homer
14	**Gm12878_HepG2**	bap/dmmpmm(Noyes_hd)/fly
	Gm12878_HepG2	MA0298.1_FZF1/Jaspar
	Gm12878_Helas	MA0354.1_PDR8/Jaspar
	Gm12878_Helas	CEBP:AP1(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer
	HepG2_K562	MA0130.1_ZNF354C/Jaspar
	HepG2_K562	RAV1(1)(AP2/EREBP)/Arabidopsis thaliana/AthaMap
	HepG2_K562	GAT3(MacIsaac)/Yeast
	HepG2_Helas	PCF/Arabidopsis-Promoters/Homer
	K562_Helas	HOXA2(Homeobox)/mES-Hoxa2-ChIP-Seq(Donaldson et al.)/Homer
	K562_Helas	MA0038.1_Gfi1/Jaspar
	K562_Helas	MA0581.1_LEC2/Jaspar
	K562_Helas	MA0362.1_RDS2/Jaspar
	H1_Helas	BH2/dmmpmm(Noyes_hd)/fly
	H1_Specific	PH0158.1_Rhox11_2/Jaspar
	Helas_Specific	MA0334.1_MET32/Jaspar
	K562_Specific	MafF(bZIP)/HepG2-MafF-ChIP-Seq(GSE31477)/Homer
	HepG2_Specific	ovo/dmmpmm(Down)/fly
	Gm12878_Specific	Rbpj1(?)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer
15	**Gm12878_Helas**	MA0373.1_RPN4/Jaspar
	Gm12878_Helas	PHD1(MacIsaac)/Yeast
	Gm12878_Helas	PH0158.1_Rhox11_2/Jaspar
	Gm12878_K562	Unknown-ESC-element(?)/mES-Nanog-

		ChIP-Seq(GSE11724)/Homer
	Gm12878_K562	MA0133.1_BRCA1/Jaspar
	Gm12878_K562	MA0347.1_NRG1/Jaspar
	Gm12878_HepG2	CHR(?)/Hela-CellCycle-Expression/Homer
	Gm12878_HepG2	RDS1(MacIsaac)/Yeast
	HepG2_H1	brk/dmmpmm(Bergman)/fly
	HepG2_H1	MA0147.2_Myc/Jaspar
	HepG2_Helas	MA0368.1_RIM101/Jaspar
	K562_H1	ABI4(2)(AP2/EREBP)/Zea mays/AthaMap
	K562_H1	Unknown6/Drosophila-Promoters/Homer
	K562_H1	MF0010.1_Homeobox_class/Jaspar
	K562_H1	MA0382.1_SKO1/Jaspar
	Gm12878_HepG2_K562_H1	MafA(bZIP)/Islet-MafA-ChIP-Seq(GSE30298)/Homer
	Gm12878_K562_H1_Helas	SOK2/SOK2_BUT14/4-SUT1(Harbison)/Yeast
	HepG2_K562_H1_Helas	MA0531.1_CTCF/Jaspar
	HepG2_H1_Helas	MA0385.1_SOK2/Jaspar
	H1_Specific	PB0113.1_E2F3_2/Jaspar
	Helas_Specific	AARE(HLH)/mES-cMyc-ChIP-Seq/Homer
	K562_Specific	PB0076.1_Sp4_1/Jaspar
	HepG2_Specific	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer
	Gm12878_Specific	YY1(Zf)/Promoter/Homer
	All	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer
	All	MA0019.1_Ddit3::Cebpa/Jaspar

Appendix – B

Table S5.1 Gm12878 centrality measures

node	degree	output	betweenness	closeness	n.closeness	Avg.Centrality
S1	24	815	58	0.01	0.00	179.40
S2	35	1822.3	254	0.00	0.00	422.26
S3	49	2424.4	384	0.01	0.00	571.48
S4	45	1979.7	615	0.01	0.00	527.94
S5	27	678	79	0.01	0.00	156.80
S6	32	1163.4	140	0.01	0.00	267.08
S7	62	3198.2	618	0.02	0.00	775.64
S8	65	3503.9	1297	0.02	0.00	973.18
S9	51	2809.8	995	0.02	0.00	771.16
S10	58	3600.2	1196	0.03	0.00	970.85
S11	49	2450.3	532	0.02	0.00	606.27
S12	52	2672	304	0.02	0.00	605.60
S13	40	1986.4	1755	0.04	0.00	756.29
S14	53	2761.9	421	0.03	0.00	647.19
S15	52	2617.6	520	0.02	0.00	637.92

