

Investigating how chromatin regulates gene expression and cellular processes

By
Vadnala Rakesh Netha
LIFE10201604002

The Institute of Mathematical Sciences, Chennai

A thesis submitted to the
Board of Studies in Life Sciences
In partial fulfillment of requirements
For the Degree of
DOCTOR OF PHILOSOPHY

of
HOMI BHABHA NATIONAL INSTITUTE

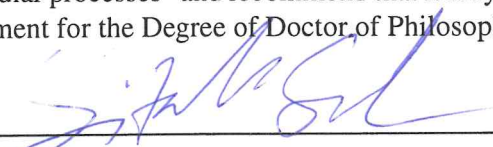


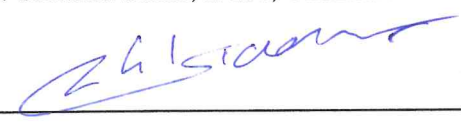
July, 2022

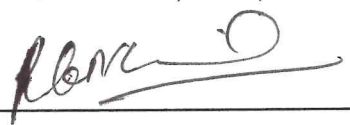
Homi Bhabha National Institute

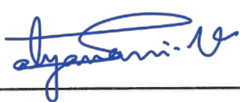
Recommendations of the Viva Voce Board


As members of the Viva Voce Board, we certify that we have read the dissertation prepared by Vadnala Rakesh Netha entitled “Investigating how chromatin regulates gene expression and cellular processes” and recommend that it maybe accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.



_____ Date: 05-04-2023
Chair - Prof. Sitabhra Sinha, IMSc, Chennai.


_____ Date: 05-04-2023
Guide/Convener - Prof. Rahul Siddharthan, IMSc, Chennai.


_____ Date: 05-04-2023
Examiner - Prof. Rakesh Mishra, TIGS, Bangalore.


_____ Date: 05-04-2023
Member 1 - Prof. Satyavani Vemparala, IMSc, Chennai.


_____ Date: 05-04-2023
Member 2 - Prof. Areejit Samal, IMSc, Chennai.



_____ Date: 05-04-2023
Member 3 - Prof. Leelavati Narlikar, IISER, Pune.

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

Date: 05-04-2023

Place: IMSc, Chennai


Guide - Prof. Rahul Siddharthan

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.



Vadnala Rakesh Netha

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.

V. Rakesh Netha

Vadnala Rakesh Netha

List of Publications arising from the thesis

Journal

1. Published

- (a) Narayanan A, Vadnala RN, Ganguly P, Selvakumar P, Rudramurthy SM, Prasad R, Chakrabarti A, Siddharthan R, Sanyal K. 2021. Functional and comparative analysis of centromeres reveals clade-specific genome rearrangements in *Candida auris* and a chromosome number change in related species. *mBio* 12:e00905-21.

2. Communicated

- (a) Rakesh Netha Vadnala, Sridhar Hannenhalli, Leelavati Narlikar, Rahul Siddharthan. Transcription factors organize into functional groups on the linear genome and in 3D chromatin. *bioRxiv* 2022.04.06.487423; doi 10.1101/2022.04.06.487423.

3. Other Publications

- (a) Sundar Ram Sankaranarayanan, Giuseppe Ianiri, Marco A Coelho, Md Hashim Reza, Bhagya C Thimmappa, Promit Ganguly, Rakesh Netha Vadnala, Sheng Sun, Rahul Siddharthan, Christian Tellgren-Roth, Thomas L Dawson, Joseph Heitman, Kaustuv Sanyal. Loss of centromere function drives karyotype evolution in closely related *Malassezia* species. *eLife* 9:e53944 (2020).

TO MY BELOVED PARENTS AND SISTERS

ACKNOWLEDGEMENTS

It has been a wonderful journey in my life and final destination of my PhD would not have been possible without the contribution of many people. I am greatly indebted to them for being part of it.

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Rahul Siddharthan for his valuable guidance, patience, and support both academically and personally. I have always found him to be easily approachable and his immense knowledge and experience helped me throughout my doctoral research. I always admired his passion for science and science outreach programs.

I would like to extend my gratitude to my collaborators Prof. Sridhar Hannenhalli, Prof. Leelavati Narlikar, and Prof. Kaustuv Sanyal for the opportunity to work with them and for the ideas, great discussions, and their guidance throughout my doctoral research. I feel gifted and lucky to have worked with them.

I want to thank Prof. Gautam Menon, Prof. Sitabhra Sinha, Prof. Satyavani Vemparala, Prof. S Krishnaswamy, Prof. Areejit Samal and Dr. Varuni P for their wonderful teaching, discussions, and support during my doctoral research. I would also like to thank my doctoral committee members for their guidance.

I want to thank my fellow student and postdoc collaborators Dr. Ankit Agrawal, Dr. Aswathy Narayanan, Dr. Sundar Ram Sankaranarayanan, Chandrani Kumari, and Pavitra S for wonderful discussions and for sharing their knowledge. I thank all seniors, classmates, and junior students of the computational Biology group at IMSc, Chennai for the great atmosphere and all the journal club talks.

I am really grateful to have a fantastic bunch of friends throughout my stay at IMSc.

I would like to thank Jayakumar Ravindran, Karthick Babu, Devanand, Vasana, Pritam Sen, Sathish Kumar, Sreejith, Raghavendra, Hitesh, Selvakumar, Abhijith, Gourav, and Shivani for their support, care, trips, and for wonderful memories. I am grateful to my classmates Janani, Deepika, and Vivek for the wonderful time during the course work. I have spent so much time enjoying games at IMSc, especially the memories of two cricket tournaments will remain with me forever. I enjoyed the company of Sohan, Tuhin, Nirmalya, Rajesh, Priyamvad, Pankaj, Prashanth, Srivastava, Devanand, Sourav, Bhargava, Raghavendra, Sahil, Subhankar, Sujoy, Karthick Babu, Gopal, Mahaveer, Umang, Hitesh, Sathish, Jyotijwal, Abhimanyu, Anuj, Vaibhav, Parth, Areejit, Sayantan, Chandrasekhar, Prabhat, Moovendan, Sukumar, Francis, Logu, Murugan, Velumurugan, Vishwajit, Sushant, Dhananjay, Shashikanta, and Surabhi on the cricket field. I am grateful to Arun, Kiruba, Vasana, Pritam, Abhinash, Anupam, Ramit, Amit, Pavitra, Sathish, Manas, Ravi, Apurba, and Amit Suthar for the company on the badminton court. During the very difficult time of the COVID second wave, I got some respite by spending and learning some tennis with Sathish, Karthick, Shivani, and Gourav. I am grateful to Mani, Vasana, Chandrasekhar, Akhil, and Sathish for gym workout tips. I am grateful Ujjal, Anupam, Sujoy, Ajjath, and Semanti for the memories of whatever little football I played. I have interacted and discussed various things ranging from life, politics, culture, Indian languages to movies, sports with many people on campus which I believe improved my personality. I had fun time during dinner table conversations, cricket tournaments, IMSc - CMI cricket matches, badminton tournaments, and institute trips with many of the above-mentioned fellow students including Sabiar, Sreevidya, Farheena, Arindam, Pavan, Vinay, Sunayanaa, Ajay S, Ajay Kumar, Vinod, Tanmoy, Ankur and I am grateful for them for being part of it.

I would like to thank the IMSc administration staff for being kind and for their help on many occasions. I also want to thank all other departments of system administration, library, civil, electrical, and canteen staff for making life so easy even during difficult times of floods, cyclones, and the COVID crisis.

Finally, I would like to express my deepest gratitude and am grateful to my parents and sisters for their unconditional love, care, and support throughout my life.

Contents

SYNOPSIS	xiii
-----------------	-------------

LIST OF FIGURES	xxii
------------------------	-------------

LIST OF TABLES	xxvi
-----------------------	-------------

1 Introduction	1
-----------------------	----------

1.1 Gene regulation	1
-------------------------------	---

1.2 Genome organization	4
-----------------------------------	---

1.3 Transcription factors and gene regulation	6
---	---

1.4 Experimental approaches to study genome organization and TF binding .	10
---	----

1.4.1 ChIP-seq	10
--------------------------	----

1.4.2 Chromatin conformation capture methods	12
--	----

1.5 Motivation of the study	14
---------------------------------------	----

2 ChromTogether - a method to assess significance of TF pair co-occurrence	17
---	-----------

2.1 Interaction network	18
-----------------------------------	----

2.2	Binding network	18
2.3	Randomized networks	18
3	Transcription factors arrange into functional groups on the linear genome and in 3D chromatin	23
3.1	TFs fall into two broad groups	25
3.2	Sequential co-occurrence patterns largely mirror spatial co-occurrence . .	34
3.3	Motif instances attract and repel similarly to TFBS	36
3.4	A consensus TF-TF co-occurrence network	41
3.5	The two main TF groups interact differently with proteins, DNA, and genes	41
3.5.1	PPI interactions are enriched in intra-group TF pairs in GM12878	41
3.5.2	Domain-domain interactions are enriched in attracting TF pairs .	43
3.5.3	Internal nodes in PPI pathways largely differ in Group 1 and Group 2 for GM12878	44
3.5.4	Group 1 TFs bind closer to promoters than Group 2 TFs	44
3.5.5	Group 1 TFs bind to GC rich regions while Group 2 TFs to GC poor regions	44
3.5.6	Group 1 target genes are enriched for housekeeping functions, Group 2 for tissue-specific functions	48
3.6	Motif strength correlates with spatial interaction	53
3.7	Discussion	55
4	Characterization of centromeres of <i>C. auris</i> and related species	57

4.1	Phylogeny of <i>C. auris</i> and related species	59
4.2	<i>C. auris</i> poses small GC-poor and repeat-free centromeres	60
4.3	<i>C. haemulonii</i> and related species similar centromere properties as <i>C. auris</i>	65
4.4	Discussion	65
5	Conclusion	67
A	List of data sets used in the study	69
B	Supplementary Information	77
C	Licence agreements	97
	Bibliography	102

SYNOPSIS

1. Introduction

Gene regulation is the act of controlling expression of genes present in the genome. It is a fundamental mechanism which plays important roles from responding to external stimuli to much more complex processes like cellular differentiation and morphogenesis. Gene expression, particularly in eukaryotic organisms, can be regulated at various stages such as signalling, genome reorganization, transcription, post-transcription changes, and translation. In this thesis, we study the interplay of two of the above mechanisms, chromatin structure and regulation of transcription by binding of sequence specific transcription factors, to gain insight into elements of gene regulation using data generated from state-of-the-art next-generation sequencing techniques.

In eukaryotic species, DNA is found in the form of chromatin. Chromatin is a DNA-protein complex that aids in the packaging of DNA inside the nucleus of the cell. Nucleosomes are formed when DNA polymers are wrapped around histone octamers, and these nucleosomes are condensed and folded further into more compact forms of chromosomes, which can be seen during the metaphase of the cell cycle. The chromatin exists as either open or closed structure at various locations of the genome. For a gene to be transcribed, the chromatin around it and its activating regulatory elements usually needs to be open (i.e., depleted of nucleosomes) as it allows transcription machinery to come in and bind at this open region, enabling the gene to be transcribed. The genome organi-

sation not only has open or closed chromatin regions for facilitating gene regulation, but also various functional 3D structures at various levels of the hierarchical organisation of the genome, such as functional loops between promoter and enhancer regions, enhancer clusters, transcription factories, loop domains, topologically associated domains, chromosome territories, and compartments for facilitating gene expression regulation in the chromosome. These structures also play roles in functions other than gene regulation such as genome stability, cell division, heredity, and DNA repair mechanisms. The chromosomes contain different distinct regions such as centromeres and telomeres specialized for different functions. Centromeres facilitate equal segregation of chromosomes into new daughter cells during cell division. Telomeres present at either end of chromosome protects the rest of chromosome from any damage.

Transcription factors (TFs) are a group of proteins that bind to DNA in regulatory regions of the genome, such as promoters, enhancers, insulators, and domain borders, in a sequence-specific manner and these short specific sequences are represented as sequence motifs. The binding of the activating or repressing transcription factors on DNA enables or inhibits recruitment of the RNA polymerase, often with the help of co-activators or co-repressors. These transcription factors have the ability to regulate their own gene expression as well as co-regulate several other genes at the same time. A gene can also be controlled by numerous TFs at the same time, resulting in a complex gene regulatory network. TFs can act on their target genes either independently or as part of a complex of TFs and co-factors by binding at the regulatory regions. Cooperative binding of TFs to DNA and competition among TFs for binding to same sequence or region have both been observed and are shown to play a role in gene regulation.

To date, studies mostly have concentrated on looking at the one-dimensional (1D) organization of the genome such as regulatory regions (promoters and enhancers) to find cooperative or antagonistic TF interactions in order to obtain insight into regulation. These studies view DNA as a 1D string of letters. However, genome organisation can facilitate

the combinatorial effects of TFs by allowing them to bind to regions that are sequentially distant but present in three-dimensional (3D) proximity due to the formation of the above-mentioned 3D functional units of promoter-enhancer loops, loop domains, and TADs, among others. Recent efforts in this direction have yielded encouraging results in terms of understanding TF-TF interaction networks. Malin *et al.* showed that spatially clustered enhancers containing homotypic TF motif sites experience greater *in vivo* TF occupancy than motif sites devoid of clusters highlighting the role of enhancer cluster in weak TFs binding. Another study by Ma *et al.* using bulk and single cell Hi-C data in human and mouse cell lines explored the co-localization of homotypic and heterotypic motif sites in Hi-C contact regions and thereby suggested spatial TF interaction network, TF-TF interactions from spatial clustering of motif sites for various TFs.

The layout of the thesis is as follows. In chapter 2, we present a novel method to gain insights into the TF-TF interaction network and regulatory roles they might play by considering the three-dimensional genome organisation at a large scale in a comprehensive manner, by pooling publicly available data of ChIP-seq data for various TFs, chromatin interaction data for multiple cell lines, and other publicly available data generated as part of the ENCODE project. In chapter 3, we provide various functional observations obtained using the method presented in chapter 2. In chapter 4, we look at the various properties of centromere regions of *C. auris* and related fungal species to understand the mechanism/events for the rapid emergence of *C. auris* and its different geological clades.

2. ChromTogether: assessing TF-TF co-occurrence from high-throughput data

ChIP-seq (chromatin immunoprecipitation sequencing), an assay based on next-generation-sequencing (NGS), enables us to identify genome-wide transcription factor binding positions, or to map chromatin associated with marks such as histone methylation or acetylation. To better understand the regulatory rules, it is necessary to identify TF interactions at cis-regulatory modules. Using ChIP-seq data for numerous TFs and chromatin marks, various studies have attempted to discover cis-regulatory binding modules or TF

co-localizations in these modules. The majority of efforts to investigate TF pair co-localization have treated the genome as a linear sequence. However, as observed in the formation of complexes on promoter-enhancer loops, factors and co-factors can create functional complexes including more than one genomic area in 3D despite the fact that these locations may be far away sequentially. Therefore, co-localization factors in 3D proximal regions as well as 1D sequentially contiguous regions must be investigated.

We provide a new technique for identifying statistically significant co-localizing TF pairs in 3D proximal interactions of the genome. The method takes advantage of chromatin proximity data from chromosome conformation capture (3C) techniques such as ChIA-PET and Hi-C, as well as TF binding site profiles from ChIP-seq experiments. The interactions between chromatin regions (significant interactions in the 3C experiment) are represented as an undirected graph. Overlapping regions are merged to create unique non-overlapping regions, which are nodes in this graph. Regions with high interaction degree are removed since they can be artifacts of the merge or of bulk averaging. Links joining regions less than 2kb apart are removed, leaving only long-range chromatin interactions to be considered. We call the resulting graph the “interaction network.”

Now, each of the nodes in the interaction network may be bound by zero or more TFs, as provided by ChIP-seq data. We define a co-occurrence of TFs A and B as binding of A and B on adjacent nodes of the interaction network. We construct a second network, the “binding network”, as a bipartite network linking TF nodes to region nodes. The interaction network and binding network for a toy example are illustrated in Figure 1 (a) and (b).

We want to find if pairs of TFs co-occur significantly more or less frequently than by chance. We do this by generating a large number of randomized binding networks. The binding network, which is a bipartite network of genomic areas and TFs, is randomised by reassigning TF-region links while preserving key features of the original network: the degree of each TF node is preserved, and links from TFs to regions are reassigned to

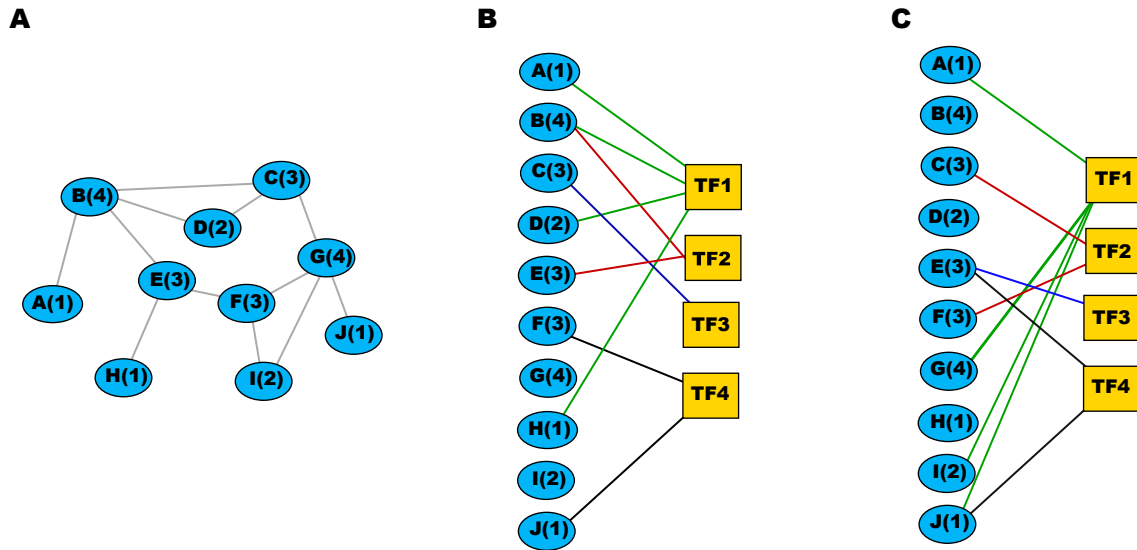


Figure 1: (A) The interaction network, where each node is a contiguous genomic region and links indicate contacts between regions as determined by Hi-C data. (B) The binding network, a bipartite network where blue nodes are regions as in (a), yellow squares are transcription factors, and links indicate binding of a TF to a region. Links from TF 1 are shown in green, from TF 2 in red, for clarity. (C) A possible randomization of links where the degree number of the bound region, as in subfigure (A), is approximately preserved.

new regions with approximately the same degree in the interaction network. If TF pairs co-occur significantly more or less frequently than in randomised networks, we call them “attractive” or “repulsive” respectively. The significance is determined by considering 1000 randomized networks and counting the fraction of these networks where the TF pair co-occurs more or less than in the real network for attraction or repulsion. These are p -values for the null hypothesis that they co-occur randomly. The Benjamini-Hochberg approach is used to correct for multiple hypothesis testing of TF pairs, yielding q -values that indicate false discovery rates.

3. Transcription factors arrange into functional groups on the linear genome and in 3D chromatin

TFs fall into two broad classes: We find that TFs segregate broadly into two groups which we call “Group 1” and “Group 2” across human cell lines, GM12878, K562, and HeLa-S3 for roughly around 60 TFs in each cell line. TF pairs within a group mostly attract

each other and avoid TFs of the other group. This is more pronounced in the GM12878 cell line; K562 does not exhibit a prominent mutually-attracting group 2, although the functional differences between the two groups are well preserved in the K562 cell line. CTCF and cohesin subunits RAD21 and SMC3 also attract each other in all the cell lines examined, but they avoid most other factors. Cohesin and CTCF have previously been identified to co-occur in spatially proximal locations and have been proposed as a mechanism for the formation of three-dimensional chromatin loops.

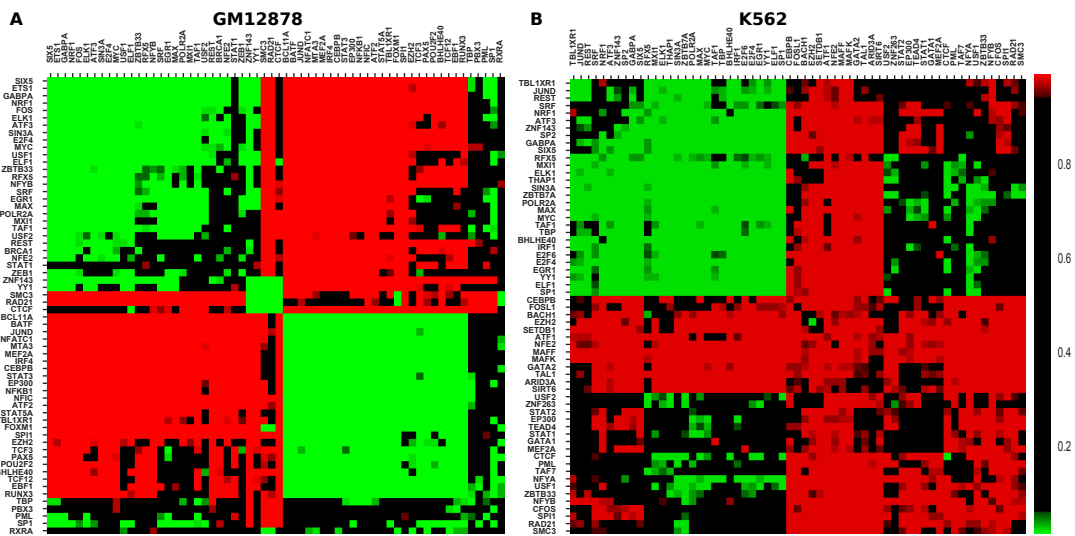


Figure 2: Clustered q -value heatmap showing attracting and repelling TF pairs in green and red respectively as described in Methods for (A) GM12878 and (B) K562 cell lines using ChIA-Pet polII data

Sequential co-occurrence patterns largely mirror spatial co-occurrence:

Since sequential combinatorial binding of TFs is a well-studied problem, we asked whether the same patterns of attraction or repulsion occur within sequentially contiguous regulatory sequence. We used the same randomization method but looked for co-occurrence within sequentially contiguous regions. Most TF pairs, it turns out, have the same qualitative behaviour as those seen in spatially proximal locations. Factors such as TBP, PML, and SP1, which are usually enriched at promoter regions, attract almost all other TFs used in the study sequentially in the GM12878 cell line, but do not show any significant behaviour in spatially proximal regions, indicating that these factors have promoter specific binding. While K562 shows only one attracting group spatially, we find that it does

exhibit two attracting groups sequentially.

Motif instances attract and repel similarly to TFBS: TFs in general bind to DNA in a sequence-specific manner and these short specific sequences are represented as sequence motifs. Motif instances of a given TF are present at several locations of genome, but TF-DNA binding depends on many factors, and only a few motif instances are bound by a TF in a given cell type. Above, we considered binding events identified via ChIP-seq experiments. Here we ask whether a similar co-occurrence pattern occurs at a motif level, even though motif instances are generally poor indicators of tissue-specific TF binding. Using 93 distinct motifs, we observe strongly similar patterns of co-occurrence of motif sites in spatially proximal regions for four cell lines i.e. GM12878, K562, HeLa-S3, and MCF7.

The two main groups of TFs interact differently with proteins, DNA, and genes: We consider physical TF-TF interactions using known protein-protein interaction (PPI) networks from the HIPPIE database. Attracting TF pairs show significant enrichment of physical interactions over repelling TF pairs. We also examined the intermediate TFs present on the shortest pathway between pairs of TFs in the PPI network. Interestingly, the frequent internal nodes are largely unique for each group. Moreover, the few TFs common to these groups tend to occur as internal nodes of internal-group TF pairs, suggesting possible roles as co-activators or co-repressors.

To further examine the possibility of undocumented protein-protein interactions, we hypothesised that if two proteins contain domains that have been shown to physically interact in existing domain-domain interaction databases, the two proteins could potentially interact through these domains. We considered such potential TF-TF interactions among the attracting and repelling TF pairs identified in our analysis. We found that attracting TF pairs have considerably more potential for TF-TF physical interaction than repelling TF pairs.

We find that group 1 TFs bind significantly closer to gene Transcription Start Sites (TSSs)

than group 2 TFs.

We identify putative target genes of each TF as genes located either within 2kbp sequentially of a TFBS for that factor, or on a spatially proximal region to that TFBS. Within each group, we consider the functions of target genes, using Gene Ontology (GO) functional enrichment analysis and disease trait enrichment analysis. GO biological process, molecular function, and GWAS disease trait enrichment analysis showed Group 1 target genes are enriched for housekeeping functions, whereas Group 2 target genes for tissue-specific functions. For example, Group 1 target genes in all cell lines GM12878, K562, HeLa-S3 showed enrichment for transcription, cell cycle, and metabolism etc. Group 2 target genes of GM12878, a lymphoblastoid cell line from a healthy donor, are enriched for immune pathways, and immune-related diseases, but this is not conserved in K562 and HeLa-S3, which are cancer cell lines, suggesting lineage-specific divergence.

4. Characterization of centromeres of *C. auris* and related species

In eukaryotes, DNA is packaged into chromosome with the help of proteins called histones. Chromosomes contain regions known as centromeres, which play an important role in faithful segregation of chromosomes into the daughter cells during cell division. Centromere function is highly conserved across all the eukaryotic organisms. These regions are also known to facilitate chromosomal rearrangements and serve to generate diversity especially in asexual organisms thereby contributing to the evolution of new karyotypes and species. Though centromere function is highly conserved across eukaryotic organisms, they exhibit greater diversity in their sequence and types: for example, they exhibit diversity in length of centromere region, repeat/transposon content, and GC-richness. In particular, budding yeast *S. cerevisiae* exhibits small “point centromeres” of ≈ 125 bp length with well-defined repeat features; multicellular eukaryotes have centromeres tens of thousands of basepairs long; and yeasts such as *C. albicans* have centromeres of intermediate length with no discernible identifying sequence features.

Candida auris is a rapidly emerging multi-drug resistant fungal pathogen which is caus-

ing systematic infections worldwide and posing threats to patients with other clinical conditions like diabetes mellitus, chronic renal disease, and, more recently, COVID-19 infections. It has evolved simultaneously into different geological clades—South Asian (clade 1), East Asian (clade 2), South African (clade 3), South American (clade 4), and a potential fifth clade from Iran. These clades are separated by tens of thousands of single nucleotide polymorphisms, suggesting rapid evolution modes in a fungal pathogen. Chromosomal rearrangements and aneuploidy are known to enhance drug resistance and virulence in primarily asexual fungi. Centromeres are susceptible to breaks in other fungal species and are most likely contribute to the diversity and rapid emergency of new clades of *C. auris*. This study, led by the Sanyal lab at JNCASR, focused on identification, characterization of centromeres and their role in karyotype diversification of *C. auris* and related fungal species through comparative genome analysis to understand the underlying mechanism/events for rapid emergence of *C. auris*. As a part of the study, I have done the bioinformatic analysis to characterize various properties of centromeres in *C. auris* and related fungal species.

We identified single copy orthologous protein sequences present in *C. auris* clades and a group of related fungal organisms belonging to Clavispora/Candida family. We used this to evaluate the phylogenetic relationship between these species.

We evaluated the genomes of all these related fungal species for GC content across their genome and observed that the centromere regions in these fungal species contain low GC content compared to the rest of genome. We also considered GC3 content, that is, the GC content of the third-position nucleotide in coding sequence, and found a dip in GC3 near all centromeres. These have been previously identified as specific centromeric markers, and served as confirmation for centromere identification in *C. auris*.

We have shown that *C. auris* lack pericentromeric heterochromatin regions by analyzing available RNA-seq data for *C. auris*. A similar result was previously reported in *C. lusitana*.

taniae.

5. Conclusion

Our tool ChromTogether has proved effective in identifying functionally different interacting groups of TFs across multiple cell lines. In future we expect to apply it to other organisms, and newer datasets, to better elucidate TF-chromatin interaction and its role in gene regulation. Such information would also be useful in improved *in silico* prediction of TF binding sites, which is an important challenge given the expense of ChIP-seq experiments, and the subject of much effort worldwide. We review the state of *in silico* TFBS prediction, and chromatin studies, and describe how our work would fit into the future of this field.

The bulk of this thesis used experimental results from the ENCODE project. Chapter 4 was a direct collaboration with an experimental lab, with relevance to disease biology. Future progress in the field will be driven by such collaborations between computational and experimental biologists and I look forward to being a part of it.

List of Figures

1	Synopsis: schematic model of ChromTogether methodology	xvii
2	Synopsis: co-occurrence heatmap of GM12878 and K562 cell lines . .	xviii
1.1	Hierarchical organization of genome in interphase of the cell cycle. . .	5
1.2	Transcription factors of eukaryotic cells.	7
1.3	Structural domains of a transcription factor.	8
1.4	Chromatin immuno precipitation sequencing(ChIP-seq) method and analysis.	11
1.5	Chromosome conformation capture techniques.	13
2.1	Schematic model of ChromTogether methodology.	19
2.2	Schematic diagram showing TF pair co-localization in 3D proximal regions.	20
3.1	Properties of GM12878 Pol II ChIA-PET interaction network.	26
3.2	Co-occurrence heatmap of GM12878 cell line.	27
3.3	Comparison of observations between our study and previous study by Ma <i>et al.</i>	29

3.4	Co-occurrence heatmap of K562 cell line.	30
3.5	Co-occurrence heatmap of HeLa-S3 cell line.	31
3.6	Co-occurrence pattern of histone marks in GM12878	33
3.7	Co-occurrence pattern in various genomic compartments of GM12878.	35
3.8	Schematic diagram showing TF pair co-localization in 1D sequentially contiguous regions.	36
3.9	Comparison of TF pair co-occurrence on linear genome and in 3D genome	37
3.10	CTCF sequence logo.	38
3.11	Co-occurrence pattern of motif sites	40
3.12	Consensus TF interaction networks	42
3.13	Internal nodes of PPI network of TFs	45
3.14	Cumulative plot of distance between TF binding sites and their nearest TSSs	46
3.15	GC content of ChIP-seq peaks regions in GM12878 cell line	47
3.16	GC content of ChIP-seq peaks regions in K562 cell line	47
3.17	GC content of ChIA-PET regions of GM12878 cell line	49
3.18	GO enrichment analysis of target genes of Group 1 and Group 2 TFs (GM12878)	50
3.19	Enriched GWAS disease terms for target genes of Group 1 and Group 2 TFs (GM12878)	51

3.20	Spatial interactions facilitate <i>in vivo</i> TF binding at weaker motif sites (GM12878)	54
3.21	Spatial interactions facilitate <i>in vivo</i> TF binding at weaker motif sites (K562)	56
4.1	Structure of chromosome	58
4.2	Phylogenetic tree of <i>C. auris</i> and related species	60
4.3	GC and GC3 content of <i>C. auris</i> clades	63
4.4	RNA-seq analysis of <i>C. auris</i>	64
4.5	GC content analysis of haemulonii complex species	66
B.1	Co-occurrence heatmap of GM12878 cell line using Hi-C.	82
B.2	Comparison of observations on Hi-C data by our method and previous study by Ma et al.	83
B.3	Comparison TF motif site co-occurrence in spatial and linear proximal regions (GM12878)	84
B.4	Comparison TF motif site co-occurrence in spatial and linear proximal regions (K562)	85
B.5	Comparison TF motif site co-occurrence in spatial and linear proximal regions (HeLa-S3)	86
B.6	Comparison TF motif site co-occurrence in spatial and linear proximal regions (MCF7)	87
B.7	Comparison of co-occurrence pattern with TFBS and motif sites (GM12878)	88
B.8	Comparison of co-occurrence pattern with TFBS and motif sites (K562)	89

B.9 GC content of ChIA-PET regions of K562 cell line	90
B.10 GC content of ChIA-PET regions of HeLa-S3 cell line	91
B.11 GC content of ChIA-PET regions of MCF7 cell line	92
B.12 Distribution of number of TFs per gene (GM12878)	93
B.13 Distribution of number of TFs per gene (K562	93
B.14 GO enriched terms of Group 1 and Group 2 TF target genes (K562) .	94
B.15 GO enriched terms of Group 1 and Group 2 TF target genes (HeLa-S3)	95

List of Tables

3.1	Domain-domain interactions are enriched among attracting TF pairs in GM12878	43
3.2	Domain-domain interactions are enriched among attracting TF pairs in K562	43
4.1	List of fungal species genome assemblies	61
A.1	List of Pol2 ChIA-PET datasets	69
A.2	List of TF ChIP-seq datasets	76
B.1	List of PWMs selected form clusters of similar motifs	81

Chapter 1

Introduction

The genetic material present in the cells of living organisms is the substance that contains the information for carrying out all the cellular and hereditary activities of the organism. It has the ability to self-replicate itself and pass it on to the next generation or to the new daughter cells. The genetic material is made up of nucleic acids (DNA in most organisms or RNA in some RNA viruses). The genetic material is present in different forms in different species. In prokaryotes, it is mostly present in the form of a single, coiled, circular chromosome, whereas in eukaryotes it exists as open DNA which along with special protein molecules forms chromatin and compacts into chromosome structures. Eukaryotic genomes are usually divided into multiple chromosomes. This organization of genetic material is vital and plays an important role in facilitating stability, genome regulation, replication, and cell cycle process, to name a few.

1.1 Gene regulation

The genome of an organism contains the information to encode various functional molecules such as RNA and protein molecules which are workhorses of the cells and carry out cellular and biological processes. But only a subset of the genes are expressed or encoded in

any given cell, in a spatiotemporal manner in accordance with the cell. For example, in multicellular organisms, all cells contain the same genome but a different set of genes are expressed in each cell type to give their characteristic phenotype. How this selective gene expression is achieved has been a fundamental question and has driven the discovery of many of the fundamental rules and mechanisms of gene regulation at play.

The genome contains a set of special functional regions known as genes, whose sequence information is passed onto the RNA molecules through the process known as transcription and this information is further passed to manufacture protein molecules which are made up of amino acids through the process of translation. This flow of information from DNA to RNA to protein molecules is known as the central dogma of molecular biology. There are various steps in the pathway leading from DNA to protein and these include processes like signaling, genome organization, transcription, post-transcriptional changes, translation, post-translational modifications, and translocation of protein molecules which can be potentially targeted for control of gene expression.

Transcription is one of the extensively studied processes and plays a major part in gene regulation. Francois Jacob and Jacques Monod, working on simple viral and bacterial systems like lambda bacteriophage and *E. coli*, discovered structural gene arrangement and switch-like behavior of gene expression which controls the initiation of transcription. For example, in *E. coli* five of the genes encoding enzymes required for the manufacture of an amino acid tryptophan are arranged next to each other on the genome and are transcribed together as one single long mRNA molecule. The expression of these five genes is switched on or off depending on the availability of amino acid tryptophan in their growth media. The molecular basis of the above mechanism includes the presence of regulatory elements, namely promoter and operator sequences. A promoter is a region usually present upstream of the gene at which the RNA polymerase bind and initiates transcription to make RNA copies. The operator, in this case, is a regulatory sequence region present within the promoter region. The tryptophan repressor protein complex in

E. coli, when tryptophan is plentiful, is bound by tryptophan, inducing a conformational change. The repressor complex can then bind to the operator sequence upstream of the tryptophan operon, blocking the transcriptional machinery and suppressing transcription.

Transcriptional repressors like the tryptophan repressor molecule in the above example or transcriptional activators which activate the switch or gene expression are examples of transcription factors that bind to the regulatory regions of the genome to activate or repress a specific set of genes. Two early discovered gene regulatory proteins or transcription factors are the lambda repressor and the Lac repressor. The lambda repressor is encoded by the virus, bacteriophage lambda. This repressor shuts off the expression of components required for the formation of new virus particles. The Lac repressor an *E. coli* bacterial protein, represses the expression of genes encoding for enzymes involved in lactose metabolism when lactose is absent in their growth medium or surrounding environment. These transcription factors which recognize special short sequences of DNA determine which subset of genes in a cell will be transcribed. Other than the ability to recognize specific sites on DNA, these transcription factors have the capacity to interact with each other in adjacent regions and act in a combinatorial manner to influence gene expression. It was also shown experimentally in bacterial systems later that they can form regulatory DNA loops through interaction between them when bound to the remote sites.

Gene regulation in eukaryotes shares many of the principles present in prokaryotes, but is much more complex and involves many other mechanisms. Eukaryotes contain regulatory elements like promoters, enhancers, and insulators to which the transcription factors bind at the transcription initiation and regulate transcription. Enhancers in eukaryotes form regulatory DNA loops with promoters and enhance transcription. One of the fundamental differences between prokaryotic and eukaryotic genomes is the presence of nucleosomes and chromatin in eukaryotes. For a gene to be transcribed, the chromatin around it and its activating regulatory elements usually need to be open (i.e., depleted of nucleosomes) as it allows transcription machinery to come in and bind at this open region, enabling

the gene to be transcribed. So, gene regulation in eukaryotes can involve additional processes which involve modifying the local chromatin state around the genes and regulatory regions.

Eukaryotic gene regulation involves a set of processes known as post-transcriptional modifications. The sequence regions of genes in eukaryotes contain non-coding sequences called introns interspersed in coding regions called exons. The intronic regions are eliminated from pre-mRNA molecules after transcription to produce mature mRNA molecules consisting of a subset of exon regions through a process known as splicing. This can give rise to a variety of alternative RNA molecules from a single gene. The interference pathway involves selective degradation of mRNA molecules through small non-coding RNA molecules such as miRNA, and siRNA and are part of post-transcriptional regulation mechanisms.

In this thesis, we study the interplay of two of the above mechanisms: chromatin structure, and regulation of transcription by binding of sequence-specific transcription factors, to gain insight into elements of gene regulation.

1.2 Genome organization

The genome in prokaryotes is usually smaller in size and is organized into single, coiled, and circular chromosomes, whereas in eukaryotic organisms it is usually large and present as multiple linear polymers called chromosomes which are organized hierarchically, and packaged into a compact and condensed state with help of proteins known as histones. For example, in humans, roughly around 3 billion nucleotide pairs are present in 23 chromosomes, which are packaged into a nucleus about 10 μm in diameter.

DNA is wrapped around an octamer of histone protein molecules (two H2A, two H2B, two H3, and two H4) to form a basic unit of DNA condensation known as a nucleosome to form a "beads on string" structure which can be visualized under a microscope. This com-

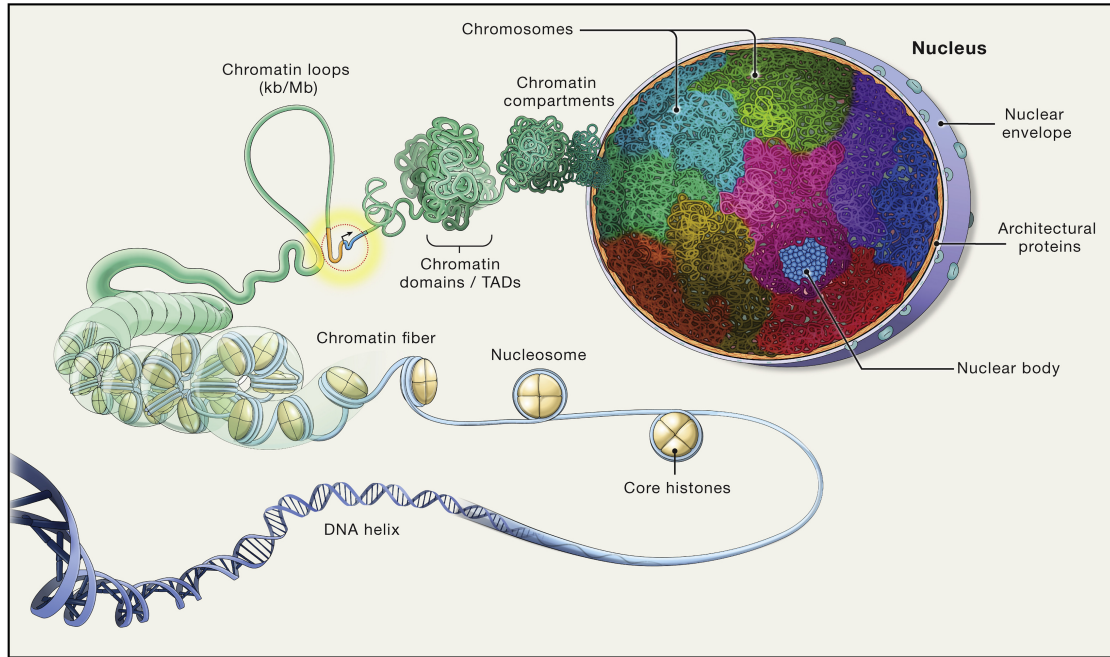


Figure 1.1: **Hierarchical organization of genome in interphase of the cell cycle.** The basic unit of chromatin organization is the nucleosome and these nucleosomes are further folded to form higher-order structures. Along with this hierarchical organization, various functional structures are formed such as compartments of euchromatin and heterochromatin, functional chromatin loops, loop domains, chromosome territories, etc. This figure is reprinted from [1] with permission from Elsevier ©(5323790188865, June 07, 2022).

plex of DNA and proteins is often referred to as chromatin. These nucleosomes further wrap to form 30-nanometer fibers consisting of an array of nucleosomes. These fibers are further condensed and compacted into the higher-order structure of chromosomes which exist or are formed during the metaphase of the cell cycle.

During the interphase stage of the cell cycle, various functional structures are formed or exist along with the hierarchical organization of the chromatin at different levels of scale or resolution. These structures include 3D chromatin loops between the regulatory regions such as loop formation between promoter and enhancer regions and protein-mediated loops such as loop formation between Polycomb repressed genes. At a lower level of resolution at megabase pairs, chromatin forms structures of domains known as topologically associated or self-interacting domains. The chromatin regions within these domains interact more often physically in 3D than with chromatin regions or sequences present outside

the domain. At an even lower level of resolution, the domains with similar epigenetic signatures form compartments with strong interdomain interactions. Further, the genome is segregated into compartments of open euchromatin or closed heterochromatin. The actively expressed genes or regulatory regions are present in open euchromatin regions of the genome, whereas the inactive genes are present in the heterochromatin region which is highly condensed. Interactions among proteins such as heterochromatin protein 1 (HP1) and DNA lead to the formation of condensates of heterochromatin structure. Chromosomes further occupy relatively compact spatially restricted regions to form chromosome territories in the nucleus and, as a result, most of the interactions of chromatin in 3D are within a chromosome with rare trans interaction between chromatin regions of different chromosomes [2, 3, 4, 5, 6, 7, 8, 9, 10].

This ordered organization of the genome and the presence of various structural elements is non-random and is different between the cell types, although variations exist between different cells of the same type. This ordered organization contains various functional elements as mentioned above and these elements and their states differ between cell types. For example, functional loops of promoter and other regulatory sequences differ between cell types, inactive genes repressed through heterochromatin formation again differ between cell types.

1.3 Transcription factors and gene regulation

Transcription factors (TFs) or gene regulatory factors are proteins that bind to DNA at regulatory regions such as promoters, enhancers, insulators, and domain boundaries in a sequence-specific manner and control the rate of transcription. These specific DNA sequences are represented by sequence motifs.

Classical biochemical, genetic and structural biology analysis *in vitro* have led to the discovery of RNA polymerases, general transcription factors, and mechanisms underly-

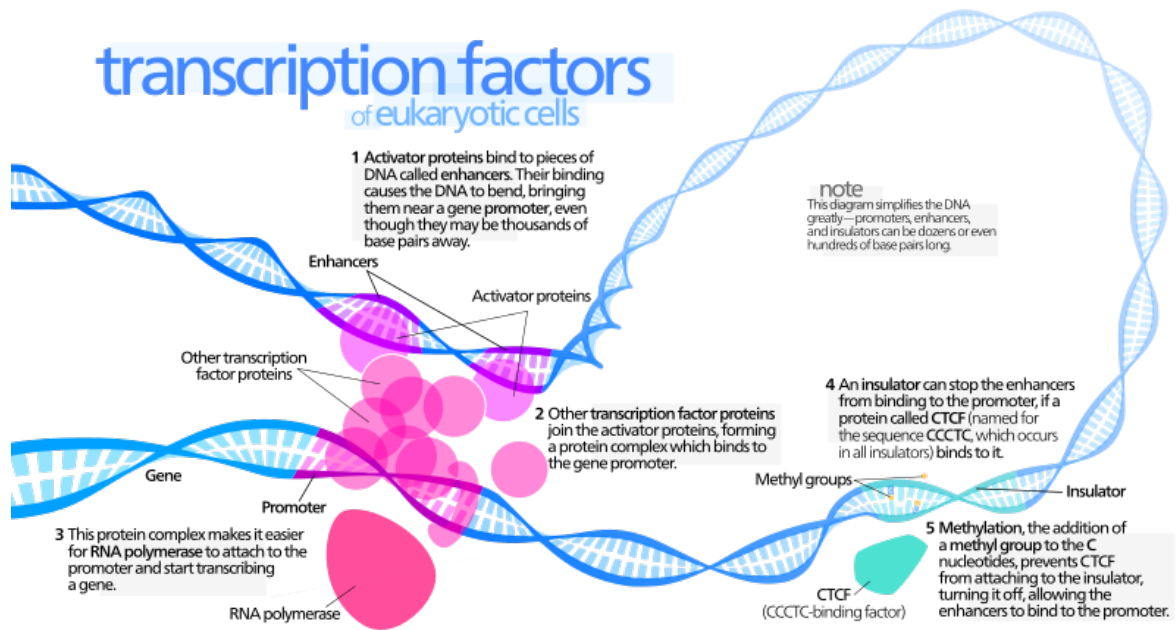


Figure 1.2: **Transcription factors of eukaryotic cells.** Transcription factors bind the regulatory regions of promoters, and enhancers in a sequence-specific manner to regulate target gene expression. They can mediate interaction with other proteins such as mediator proteins or co-factors in loop formation or help in recruitment and stabilization of RNA polymerase at the promoter or perform enzymatic activities in chromatin remodeling. This figure is reproduced from Wikipedia under Creative Commons Attribution License CC BY 3.0

ing transcription initiation and elongation [11]. RNA polymerase binds to the promoter region present upstream of the gene to initiate the transcription of the gene. RNA polymerase requires additional proteins such as the σ factor in bacterial species and many additional general transcription factors (TFIIB, TFIID, TFIIE, TFIIIF, TFIIF) for transcription initiation. These general transcription factors help in the correct positioning of RNA polymerase and allow transcription to begin and also help in the release of polymerase from the promoter region for transcription elongation [12, 13]. As seen earlier, DNA is present in the form of chromatin *in vivo* and the transcription process requires additional activator or repressor transcription factors to attract or block RNA polymerase, mediator proteins to mediate interactions between the regulatory proteins bound at distal sites and proteins to modify the local chromatin such as histone acetylase or histone deacetylase to add or remove acetylation marks on histones which can make DNA region more accessible for other regulatory proteins to bind in the region [14].

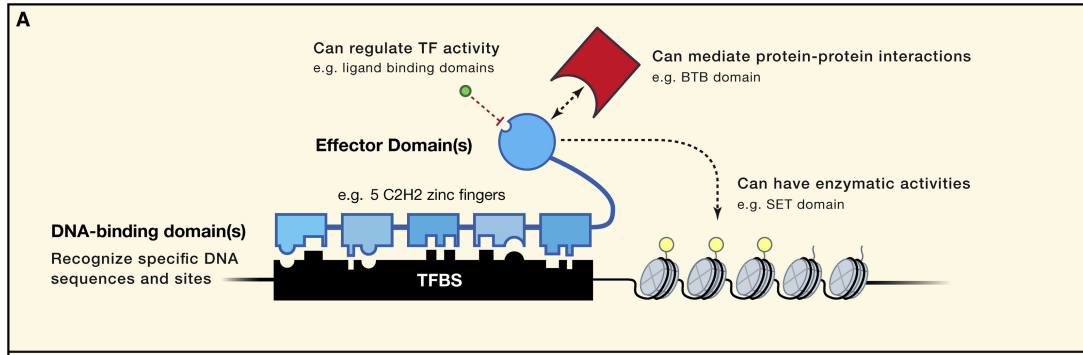


Figure 1.3: **Structural domains of a transcription factor.** Transcription factors typically contain a DNA binding domain that binds DNA in a sequence-specific manner and other activation domains through which it either interacts with other factors or performs some enzymatic activity. This figure is reprinted from [15] with permission from Elsevier ©(5324180255000, June 08, 2022).

The structures of transcription factors typically contain one DNA binding domain (DBD) and an activation domain which either interacts with other proteins or co-factor molecules or performs some enzymatic activity, and signal sensing domains to which external signal ligands bind and transmit the information to up or down-regulate gene expression. The DBD is responsible for the sequence-specific binding of the transcription factors. A variety of structural motifs or domains recognize the target sequences and are used to classify the TFs. For example, these motifs include basic helix-loop-helix, basic-leucine zipper(bZIP), homeodomains, and zinc fingers to name a few [15]. Deciphering these sequence motifs of TFs is crucial to understanding the mechanisms underlying their gene regulatory activities. Various experimental methods *in vitro* and *in vivo* were developed to identify the DNA binding ability of the proteins and their target short specific sequences. These experimental approaches include protein binding microarrays (PBMs), high-throughput in vitro selection (HT-SELEX), and chromatin immunoprecipitation sequencing (ChIP-seq) [16, 17, 18, 19, 20, 21, 22, 23, 24].

Eukaryotic species contain a large number of TFs regulating gene expression. For example, in humans, > 1600 TFs are discovered so far. These have been shown to perform roles in various biological functions. In multicellular organisms, they are involved in the development and differentiation of different cell types and are called master transcrip-

tional regulators. TFs Oct4, Nanog, Sox2 act as master regulators in embryonic stem cells (ESCs) [25, 26, 27]. They function as signal transmitters to up or down-regulate the expression of target genes in response to intercellular signals or environmental cues. TFs such as tumor suppressors or proto-oncogenes help in regulating the cell cycle. TF genes are shown to be associated with many diseases and phenotypes.

TFs can regulate the expression of their own gene as well as other genes present in the genome. At the same time, many TFs can regulate a single gene's expression. All this results in a very complex gene regulatory network. A gene regulatory network is a network of genes and TFs that shows connections between TFs and their target genes. Studies on gene regulatory networks have shown the occurrence of recurrent network motif structures in them. Some of these network motifs include negative autoregulation (TF down-regulates its own gene expression), positive autoregulation (TF enhances its own gene expression), feed-forward loops, etc [28]. The presence of these network motifs has been identified in organisms ranging from bacteria, and yeasts to plants and animals and further small synthetic gene networks are designed to perform simple functions [29, 30, 31, 32, 33, 34, 35].

Although the TFs in prokaryotes can bind to DNA individually and can affect target gene expression, in eukaryotes it is much more complex and TFs often employ their effect through complexes of multiple TFs. Many TFs have been shown to form homodimers or heterodimers with other TFs and bind DNA. Enhancers usually contain multiple motifs of different TFs and the binding of different TFs in these regulatory regions is reflected in the enhancer functions. Combinations of TFs may act in an additive, synergistic, or redundant manner, highlighting the importance of combinatorial TF regulation [36, 37, 38, 39, 40, 41, 42, 43, 44, 45]. TFs can cooperate or compete with each other for DNA binding. Structural analysis of protein-DNA and protein-protein interactions have been studied to understand the mechanism of co-operative binding TFs on DNA. These involve either direct protein-protein interaction before binding to DNA or cooperation mediated

by DNA[46]. The latest advances in experimental methods like ChIP-seq allow the study of such combinations of TFs from a statistical viewpoint. These studies look for TF pairs with a high number of binding sites or motif site co-localization occurring more than randomly expected in the genome [47, 48, 49, 50, 51, 52, 53, 54].

1.4 Experimental approaches to study genome organization and TF binding

Recent advances in sequencing technologies and computational analysis have led to developments of various experimental methods to study various aspects of the genome such as genome architecture, TF binding or epigenetic mark profiles, gene expression quantification, and genome variants [55, 56]. The rapid decline in sequencing costs has led to the generation of vast amounts of data by individual laboratories as well as consortiums such as Human Genome Project, ENCODE, GTEx, The Human Cell Atlas, and The Cancer Genome Atlas Program, to understand various biological aspects [57, 58, 59, 60]. These experimental methods typically involve sequencing the elements of DNA associated with a biochemical reaction/event and identifying the positions of regions in the reference genome.

1.4.1 ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a sequencing methodology to profile the binding locations in the genome for any given protein of interest such as TFs or genomic regions associated with epigenetic marks such as histone methylation or acetylation patterns [24, 61, 62].

The protein of interest is covalently cross-linked to the DNA in living cells, the cells are broken apart, and the DNA is digested or fragmented into small pieces using restriction

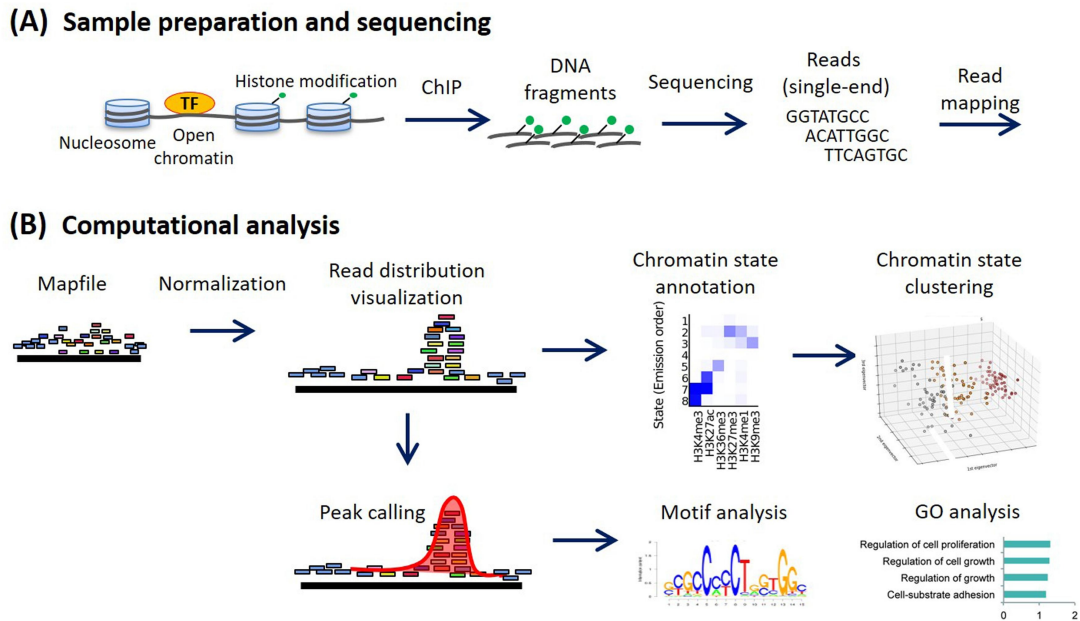


Figure 1.4: **Chromatin immuno precipitation sequencing(ChIP-seq) method and analysis.** The figure shows the details of various steps involved in the ChIP-seq method starting from sample preparation, sequencing the DNA elements bound by the TF, and computational methods to detect the positions of genome bound by the TF. This figure is reproduced from [62] under Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND)

enzymes or by mechanical forces. Antibodies against the protein of interest are used to purify pieces of the DNA to which the protein is covalently cross-linked in the cell. Later, these regions are sequenced and referred to as reads. These reads are mapped to the reference genome. This mapped read information is used to identify significantly enriched regions of the genome where the given protein is bound *in vivo* or epigenetic modifications are present.

The computational analysis includes mapping the sequence reads to the genome (called “read mapping” [63, 64]), and identifying significantly enriched loci of protein binding from the mapped reads (“peak calling” [65, 66, 67, 68]). These loci or peak regions are used for functional analysis of motif finding, identifying the chromatin state of these regions for various epigenetic marks or functional genomic elements [69, 70, 71].

1.4.2 Chromatin conformation capture methods

Chromatin conformation capture (3C) methods are a set of methods based on chemical crosslinking of chromatin and are used to study the 3D functional elements and genome organization of the chromatin [72, 73].

The chromatin in the nucleus of the cells is crosslinked *in vivo* and fragmented into smaller pieces using a restriction enzyme which are then re-ligated to each other using a ligase enzyme. This re-ligation step can form a new chimeric product between the regions of the genome which are proximal to each other in 3D space but may be present far apart on the linear genome sequence. These re-ligated sequences are collected and sequenced. These sequenced reads are aligned to the reference genome to identify parts or locations of the genome from which these sequences have originated. These sequence reads are further processed to identify significantly interacting or spatially proximal regions present in the nucleus.

The difference between different 3C-based methods is mainly in terms of the characterization of a number of chromatin interactions. For example, the 3C method is used to characterize whether two genomic regions of interest are present spatially proximal. 4C allows probing a fragment with all unknown fragments, 5C characterizes all interactions within regions of a given domain of the genome. Hi-C method is used in a high throughput manner to characterize all possible pairs of fragments of the genome for spatial proximity. Other variants such as ChIA-PET and CHIP-loop enrich all interactions which involve a specific protein of interest by incorporating an additional step of immunoprecipitation for the protein [74, 75, 76, 77, 78].

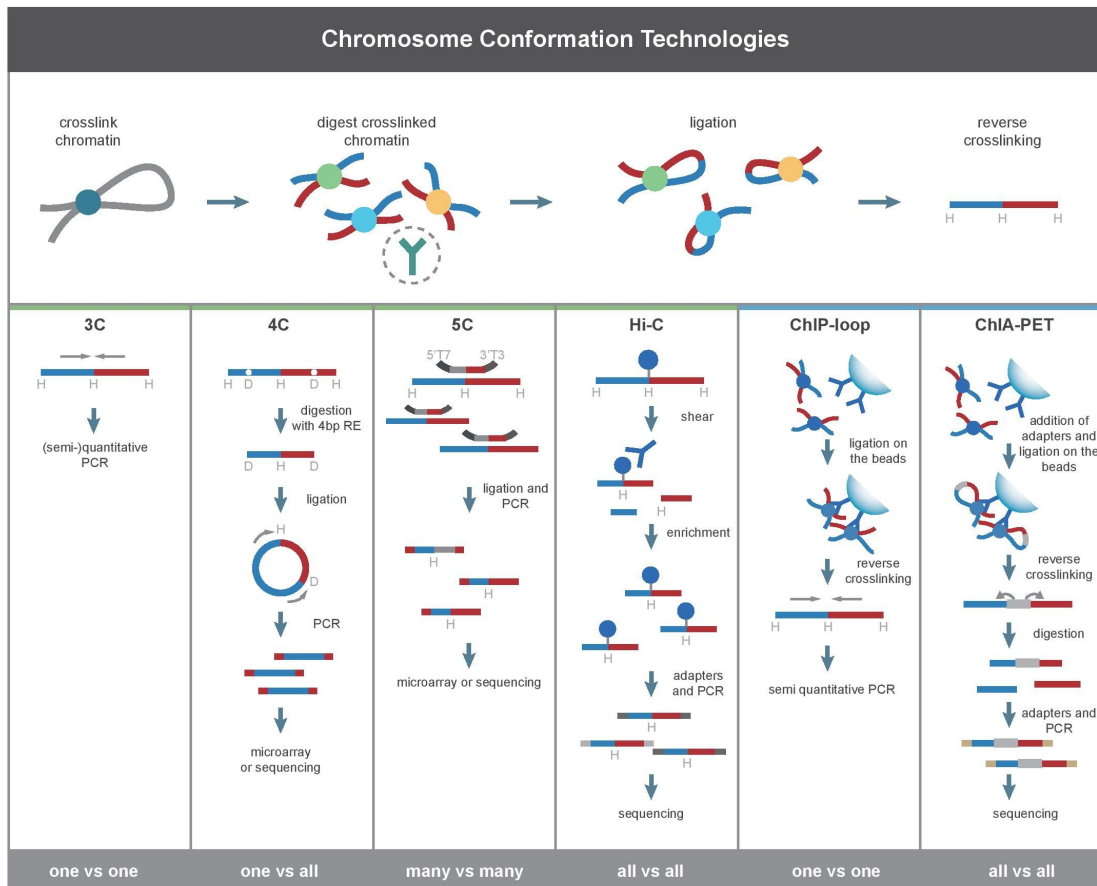


Figure 1.5: Chromosome conformation capture techniques. The figure shows the basic steps involved in chromatin conformation capture (3C) methods and difference between various 3C derived methods (4C, 5C, Hi-C, ChIP-loop, and ChIA-PET). This figure is reproduced from [72] under Creative Commons Attribution License CC BY-SA 4.0

1.5 Motivation of the study

In chapter 2 and chapter 3 of the thesis, we address the question of transcriptional regulation by transcription factors through their interaction mediated by the three-dimensional organization of the genome.

In eukaryotes as mentioned above, transcription is often regulated by a combination of transcription factors. These TFs bind to cis-regulatory regions, i.e. promoters and enhancers. Enhancers regions which are far away from the target gene location on linear DNA sequence level are brought into proximity of the target genes through proteins such as mediators. Multiple TFs regulating the target gene bind either promoter regions proximal to the gene location, or enhancers which are brought in spatial proximity, and work in a combinatorial manner in transcriptional regulation. Although researchers in the field have tried to explore combinations of TFs working together, studies so far are either limited to looking at a limited number of TFs or only looking at their co-localization on the region of 1D linear DNA sequence. Overlap of ChIP-seq peaks of TFs has been used extensively to identify the cis-regulatory elements in a given cell type [51, 52, 53, 54, 57, 79, 80, 81]. But recent studies on functional elements of the 3D organization suggest they might play an equally important role as linear genomic DNA sequence elements in gene regulation. 3D functional chromatin architecture potentially can serve as a medium for different TFs to interact in a combinatorial manner and perform their regulatory roles.

TFs bind to DNA in a sequence-specific manner and these are represented as sequence motifs. These motif sequences for any of the TF may be present at many regions in the genome, but the TF binds to only a few of these regions in *in vivo*. Although experimental approaches such as ChIP-seq and ChIP-Exo are routinely used to identify these *in vivo* binding sites, it is laborious and expensive to perform such experiments for all the TFs across all the cell types. Therefore, researchers have used computational models to predict these *in vivo* binding sites of TF using information such as binding motifs, lo-

cal DNA structural information, accessibility of the chromatin, association of chromatin region with various epigenetic marks, and gene expression data [82, 83, 84, 85, 86]. It is important to study whether the 3D organization or 3D functional chromatin structures contribute to *in vivo* TF binding and incorporate such information in TF *in vivo* binding prediction models to improve their performance.

Chapter 2

ChromTogether - a method to assess significance of TF pair co-occurrence

In eukaryotes, TFs mostly work together with other factors or form complexes to regulate gene expression. Though efforts have been made to discover such TF interactions, these studies were either limited to a few TFs or have looked at TF-TF interactions in linear genomic regions. We know, however, that enhancers, which are sequentially distant from genes and promoters, are brought into proximity with these regions via the 3D arrangement of chromatin. To investigate such TF pair interactions facilitated through the genome's 3D organization, we developed a method, "ChromTogether", to find TF pairs that either significantly co-localize or avoid each other in spatially proximal chromatin regions.

The method relies on two types of data: chromatin organization information, and TF binding information. Information on chromatin organization or spatial proximity of chromatin regions is taken from chromosome conformation capture (3C) related experiments such as ChIA-PET, Hi-C, and Capture Hi-C. The TF or epigenetic mark genome profile is taken from ChIP-seq experiments. These two data types are represented in terms of two networks which we call the "interaction network" and the "binding network".

2.1 Interaction network

The interaction network is a network of individual chromatin/genomic regions of the genome derived from the chromatin conformation capture derived experiments. The links or edges between these nodes in the network represent spatial proximity or physical interactions captured by the chromatin conformation capture experiments. First, the chromatin interaction data is pre-processed to identify unique non-overlapping individual genomic regions. The individual chromatin fragments or regions are merged if there is any overlap between the two regions. Then, the chromatin interactions or pairs occurring within 2kbp of each other in the genome are filtered out to obtain a set of long-range interaction pairs. Further, we also remove regions with an unrealistically high degree (>20) or length (>30 kbp) as these are likely the result of the merging step or from the experiment reporting a cell population average. A typical interaction network is illustrated in Figure 2.1 (A).

2.2 Binding network

We next look at TF-binding information from ChIP-seq data for the same cell type. We construct a corresponding “binding network” which contains all the region nodes from the above interaction network as well as additional TF nodes. An edge in this network connects a TF node to a region node if there is evidence of the TF binding to that region in a ChIP-seq experiment. The binding network is illustrated in Figure 2.1(B).

2.3 Randomized networks

The goal of the method is to assess TF pairs that co-occur significantly more or less than random on the interaction network. For this purpose, all instances of TFs i and j bind-

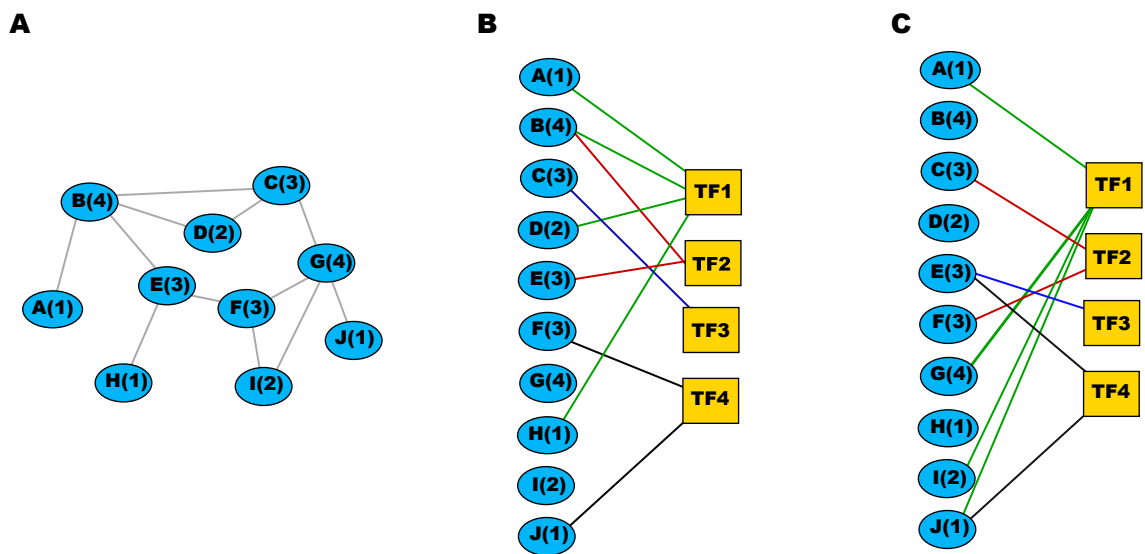


Figure 2.1: **Schematic model of ChromTogether methodology.**(A) The interaction network, where each node is a contiguous genomic region and links indicate contacts between regions as determined by chromatin interaction data. The degree of each region in the region-region interaction network is shown in brackets. (B) The binding network, a bipartite network where blue nodes are regions as in (A), yellow squares are TFs, and links indicate binding of a TF to a region. Links from TF1 are shown in green, from TF2 in red, for clarity. (C) A possible randomization of links where each TF-region link is randomly reassigned from the TF to another (possibly the same) region such that the region-region interaction degree of the bound region, in brackets, is approximately preserved.