Table S5.2 H1hesc centrality measures

node	degree	output	betweenness	closeness	n.closeness	Avg.Centr
S1	23	819.3	70	0.00	0.00	182.461
S2	39	2121.2	920	0.01	0.00	616.042
S3	41	1822.6	259	0.01	0.00	424.521
S4	41	1739	935	0.01	0.00	543.002
S5	39	1506.8	227	0.01	0.00	354.562
S6	32	1167	150	0.01	0.00	269.802
S7	56	2811	557	0.02	0.00	684.804
S8	45	2192.4	1049	0.02	0.00	657.285
S9	48	2595.3	849	0.02	0.00	698.465
S10	48	2791.4	1409	0.03	0.00	849.686
S11	48	2356.6	646	0.02	0.00	610.125
S12	46	2195.8	265	0.02	0.00	501.363
S13	46	2419.8	232	0.03	0.00	539.567
S14	34	1397.6	1340	0.02	0.00	554.325
S15	49	2424.6	495	0.02	0.00	593.725

Table S5.3 Helas centrality measures

node	degree	output	betweenness	closeness	n.closeness	Avg.Centra
S1	27	1071	121	0.01	0.00	243.8014
S2	35	1846	849	0.00	0.00	546.0011
S3	50	2612	350	0.00	0.00	602.401
S4	62	3294	1327	0.01	0.00	936.6014
S5	30	1800	132	0.01	0.00	392.4011
S6	28	1005	132	0.01	0.00	233.0014
S7	43	2020	357	0.02	0.00	484.004
S8	45	2133	1263	0.02	0.00	688.2042
S9	43	2135	1081	0.02	0.00	651.8045
S10	51	2950	1146	0.03	0.00	829.4063
S11	54	2979	658	0.02	0.00	738.2052
S12	52	2627	356	0.02	0.00	607.0033
S13	22	608	938	0.03	0.00	313.6067
S14	60	3303	544	0.03	0.00	781.4058
S15	47	2301	434	0.02	0.00	556.4046

Table S5.4 HepG2 centrality measures

node	degree	output	betweenness	closeness	n.closeness	Avg Cent
S1	32	1531.2	180	0.01	0.00	348.64
S2	37	2009	1332	0.01	0.00	675.6
S3	44	2141.1	298	0.01	0.00	496.62
S4	44	1901.3	1139	0.01	0.00	616.86
S5	29	857.4	107	0.01	0.00	198.68
S6	34	1335.1	186	0.01	0.00	311.02
S7	56	2847.8	724	0.02	0.00	725.56
S8	51	2584	1107	0.02	0.00	748.4
S9	53	3058.4	1009	0.02	0.00	824.08
S10	53	3260.4	1829	0.03	0.00	1028.5
S11	48	2366.1	554	0.02	0.00	593.63
S12	47	2250.8	258	0.02	0.00	511.16
S13	49	2567.5	348	0.03	0.00	592.91
S14	59	3207	1864	0.03	0.00	1026
S15	49	2503.9	461	0.02	0.00	602.78

Table S5.5 K562 centrality measures

node	degree	output	betweenness	closeness	n.closeness	Avg.Cent
S1	23	722	118	0.01	0.00	172.601
S2	36	1814.8	308	0.00	0.00	431.761
S3	40	1770	266	0.01	0.00	415.201
S4	43	1911.8	920	0.01	0.00	574.962
S5	30	901.9	146	0.01	0.00	215.582
S6	31	1072.8	147	0.01	0.00	250.162
S7	44	2002.7	358	0.02	0.00	480.944
S8	53	2727	1332	0.02	0.00	822.405
S9	44	2331.4	991	0.02	0.00	673.285
S10	47	2791	734	0.03	0.00	714.406
S11	56	2961.6	1435	0.02	0.00	890.525
S12	52	2581.8	355	0.02	0.00	597.763
S13	37	1726.9	410	0.03	0.00	434.786
S14	59	3133.3	566	0.02	0.00	751.665
S15	48	2347.7	1282	0.02	0.00	735.545