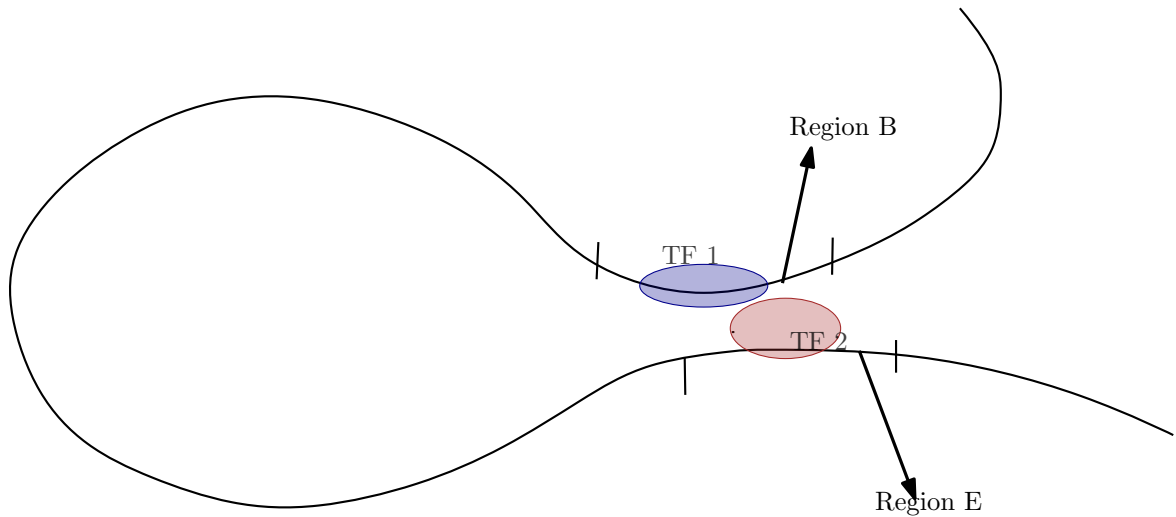


Figure 2.2: **Schematic diagram showing TF pair co-localization in 3D proximal regions.** This cartoon illustrates one instance of two TFs 1 and 2 co-localizing in spatially proximal chromatin regions B and E. In ChromTogether method, all such instances of any TF pair co-localizations are counted and significance of count is estimated for the TF pair co-occurrence using the randomized networks.

ing to adjacent nodes in the interaction network are considered as co-occurrences of the two TFs. The cartoon figure 2.2 illustrates one instance where two TFs co-localize in spatial proximal regions. We estimate the significance of these co-occurrences by simulating random networks whose construction is described below. We define co-occurrence that is significantly more frequent than random as “attraction” and co-occurrence that is significantly less frequent than random as “repulsion”.

Our randomized networks are constructed to preserve the essential properties of the original data. We view the binding network as a bipartite network, with nodes representing regions in the interaction network linked to nodes representing TFs in the ChIP-seq data. Region-region links are not included in the binding network, but the region-region interaction degree of each region is known and the node corresponding to each region is labeled, in the binding network, with its region-region interaction degree. We generate random networks by reassigning these TF-region links with the constraint that each TF-region link is assigned from the same TF to a region with a similar region-region degree. Specifically: we reassign a TF link to a region x of degree ≤ 5 to a random region y with

exactly the same region-region degree as region x ; a region x with degree between 5 and 10 to a random region y with a degree within ± 2 of region x 's degree; and a region x with degree >10 to a random region of degree >10 . This ensures that in the randomized bipartite network, each TF node has the same out-degree as originally and that the interaction-network degree distribution for TF target regions is approximately conserved. This is illustrated in figure 2.1(C).

We construct 1000 such randomized networks. From these, we calculate two p -values for each pair of TFs: p_{ij} is the p -value for TFs i and j co-occurring equally or more often than in the real data under the null that they co-occur randomly, calculated as the fraction of random networks where this happens; and similarly p'_{ij} is the p -value for i and j co-occurring equally or *less* often than in the real data. To correct for multiple pairwise comparisons, we use the full list of p -values to calculate q -values q and q' , which indicate false discovery rates, following the Benjamini-Hochberg procedure [87]. Small values of q , and q' , respectively indicate TF pairs co-occurring significantly more often, and less often than chance. Finally, we show the significant q and q' values (less than 0.05) on a heatmap, in green and red respectively. In order to show both attraction and repulsion on the same heatmap, we plot q in green and $\tilde{q}_{ij} = 1 - q'_{ij}$, $\tilde{q}_{ij} \geq 0.95$ in red (with the brightest reds being the largest, i.e. most significant values). As a control, we selected one of the random networks as a real network and compared it with all other random networks which results in no significant attraction or repulsion of TF pairs.

The list of publicly available data sets of chromatin interaction data of Pol II ChIA-PET of various cell lines and TF binding data from ChIP-seq experiments which were used to identify the significant co-occurring TF pairs using the ChromTogether method is provided in tables A.1 and A.2 respectively.

Chapter 3

Transcription factors arrange into functional groups on the linear genome and in 3D chromatin

Transcription factors (TFs) in eukaryotes bind DNA and regulate gene expression in combination with other factors, co-factors, or architectural regulators. Most often, these combinatorial interactions are context dependent: these associations or interactions differ from cell type to cell type, and different complexes exist at different stages within a cell type. Identification of these pairs of TF interactions in a given context of the cell is a well studied problem. Previous researchers have extensively studied the problem to understand the regulatory aspects of gene expression. Recent advances of sequencing technologies have further fuelled such studies. But most of these studies so far have only considered the one-dimensional aspect of the genome, that is, the linear sequence of the genome. The three dimensional structure of the genome is also extensively studied in recent times and these studies suggest the 3D aspect of the genome plays an equally important role, if not more so, in gene regulation. To this end, we have developed a method to identify significantly co-localized pairs of TFs in 3D chromatin interactions which has been described

in detail in chapter 2 of the thesis.

In this chapter, we present observations from our analysis looking for pairs of TFs co-occurring significantly both on the linear genome as well as in 3D chromatin interactions. We also performed various functional analyses on the observed co-occurring TF pairs, which are described in detail in later sections of this chapter. We performed the analysis on four different human cell lines namely, GM12878, K562, HeLa-S3, and MCF7. GM12878 cell line is a lymphoblastoid cell line which has a normal karyotype, whereas the other three cell lines K562, HeLa-S3, and MCF7 are cancerous cell lines derived from lymphoblasts from the bone marrow of a chronic myelogenous leukemia patient, cervical cancer cells, and breast cancer cells respectively.

We have used Pol II ChIA-PET data for inferring long-range 3D chromatin interactions for our analysis. ChIA-PET is a chromatin conformation capture derived high throughput technique used to detect the chromatin interaction associated or facilitated by any protein of interest [78, 88, 89]. Pol II is an RNA polymerase enzyme that is required for the production of mRNA molecules in the nucleus. Pol II is usually present at the active gene promoter elements, and is known also to bind to enhancer elements. Enrichment of Pol II through immunoprecipitation in ChIA-PET essentially captures promoter-promoter or promoter-enhancer loop interactions which exist in the cell. The publicly available Pol II ChIA-PET data sets for various cell lines used in the study are given in table A.1.

For TF binding information, we have used ChIP-seq data generated for various TFs as a part of a larger study by ENCODE consortium[57]. We used these uniformly processed TF ChIP-seq by ENCODE consortium. The TFs considered for each of the cell lines in the study are provided in table A.2.

3.1 TFs fall into two broad groups

The interaction network of chromatin interactions using Pol II ChIA-PET data for GM12878, K562, and HeLa-S3 and the binding network using available ChIP-seq data of TFs for the same cell line are generated as described in chapter 2 of the thesis. The Interaction network constructed from Pol II ChIA-PET data of GM12878 contains 76083 nodes, 79385 edges, and 18125 isolated connected components. The binding network constructed using above 76083 chromatin region nodes from the interaction network and 62 TFs of the GM12878 resulted in 478406 edges in the binding network.

We explored the co-binding of TFs using the method provided in chapter 2 of the thesis. Briefly, the method uses the above mentioned interaction network and binding network to count instances of TF pairs co-localized in 3D chromatin interactions. Further, the count is compared with a similar count of co-localization in 1000 randomly generated networks to calculate p -values of co-occurrence for each possible TF pair. From this list of p -values, q -values were calculated using the Benjamini-Hochberg method[87]. The resulting matrix of q -values was clustered along both the rows and columns simultaneously. The significant attracting and avoiding TF pairs are shown in green and red in a single heatmap. The co-occurrence heatmaps for the GM12878, K562, and HeLa-S3 for TFs with binding profile data available publicly are shown in figures 3.2, 3.4, and 3.5 respectively.

In the GM12878 cell line, TFs segregate into two groups (Figure 3.2). TFs in each group tend to attract other members of the same group (green, top left, and bottom right of heatmap) but repel members of the other group (red, top-right, bottom-left). We call the top-right group “Group 1” and the bottom-right group “Group 2”.

A similar study, using a different methodology, was performed by Ma *et al.*[90] using Hi-C data of GM12878 cell line. Figure 3.3 compares the results of the two studies; with very few exceptions, our results are consistent with those of Ma *et al.*.

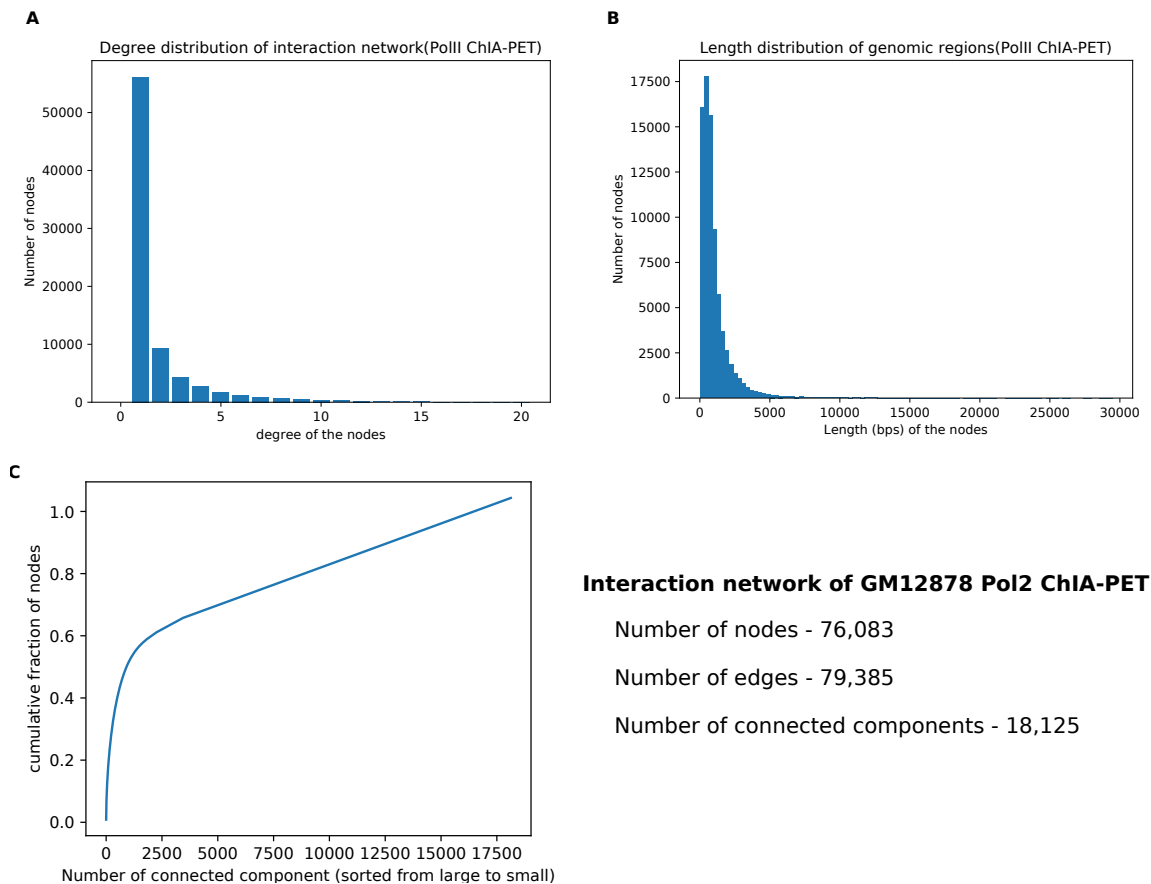


Figure 3.1: Properties of GM12878 Pol II ChIA-PET interaction network. (A) shows the degree distribution of nodes in the interaction network. (B) Distribution of the genomic lengths of nodes in base pairs (bps) shows most of regions are within 5kbp length. (C) shows the number of isolated components and cumulative fraction of nodes in the components (The isolated components are sorted in their node size from high to low).

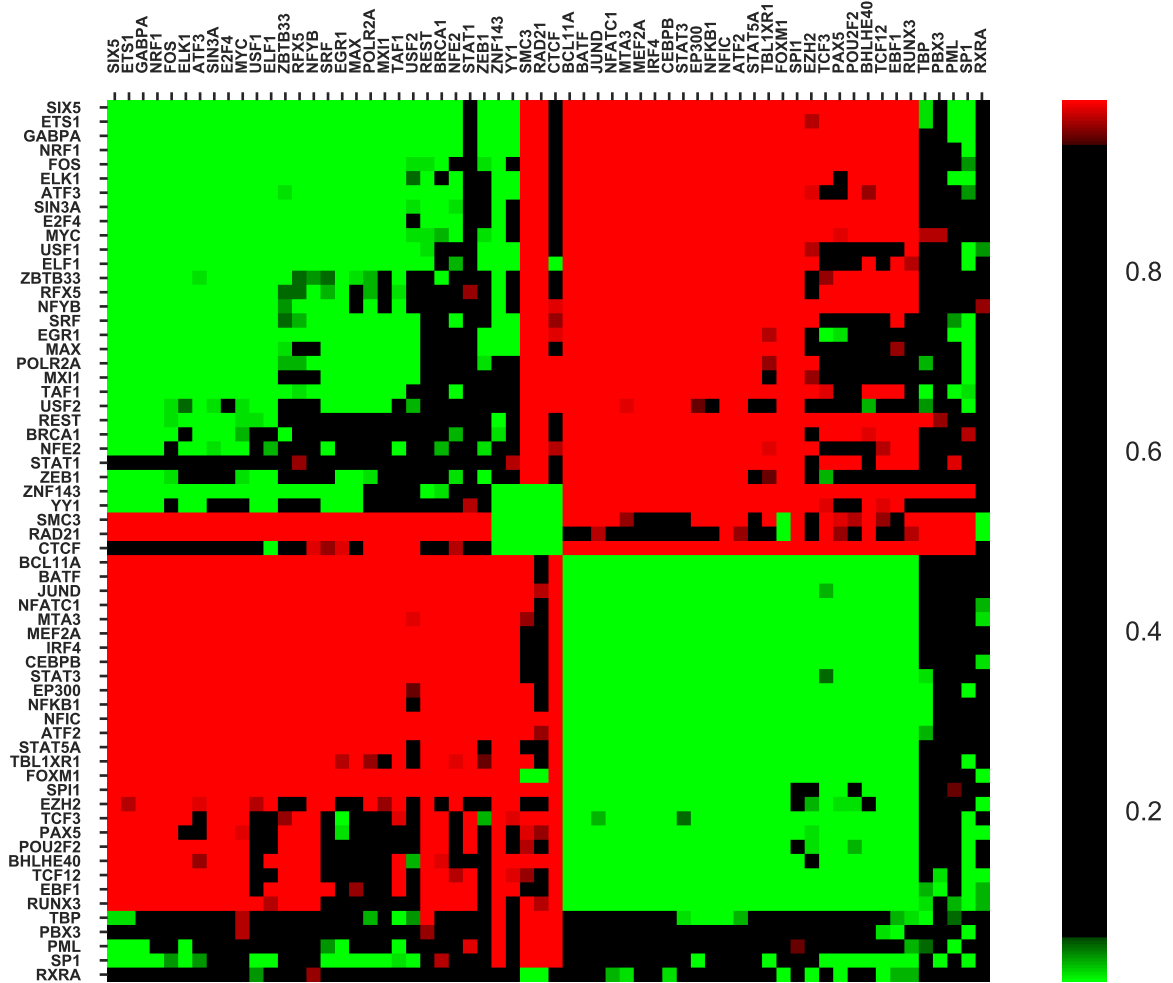


Figure 3.2: **Co-occurrence heatmap of GM12878 cell line.** The heatmap shows attracting and repelling TF pairs in green and red respectively as described in Methods. The q -value matrix is clustered simultaneously along rows and columns to identify TFs with similar pattern of co-occurrence with other factors.

We also performed TF co-occurrence analysis using our method on the same Hi-C data set (GEO GSE63525) used in the Ma *et al.* study. First, we identified significant long-range 3D chromatin interactions from this dataset using FitHiC2 tool[91]. Using these significant 3D chromatin interactions, we applied our method to identify pairs of TFs that either significantly attract or avoid each other. The resulting heatmap is in figure B.1. Once again we observe results are consistent with Ma *et al.* for the common factors between the two studies (figure B.2). However, differences exist between the co-occurrence pattern obtained using ChIA-PET and Hi-C datasets. We argue that the ChIA-PET data is more meaningful for the following reason: TFs bind at numerous loci on the genome independent of their interaction with promoters. However, functionally relevant binding of TFs, which directly regulates the expression of a target gene, must bind to a functional promoter or enhancer and physically interact with a promoter. ChIA-PET tagged with Pol-II therefore may show more functionally relevant regions. In addition, ChIA-pet data has higher resolution (1kb, versus 5kb for Hi-C).

In the K562 cell line, we do not observe two distinct classes of attracting TFs as in GM12878; instead, one attracting group emerges on the top-left, and most other TF pairs repel. However, as discussed further below, many of the characteristics of Group 1 from GM12878 are retained in the one attracting group in K562. We further discuss the differences between Group 1 and Group 2 in later sections.

In the HeLa-S3 cell line, we observe again two distinct groups of attracting TFs as in GM12878, but the number of factors in Group 2 is less compared to the number of factors in Group 1.

Notably, in the GM12878 cell line, CTCF and cohesin subunits SMC3 and RAD21 almost universally repel other TFs and mostly co-occur among themselves. In K562 and HeLa-S3 cohesin subunits, SMC3 and RAD21 attract each other. Cohesin and CTCF have previously been reported to co-occur in spatially proximal regions and this co-occurrence is believed to be critical for the formation of 3D chromatin loops[3, 92]. Pairs of CTCF

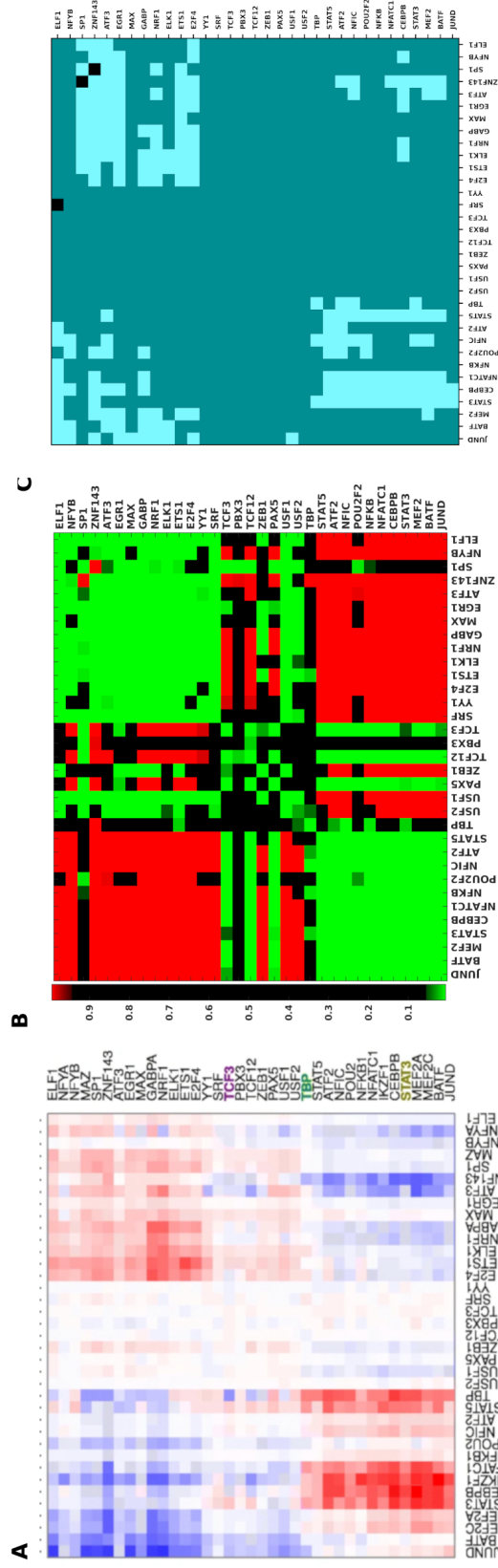


Figure 3.3: **Comparison of observations between our study and previous study by Ma *et al.*** (A) The heatmap from the earlier study by Ma *et al.* [90] ((reproduced under licence CC-BY-4.0)) showing attracting TF pairs in red and repelling TF pairs in blue. (B) The heatmap shows the attracting and repelling pairs in green and red respectively for common TFs used in both studies, using the method we proposed in methods section. (C) The heatmap shows a qualitative comparison of the two studies as follows: bright blue = both significant, in agreement; black = both significant, in disagreement (one showing attraction, the other repulsion); dark blue = one or both insignificant.

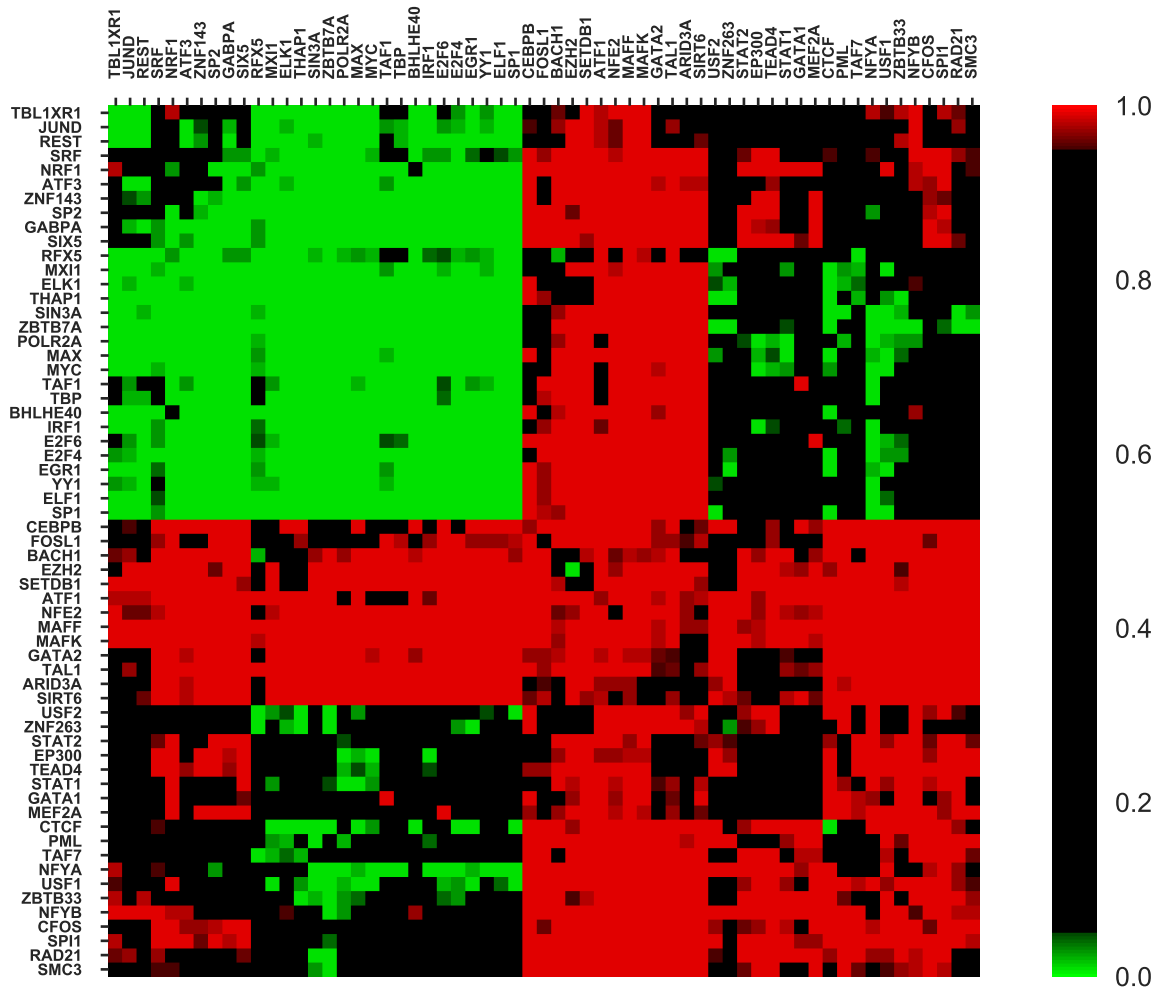


Figure 3.4: **Co-occurrence heatmap of K562 cell line.** Clustered q-value heatmap shows only group of TFs attract among themselves in K562 cell line.

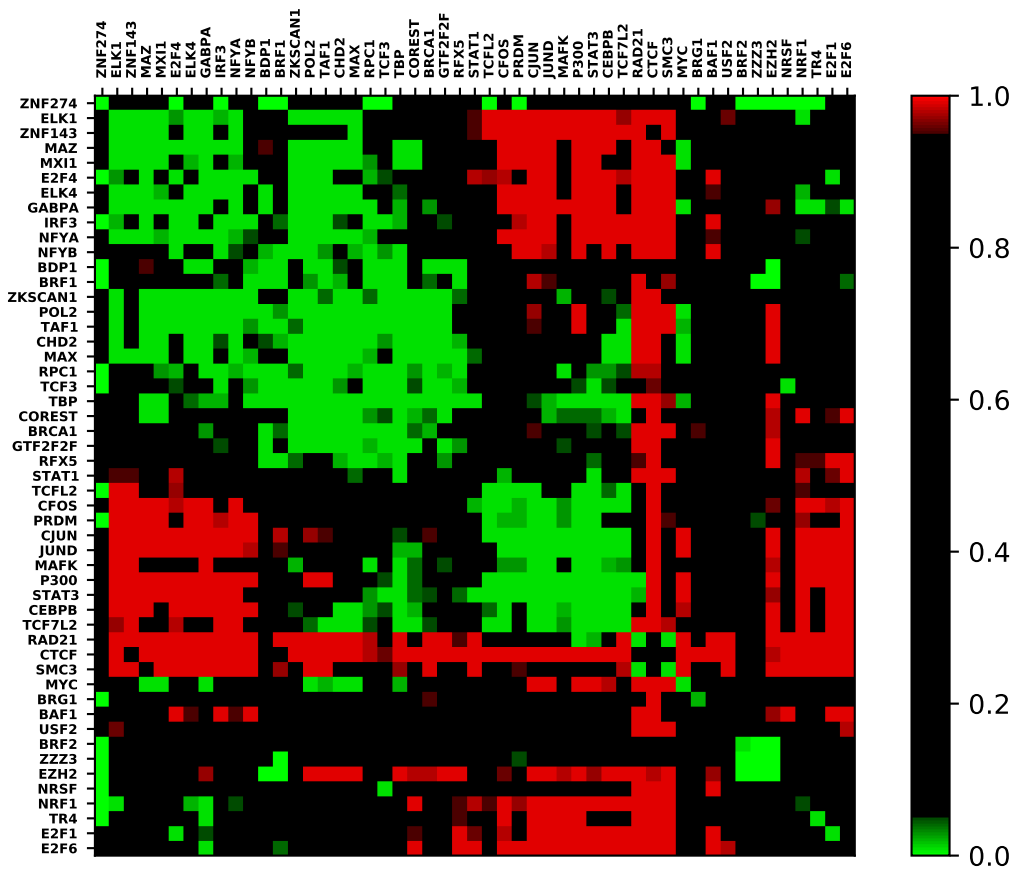


Figure 3.5: **Co-occurrence heatmap of HeLa-S3 cell line.** In HeLa-s3, once again two groups of TFs are identified, though the number of TFs in Group 2 is very less.

motifs are found in a divergent orientation near cohesin, and the chromatin “loop extrusion model” has been proposed for chromatin organization via “topologically associated domains” (TADs). In this model, a loop of DNA is pushed through a cohesin ring, until it is hindered by CTCF molecules bound at the motif sites. This has been recently observed *in vitro* [93]. Our observations support the proposed interplay of cohesin and CTCF, and further suggest that the binding regions of cohesin in particular tend not to be in close spatial proximity with either promoters or other distal regulatory regions where other TFs tend to bind. Figure 3.9, in the next section, suggests that CTCF and cohesin subunits, in addition to being in spatial proximity to each other, are also in close sequential proximity with each other in both GM12878 and K562 cell lines.

In addition to TFs, we also considered histone marks in our GM12878 analysis. The resulting attracting and repelling marks and TFs are shown in figure 3.6. We observe that histone marks associated with active promoters or enhancers i.e. H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K9ac significantly co-occur among themselves even in spatially proximal regions similar to what has been reported in various previous studies based on sequentially contiguous regions of the genome [94, 95, 96, 97, 98] and also co-occur with factors TBP, PML, and SP1 which are usually associated with the promoter regions. Also, these histone marks significantly repel their antagonistic marks H3K27me3 and H3K9me3.

Previous chromatin interaction studies have shown that the genome is primarily partitioned into distinct two compartments A and B, characterized by open and closed chromatin respectively [3, 77, 99, 100]. It would be interesting to know if these compartments have any relation, or more importantly if they contribute to the segregation of TFs that we observed in our investigation. To this end, we investigated the co-occurrence of TFs in the A and B compartments separately by classifying the spatial chromatin interactions falling into these compartments. But we do not observe any such correlation of genome compartments with clusters of TFs in terms of their co-occurrence. The TF pair co-occurrence

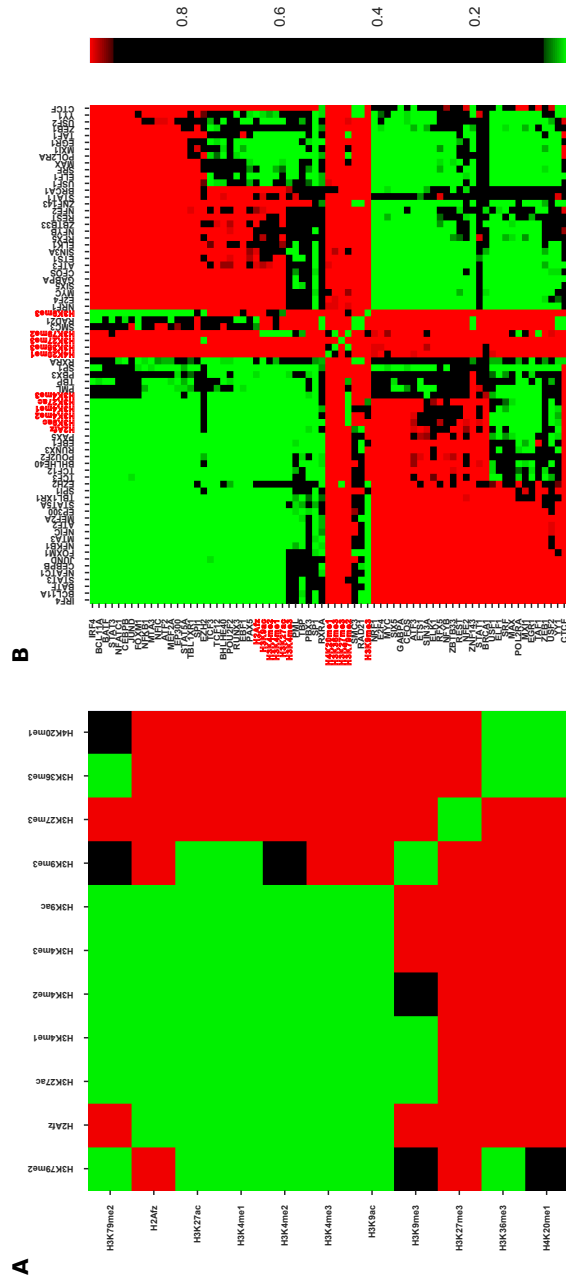


Figure 3.6: **Co-occurrence pattern of histone marks in GM12878**(A) Clustered q-value heatmap similar to figure 3.2 for various histone marks. (B) shows the co-occurrence pattern for all the TFs along with the histone marks (labelled in red).

behaviour is given in figure 3.7A and 3.7B for A and B compartments respectively. The co-occurrence behavior is similar in both compartments, but the numbers of significant attracting or avoiding TF pairs are fewer in the B compartment, as expected, it is generally inactive with a closed chromatin structure.

3.2 Sequential co-occurrence patterns largely mirror spatial co-occurrence

Prior to the advent of high-throughput chromatin interaction experiments such as Hi-C, co-occurrence of transcription factors was studied in sequentially contiguous regions such as promoters and enhancers [51, 52, 53, 54, 57, 79, 80, 81]. We asked whether the same patterns of attraction or repulsion occur within a contiguous regulatory sequence. We repeated the analysis using the same regions from the ChIA-PET data, and the same randomization method, but this time considering co-occurrence within rather than across interacting regions. Figure 3.8 show schematic diagram of two TFs co-localized on sequentially contiguous 1D region of the chromatin. It turns out that most TF pairs exhibit the same qualitative behavior as in spatially proximal regions; however, a few TF pairs show different qualitative behavior (figure 3.9). These include interactions across groups: in particular, factors of Group 1 (SRF, EGR1, MAX, MXI1, TAF1) attract factors of Group 2 (TCF3, PAX5, POU2F2, BHLHE40, TCF12), visible as a green block in figure 3.9A (middle). Their pairwise spatial interactions are insignificant but they show significant co-occurrence in sequentially contiguous regions.

In GM12878, factors such as TBP, PML, and SP1 which are usually enriched at the promoter regions attract almost all other TFs used in the study sequentially but do not show any significant behavior in spatially proximal regions, highlighting the promoter-specific binding of these factors. In the K562 cell line, as mentioned in the previous section, we observed only one main cluster of TFs attracting each other, but the other cluster of TFs,

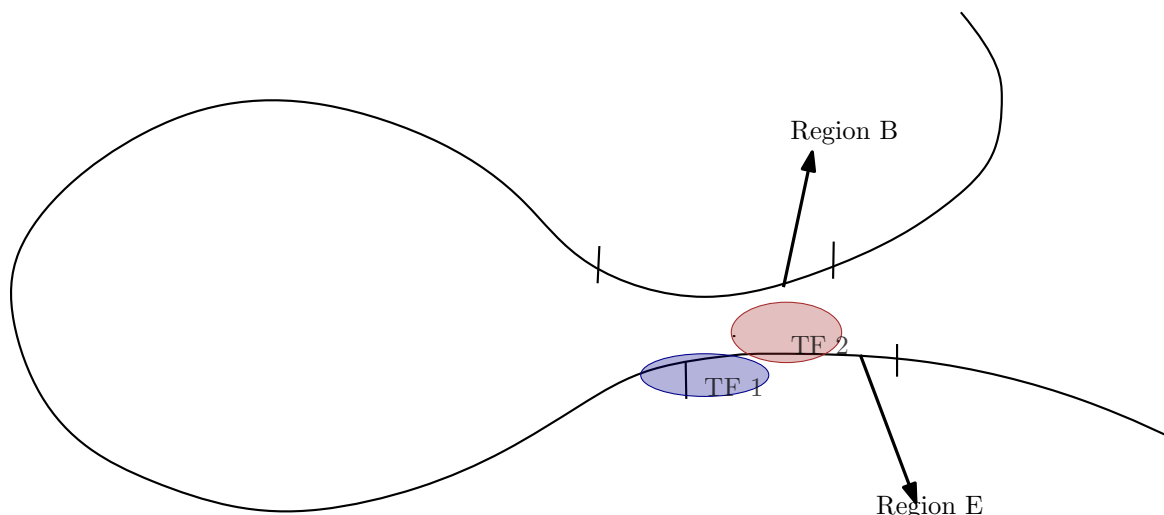


Figure 3.8: **Schematic diagram showing TF pair co-localization in 1D sequentially contiguous regions.** This cartoon illustrates one instance of two TFs 1 and 2 co-localizing in 1D linear genomic region E. The ChromTogether method is slightly modified to count all such instances of co-localization for a pair of TFs in linear genomic regions and significance is estimated comparing the similar co-localization count in randomized networks.

which do not attract spatially, do in many cases attract in sequentially contiguous regions. Overall, the analysis in spatial and sequential regions and comparison between them substantiate the evidence of two clusters of TFs present and suggest their potential role in the regulatory aspects.

3.3 Motif instances attract and repel similarly to TFBS

Transcription factors bind to DNA at regulatory regions of the genome in a sequence-specific manner, that is, they have a binding preference to a specific set of DNA sequences known as motif sequences. These sequences are usually short, that is 6-20 bp. These DNA sequences are represented traditionally by consensus sequences or by position weight matrices (PWM) also known as position-specific scoring matrices (PSSM). A PWM W is a $L \times 4$ matrix whose element $W_{i\alpha}$, where i is the position and α is a nucleotide, indicates the probability of finding the nucleotide α at position i of a putative binding site. These PWMs are conveniently visualized using sequence logos. Figure 3.10 show the sequence

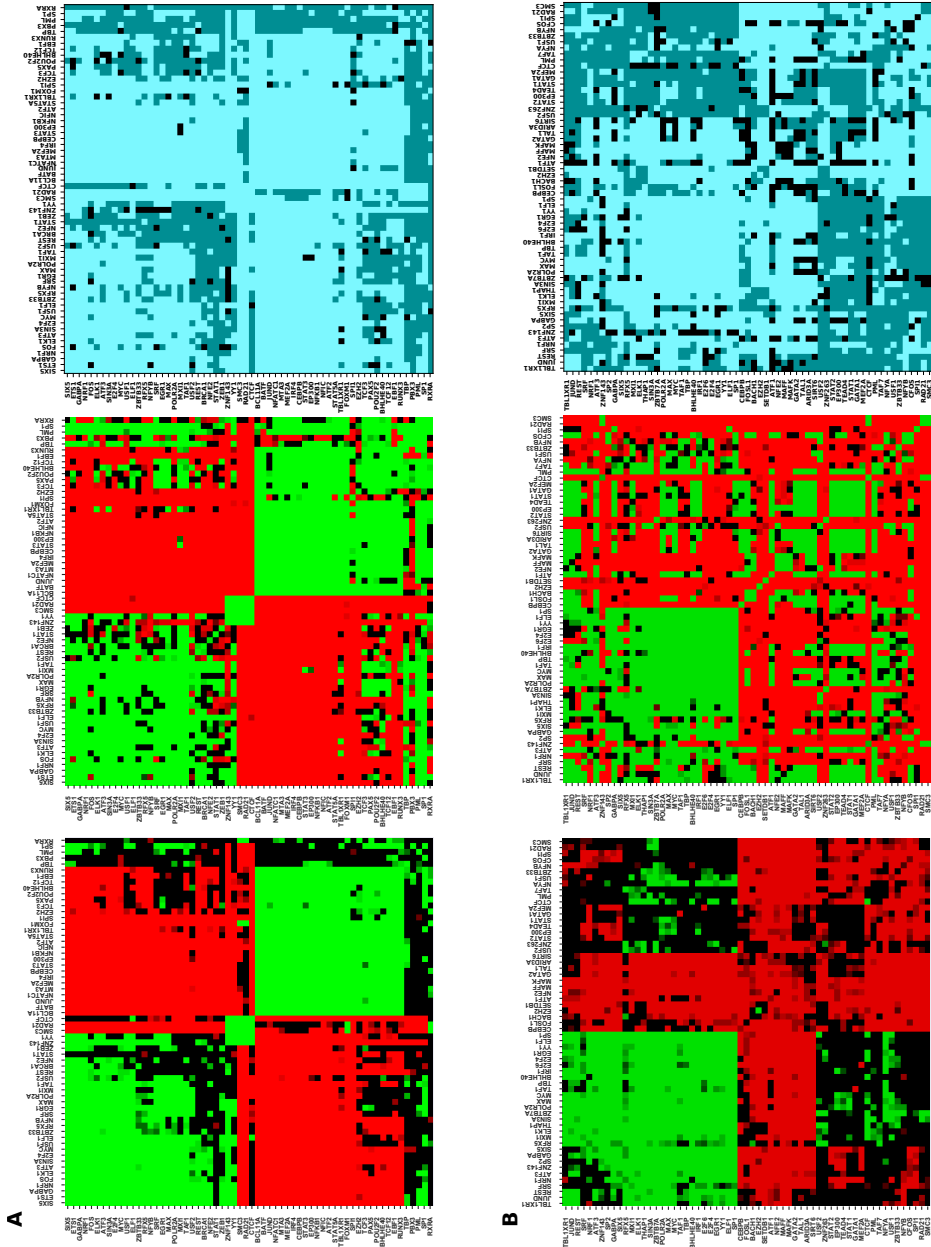


Figure 3.9: Comparison of TF pair co-occurrence on linear genome and in 3D genome. Top and bottom panels show the co-occurrence pattern in spatially proximal and sequentially contiguous regions for GM12878 and K562 cell lines respectively. The left heatmap in both panels corresponds to spatial co-occurrence, the middle to sequential co-occurrence, while the right heatmap is a comparison of these, coded as follows: both significant, in agreement: bright blue; in disagreement, both insignificant: black; one or both insignificant: dark blue.

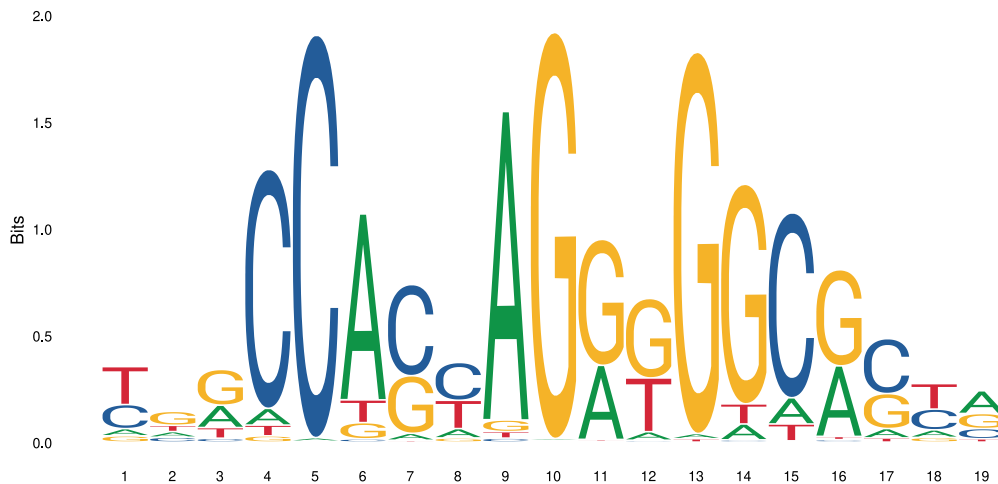


Figure 3.10: **CTCF sequence logo.** Sequence logos represents probability or frequency of each nucleotide at specific position of DNA. The height of nucleotide at each positions is given in terms of its information content.

logo for the factor CTCF. In the logo, at each position i , the total height of the column of letters is the information content at that column, defined as $I = 2 + \sum_{\alpha=A,C,G,T} W_{i\alpha} \log W_{i\alpha}$, which is maximum (2 bits) if only one nucleotide is possible, and minimum (0 bits) if all four nucleotides are equally probable. The relative heights are proportional to the probabilities of nucleotides at that position.

Traditionally, TF binding sequences are discovered *in vitro* through experimental approaches such as high throughput DNA microarrays, SELEX (Systematic Evolution of Ligands by Exponential Enrichment), and more recently, ChIP-seq. From these binding regions, individual motif instances are determined, which are statistically over-represented short sequences within these binding regions. These motif instances are aligned and position-specific nucleotide counts are used to construct a PWM.

Motif for TFS in various species, discovered over the years, are cataloged in databases such as JASPAR [101], TRANSFAC [102], HT-SELEX [103], Factorbook [104], UniPROBE [105], and CisBP [106]. Once a TF motif sequence is known, it can be used to scan and predict putative binding sites for the TF on the genome.

Our results above from the TF ChIP-seq binding data show considerable similarities and

differences across cell lines in the pattern of significant co-occurrence or avoidance in both spatially proximal and sequential contiguous regions. TFs generally bind to the genomic regions in a sequence-specific manner and these short specific sequences are represented as sequence motifs. Motif instances of a given TF are present at several locations of the genome, but only at a few of those sites does the TF bind in a given cell type depending on additional factors. We have so far considered binding events identified via ChIP-seq experiments. Here we ask whether a similar co-occurrence pattern occurs at a motif level, even though motif instances are generally poor indicators of tissue-specific TF binding.

Motif information is available for a large number of TFs beyond the ones for which ChIP-seq data is available. We can study motif co-occurrence in various cell lines. While motif matches (unlike ChIP-seq peaks) are independent of cell lines, the chromatin contact information is cell line-dependent. We predict motif instances with FIMO[107] using motifs for TFs from the JASPAR database [101] as bundled with the MEME suite[108], specifically the file `JASPAR2018_CORE_vertbrates_redundant.meme`. Rather than use all 719 motifs in that file, we identified similar motifs using TOMTOM[109] and clustered them into 93 groups, in each of which we picked the most informative motif as a representative. The selected TFs and their motif ids are given in supplementary material table B.1. The motif information score is calculated as : $I = \sum_i (2 + \sum_{\alpha=A,C,G,T} W_{i\alpha} \log_2 W_{i\alpha})$ where $W_{i\alpha}$ is the PWM probability for position i and nucleotide α .

Using these 93 distinct motifs, we observe strongly similar patterns of co-occurrence of motif sites in spatially proximal regions for four cell lines: GM12878(A), K562(B), MCF-7(C), and HeLa-S3(D), as shown in figure 3.11.

Motif co-occurrence in sequentially contiguous regions too is broadly similar to the spatially proximal regions, in all the four cell lines examined. This can be seen in figures B.3, B.4, B.6 and B.5 for GM12878, K562, MCF7 and HeLa-S3 cell lines respectively in appendix A. However, there are many more significant examples both of attraction and of repulsion in sequentially contiguous regions, suggesting perhaps a greater degree of

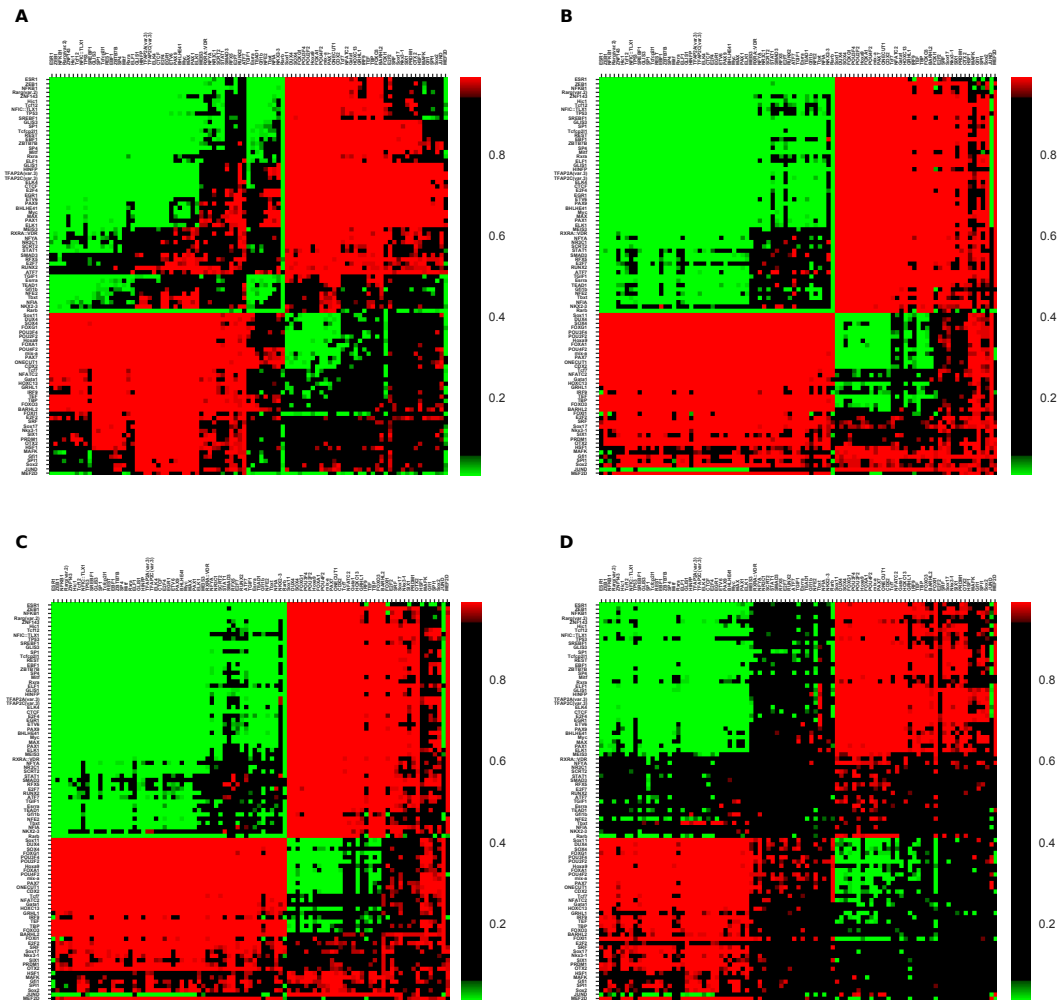


Figure 3.11: **Co-occurrence pattern of motif sites.** Comparison of motif co-occurrence in four cell lines: (A) GM12878, (B) K562, (C) MCF7 and (D) HeLa-S3. For ease of comparison, the order of factors from the clustering in GM12878 is used in all four subfigures

combinatorial control within promoters and enhancers.

3.4 A consensus TF-TF co-occurrence network

Using the co-occurrence patterns for binding events and motifs that we see in spatially and sequentially contiguous regions, we propose a consensus network for TF pair co-occurrence. We use both motif and experimental (ChIP-seq) binding events, and both spatially proximal and sequentially contiguous regions, giving us four datasets; and we require that TF pairs attract in at least two of these four datasets. The consensus network is shown in figure 3.12(A) and 3.12(B) for GM12878 and K562 cell lines respectively.

3.5 The two main TF groups interact differently with proteins, DNA, and genes

In GM12878 TFs separate into two groups that attract within a group, but repel across groups, with some exceptions. In K562 there is one group that mutually attracts (and largely mirrors Group 1 in GM12878), and another that repels most other TFs in our dataset. The two groups in each case exhibit differences in their interactions with other proteins, DNA binding locations, and downstream gene targets.

3.5.1 PPI interactions are enriched in intra-group TF pairs in GM12878

We examined physical interactions of TF-TF pairs using the Human Integrated Protein-Protein Interaction Reference (v2.2) database (HIPPIE) [110]. We observed enrichment of physical interaction among attracting TF pairs (92 out of 539 i.e. ~ 0.17) compared to avoiding TF pairs (40 out of 523 i.e. ~ 0.07) indicating that attracting TF pairs are significantly more likely to interact physically than avoiding pairs ($p = 1.98 \times 10^{-6}$,

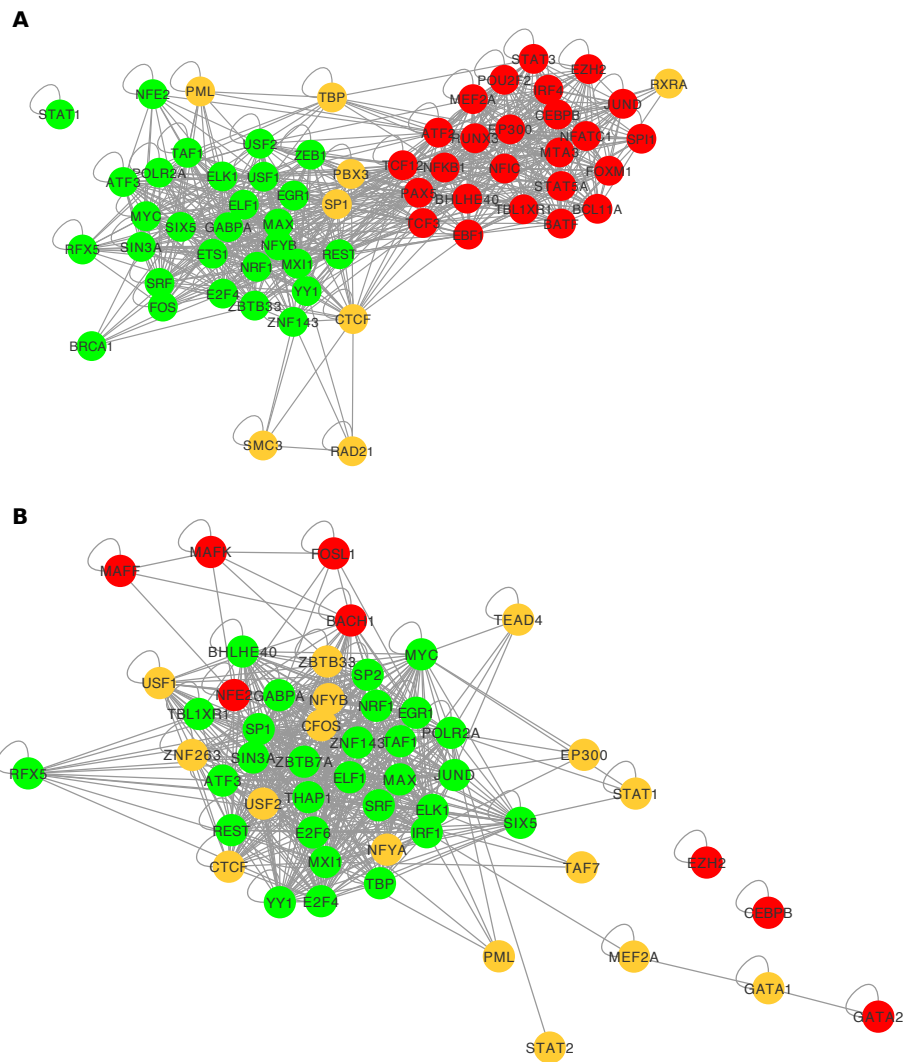


Figure 3.12: **Consensus TF interaction networks.** High confidence consensus networks built using the TF pair co-occurrence observations from all four methods (spatial or sequential, binding or motif instance) for (A) GM12878 and (B) K562 cell lines.

hypergeometric test). There is no significant difference between Group 1 TF pairs and Group 2 TF pairs (interacting fraction ~ 0.15 and ~ 0.18 respectively).

3.5.2 Domain-domain interactions are enriched in attracting TF pairs

To further examine the possibility of undocumented protein-protein interactions, we considered the domain structures of the TFs and looked for possible domain-domain interactions among all TF pairs present in the study, using the database of three-dimensional interacting domains (3did) [111]. Any pair of TFs containing respective interacting domains were considered “potentially interacting”. Attracting TF pairs in both cell lines show significant enrichment over avoiding TF pairs for potential physical interaction (tables 3.1 and 3.2).

	Attracting TF pairs	Avoiding TF pairs	Total
potential physical interactions	131	61	192
No potential physical interactions	598	718	1316
Total	729	779	1508

Table 3.1: **Domain-domain interactions are enriched among attracting TF pairs in GM12878.** The number of possible physical interactions from the literature of domain-domain physical interactions among attracting and avoiding TF pairs for GM12878 cell line. Attracting pairs are significantly more likely to have possible domain-domain interactions ($p = 2.17 \times 10^{-9}$, hypergeometric test)

	Attracting TF pairs	Avoiding TF pairs	Total
potential physical interactions	79	24	103
No potential physical interactions	369	517	886
Total	448	541	989

Table 3.2: **Domain-domain interactions are enriched among attracting TF pairs in K562.** Similar to table 3.1, for K562 cell line. Attracting pairs are significantly more likely to have possible domain-domain interactions ($p = 7.65 \times 10^{-12}$, hypergeometric test)

3.5.3 Internal nodes in PPI pathways largely differ in Group 1 and Group 2 for GM12878

We next looked at TFs that may be binding indirectly, via co-factors. We considered the shortest PPI pathway for each pair of TFs. Specifically, we asked what TFs (not necessarily in our list of 62) occur in the shortest interaction pathway between that pair of TFs, and evaluated the frequency of occurrence of each such “internal node”. We only considered annotated TFs, and not other proteins (figure 3.13). Interestingly, the frequent internal nodes are largely unique for each group. Moreover, the few TFs common to these groups tend to occur as internal nodes of internal-group TF pairs, suggesting possible roles as co-activators or co-repressors. Taken together, this analysis suggests that the two groups have distinct structural underpinnings.

3.5.4 Group 1 TFs bind closer to promoters than Group 2 TFs

Figure 3.14 plots the cumulative distribution of ChIP-seq peak distance from the nearest transcription start site (TSS) for Group 1, Group 2 and ungrouped TFs, in (A) GM12878 and (b) K562. Group 1 TFs tend to bind closer to the TSS than Group 2 or ungrouped TFs. This is consistent with what Ma *et al*[90] report.

3.5.5 Group 1 TFs bind to GC rich regions while Group 2 TFs to GC poor regions

Previous studies have shown different classes of regulatory elements with different compositions of nucleotides. For example, promoters are classified into GC-rich sequences (containing a high fraction of guanine (G) and cytosine (C) nucleotides) or AT-rich sequences (containing a high fraction of adenine (A) and thymine (T) nucleotides). They further suggested a relationship between these different promoters associated with differ-

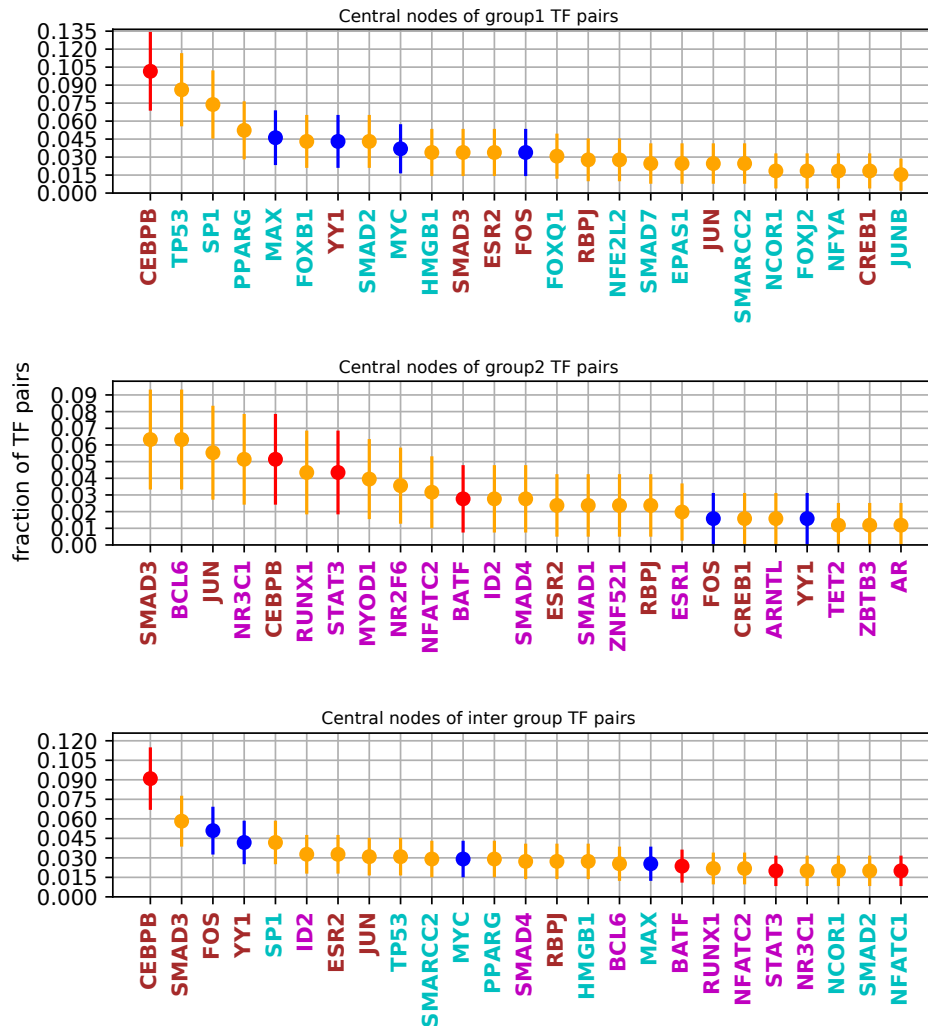


Figure 3.13: **Internal nodes of PPI network of TFs.** The TFs most frequently occurring as internal nodes in the shortest PPI pathway between TF pairs, for pairs from Group 1 (top), Group 2 (middle), and cross-group (bottom panel). The blue, red markers are Group 1, Group 2 TFs respectively and yellow are TFs not present in our co-occurrence data. TFs in cyan text occur as central nodes within Group 1 only, in magenta are central nodes within Group 2 only, and brown occur as central nodes in both groups.

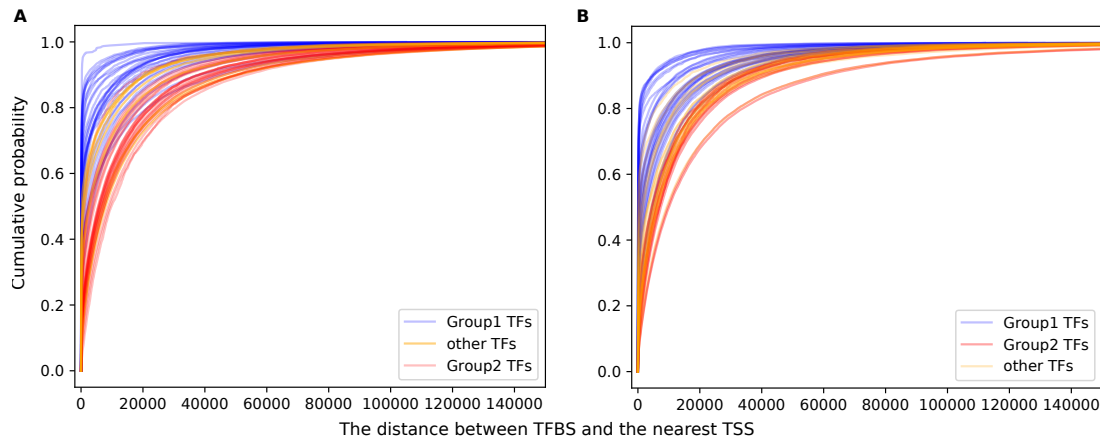


Figure 3.14: Cumulative plot of distance between TF binding sites and their nearest TSSs. (A) The distance between TFBS and the nearest TSS, and the cumulative probability for sites to occur within that distance, for Group 1 TFs (blue), Group 2 TFs (red) and other TFs (yellow), in (A) GM12878 and (B) K562.

ent classes of genes. Recent studies have also suggested enhancers regulate promoters with similar GC composition [112, 113]. We asked whether there are any such differences in nucleotide compositions of regions in which the two groups of TFs bind on the genome, to see if these two classes of TFs emerge as a consequence of TF preference for DNA sequences with different nucleotide compositions. To this end, we have looked at the GC composition of ChIP-seq peaks of each TF. We observe that the two classes of TFs show a clear difference in the GC composition of their binding sequence. Figures 3.15 and 3.16 show average GC content of ChIP-seq peak regions for each of the factors considered in the study for GM12878 and K562 cell lines respectively. Since the two groups of TFs are observed on 3D long-range interactions as well as on linear genomic elements, our observation further supports the idea that regulatory elements with similar nucleotide composition interact physically and have common regulatory mechanisms.

We also further looked at the distribution of GC content of individual Pol II ChIA-PET regions and compared it with the distribution of regions bound by a given TF or with TF motif instance. Figure 3.17 shows distribution of GC fraction of Pol II ChIA-PET regions with motif instance for each factor considered in the study for GM12878 cell line. Once again, the two groups of factors clearly show differences in GC composition. Similar

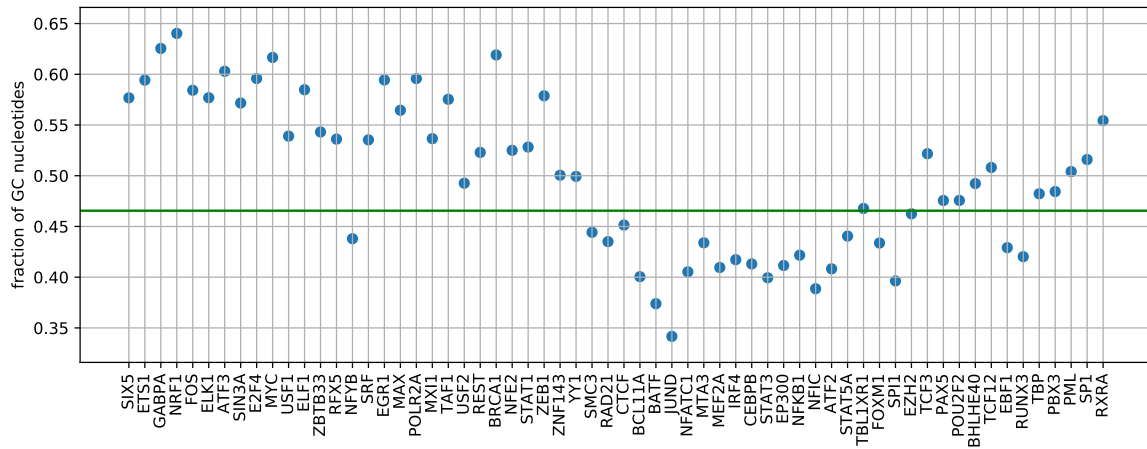


Figure 3.15: **GC content of ChIP-seq peaks regions in GM12878 cell line.** Plot shows the average GC content of ChIP-seq peaks regions for each of the factor considered in the study of GM12878 cell line. The order of the TFs arranged is same as shown in clustered q-value heatmap 3.2. The green shows the median value of GC content of all the factor ChIP-seq regions. The average GC values of Group 1 TFs are higher the median value, whereas for Group 2 TFs it is lower than the median value.

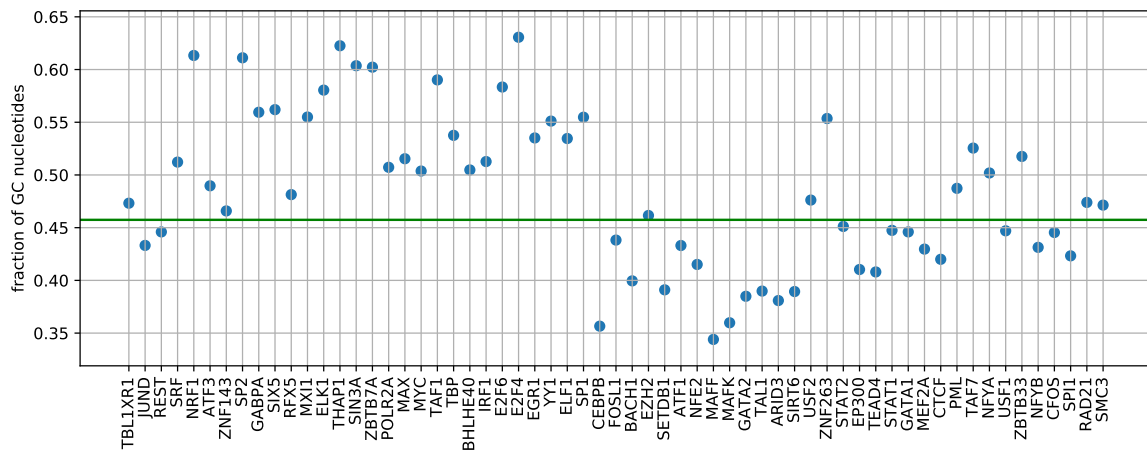


Figure 3.16: **GC content of ChIP-seq peaks regions in K562 cell line.** Plot shows the average GC content of ChIP-seq peaks regions for each of the factor considered in the study of K562 cell line. The order of the TFs arranged is same as shown in clustered q-value heatmap 3.4. The green shows the median value of GC content of all the factor ChIP-seq regions. The average GC values of Group 1 TFs are higher the median value, whereas for Group 2 TFs it is lower than the median value.

plots for other cell lines K562, HeLa-S3, and MCF7 are given in figures [B.9](#), [B.10](#), [B.11](#) respectively in appendix [A](#).

3.5.6 Group 1 target genes are enriched for housekeeping functions, Group 2 for tissue-specific functions

We identify putative target genes of each TF as genes located either within 2kbp sequentially of a TFBS for that factor, or on a spatially proximal region to that TFBS. Within each group, We consider the functions of target genes, using Gene Ontology (GO) functional enrichment analysis and disease trait enrichment analysis.

Based on the distributions of putative TF regulators per gene (shown in figure [B.12](#) in appendix [A](#)), genes that are targets for at least five TFs from one group, and at most two TFs from the other group, are taken to be target genes for the former group. We perform GO term enrichment analysis for biological processes, and molecular functions using DAVID [[114](#)] with a background of all human genes. We find that Group 1 target genes are enriched for housekeeping functions such as transcription, cell cycle, transport, and metabolism, while Group 2 target genes are enriched for immune/inflammatory response (figure [3.18A](#)).

We specifically assessed whether Group 1 target genes are enriched for housekeeping genes using a database of known housekeeping genes [[115](#)]. Group 1 target genes are significantly enriched over Group 2 target genes for housekeeping genes (hypergeometric test p -value 2.7×10^{-9}). The enrichment is not highly sensitive to the target gene selection criteria (for example, requiring 3 TFs from one group and 0 from the other yields similar results).

The biological process enrichment is corroborated by molecular function enrichment. The molecular functions of Group 1 target gene involve mostly DNA binding, metal ion binding events, while Group 2 genes involves in receptor activity of chemokines and cytokines,

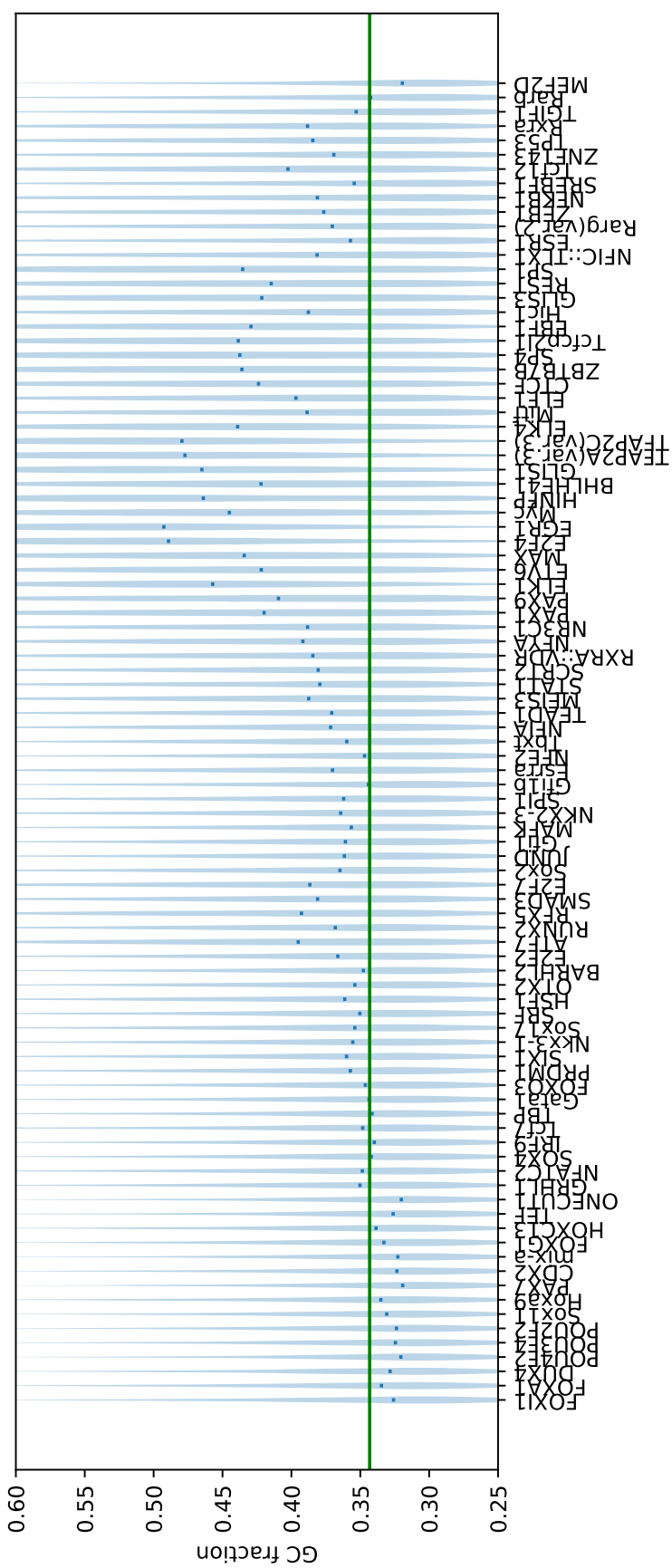


Figure 3.17: **GC content of ChIA-PET regions of GM12878 cell.** The violin plot of each factor show the distribution of GC fraction values of ChIA-PET regions with presence of TF motif. The green line shows the median GC fraction value for the set of all ChIA-PET regions.

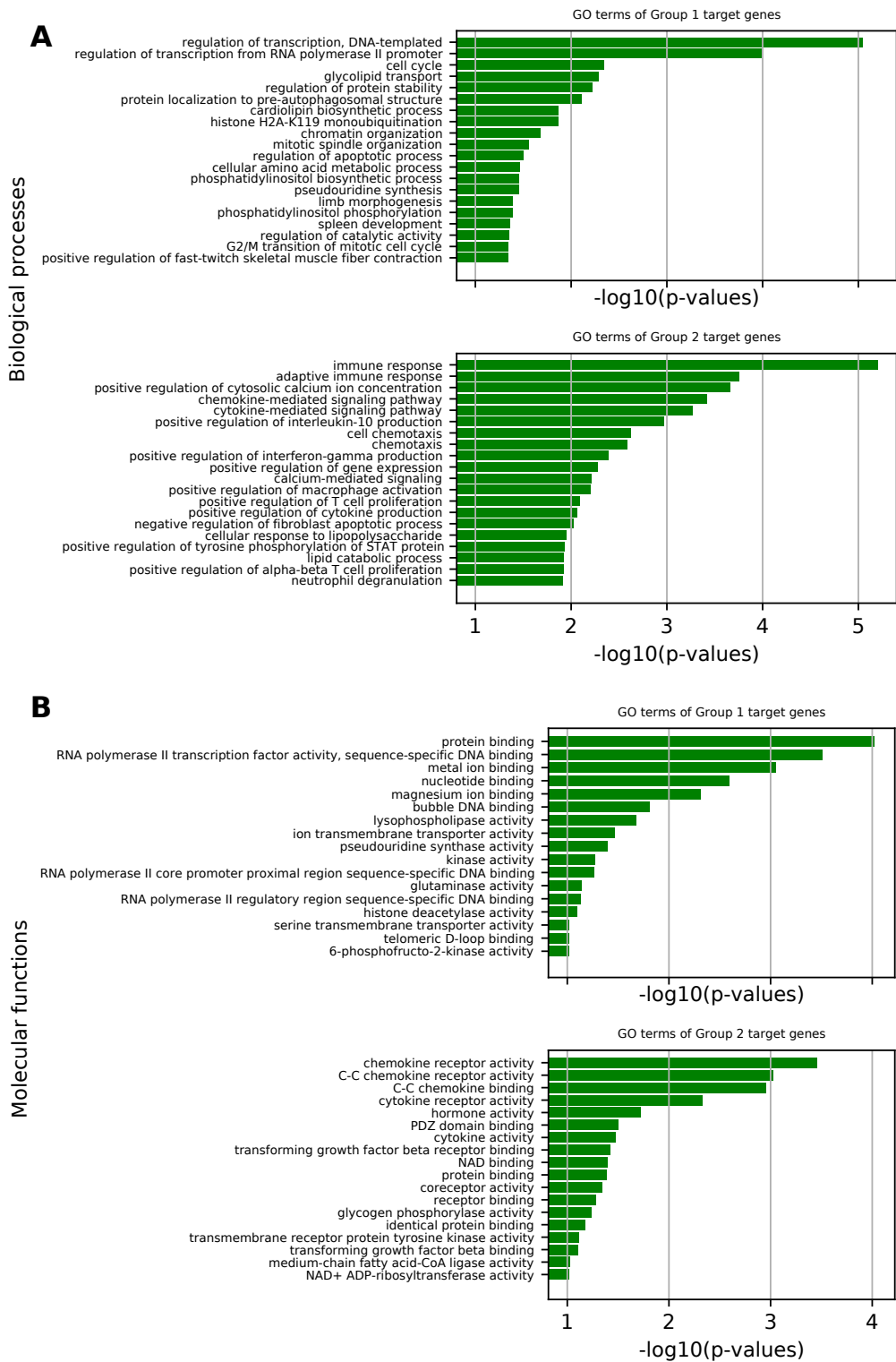


Figure 3.18: **GO enrichment analysis of target genes of Group 1 and Group 2 TFs (GM12878).** (A) shows the enriched biological process GO terms for the Group 1 and Group 2 target genes. (B) Similarly, enriched GO molecular function terms are shown for Group 1 and Group 2 target genes.

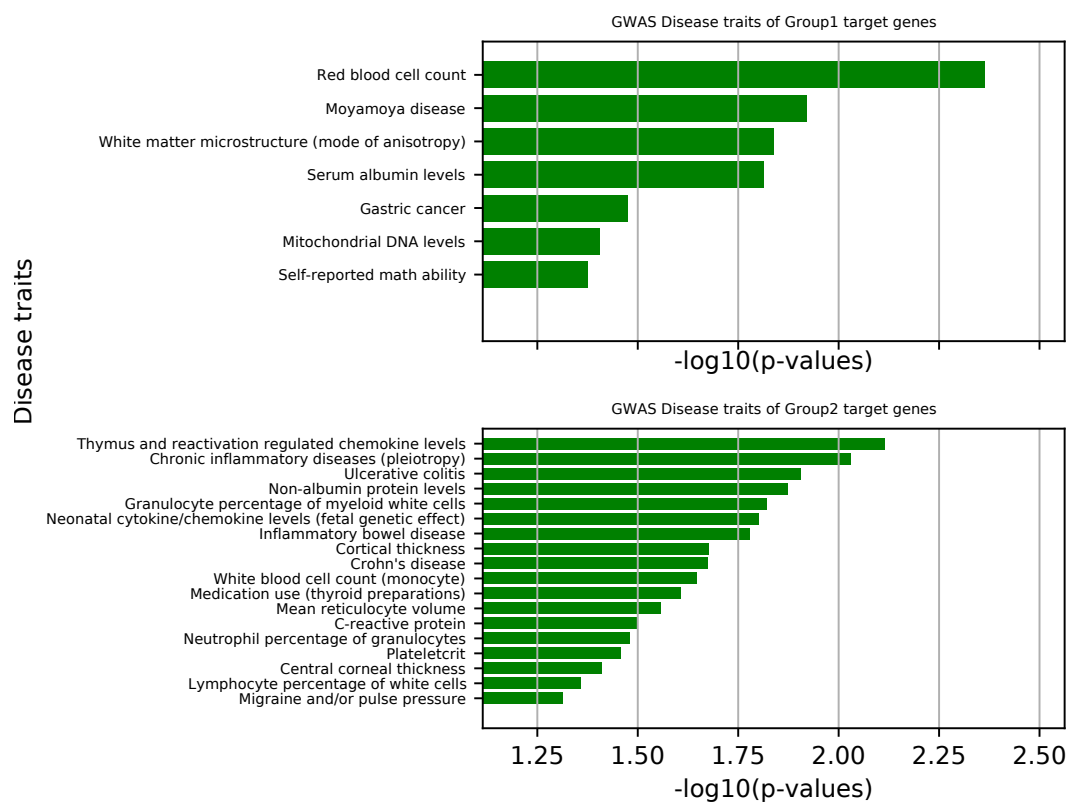


Figure 3.19: **Enriched GWAS disease terms for target genes of Group 1 and Group 2 TFs (GM12878).** The plot shows enriched GWAS disease or phenotype terms for target genes identified for Group 1 and Group 2 TFs of GM12878 cell line.

involved in immune responses (figure 3.18B).

Next, we assessed the group-specific target genes for association with disease traits from the GWAS catalog [116], only considering the traits associated with at least 5 genes. The target genes of each group were examined for significant overlap with these traits (Fisher exact test $p < 0.05$). Consistently, Group 2 targets were enriched for inflammatory disease, immunological diseases involving lymphocyte, leukocytes, and neutrophils, while Group 1 genes are not particularly enriched for any particular diseases (figure 3.19). Enrichment of immune response and traits in Group 2 targets is especially interesting considering that GM12878 is a lymphocytic cell line.

We performed a similar analysis for K562 cell line treating the large group of mutually attracting TFs as Group 1 and the remainder as Group 2. Again, Group 1 target genes are enriched for housekeeping functions such as transcription, metabolism, and transport (hypergeometric test p -value 1.4×10^{-3}). Group 2 target genes are not particularly enriched for pathway or functions (figure B.14 in appendix A).

Finally, in the HeLa-S3 cell line we see a similar segregation into two groups as in GM12878, but Group 1 is much larger than Group 2 (figure 3.5). Again, Group 1 is significantly enriched for housekeeping functions ($p = 4.9 \times 10^{-17}$) (figure B.15). Group 2 is enriched for protein ubiquitination and various metabolic processes, but with little overlap to Group 2 of GM12878.

Both the similarity and differences across the cell lines in terms of TF groupings and their target functions are worth noting. GM12878 is a lymphoblastoid cell line derived from the blood of a healthy female donor, K562 is lymphoblasts isolated from the bone marrow of a chronic myelogenous leukemia patient, and HeLa-S3 is a cervix carcinoma cell line. While Group 1 TF targets in all three cell lines are enriched for house keeping functions, Group 2 of TFs is most prominent in GM12878, derived from a healthy donor, and their targets are enriched for immune function. In contrast, leukemia-derived K562 lacks a functionally coherent Group 2 TFs, which may suggest a loss of lineage-specification in

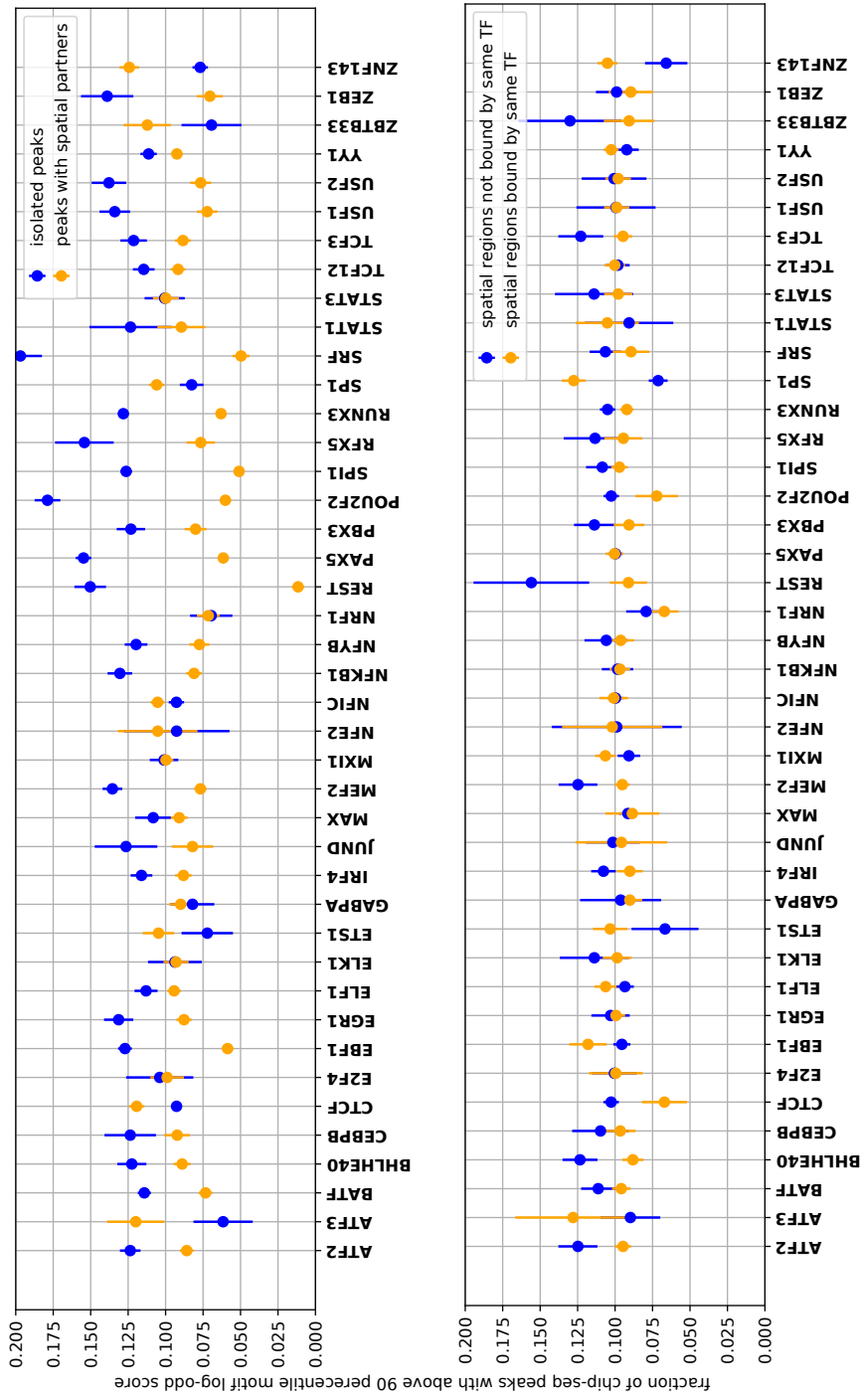
leukemic lymphoblasts. HeLa-S3 has a smaller Group 2 than GM12878 with a divergent set of, potentially lineage-specific, target functions.

3.6 Motif strength correlates with spatial interaction

Previous studies have demonstrated the significance of long-range spatially proximal regions in aiding *in vivo* TF binding at weaker motif sites by containing several homotypic motif sites in spatially clustered genomic regions [90, 117]. We extend the investigation to spatial interactions with other genomic regions, regardless of whether there are TFs binding to those other regions.

For each TF we consider the most informative PWM available in JASPAR[101]. Individual binding sites may have stronger or weaker matches to the PWM, which is an estimate of their binding affinity to the TF. We use the log likelihood ratio (LLR) as the “motif score” for a site, a measure of the likely “strength” of binding at that site. Given a sequence S of length L and a PWM W of the same length, the $LLR = \log(P(S|W)/P(S|B)) = \sum_{i=1}^L \log W_{iS_i} - \log B_{S_i}$ where B is a background model for the sequence (probability 0.25 per nucleotide). For each TF, we consider peaks containing sites whose motif scores are among among the top 10% as “peaks with strong motifs”. We then compare the fraction of peaks with strong motifs between two groups of peaks—those that are isolated (occurring in regions without any spatial chromatin interactions) and those in spatially interacting regions.

We observe for almost all factors, spatially isolated TFBS exhibit a significantly greater fraction of strong motifs (Figure 3.20A). If we consider whether the spatially interacting region is bound by the same TF or is not (but possibly bound by another TF), in the majority of cases binding by the same TF is associated with a weaker motif, but this is highly variable (Figure 3.20A). A similar trend is observed in K562 cell line (figure 3.21). This suggests that spatial interactions enable binding by TFs to weaker motifs, even in the



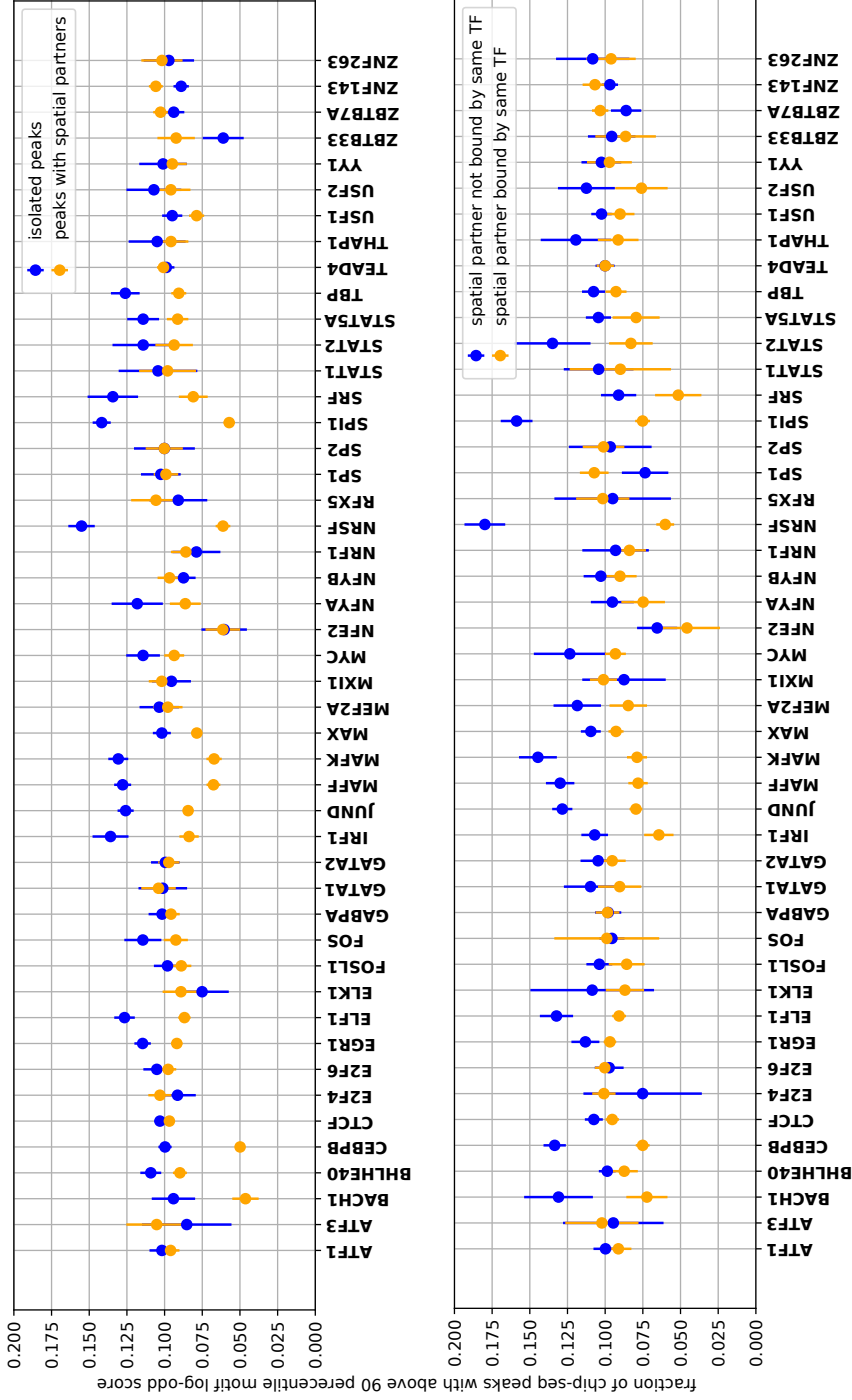
Transcription factors

Figure 3.20: **Spatial interactions facilitate *in vivo* TF binding at weaker motif sites (GM12878).** Spatially isolated ChIP-seq peaks mostly have a higher fraction of strong motifs than peaks within spatial clusters (top panel). For peaks within spatial clusters, when the same TF binds to a spatially-interacting region, the associated motif is weaker in 24 out of 42 cases (bottom panel).

absence of homotypic binding or the presence of heterotypic binding as was suggested previously.

3.7 Discussion

Our method ChromTogether recovers previously reported results of Ma *et al.*[90] in the GM12878 cell line, in both the genome-wide Hi-C data that they used and in ChIA-PET Pol-II data that is more specific to promoter-enhancer interactions. Our work thus serves as an independent validation, via a different method, of their results. We study a larger number of transcription factors, a greater number of cell lines, and sequential co-occurrence, as well as motif co-occurrence. We uncover numerous functional implications of the association of TFs into attracting and repelling groups. Attractive TF pairs exhibit significantly more physical interactions suggesting an underlying mechanism. The two TF groups differ significantly in their genomic and network properties, as well in their function—while one group regulates housekeeping function, the other potentially regulates lineage-specific functions, that are disrupted in cancer. Weaker binding sites tend to occur in spatially interacting regions of the genome. Our results provide novel insights into how chromatin regulates the genome, and suggests a complex pattern of spatial cooperativity of TFs that has evolved with the genome to support housekeeping and lineage-specific functions.



Transcription factors

Figure 3.21: Spatial interactions facilitate *in vivo* TF binding at weaker motif sites (K562). Similar to GM12878 cell line, Spatially isolated ChIP-seq peaks mostly have a higher fraction of strong motifs than peaks within spatial clusters in K562 cell line.

Chapter 4

Characterization of centromeres of

C. auris and related species

In eukaryotes, DNA is tightly packaged into structures called chromosomes with the help of proteins called histones. These tightly packed chromosomes can be visualized under the microscope when a cell undergoes cell division. Chromosomes play an essential role in various cellular activities such as cell division, heredity, mutation, and repair. They contain various functional parts such as centromeres and telomeres. Centromeres play an important role in the faithful segregation of chromosomes into the daughter cells during cell division. Telomeres are present at the end of the chromosomes, protect the end of the chromosome from digestion by nuclease enzymes, and prevent random chromosomal fusions.

Centromere function is highly conserved across all eukaryotic organisms. It serves as the site for the assembly of the kinetochore structure, which is linked to microtubules emanating from spindle fibers during cell division, aiding faithful segregation of duplicated chromosomes into new daughter nuclei. These regions are also known to facilitate chromosomal rearrangements and serve to generate diversity, especially in asexual organisms thereby contributing to the evolution of new karyotypes and species. Though centromere

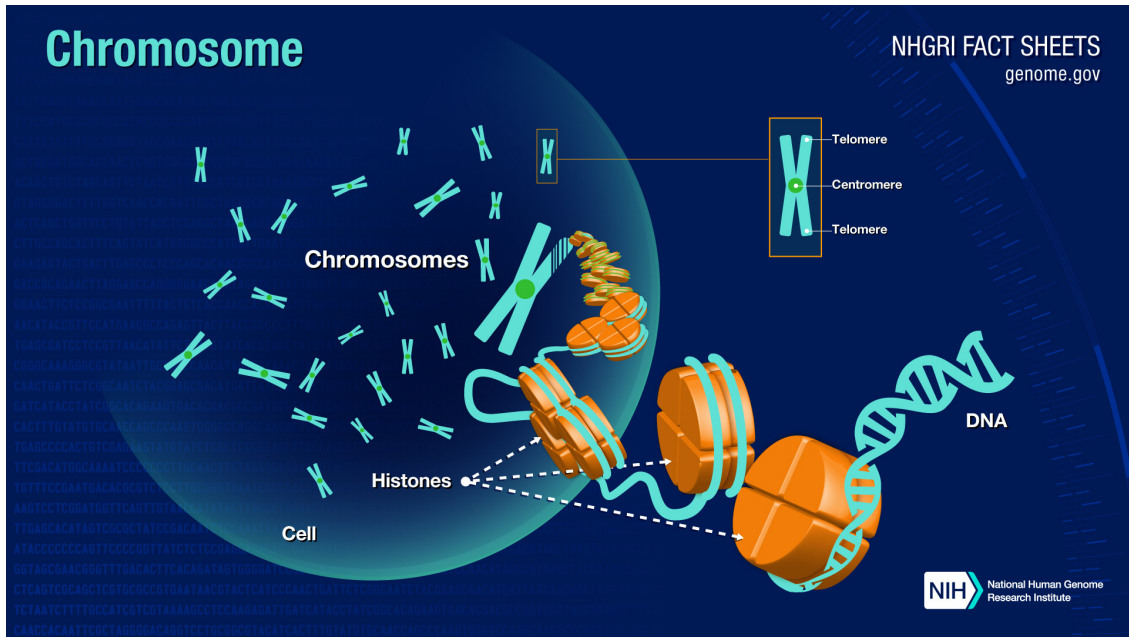


Figure 4.1: **Structure of chromosome.** The Image shows the structure of chromosome visible during the cell division. The DNA present in eukaryotes usually is packed into multiple linear chromosomes. (Courtesy: National Human Genome Research Institute, <https://www.genome.gov/about-genomics/fact-sheets/Chromosomes-Fact-Sheet>)

function is highly conserved across eukaryotic organisms, they exhibit greater diversity in their sequence and types: for example, they exhibit diversity in length of centromere region, repeat/transposon content, and GC-richness. In particular, budding yeast *S. cerevisiae* exhibits small “point centromeres” of $\approx 125\text{bp}$ length with well-defined repeat features; multicellular eukaryotes have centromeres tens of thousands of base pairs long; yeasts such as *C. albicans* have centromeres of intermediate length with no discernible identifying sequence features.

Candida auris is a rapidly emerging multi-drug resistant fungal pathogen which is causing systemic infections worldwide and posing threats to patients with other clinical conditions like diabetes mellitus, chronic renal disease, and, more recently, COVID-19 infections [118, 119, 120, 121, 122, 123, 124]. It has evolved simultaneously into different geological clades—South Asian (clade 1), East Asian (clade 2), South African (clade 3), South American (clade 4), and a potential fifth clade from Iran. These clades are separated by tens of thousands of single nucleotide polymorphisms, suggesting rapid evolution

modes in a fungal pathogen. Chromosomal rearrangements and aneuploidy are known to enhance drug resistance and virulence in primarily asexual fungi. Centromeres are susceptible to breaks in other fungal species and are likely to contribute to the diversity and rapid emergence of new clades of *C. auris*. This study, led by the Sanyal lab at JNCASR, focused on the identification, and characterization of centromeres and their role in karyotype diversification of *C. auris* and related fungal species through comparative genome analysis to understand the underlying mechanism/events for the rapid emergence of *C. auris*. As a part of the study, I have done the bioinformatic analysis to characterize various properties of centromeres in *C. auris* and related fungal species.

4.1 Phylogeny of *C. auris* and related species

We inferred the phylogeny among the *C. auris* and closely related fungal species given in table 4.1. For species that do not have any available functional annotation, the genomes were annotated using MAKER(2.31.10) [125] directly from homology evidence of closely related species RNA sequences and protein evidence. Clusters of orthologous proteins were identified using OrthoMCL v2.0.9 [126] for the above fungal species. Further, single copy ortholog proteins present in all species from the above-identified clusters are used for phylogenetic tree construction. The single-copy ortholog proteins are aligned using Clustal Omega [127]. All the alignments for all genes were concatenated including gaps for each species, and then columns containing gaps were removed. The resulting sequences were used for the phylogenetic tree construction. The tree was constructed using MrBayes(v3.2.5) [128] and drawn using FigTree (v1.4.4).

The phylogeny analysis revealed *C. auris* and its clades are phylogenetically similar to other multi-drug resistant, pathogenic species *C. haemulonii*, *C. heveicola*, *C. pseudo-haemulonii*, and *C. duobushaemulonii* which each have seven chromosomes and similar genomic size as *C. auris*. These pathogenic species are part of the Clavispora/Candida

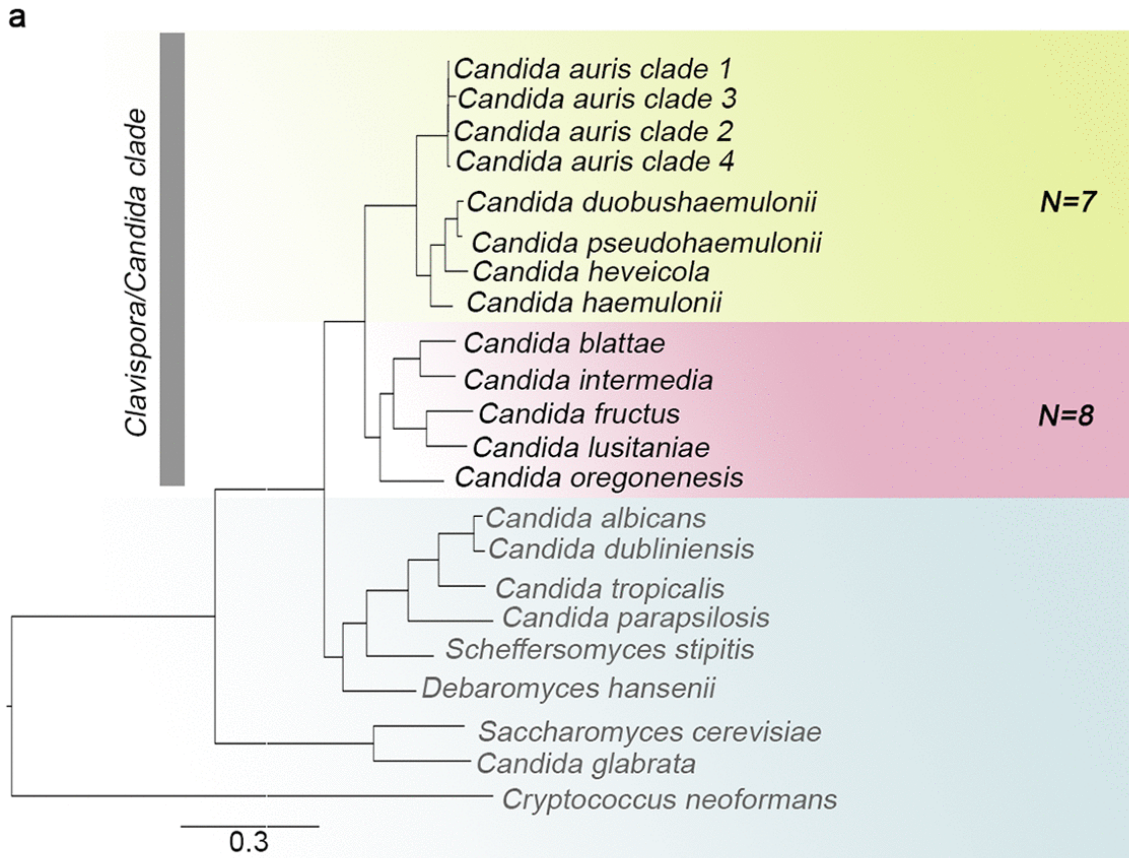


Figure 4.2: **Phylogenetic tree of *C. auris* and related species.** Phylogenetic tree depicting the relatedness of *C. auris* geographical clades and other member species of the Clavispora/Candida clade. Other species in Ascomycota with characterized/predicted centromeres are shown.

clade which includes pathogens *C. fructus* and *C. lusitaniae* which have eight chromosomes each.

4.2 *C. auris* poses small GC-poor and repeat-free centromeres

Genome sequence analysis of the centromeres of *C. auris* shows that the seven centromeres present are unique in their sequences and lack the presence of any DNA sequence motif, except poly(A) and poly(T) stretches. The presence of poly(A) and poly(T)

Species	source of genome assembly
<i>S. cerevisiae</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2
<i>C. albicans</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000182965.3
<i>C. dubliniensis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026945.1
<i>C. glabrata</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002545.3
<i>C. tropicalis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006335.3
<i>D. hansenii</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006445.2
<i>L. elongisporus</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149685.1/
<i>C. parapsilosis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_000182765.2
<i>C. orthopsilosis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000315875.1
<i>S. stipitis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000209165.1
<i>C. lusitaniae</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000003835.1/
<i>C. fructus</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_003707795.1/
<i>C. auris clade 1</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_002759435.2
<i>C. haemulonii</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_002926055.2
<i>C. duobushaemulonii</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_002926085.2
<i>C. pseudohaemulonii</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_003013735.1
<i>C. intermedia</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_900106115.1
<i>C. neoformans</i>	https://www.ncbi.nlm.nih.gov/assembly/GCF_000091045.1/
<i>C. auris clade 2</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_003013715.2/
<i>C. auris clade 3</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_005234155.1/
<i>C. auris clade 4</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_008275145.1/
<i>C. heveicola</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_003708405.1/
<i>C. oregonensis</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_003707785.2/
<i>C. blattae</i>	https://www.ncbi.nlm.nih.gov/assembly/GCA_003706955.2/

Table 4.1: **List of fungal species genome assemblies.** The table provides all the fungal species considered in the phylogenetic analysis and their available genome assemblies and annotation files.

stretches led us to analyze the GC content of these centromeres. The centromere regions of all seven chromosomes overlap with minima of the GC content of the chromosomes. We analyzed the GC content through two sequence features, that is, GC content and GC3 content. GC content is calculated as a percentage of G or C nucleotide along the length of the genome with a sliding window of 5000bp and a step size of 1000bps. GC3 content is the percentage of G or C nucleotide at the third base of a codon in the coding region sequences of annotated ORFs except for stop codons, for which a moving average is calculated for every 10 adjacent ORFs. All the centromere regions of *C. auris* and its clades overlap with the GC and GC3 troughs and are shown in figure 4.3. It has been previously observed that centromeres of *C. lusitaniae*[129] and *M. sympodialis*[130] coincide with GC and GC3 troughs.

All the seven centromere sequences occupy the entire ORF-free region. This is similar to *C. lusitaniae*, which also occupies the entire ORF-free regions and also lacks pericentric heterochromatin regions. We have used publicly available RNA-seq data sets for *C. auris* (SRR6099290, SRR6099291, SRR6099292, SRR6099293) for analyzing transcriptional status of centromeres. The raw sequence reads were aligned to the reference genome using HISAT2 (v2.1.0) [131]. The aligned reads are then graphically visualized in the IGV to analyze gene expression levels at/around the centromeres on different chromosomes. For studying the transcriptional status of ORFs overlapping with or flanking the centromeres, the abundance of annotated transcripts was quantified using the pseudo alignment program kallisto (v0.46.1) [132]. The expression of genes around/overlapping the centromere in TPM (transcripts per million) was compared to the global gene expression level. RNA-seq analysis revealed that the gene expression of centromere neighboring ORFs is not suppressed and shows that *C. auris*, just like *C. lusitaniae* possess pericentric heterochromatin deficient centromeres. Figure 4.4 shows the expression of ORFs flanking the centromere regions is comparable to the global gene expression.

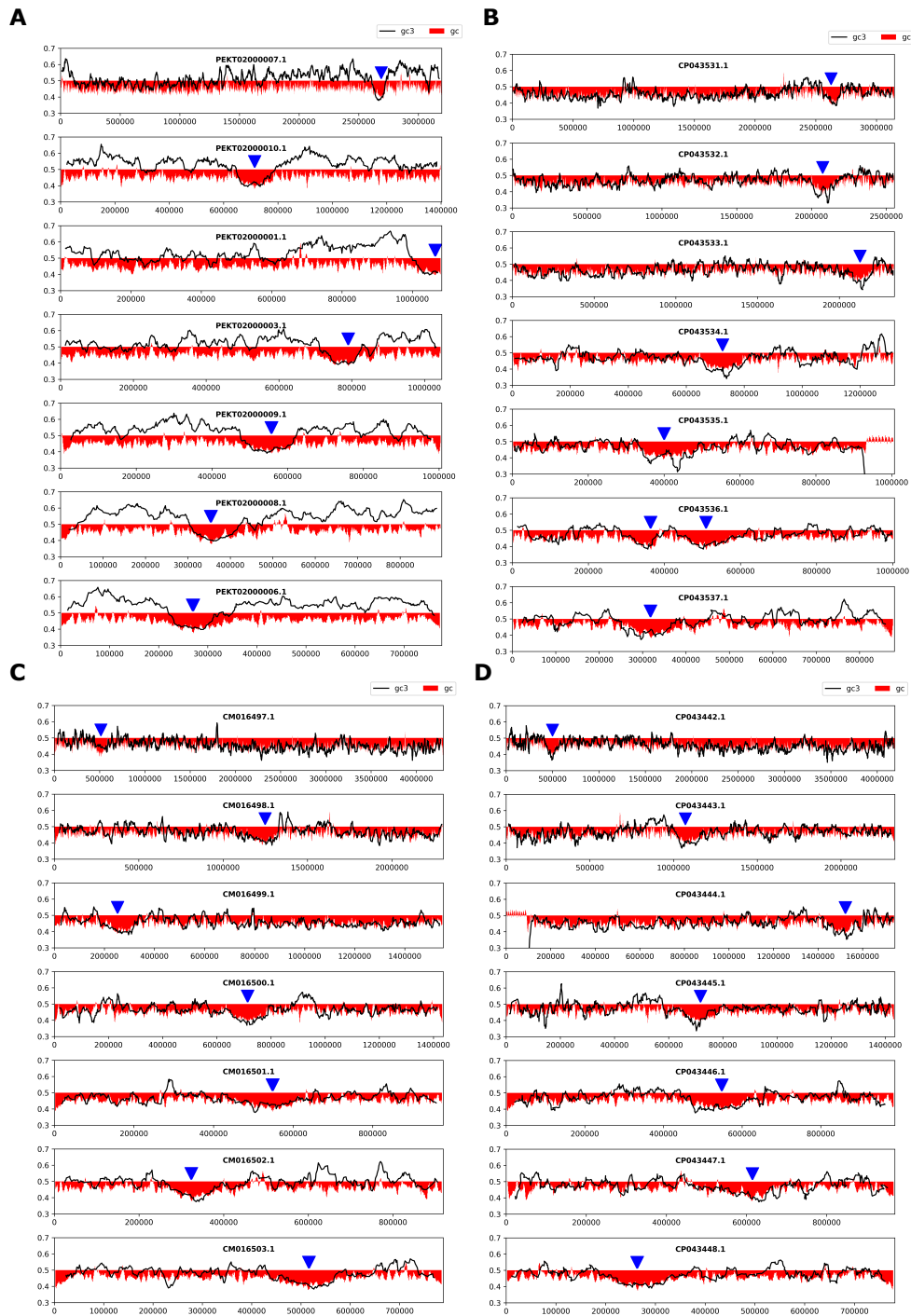


Figure 4.3: **GC and GC3 content of *C. auris* clades.** Centromere positions (blue triangles) overlap with GC (red) and GC3 (black) minima across all scaffolds for *C. auris* clade 1 in (A), clade2 in (B), clade3 in (C), and clade4 in (D) respectively. Coordinates are shown on the x axis, and the %GC is shown on the y axis. The red color bars show the %GC by depicting the amount of deviation from 50% GC (above midline if values are >50% and below if values are <50%)

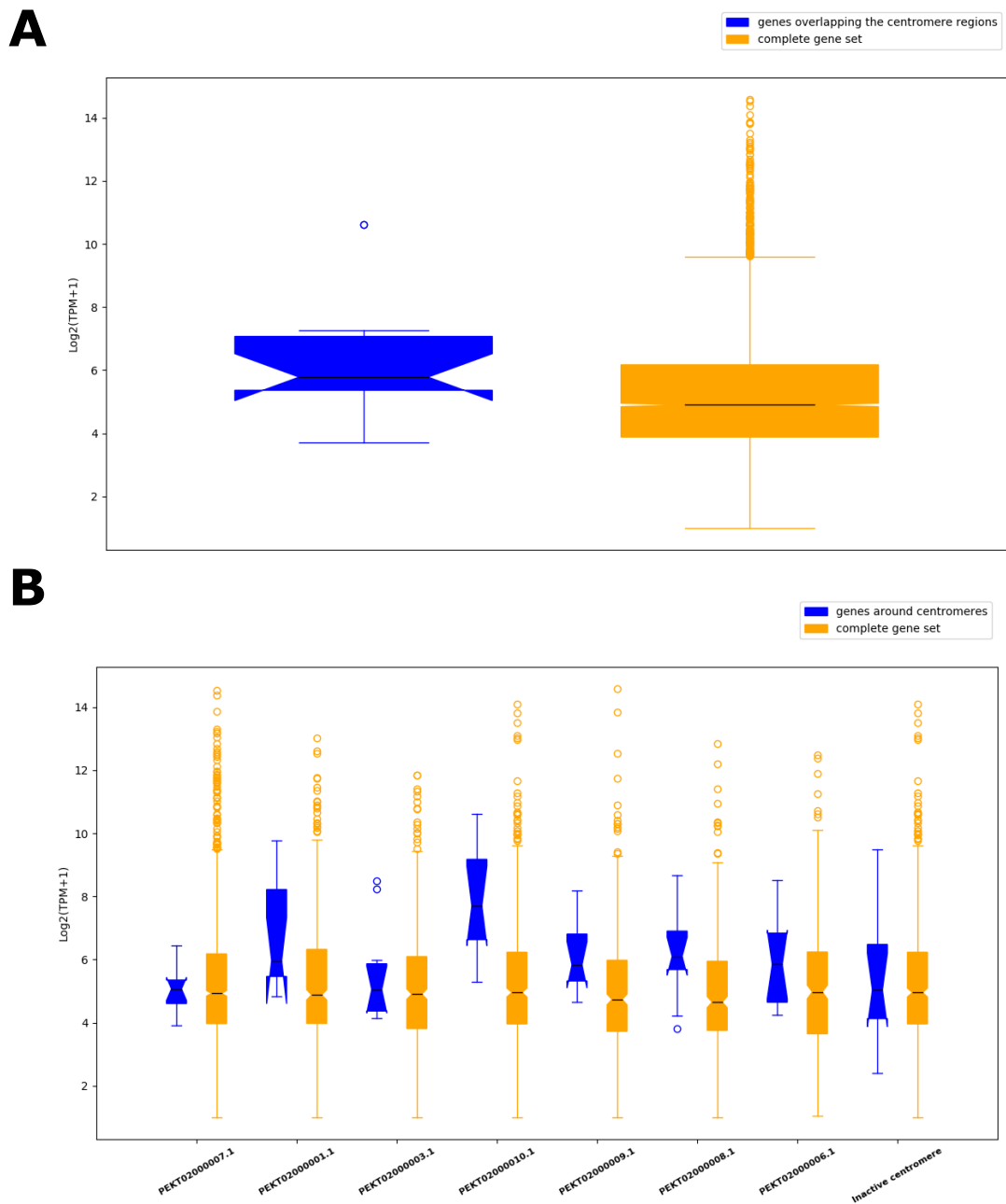


Figure 4.4: **RNA-seq analysis of *C. auris*.** (A) The box plot compares expression of ORFs flanking the centromeres of *C. auris* and global expression of all the ORFs. (B) The box plot compares the expression of ORFs flanking centromere and all the ORFs present in each of the scaffold/chromosomes. The above boxplot shows the expression of flanking ORFs around centromere are not suppressed and their expression comparable to the global gene expression.

4.3 *C. haemulonii* and related species similar centromere properties as *C. auris*

The size of *C. auris* is similar to the size of genomes of phylogenetically related, multi-drug resistant, pathogen species *C. haemulonii*, *C. duobushaemulonii*, and *C. pseudo-haemulonii*. Predicted centromere regions in these species using gene synteny conservation around centromeres compared to *C. auris* shows similar properties such as GC-poor, ORF-free centromeres. The centromeres overlap with GC and GC3 minima (figure4.5). Further, these regions are also free of ORFs similar to *C. auris*.

Similarly, we find the centromeres of species in the Clavispora/Candida clade predicted with gene synteny conservation around centromeres of phylogenetically related *C. auris*, *C. lusitanae* are GC-poor and ORF free regions.

4.4 Discussion

This study, led by Sanyal lab at JNCASR, Bangalore, in which I participated, uses functional and comparative analysis of genomes to identify the centromere landscape of the Clavispora/Candida clade which constitutes of small, GC-poor, ORF free regions, devoid of pericentric heterochromatin regions, motifs, and repeats. The species in the clade can be classified into species with seven or eight chromosomes. The study proposes inactivation of one of the centromeres in a common ancestor has led to the rapid emergence of multiple species with seven chromosomes.

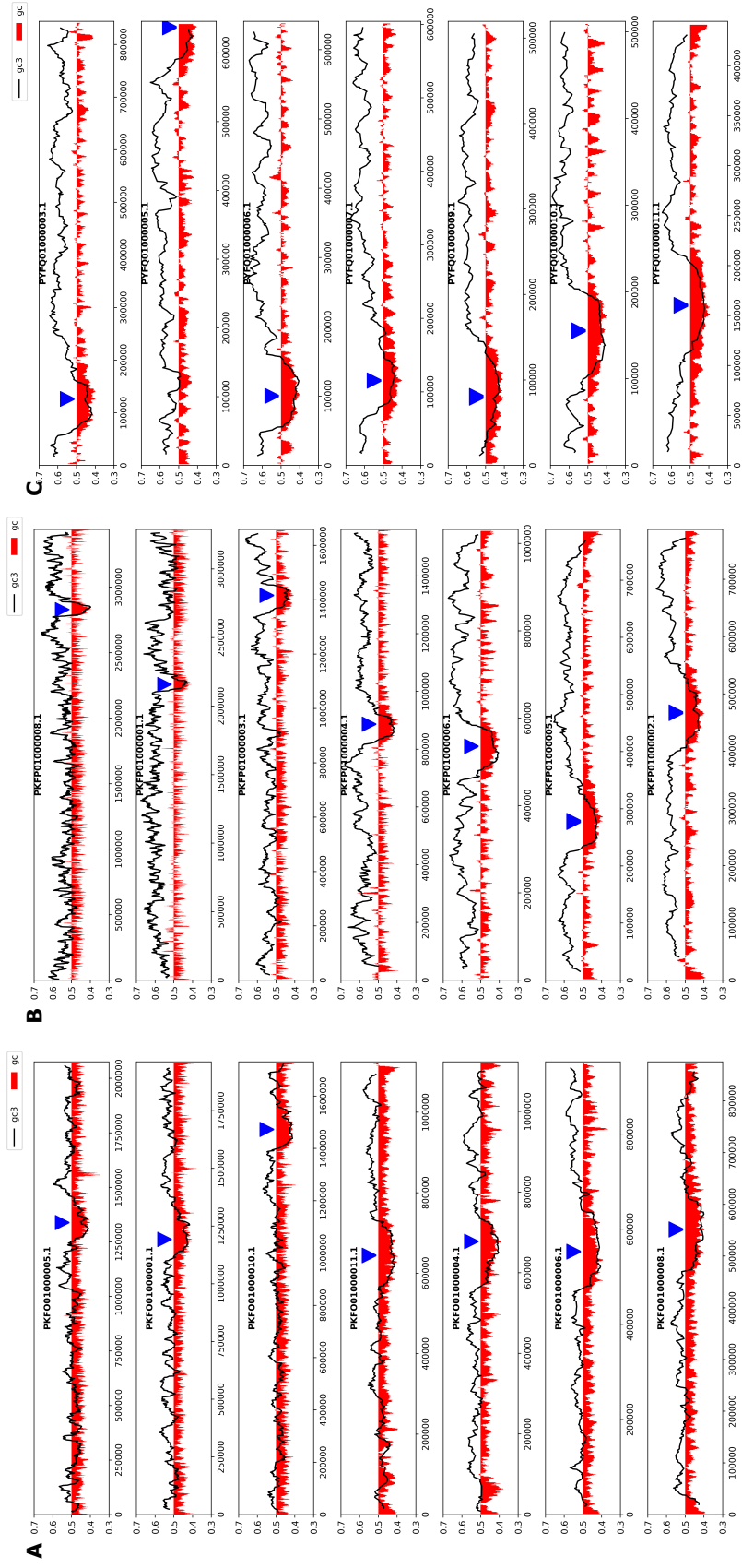


Figure 4.5: GC content analysis of haemulonii complex species. The plots show the GC (red) and GC3 (black) content minima in (A) *C. haemulonii*, (B) *C. pseudohaemulonii*, and (C) *C. duobushaemulonii* overlapping with the centromere regions (blue triangles). Coordinates are shown on the x axis, and the %GC is shown on the y axis. The red color bars show the %GC by depicting the amount of deviation from 50% GC (above midline if values are >50% and below if values are <50%)

Chapter 5

Conclusion

In this thesis, we presented a method developed to identify pairs of TFs that significantly attract or avoid each other through the long-range 3D chromatin interactions. The method also includes module to identify TF pair interactions on linear genomic regions. The method has been expanded to handle either with experimentally determined transcription factor binding sites (TFBS) or with computational predictions of motif sites. The associated tool “[ChromTogether](#)” is open source and is available for public use.

Our method has proved effective in identifying TF interaction known already in the literature. We further identified functionally different interacting groups of TFs across multiple cell lines used in the study namely Gm12878, K562, HeLa-S3, and MCF7. These groups of TFs attracting other member within a group and avoiding members of other groups are mostly similar even if one looks at TF co-localization on linear sequence elements or at motif level instead of TFBS evidences. This suggests that the 3D organization of chromatin has an important role in gene regulation. Across cell lines, we observe one group of TFs which regulate house-keeping genes which bind mostly to GC-rich promoter regions proximal to the TSSs. Whereas members of other group bind either promoter or enhancers which are AT-rich regions. In GM12878 cell line, we find the second group of TFs regulate genes involved in cell-lineage specific functions.

Our analysis further suggest the role of spatial organization of chromatin in facilitating in vivo TF binding. We believe the information of TF co-localization in spatial organization would be useful to improve and further would like to use it *in silico* prediction of TF binding sites, which is an important challenge given the expense of ChIP-seq experiments, and the subject of much effort worldwide. We would also like to expand our analysis to many more cell types, other organisms to gain further knowledge and understand regulatory aspects. Specifically, we are interested in the evolution of different functional groups of TFs during developmental stages.

In a separate piece of work, we analyse the phylogenetics, karyotypes and centromeres of four clades of the pathogenic yeast *Candida auris*. This was a larger project where we contributed bioinformatic analysis, specifically, studying the phylogeny of various *Candida* species; analysing the GC and GC3 content of the *C. auris* genome and confirming that the centromeres occur in GC and GC3 troughs; and analysing the transcription of ORFs surrounding centromeres, finding that their expression is not suppressed compared to the rest of the genome. This work was the subject of a commentary by Rusche [133] who highlights the remarkable stability of *C. auris* centromeres that have maintained their syntenic positions for over 100 million years despite any clear sequence identifier for centromeric position (other than GC/GC3 content), and also the centromere-adjacent rearrangements that could drive new traits as well as speciation.

The major work in this thesis was based on published, publicly-available high-throughput data (ChIP-seq, Hi-C, ChIA-PET etc), while chapter 4 described a close collaboration with an experimental group. The close interplay between computational techniques and algorithms and real biological data has been and will remain an enriching and rewarding experience.

Appendix A

List of data sets used in the study

Cell line	Data Set	Accession Number
GM12878	ChIA-PET (RNAPII)	GSM1872887
K562	ChIA-PET (RNAPII)	GSM970213
MCF7	ChIA-PET (RNAPII)	GSM970209
HeLa-S3	ChIA-PET (RNAPII)	GSM1872889

Table A.1: **List of POI2 ChIA-PET datasets.** Accession IDs of chromatin interaction data sets used in the study

GM12878	
ATF2	wgEncodeEH002306
ATF3	wgEncodeEH001562
BATF	wgEncodeEH001479
BCL11A	wgEncodeEH001486
BHLHE40	wgEncodeEH002025
BRCA1	wgEncodeEH001830
CEBPB	wgEncodeEH003212
CTCF	wgEncodeEH001851, wgEncodeEH000029, wgEncodeEH000394, wgEncodeEH000532
E2F4	wgEncodeEH002867

EBF1	wgEncodeEH001832
EGR1	wgEncodeEH002328
ELF1	wgEncodeEH001617
ELK1	wgEncodeEH002851
EP300	wgEncodeEH002824, wgEncodeEH002037
ETS1	wgEncodeEH001564
EZH2	wgEncodeEH002411
FOS	wgEncodeEH000622
FOXM1	wgEncodeEH002529
GABPA	wgEncodeEH001462
IRF4	wgEncodeEH001484
JUND	wgEncodeEH000639
MAX	wgEncodeEH002806
MAZ	wgEncodeEH002852
MEF2A	wgEncodeEH001565
MTA3	wgEncodeEH002329
MXI1	wgEncodeEH002026
MYC	wgEncodeEH000547
NFATC1	wgEncodeEH002307
NFE2	wgEncodeEH001808
NFIC	wgEncodeEH002343
NFYB	wgEncodeEH002065
NRF1	wgEncodeEH001846
PAX5	wgEncodeEH001489, wgEncodeEH001495
PBX3	wgEncodeEH001477
PML	wgEncodeEH002308
POLR2A	wgEncodeEH000626, wgEncodeEH001517
POU2F2	wgEncodeEH001475

RAD21	wgEncodeEH000749
REST	wgEncodeEH002314
RFX5	wgEncodeEH001810
RUNX3	wgEncodeEH002330
RXRA	wgEncodeEH001541
SIN3A	wgEncodeEH002868
SIX5	wgEncodeEH001542
SMC3	wgEncodeEH001833
SP1	wgEncodeEH001496
SPI1	wgEncodeEH001476
SRF	wgEncodeEH001464
STAT1	wgEncodeEH001852
STAT3	wgEncodeEH001811
STAT5A	wgEncodeEH002321
TAF1	wgEncodeEH001478
TBL1XR1	wgEncodeEH002853
TBP	wgEncodeEH001798
TCF12	wgEncodeEH001485
TCF3	wgEncodeEH002315
USF1	wgEncodeEH001468
USF2	wgEncodeEH001812
YY1	wgEncodeEH000695, wgEncodeEH001657
ZBTB33	wgEncodeEH001488
ZEB1	wgEncodeEH001645
ZNF143	wgEncodeEH001853
K562	
ARID3A	wgEncodeEH002861
ATF1	wgEncodeEH002865

ATF3	wgEncodeEH000700
BACH1	wgEncodeEH002846
BHLHE40	wgEncodeEH001857
CEBPB	wgEncodeEH001821
CTCF	wgEncodeEH002279, wgEncodeEH000042
CTCF	wgEncodeEH002797
E2F4	wgEncodeEH000671
E2F6	wgEncodeEH000676, wgEncodeEH001598
EGR1	wgEncodeEH001646
ELF1	wgEncodeEH001619
ELK1	wgEncodeEH003356
EP300	wgEncodeEH002834, wgEncodeEH002086
EZH2	wgEncodeEH002089
FOS	wgEncodeEH001207, wgEncodeEH000619
FOSL1	wgEncodeEH001637
GABPA	wgEncodeEH001604
GATA1	wgEncodeEH000638
GATA2	wgEncodeEH000683
GATA2	wgEncodeEH001208, wgEncodeEH001576
IRF1	wgEncodeEH001866, wgEncodeEH002798, wgEncodeEH002799
JUND	wgEncodeEH002164
MAFF	wgEncodeEH002804
MAFK	wgEncodeEH001844
MAX	wgEncodeEH002869
MEF2A	wgEncodeEH001663
MXI1	wgEncodeEH001827
MYC	wgEncodeEH002800, wgEncodeEH001867

NFE2	wgEncodeEH000624
NFYA	wgEncodeEH002021
NFYB	wgEncodeEH002024
NRF1	wgEncodeEH001796
PML	wgEncodeEH002320
POLR2A	wgEncodeEH000704, wgEncodeEH000727
RAD21	wgEncodeEH001585, wgEncodeEH000649
REST	wgEncodeEH001638
RFX5	wgEncodeEH002033
SETDB1	wgEncodeEH000677
SIN3AK20	wgEncodeEH001607
SIRT6	wgEncodeEH000681
SIX5	wgEncodeEH001483
SMC3	wgEncodeEH001845
SP1	wgEncodeEH001578
SP2	wgEncodeEH001653
SPI1	wgEncodeEH001482
SRF	wgEncodeEH001600
STAT1	wgEncodeEH000760, wgEncodeEH000761
STAT2	wgEncodeEH000665
STAT5A	wgEncodeEH002347
TAF1	wgEncodeEH001582
TAF7	wgEncodeEH001654
TAL1	wgEncodeEH001824
TBL1XR1	wgEncodeEH002848, wgEncodeEH002849
TBP	wgEncodeEH001825
TEAD4	wgEncodeEH002333
THAP1	wgEncodeEH001655

USF1	wgEncodeEH001583
USF2	wgEncodeEH001797
YY1	wgEncodeEH000684, wgEncodeEH001584, wgEncodeEH001623
ZBTB33	wgEncodeEH001569
ZBTB7A	wgEncodeEH001620
ZNF143	wgEncodeEH002030
ZNF263	wgEncodeEH000630
HeLa-S3	
BRCA1	wgEncodeEH001814
BRF1	wgEncodeEH000764
BRF2	wgEncodeEH000765
CEBPB	wgEncodeEH001815
CHD2	wgEncodeEH002027
CTCF	wgEncodeEH000398, wgEncodeEH000541
E2F1	wgEncodeEH000688
E2F1	wgEncodeEH000699
E2F4	wgEncodeEH000689
E2F6	wgEncodeEH000692
ELK1	wgEncodeEH002864
ELK4	wgEncodeEH001753
EP300	wgEncodeEH001820
EZH2	wgEncodeEH003086
FAM48A	wgEncodeEH001855
FOS	wgEncodeEH000647
GABPA	wgEncodeEH001504
GTF2F1	wgEncodeEH001816
IRF3	wgEncodeEH001788

JUN	wgEncodeEH000746
JUND	wgEncodeEH000745
MAFK	wgEncodeEH002856
MAX	wgEncodeEH002830
MAZ	wgEncodeEH002855
MXI1	wgEncodeEH001826
MYC	wgEncodeEH000542, wgEncodeEH000648
NFYA	wgEncodeEH002066
NFYB	wgEncodeEH002067
NR2C2	wgEncodeEH000687
NRF1	wgEncodeEH000723
POLR2A	wgEncodeEH001474, wgEncodeEH001838
PRDM1	wgEncodeEH001817
RAD21	wgEncodeEH001789
RCOR1	wgEncodeEH002844
REST	wgEncodeEH001629
RFX5	wgEncodeEH001818
RPC155	wgEncodeEH000766
SMC3	wgEncodeEH001839
STAT1	wgEncodeEH000614
STAT3	wgEncodeEH001799
TAF1	wgEncodeEH001505
TBP	wgEncodeEH001790
TCF7L2	wgEncodeEH002069, wgEncodeEH002813
USF2	wgEncodeEH001819
ZKSCAN1	wgEncodeEH002857
ZNF143	wgEncodeEH002028
ZNF274	wgEncodeEH001763

<i>ZZZ3</i>	wgEncodeEH001872
-------------	------------------

Table A.2: **List of TF ChIP-seq datasets.** Accession IDs of the uniformly processed chip-seq peaks generated by ENCODE project used in the study

Appendix B

Supplementary Information

Primarily, we have used Pol II ChIA-PET datasets for our analysis across four cell lines GM12878, K562, HeLa-S3, and MCF7. But for comparison with the results of the previous study by Ma *et al.*, we provide the results obtained using our method on Hi-C data used by the previous study. The following figures [B.1](#) and [B.2](#) depict the co-occurrence pattern obtained by our method and comparison with the results of a previous study respectively.

We observe similar co-occurrence pattern of motif sites on linear genome and in 3D interaction just as we did using TFBS provided in the chapter 3 of the thesis. Figures [B.3](#), [B.4](#), [B.5](#), and [B.6](#) show the similarity in their co-occurrence pattern obtained using the motif instance for GM12878, K562, HeLa-S3 and MCF7 cell lines respectively. We also provide comparison of co-occurrence patterns for common factors from all four models (TFBS on linear genome and 3D interactions; motif instances on linear genome and 3D interactions) for GM12878 and K562 cell lines in figure [B.7](#) and [B.8](#) respectively.

The table [B.1](#) gives the selected PWMs selected from clusters of similar PWMs of various transcription factors collected from JASPAR database [[101](#)].

Transcription factor	JASPAR Motif ID
----------------------	-----------------

Tbxt	MA0009.1
ELK1	MA0028.2
Gata1	MA0035.3
Gfi1	MA0038.1
FOXI1	MA0042.1
MAX	MA0058.3
NFYA	MA0060.3
RXRA::VDR	MA0074.1
ELK4	MA0076.2
Sox17	MA0078.1
SP1	MA0079.3
SPI1	MA0080.4
SRF	MA0083.3
ZNF143	MA0088.2
TEAD1	MA0090.1
ZEB1	MA0103.2
NFKB1	MA0105.4
TP53	MA0106.1
TBP	MA0108.1
ESR1	MA0112.3
NR3C1	MA0113.3
NFIC::TLX1	MA0119.1
Nkx3-1	MA0124.2
HINFP	MA0131.2
STAT1	MA0137.2
REST	MA0138.1
CTCF	MA0139.1

Sox2	MA0143.3
Tcfcp2l1	MA0145.1
Myc	MA0147.1
FOXA1	MA0148.3
NFATC2	MA0152.1
EBF1	MA0154.3
FOXO3	MA0157.2
EGR1	MA0162.3
CDX2	MA0465.1
DUX4	MA0468.1
E2F4	MA0470.1
ELF1	MA0473.1
Gfi1b	MA0483.1
HSF1	MA0486.2
JUND	MA0491.1
MAFK	MA0496.2
POU2F2	MA0507.1
PRDM1	MA0508.2
RFX5	MA0510.2
RUNX2	MA0511.2
Rxra	MA0512.2
Tcf12	MA0521.1
Esrra	MA0592.2
Hoxa9	MA0594.1
SREBF1	MA0595.1
FOXG1	MA0613.1
Mitf	MA0620.1
mix-a	MA0621.1

BARHL2	MA0635.1
BHLHE41	MA0636.1
ETV6	MA0645.1
GRHL1	MA0647.1
IRF9	MA0653.1
NFIA	MA0670.1
NKX2-3	MA0672.1
ONECUT1	MA0679.1
PAX7	MA0680.1
POU4F2	MA0683.1
SP4	MA0685.1
ZBTB7B	MA0694.1
OTX2	MA0712.1
GLIS1	MA0735.1
GLIS3	MA0737.1
Hic1	MA0739.1
SCRT2	MA0744.1
E2F7	MA0758.1
Tcf7	MA0769.1
MEF2D	MA0773.1
MEIS3	MA0775.1
PAX1	MA0779.1
PAX9	MA0781.1
POU3F4	MA0789.1
SMAD3	MA0795.1
TGIF1	MA0796.1
TFAP2C(var.3)	MA0815.1
ATF7	MA0834.1

NFE2	MA0841.1
TEF	MA0843.1
Rarb	MA0857.1
Rarg(var.2)	MA0860.1
E2F2	MA0864.1
SOX4	MA0867.1
Sox11	MA0869.1
TFAP2A(var.3)	MA0872.1
HOXC13	MA0907.1
SIX1	MA1118.1

Table B.1: **List of PWMs selected form clusters of similar motifs.** The selected motifs after clustering for similar motifs using TOMTOM

The two groups of factors observed from motif site co-occurrence also show differences in their occurrence in regions with different GC content. The following violin plots [B.9](#), [B.10](#), [B.11](#) show the one group of factor are present in ChIA-PET regions with low GC fraction whereas the other group factors are present in regions with higher than median GC fraction in K562, HeLa-S3, and MCF7 respectively.

The figures [B.14](#) and [B.15](#) show enriched GO terms, and GWAS disease/phenotype terms for the target genes of Group 1 and Group 2 TFs in K562 and HeLa-S3 cell lines respectively.

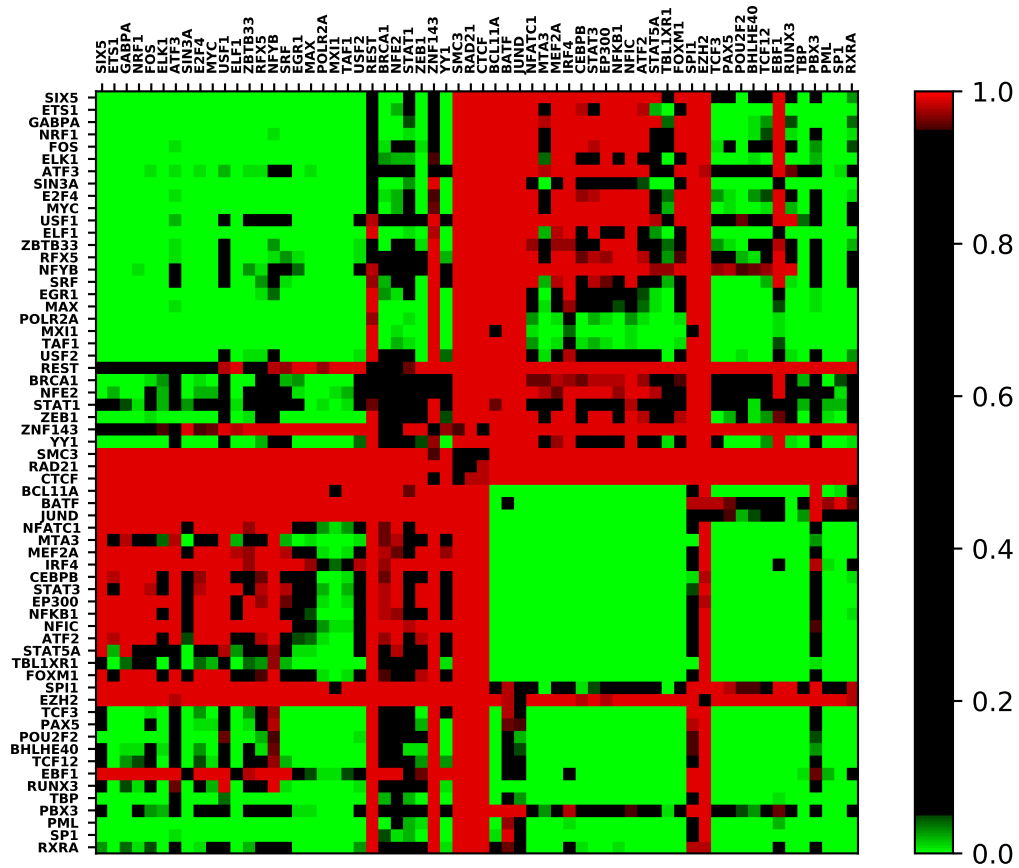


Figure B.1: **Co-occurrence heatmap of GM12878 cell line using Hi-C.** The heatmap shows the co-occurrence pattern of TF pairs obtained using the Hi-C data used in the previous study by Ma *et al.* The order of TFs is same as used in figure 3.2 and the co-occurrence is mostly similar to the pattern obtained using Pol II ChIA-PET data.

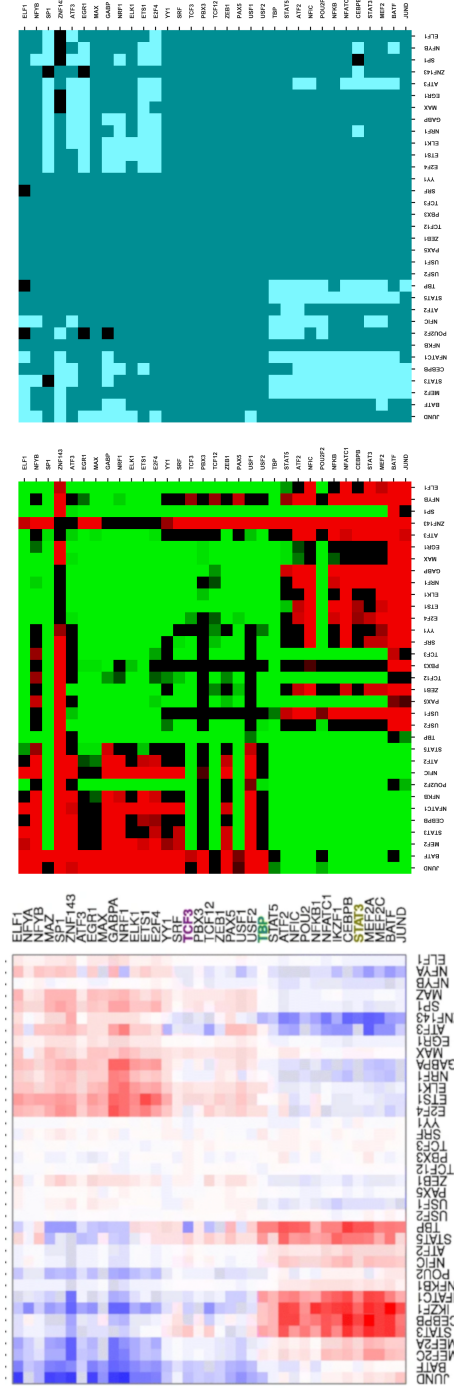


Figure B.2: **Comparison of observations on Hi-C data by our method and previous study by Ma et al.** (A) The heatmap from the earlier study by Ma et al. [90] (reproduced under licence CC-BY-4.0) showing attracting TF pairs in red and repelling pairs in blue. (B) The heatmap shows the attracting and repelling pairs in green and red respectively for common TFs used in both studies, using the method we proposed in methods section on Hi-C data. (C) The heatmap shows a qualitative comparison of the two studies as follows: bright blue = both significant, in agreement; black = both significant, in disagreement, the other repulsion, the other attraction, the other repulsion); dark blue = one or both insignificant.

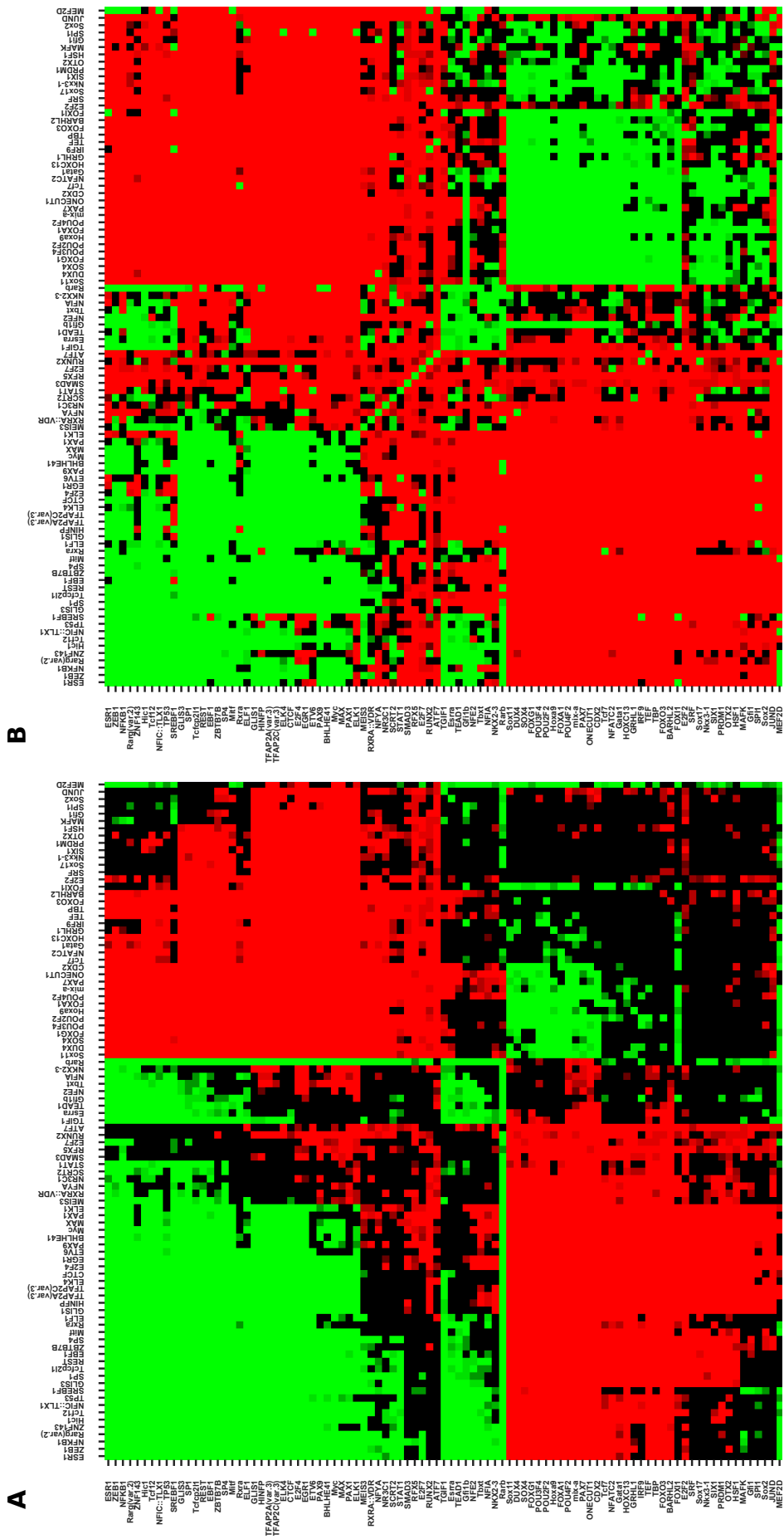


Figure B.3: Comparison TF motif site co-occurrence in spatial and linear proximal regions (GM12878). (A) The co-occurrence of TF motifs sites in spatial proximal regions and (B) in sequential contiguous regions of GM12878 cell line

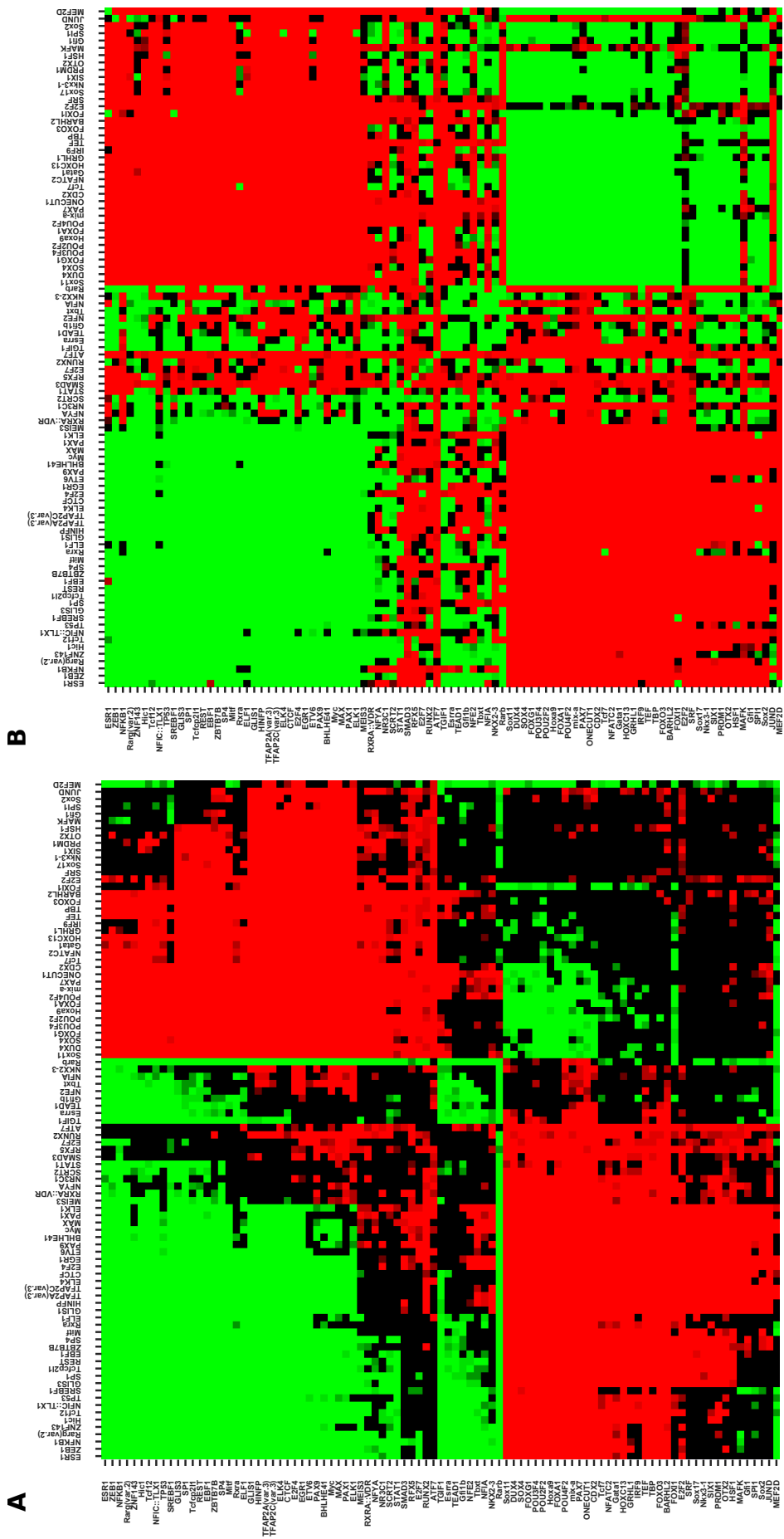


Figure B.4: Comparison TF motif site co-occurrence in spatial and linear proximal regions (K562). (A) The co-occurrence of TF motifs sites in spatial proximal regions and (B) in sequential contiguous regions of K562 cell line

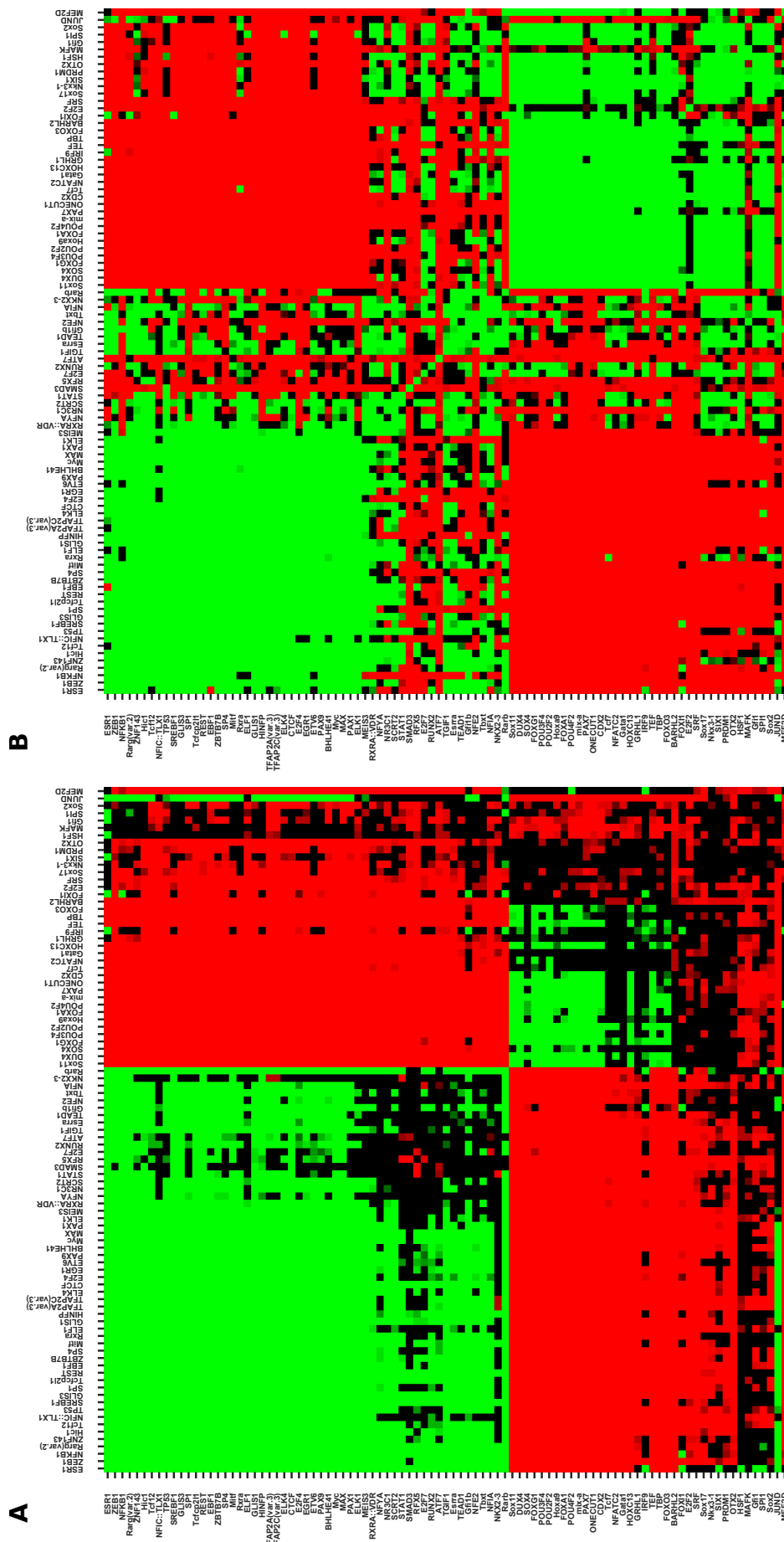


Figure B.6: Comparison TF motif site co-occurrence in spatial and linear proximal regions (MCF7). (A) The co-occurrence of TF motifs sites in spatial proximal regions and (B) in sequential contiguous regions of MCF7 cell line

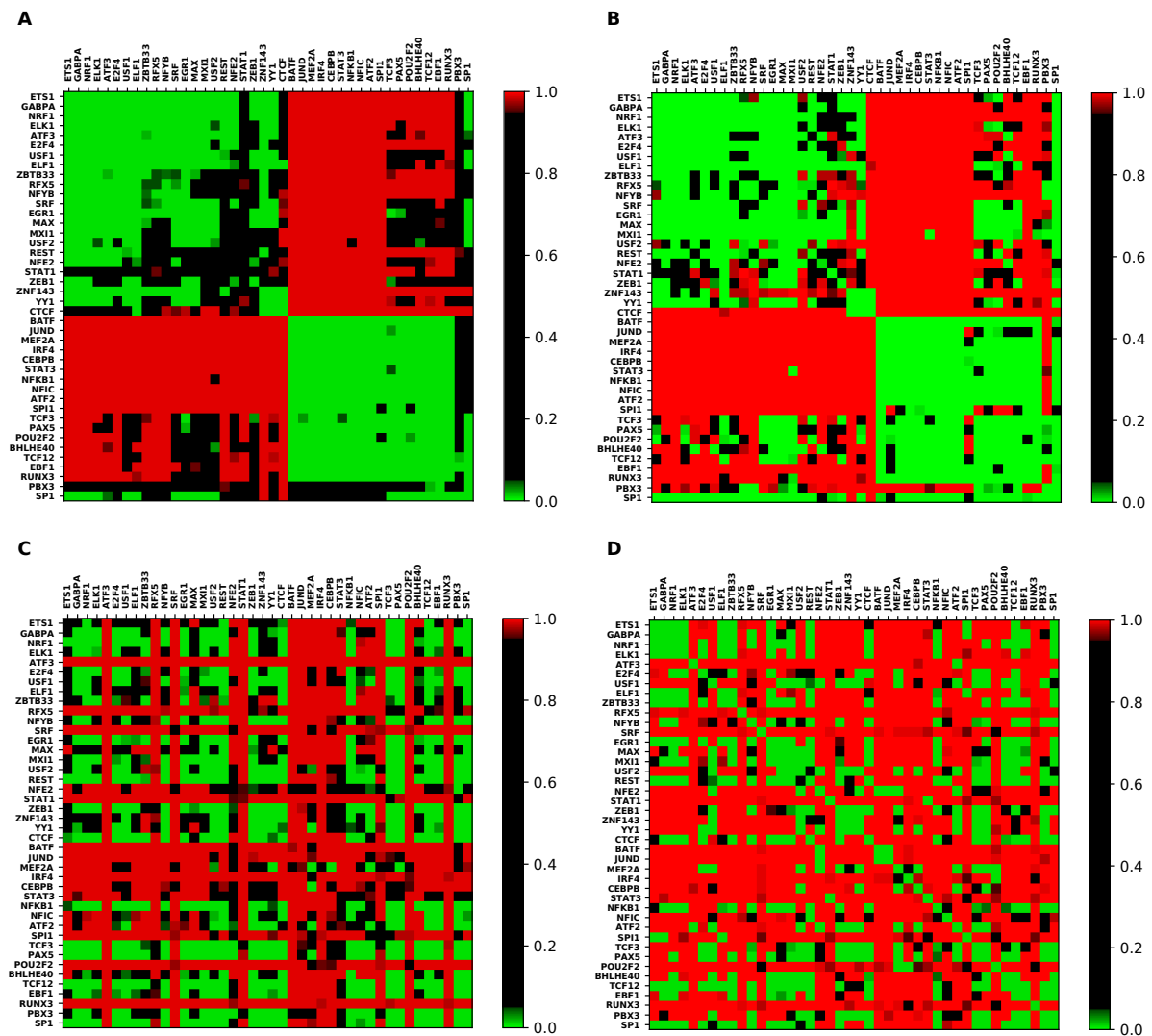


Figure B.7: Comparison of co-occurrence pattern with TFBS and motif sites (GM12878). The figure shows all the four different q-value heatmaps of GM12878 cell line. (A) using TF chip-seq peaks in spatial proximal regions, (B) using TF chip-seq peaks in sequential contiguous regions, (C) using TF motif sites in spatial proximal regions, and (D) using TF motif sites in sequential contiguous regions.

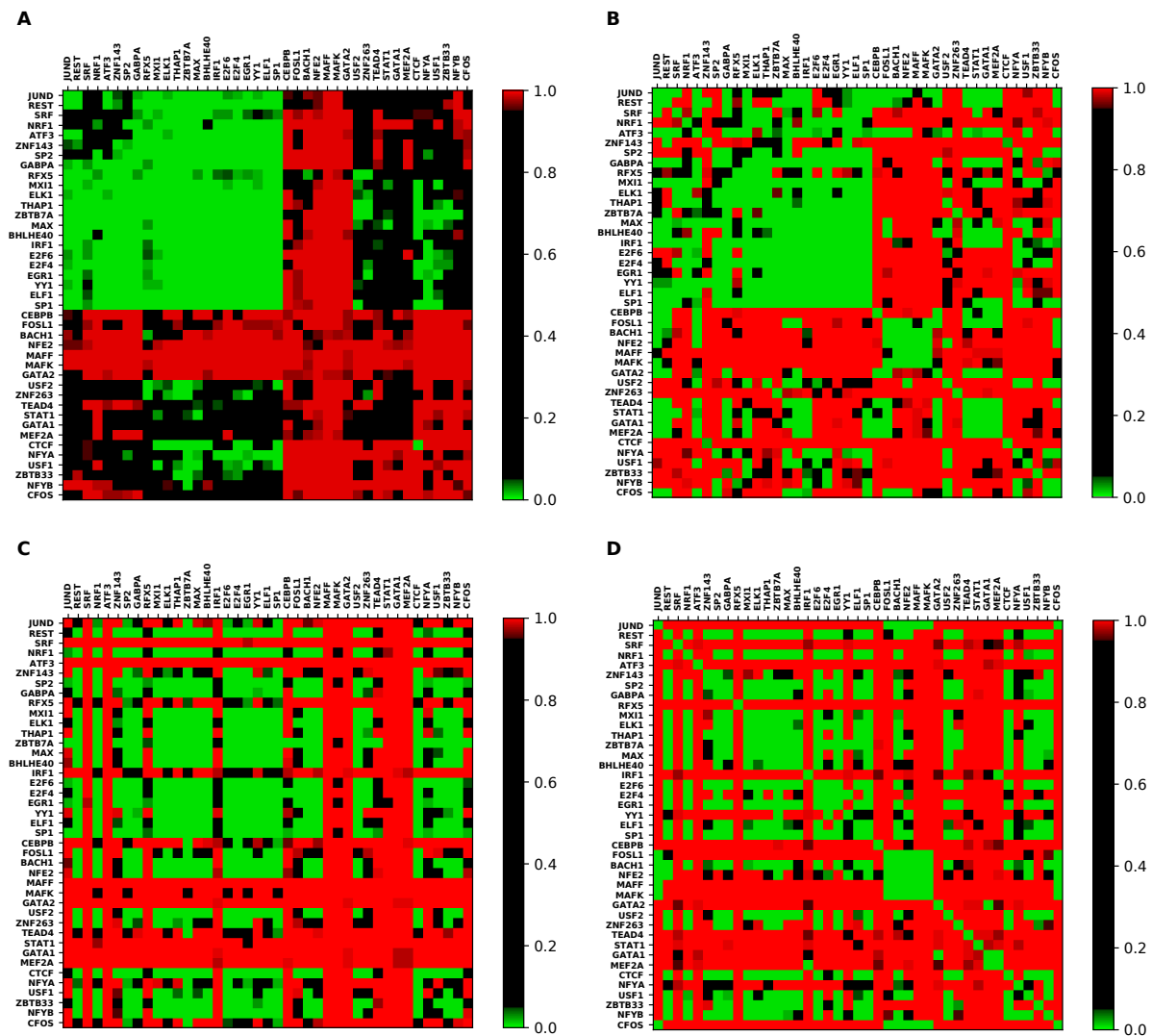


Figure B.8: Comparison of co-occurrence pattern with TFBS and motif sites (K562). The figure shows all the four different q-value heatmaps of K562 cell line. (A) using TF chip-seq peaks in spatial proximal regions, (B) using TF chip-seq peaks in sequential contiguous regions, (C) using TF motif sites in spatial proximal regions, and (D) using TF motif sites in sequential contiguous regions.

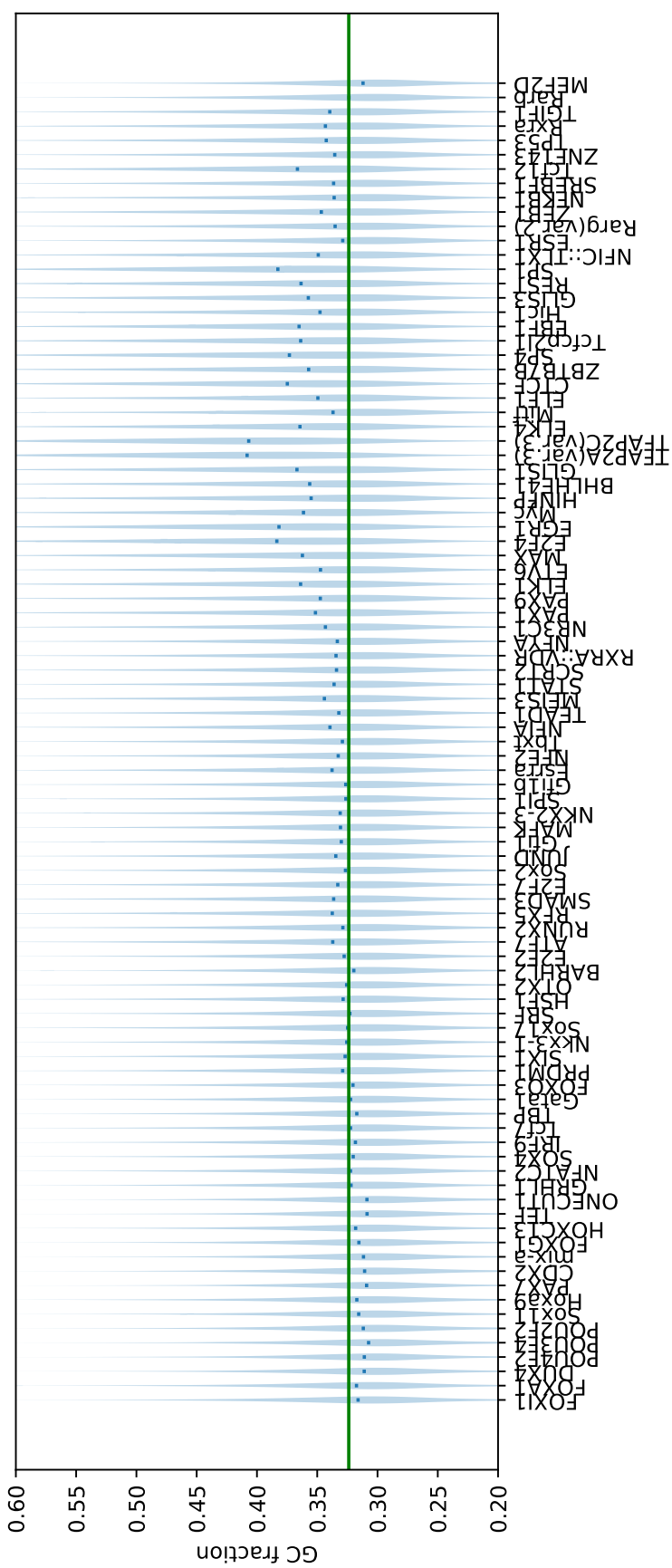


Figure B.9: **GC content of ChIA-PET regions of K562 cell line.** The violin plot of each factor show the distribution of GC fraction values of ChIA-PET regions with presence of TF motif. The green line shows the median GC fraction value for the set of all ChIA-PET regions.

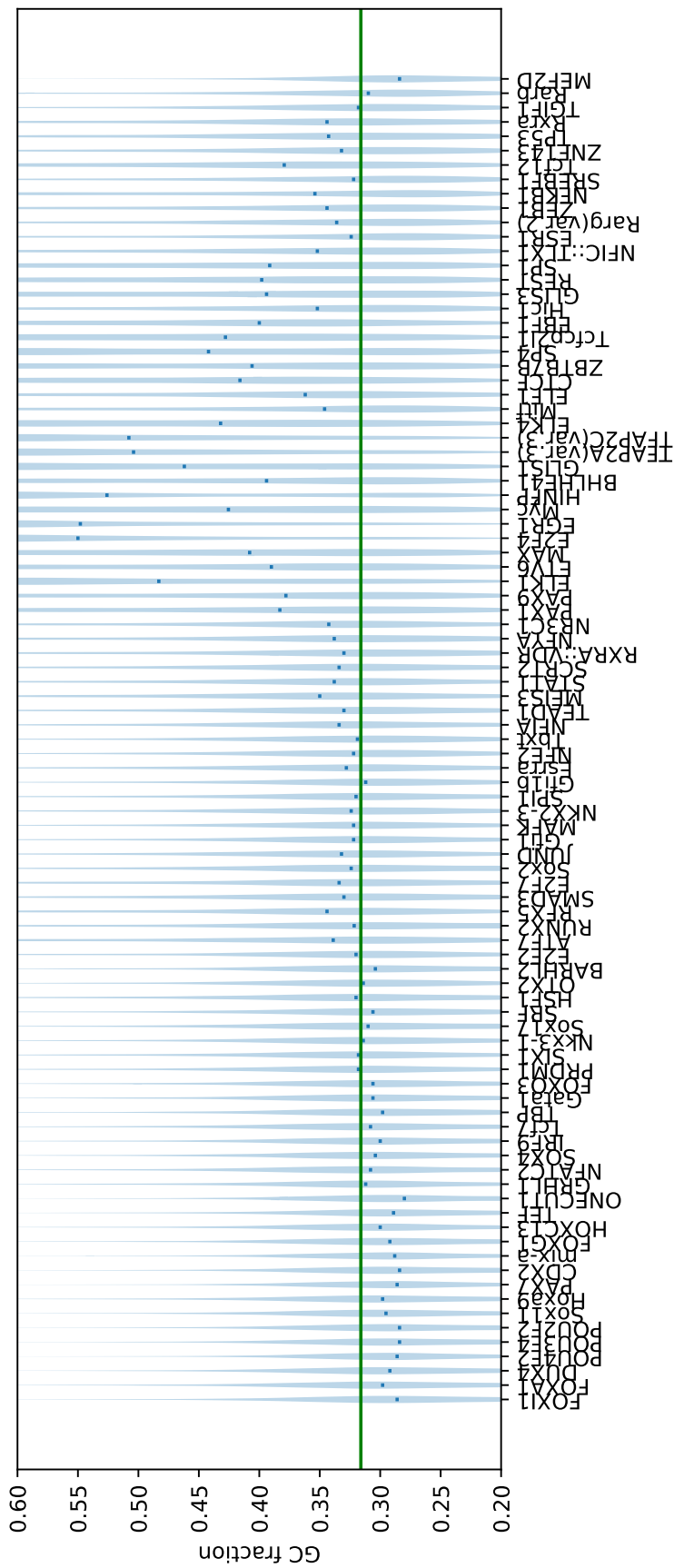


Figure B.10: **GC content of ChIA-PET regions of HeLa-S3 cell line.** The violin plot of each factor show the distribution of GC fraction values of ChIA-PET regions with presence of TF motif. The green line shows the median GC fraction value for the set of all ChIA-PET regions.

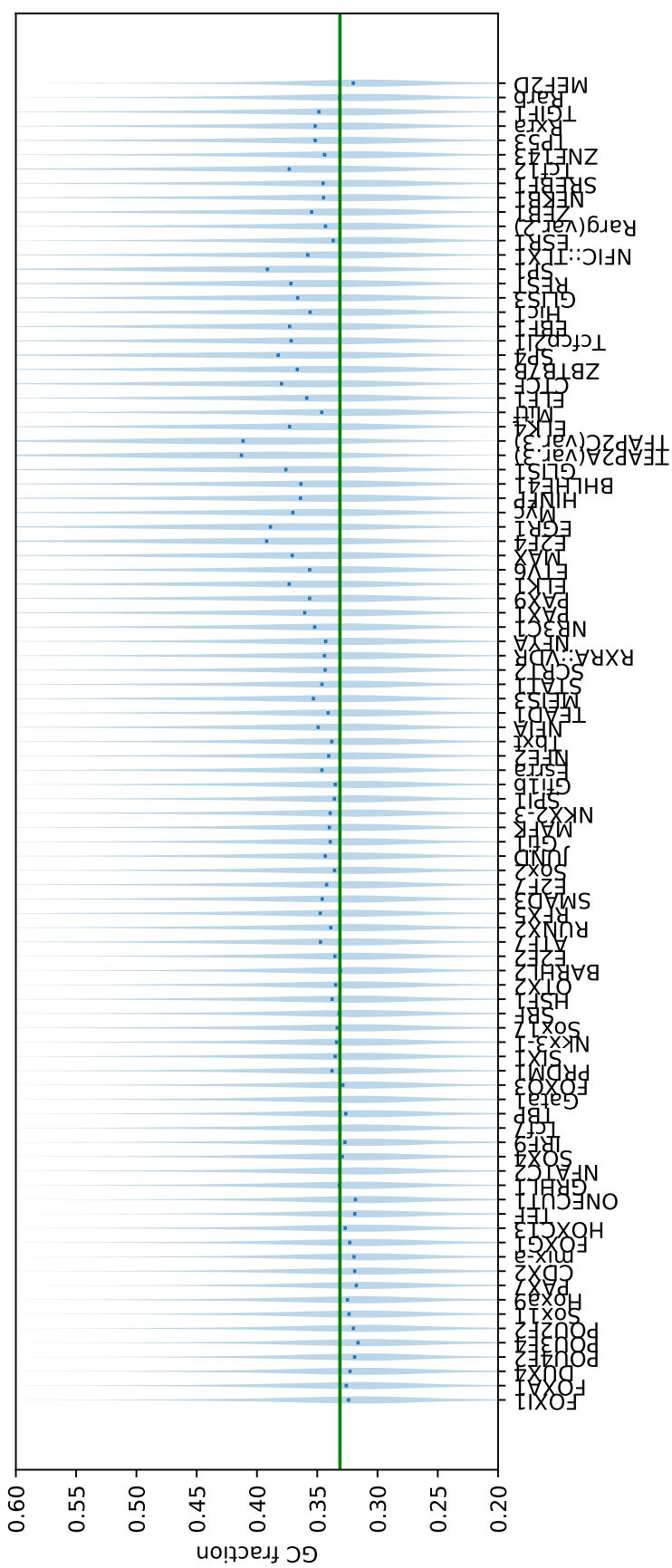


Figure B.11: **GC content of ChIA-PET regions of MCF7 cell line.** The violin plot of each factor show the distribution of GC fraction values of ChIA-PET regions with presence of TF motif. The green line shows the median GC fraction value for the set of all ChIA-PET regions.

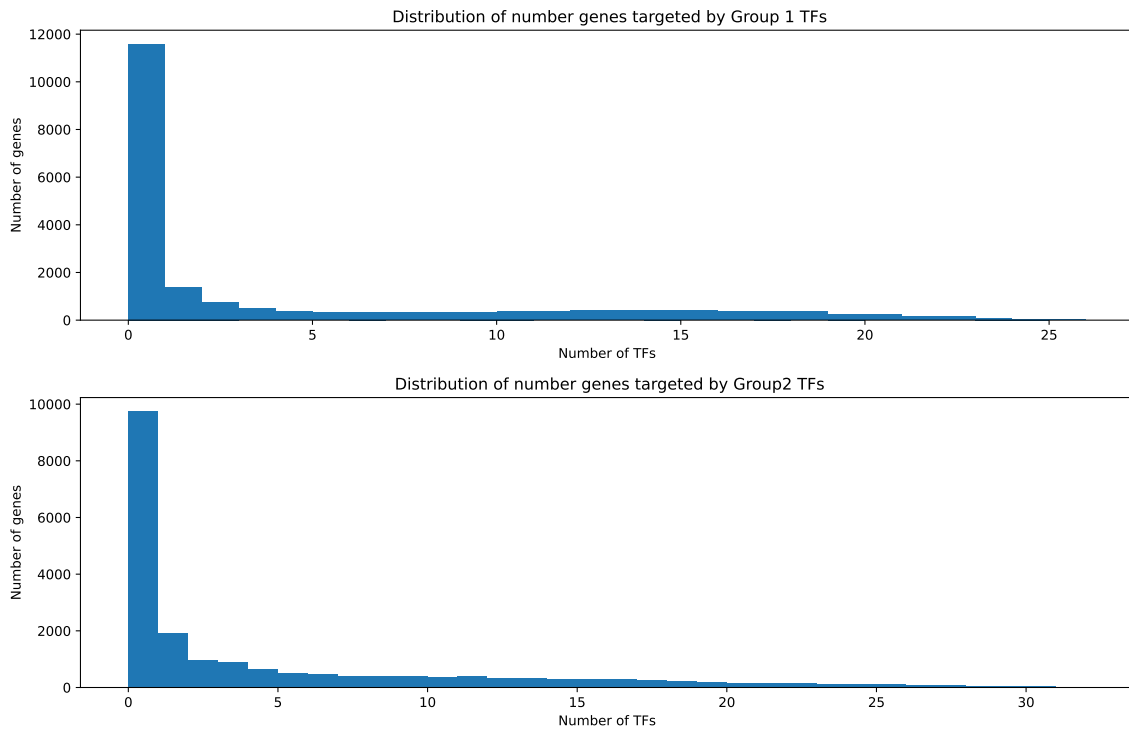


Figure B.12: **Distribution of number of TFs per gene (GM12878).** Distribution of number of putative TF regulators per gene in GM12878 cell line.

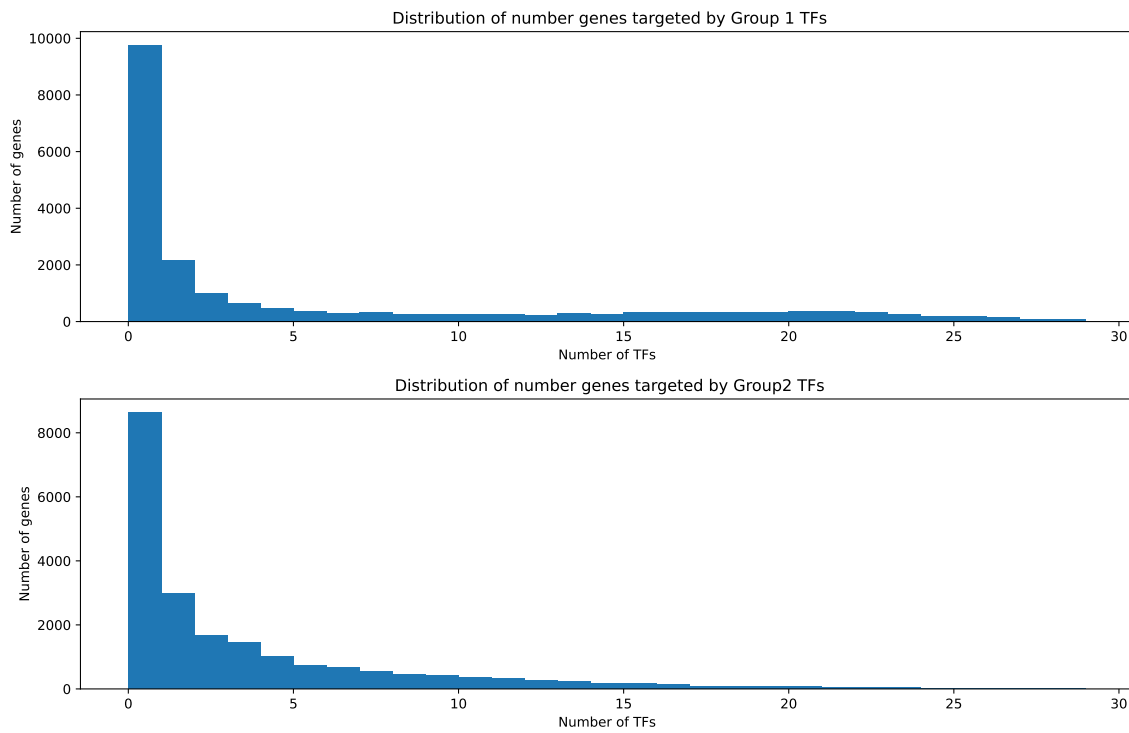


Figure B.13: **Distribution of number of TFs per gene (K562).** Distribution of number of putative TF regulators per gene in K562 cell line.

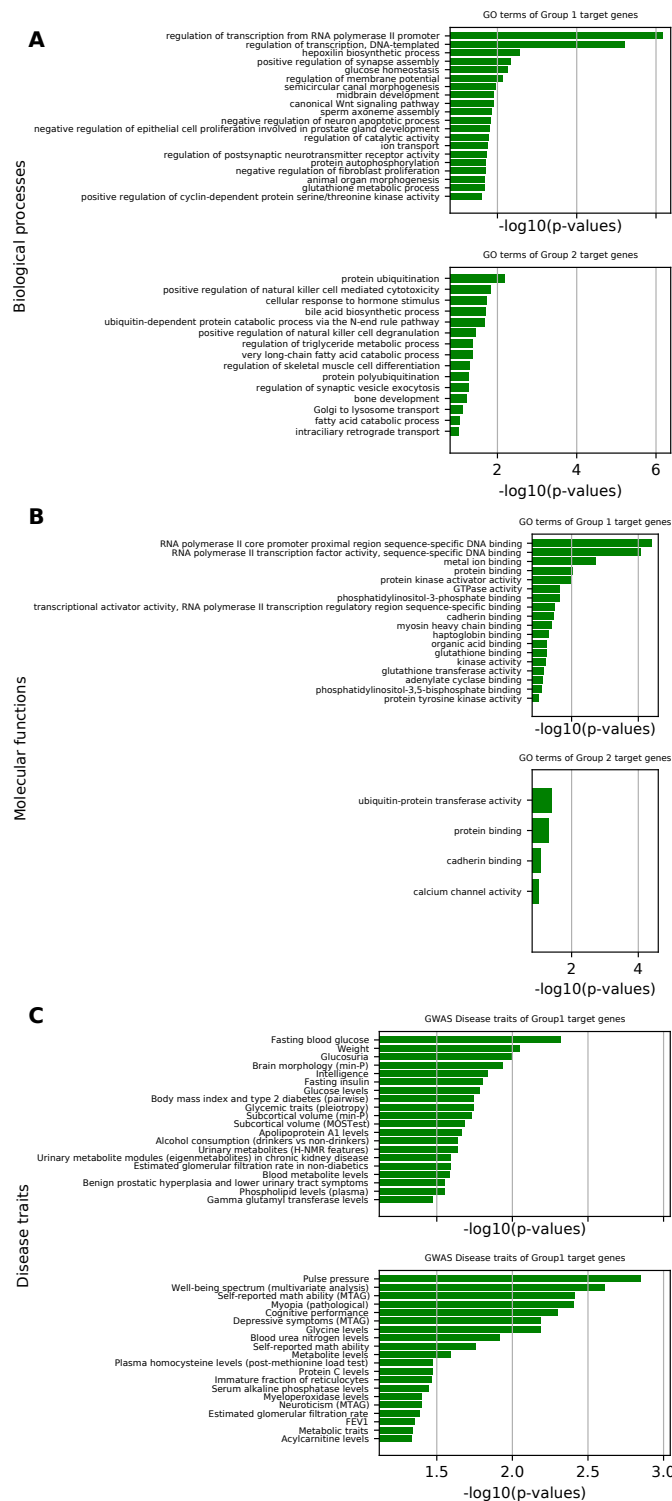


Figure B.14: GO enriched terms of Group 1 and Group 2 TF target genes (K562). The plots show enriched biological processes, molecular functions, GWAS disease traits terms for target genes of Group 1 and Group 2 TFs in K562 cell line.

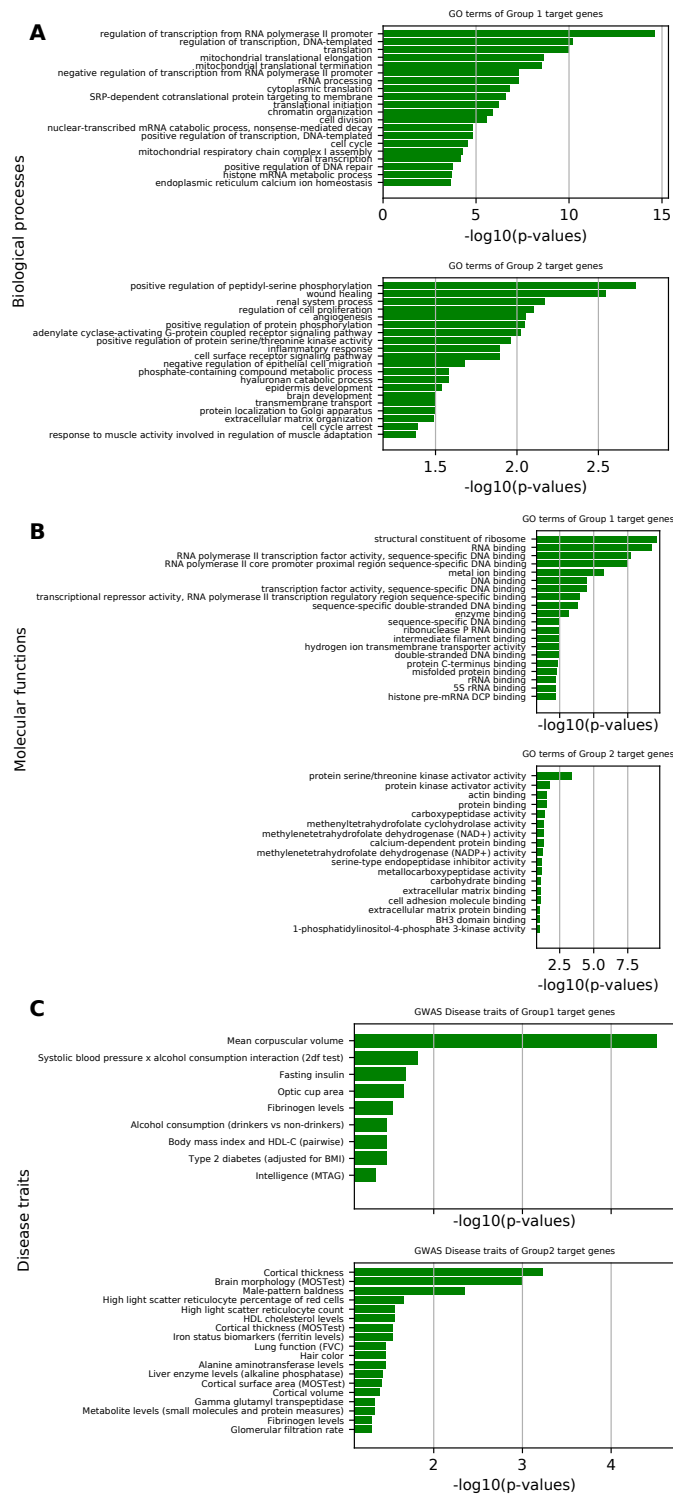


Figure B.15: GO enriched terms of Group 1 and Group 2 TF target genes (HeLa-S3). The plots show enriched biological processes, molecular functions, GWAS disease traits terms for target genes of Group 1 and Group 2 TFs in HeLa-S3 cell line.

Appendix C

Licence agreements

- The following two are licence agreement of figures [1.1](#) and [1.3](#) respectively from Elsevier.

**ELSEVIER LICENSE
TERMS AND CONDITIONS**

Jun 07, 2022

This Agreement between Mr. Rakesh Netha Vadnala ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5323790188865
License date	Jun 07, 2022
Licensed Content Publisher	Elsevier
Licensed Content Publication	Cell
Licensed Content Title	The Self-Organizing Genome: Principles of Genome Architecture and Function
Licensed Content Author	Tom Misteli
Licensed Content Date	Oct 1, 2020
Licensed Content Volume	183
Licensed Content Issue	1
Licensed Content Pages	18
Start Page	28
End Page	45

Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Investigating how chromatin regulates gene expression and cellular processes
Institution name	The Institute of Mathematical Sciences
Expected presentation date	Jul 2022
Portions	Figure 1
Requestor Location	Mr. Rakesh Netha Vadnala IMSc IV cross road, Taramani Chennai, 600113 India Attn: Mr. Rakesh Vadnala
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking

**ELSEVIER LICENSE
TERMS AND CONDITIONS**

Jun 08, 2022

This Agreement between Mr. Rakesh Netha Vadnala ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5324180255000
License date	Jun 08, 2022
Licensed Content Publisher	Elsevier
Licensed Content Publication	Cell
Licensed Content Title	The Human Transcription Factors
Licensed Content Author	Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, Matthew T. Weirauch
Licensed Content Date	Feb 8, 2018
Licensed Content Volume	172
Licensed Content Issue	4
Licensed Content Pages	16
Start Page	650
End Page	665
Type of Use	reuse in a thesis/dissertation

Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Investigating how chromatin regulates gene expression and cellular processes
Institution name	The Institute of Mathematical Sciences
Expected presentation date	Jul 2022
Portions	Figure 1A
Requestor Location	Mr. Rakesh Netha Vadnala IMSc IV cross road, Taramani Chennai, 600113 India Attn: Mr. Rakesh Vadnala
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

GENERAL TERMS

Bibliography

- [1] Tom Misteli. The self-organizing genome: Principles of genome architecture and function. *Cell*, 183(1):28–45, 2020.
- [2] Wendy A Bickmore. The spatial organization of the human genome. *Annual review of genomics and human genetics*, 14:67–84, 2013.
- [3] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.
- [4] Horng D Ou, Sébastien Phan, Thomas J Deerinck, Andrea Thor, Mark H Ellisman, and Clodagh C O’shea. Chromemnt: Visualizing 3d chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349):eaag0025, 2017.
- [5] Tom Misteli. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800, 2007.
- [6] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292–301, 2001.
- [7] Thomas Cremer, Marion Cremer, Steffen Dietzel, Stefan Müller, Irina Solovei, and Stanislav Fakan. Chromosome territories—a functional nuclear landscape. *Current opinion in cell biology*, 18(3):307–316, 2006.

- [8] Tom Sexton, Heiko Schober, Peter Fraser, and Susan M Gasser. Gene regulation through nuclear organization. *Nature structural & molecular biology*, 14(11):1049–1055, 2007.
- [9] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.
- [10] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L Van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.
- [11] Lis J. T. A 50 year history of technologies that drove discovery in eukaryotic transcription regulation. *Nature structural & molecular biology*, 26(9):777–782, 2019.
- [12] D. B. Nikolov and S. K Burley. Rna polymerase ii transcription initiation: a structural view. *Proceedings of the National Academy of Sciences of the United States of America*, 94(1):15–22, 1997.
- [13] Green MR Maston GA, Evans SK. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7, 2006.
- [14] Latchman DS. Transcription factors: an overview. *Int J Exp Pathol*, 74(5):417–422, 1993.
- [15] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.
- [16] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpää, et al. Mul-

- tiplied massively parallel select for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873, 2010.
- [17] Arttu Jolma, Jian Yan, Thomas Whittington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.
- [18] José M Franco-Zorrilla, Irene López-Vidriero, José L Carrasco, Marta Godoy, Pablo Vera, and Roberto Solano. Dna-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, 111(6):2367–2372, 2014.
- [19] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, 2014.
- [20] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–1812, 2012.
- [21] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays. *Nature genetics*, 36(12):1331–1339, 2004.
- [22] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–1435, 2006.

- [23] Artem Zykovich, Ian Korf, and David J Segal. Bind-n-seq: high-throughput analysis of in vitro protein–dna interactions using massively parallel sequencing. *Nucleic acids research*, 37(22):e151–e151, 2009.
- [24] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.
- [25] Huck-Hui Ng and M Azim Surani. The transcriptional and signalling networks of pluripotency. *Nature cell biology*, 13(5):490–496, 2011.
- [26] Richard A Young. Control of the embryonic stem cell state. *Cell*, 144(6):940–954, 2011.
- [27] Stuart H Orkin and Konrad Hochedlinger. Chromatin connections to pluripotency and cellular reprogramming. *cell*, 145(6):835–850, 2011.
- [28] U. Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8:450–461, 2007.
- [29] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.
- [30] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.
- [31] Markus J Herrgård, Markus W Covert, and Bernhard Ø Palsson. Reconstruction of microbial transcriptional regulatory networks. *Current opinion in biotechnology*, 15(1):70–77, 2004.
- [32] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.
- [33] Hong-Wu Ma, Bharani Kumar, Uta Ditges, Florian Gunzer, Jan Buer, and An-Ping Zeng. An extended transcriptional regulatory network of escherichia coli and

- analysis of its hierarchical structure and network motifs. *Nucleic acids research*, 32(22):6643–6649, 2004.
- [34] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *science*, 298(5594):799–804, 2002.
- [35] Stuart A Kauffman et al. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.
- [36] Yitzhak Pilpel, Priya Sudarsanam, and George M Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153–159, 2001.
- [37] Derek Y Chiang, Alan M Moses, Manolis Kellis, Eric S Lander, and Michael B Eisen. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biology*, 4(7):1–19, 2003.
- [38] Debopriya Das, Nilanjana Banerjee, and Michael Q Zhang. Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences*, 101(46):16234–16239, 2004.
- [39] Huai-Kuang Tsai, Henry Horng-Shing Lu, and Wen-Hsiung Li. Statistical methods for identifying yeast cell cycle transcription factors. *Proceedings of the National Academy of Sciences*, 102(38):13532–13537, 2005.
- [40] Nilanjana Banerjee and Michael Q Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic acids research*, 31(23):7024–7031, 2003.

- [41] Mamoru Kato, Naoya Hata, Nilanjana Banerjee, Bruce Futcher, and Michael Q Zhang. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome biology*, 5(8):1–13, 2004.
- [42] Andrew D Smith, Pavel Sumazin, Debopriya Das, and Michael Q Zhang. Mining chip-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, 21(suppl_1):i403–i412, 2005.
- [43] Nobuyoshi Nagamine, Yuji Kawada, and Yasubumi Sakakibara. Identifying cooperative transcriptional regulations using protein–protein interactions. *Nucleic acids research*, 33(15):4828–4837, 2005.
- [44] Franziska Reiter, Sebastian Wienerroither, and Alexander Stark. Combinatorial function of transcription factors and cofactors. *Current opinion in genetics & development*, 43:73–81, 2017.
- [45] J Omar Yáñez-Cuna, Cosmas D Arnold, Gerald Stampfel, Łukasz M Boryń, Daniel Gerlach, Martina Rath, and Alexander Stark. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome research*, 24(7):1147–1156, 2014.
- [46] Ekaterina Morgunova and Jussi Taipale. Structural perspective of cooperative transcription factor binding. *Current opinion in structural biology*, 47:1–8, 2017.
- [47] Alena Myšičková and Martin Vingron. Detection of interacting transcription factors in human tissues using predicted dna binding affinity. *BMC genomics*, 13(1):1–12, 2012.
- [48] Majid Kazemian, Hannah Pham, Scot A Wolfe, Michael H Brodsky, and Saurabh Sinha. Widespread evidence of cooperative dna binding by transcription factors in drosophila development. *Nucleic acids research*, 41(17):8237–8252, 2013.

- [49] Lisbeth Carstensen, Albin Sandelin, Ole Winther, and Niels R Hansen. Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):1–19, 2010.
- [50] Maria D. Chikina and Olga G. Troyanskaya. An effective statistical evaluation of ChIP-seq dataset similarity. *Bioinformatics*, 28(5):607–613, 2012.
- [51] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.
- [52] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.
- [53] Stefania Salvatore, Knut Dagestad Rand, Ivar Grytten, Egil Ferkingstad, Diana Domanska, Lars Holden, Marius Gheorghe, Anthony Mathelier, Ingrid Glad, and Geir Kjetil Sandve. Beware the jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis. *Briefings in bioinformatics*, 21(5):1523–1530, 2020.
- [54] Dan Xie, Alan P Boyle, Linfeng Wu, Jie Zhai, Trupti Kawli, and Michael Snyder. Dynamic trans-acting factor colocalization in human cells. *Cell*, 155(3):713–724, 2013.
- [55] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. Overview of next-generation sequencing technologies. *Current protocols in molecular biology*, 122(1), 2018.
- [56] Aquillah M. Kanzi, James Emmanuel San, Benjamin Chimukangara, Eduan Wilkinson, Maryam Fish, Veron Ramsuran, and Tulio de Oliveira. Next generation

- sequencing and bioinformatics analysis of family genetic inheritance. *Frontiers in Genetics*, 11, 2020.
- [57] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.
- [58] Lonsdale J, Thomas J, Salvatore M, and et al. The genotype-tissue expression (gtex) project. *Nat Genet*, 45:580–585, 2013.
- [59] Regev A, Teichmann SA, Lander ES, and et al. The human cell atlas. *Elife*, 6(e27041), 2017 Dec 5.
- [60] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, and et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45:1113–1120, 2013.
- [61] Anton Valouev, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5:829 – 834, 2008.
- [62] Ryuichiro Nakato and Toyonori Sakata. Methods for chip-seq analysis: A practical workflow and advanced applications. *Methods*, 187:44–53, 2021. Advance Epigenetics Methods in Biomedicine.
- [63] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.
- [64] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):1–10, 2009.

- [65] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.
- [66] Reuben Thomas, Sean Thomas, Alisha K Holloway, and Katherine S Pollard. Features that define the best chip-seq peak calling algorithms. *Briefings in bioinformatics*, 18(3):441–450, 2017.
- [67] Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in chip-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics*, 18(2):279–290, 2017.
- [68] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC genomics*, 10(1):1–15, 2009.
- [69] Bingqiang Liu, Jinyu Yang, Yang Li, Adam McDermaid, and Qin Ma. An algorithmic perspective of de novo cis-regulatory motif finding based on chip-seq data. *Briefings in bioinformatics*, 19(5):1069–1081, 2018.
- [70] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS computational biology*, 9(11):e1003326, 2013.
- [71] Morgane Thomas-Chollier, Andrew Hufton, Matthias Heinig, Sean O’keeffe, Nassim El Masri, Helge G Roeder, Thomas Manke, and Martin Vingron. Transcription factor binding predictions using trap for the analysis of chip-seq data and regulatory snps. *Nature protocols*, 6(12):1860–1869, 2011.
- [72] Guoliang Li, Liuyang Cai, Huidan Chang, Ping Hong, Qiangwei Zhou, Ekaterina V Kulakova, Nikolay A Kolchanov, and Yijun Ruan. Chromatin interaction analysis

- with paired-end tag (chia-pet) sequencing technology and application. *BMC genomics*, 15(12):1–10, 2014.
- [73] Elzo De Wit and Wouter De Laat. A decade of 3c technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.
- [74] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.
- [75] Zhihu Zhao, Gholamreza Tavossidana, Mikael Sjölander, Anita Göndör, Piero Mariani, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, et al. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, 38(11):1341–1347, 2006.
- [76] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.
- [77] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.
- [78] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yussouf Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, et al. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

- [79] Zhaohui S Qin, Lee Ann McCue, William Thompson, Linda Mayerhofer, Charles E Lawrence, and Jun S Liu. Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nature biotechnology*, 21(4):435–439, 2003.
- [80] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proceedings of the National Academy of Sciences*, 99(2):757–762, 2002.
- [81] Mirko Ronzio, Federico Zambelli, Diletta Dolfini, Roberto Mantovani, and Giulio Pavesi. Integrating peak colocalization and motif enrichment analysis for the discovery of genome-wide regulatory modules and transcription factor recruitment rules. *Frontiers in Genetics*, 11, 2020.
- [82] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, 2011.
- [83] Jason Piper, Markus C Elze, Pierre Cauchy, Peter N Cockerill, Constanze Bonifer, and Sascha Ott. Wellington: a novel method for the accurate identification of digital genomic footprints from dnase-seq data. *Nucleic acids research*, 41(21):e201–e201, 2013.
- [84] Sunil Kumar and Philipp Bucher. Predicting transcription factor site occupancy using dna sequence intrinsic and cell-type specific chromatin features. *BMC bioinformatics*, 17(1):41–50, 2016.
- [85] Qian Qin and Jianxing Feng. Imputation for transcription factor binding predictions based on deep learning. *PLoS computational biology*, 13(2):e1005403, 2017.

- [86] J. Keilwagen, S. Posch, and J. Grau. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol*, 20(9), 2019.
- [87] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- [88] G Li, L Cai, H Chang, and et al. Chromatin interaction analysis with paired-end tag (chia-pet) sequencing technology and application. *BMC Genomics*, 15(S11), 2014.
- [89] Guoliang Li, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, et al. Chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, 11(2):1–13, 2010.
- [90] Xiaoyan Ma, Daphne Ezer, Boris Adryan, and Tim J Stevens. Canonical and single-cell hi-c reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome biology*, 19(1):1–23, 2018.
- [91] Sourya Bhattacharyya Arya Kaul and Ferhat Ay. Identifying statistically significant chromatin contacts from hi-c data with fithic2. *Nature Protocols*, 15:991–1012, 2020.
- [92] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, et al. Ctf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.
- [93] Iain F Davidson, Benedikt Bauer, Daniela Goetz, Wen Tang, Gordana Wutz, and Jan-Michael Peters. Dna loop extrusion by human cohesin. *Science*, 366(6471):1338–1345, 2019.

- [94] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903, 2008.
- [95] Gangning Liang, Joy C. Y. Lin, Vivian Wei, Christine Yoo, Jonathan C. Cheng, Carvell T. Nguyen, Daniel J. Weisenberger, Gerda Egger, Daiya Takai, Felicidad A. Gonzales, and Peter A. Jones. Distinct localization of histone h3 acetylation and h3-k4 methylation to the transcription start sites in the human genome. *Proceedings of the National Academy of Sciences*, 101(19):7357–7362, 2004.
- [96] Bradley E. Bernstein, Emily L. Humphrey, Rachel L. Erlich, Robert Schneider, Peter Bouman, Jun S. Liu, Tony Kouzarides, and Stuart L. Schreiber. Methylation of histone h3 lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences*, 99(13):8695–8700, 2002.
- [97] D Schubeler, DM MacAlpine, D Scalzo, C Wirbelauer, and C Kooperberg. van, lf, gottschling, de, o’neill, lp, turner, bm et al.(2004) the histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*, 18:1263–1271, 2004.
- [98] Dmitry K. Pokholok, Christopher T. Harbison, Stuart Levine, Megan Cole, Nancy M. Hannett, Tong Ihn Lee, George W. Bell, Kimberly Walker, P. Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K. Gifford, and Richard A. Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–527, 2005.
- [99] Jean-Philippe Fortin and Kasper D Hansen. Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16(1):1–23, 2015.

- [100] Michael H. Nichols and Victor G. Corces. Principles of 3d compartmentalization of the human genome. *Cell Reports*, 35(13):109330, 2021.
- [101] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94, 2004.
- [102] Matys V, Fricke E and Geffers R, Gössling E, Haubrock M, Hehl Rand Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, and Wingender E. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–8, 2003 Jan 1.
- [103] Jagannathan V, Roulet E, Delorenzi M, and Bucher P. Htpselex—a database of high-throughput selex libraries for transcription factor binding sites. *Nucleic Acids Res*, 34(Database issue):D90–4, 2006 Jan 1.
- [104] Henry E Pratt, Gregory R Andrews, Nishigandha Phalke, Jack D Huey, Michael J Purcaro, Arjan van der Velde, Jill E Moore, and Zhiping Weng. Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Research*, 50(D1):D141–D149, 11 2021.
- [105] Hume MA, Barrera LA, Gisselbrecht SS, and Bulyk ML. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein-dna interactions. *Nucleic Acids Research*, 2014.
- [106] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, and Hughes TR. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–43, 2014.

- [107] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.
- [108] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.
- [109] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8(2):R24, 2007.
- [110] Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 10 2016.
- [111] Roberto Mosca, Arnaud Céol, Amelie Stein, Roger Olivella, and Patrick Aloy. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(D1):D374–D379, 09 2013.
- [112] M Zabidi, C Arnold, K Schernhuber, and et al. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518:556–559, 2015.
- [113] Charles-Henri Lecellier, Wyeth W Wasserman, and Anthony Mathelier. Human Enhancers Harboring Specific Sequence Composition, Activity, and Genome Organization Are Linked to the Immune Response. *Genetics*, 209(4):1055–1071, 05 2018.
- [114] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.
- [115] Bidossessi Wilfried Hounkpe, Francine Chenou, Franciele de Lima, and Erich Vinicius De Paula. Hrt atlas v1. 0 database: redefining human and mouse

- housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. *Nucleic acids research*, 49(D1):D947–D955, 2021.
- [116] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sallis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.
- [117] Justin Malin, Daphne Ezer, Xiaoyan Ma, Steve Mount, Hiren Karathia, Seung Gu Park, Boris Adryan, and Sridhar Hannenhalli. Crowdsourcing: Spatial clustering of low-affinity binding sites amplifies in vivo transcription factor occupancy. *BioRxiv*, page 024398, 2015.
- [118] Kazuo Satoh, Koichi Makimura, Yayoi Hasumi, Yayoi Nishiyama, Katsuhisa Uchida, and Hideyo Yamaguchi. *Candida auris* sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a japanese hospital. *Microbiology and immunology*, 53(1):41–44, 2009.
- [119] Soraya E Morales-López, Claudia M Parra-Giraldo, Andrés Ceballos-Garzón, Heidy P Martínez, Gerson J Rodríguez, Carlos A Álvarez-Moreno, and José Y Rodríguez. Invasive infections with multidrug-resistant yeast *Candida auris*, colombia. *Emerging infectious diseases*, 23(1):162, 2017.
- [120] Alba Ruiz-Gaitán, Ana M Moret, María Tacias-Pitarch, Ana I Aleixandre-López, Héctor Martínez-Morel, Eva Calabuig, Miguel Salavert-Lletí, Paula Ramírez, José L López-Hontangas, Ferry Hagen, et al. An outbreak due to *Candida auris* with prolonged colonisation and candidaemia in a tertiary care european hospital. *Mycoses*, 61(7):498–505, 2018.
- [121] Silke Schelenz, Ferry Hagen, Johanna L Rhodes, Alireza Abdolrasouli, Anuradha Chowdhary, Anne Hall, Lisa Ryan, Joanne Shackleton, Richard Trimlett, Jacques F

- Meis, et al. First hospital outbreak of the globally emerging candida auris in a european hospital. *Antimicrobial Resistance & Infection Control*, 5(1):1–7, 2016.
- [122] Snigdha Vallabhaneni, A Kallen, S Tsay, N Chow, R Welsh, J Kerins, SK Kemble, M Pacilli, SR Black, E Landon, et al. Investigation of the first seven reported cases of candida auris, a globally emerging invasive, multidrug-resistant fungus- united states, may 2013-august 2016. *American journal of transplantation: official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, 17(1):296–299, 2017.
- [123] Janaina de Cássia Orlandi Sardi, Diego Romário Silva, Maria José Soares Mendes-Giannini, and Pedro Luiz Rosalen. Candida auris: epidemiology, risk factors, virulence, resistance, and therapeutic options. *Microbial pathogenesis*, 125:116–121, 2018.
- [124] Jose Y Rodriguez, Patrice Le Pape, Olga Lopez, Kelin Esquea, Anny L Labiosa, and Carlos Alvarez-Moreno. Candida auris: a latent threat to critically ill patients with coronavirus disease 2019. *Clinical Infectious Diseases*, 73(9):e2836–e2837, 2021.
- [125] Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, and Yandell M. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, 18(1):188–96, 2008.
- [126] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.
- [127] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

- [128] Fredrik Ronquist and John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 08 2003.
- [129] Shivali Kapoor, Lisha Zhu, Cara Froyd, Tao Liu, and Laura N Rusche. Regional centromeres in the yeast *Candida lusitanae* lack pericentromeric heterochromatin. *Proceedings of the National Academy of Sciences*, 112(39):12139–12144, 2015.
- [130] Yafeng Zhu, Pär G Engström, Christian Tellgren-Roth, Charles D Baudo, John C Kennell, Sheng Sun, R Blake Billmyre, Markus S Schröder, Anna Andersson, Tina Holm, et al. Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of *Malassezia sympodialis*. *Nucleic Acids Research*, 45(5):2629–2643, 2017.
- [131] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.
- [132] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.
- [133] Laura N Rusche. Stable positions of epigenetically inherited centromeres in the emerging fungal pathogen *Candida auris* and its relatives. *Mbio*, 12(4):e01036–21, 2021.