Unraveling Cellular Heterogeneity and Phenotypic Drug Responses Using Chromatin Profiles

By

Neetesh

(PhD17205)

Under the Supervision of Dr. Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

August, 2022

Unraveling Cellular Heterogeneity and Phenotypic Drug Responses Using Chromatin Profiles

By

Neetesh

(PhD17205)

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree Of

Doctor of Philosophy

Under the Supervision of Dr. Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

August, 2022

Certificate

This is to certify that the thesis entitled "**Unraveling Cellular Heterogeneity and Phenotypic Drug Responses Using Chromatin Profiles**" being submitted by **Mr. Neetesh** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work, carried out by him under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2022

Dr. Vibhor Kumar

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

DEDICATION

This Thesis is Dedicated to
My Father,
Late S. S. Pandey

Acknowledgements

“None of us got to where we are alone. Whether the assistance we received was obvious or subtle, acknowledging someone's help is a big part of understanding the importance of saying thank you.”

Harvey Mackay

I bow down before ***"Almighty God,"*** *who has given me unending blessings, courage, perseverance, and knowledge to accomplish this task. He has also always blessed me with the right kinds of people, who have continuously supported, motivated, and inspired me during this journey.*

As my doctoral journey neared conclusion, it was a wonderful occasion to reflect on all the people that helped, believed in, and put out immense effort to make my trip a success. This portion of my work is not simply an acknowledgement; it also expresses my gratitude to everyone who helped me turn my aspirations into reality.

First and foremost, I am highly thankful to my supervisor ***Dr. Vibhor Kumar*** *for believing in me and enabling the completion of this doctoral thesis out of a deep sense of respect and gratitude. All of the work's stages benefited greatly from his warm encouragement, wise counsel, ongoing support, and inspiration. My doctorate study at IIIT Delhi was made possible by his own generosity. Every conversation I have with him inspires me to keep learning, develop, and improve society. He maintains a welcoming and comfortable work climate that fosters successful cooperation, and for that I want to compliment him.*

I am grateful to ***Prof. GPS Raghava****, Head, Department of Computational Biology, IIIT Delhi for his support, cooperation and providing all the necessary facilities within the department to carryout research assignments. I also pay my sincere thanks to IIIT founder Director,* ***Prof. Pankaj Jalote.*** *for establishing essential amenities, research funding and infrastructure for pursuing the research work. I would also want to express my gratitude to IIIT Director* ***Prof. Ranjan Bose*** *for his support and efforts at the facility and funding for the research work.*

I am highly thankful and indebted committee members ***Dr. Debarka Sengupta, Dr. Jaspreet Kaur Dhanjal*** *for their invaluable support and priceless suggestions throughout the degree. Also, thanks to all the faculty members* ***Dr. Sriram K., Dr. Ganesh Bagler, Dr. Angshul***

Majumdar, Dr. Subhadip Raychaudhuri, Dr. Tavpritesh Sethi, Dr. N. Arul Murugan, Dr. Gaurav Ahuja, Dr. Arjun Ray *for their constant support and help.*

Finally, I want to thank the administrative staff at IIIT-D, particularly ***Dr. Ravi Bhasin, Mrs. Priti Patel, Ms. Shipra Jain, and Ms. Anshu Dureja,*** *for always being available to answer our questions and handle our academic issues promptly. A special thanks to* ***Mr. Adarsh Aggarwal*** *for IT and computing support. I am also grateful to IIIT-D for providing first-rate infrastructure and facilities.*

Friends are real treasures of life, who always stand by your side in your brightest and darkest phases. I owe my special thanks to all the lab members and friends ***Madhu Sharma, Shreya Mishra, Omkar Chandra R., Dr. Smriti Chawla*** *for their unconditional and invaluable support. Then I want to also thank my senior and friend* ***Mr. Indra Prakash Jha*** *for his support and guidance. The warmest feeling I get is when I thank my adorable lab juniors-* ***Niharika, Durjay, Biswarup, Jaidev and Ramsha****.*

I am also grateful to friends ***Raghav Awasthi, Sarita Poonia, Shiju S, Shruti, Pradeep Singh, Samriddhi, Vivek Ruhela, Sumeet Patiyal, Rachesh Sharma, Chitrita Goswami, Nishant, Shubham, Ritu, Pooja, Akshaya, Anjali, Gayatri, Divyanshu Srivastava, Abhishek, Rohit, Meghal Dani, Ridam Pal, Dr. Rajesh, Dr. Sherry Bhalla, Dr. Shalini Sharma, Dr. Dilraj Kaur, Dr. Krishan Gupta, Dr. Aanchal Mongia, Harika,*** *Each one of you taught me something unique at the personal and professional front.*

My friends since forever: ***Dr. Arvind Yadav, Ms. Priyadarshini Rai, Dr Anant Rai, Dr. Rajesh Kumar Pathak, Dr. Piyush Aggarwal, Dr. Indra Singh and Dr. Himanshu*** *with whom I've had several sessions about almost everything under the sun. This journey would have been incomplete without either of your support.*

No words can ever be strong enough to explain the gratitude to my mother, ***Mrs. Nirmala Pandey*** *for her unconditional love, support and believing in me throughout my PhD journey. I am very thankful to my all sisters and brother in laws* ***Shashee & Ram, Kumud & Anup, Priyanka & Pritan, Shalini & Praveen and my brother Dr. Rajesh Pandey*** *who always support me in all possible ways. This acknowledgement would not be complete without expressing gratitude to one and only loving person, my late* ***father Sabha Shankar Pandey****, who might be* *watching me from heaven. What I am today is just because of them. This acknowledgement section is very small to describe his kindness, support, and showering of love.*

In the end, I want to express my gratitude and respect to those people who I am unable to name here but who were nonetheless important to our journey and will never be forgotten.

Neetesh

(Neetesh)

ABSTRACT

For effective treatment regimens, decisions should be based on specific genetic variability present across different human body cells by taking advantage of already accessible large-scale omics data like genomics, epigenomics, proteomics, and metabolomics databases. As of lately, cellular heterogeneity in phenotypic conditions (like cancer, neurodegenerative diseases, bone disease, metabolic disorders, and immune-related disorders) is inferred using genomic and epigenetic biomarkers for clinical diagnosis, patient stratification, prognosis and treatment monitoring.

For understanding regulatory changes due to disease and external stimuli in a cell, it is important to consider the role of chromatin structures as it is the regulation of the expression of the genes. But current existing datasets about chromatin interaction are derived from only a few cell-types, thereby providing limited insights for many cell-types. "Single-cell open-chromatin profiles" can be used to infer the pattern of chromatin-interaction in a cell-type. To study chromatin-interaction data for more cell-types, we developed a method called as "single-cell epigenome-based chromatin-interaction analysis (scEChIA)" that utilizes imputation of read-counts and refined L1 regularization for predicting interactions among genomic sites using "single-cell open-chromatin profiles". Unlike other methods scEChiA is not biased for only short-range interaction but it opens avenues for studying long-range chromatin interaction by using "single-cell open-chromatin profile". Using scEChIA, to predict chromatin interaction using "single-cell open-chromatin profile" of seven human brain cell types lead to identification of almost 0.7 million cis-regulatory interactions. Further analysis helped in finding the cell-type where there could be a connection to the known expression quantitative trait locus (eQTL) and their target genes the human brain. It also lead to the identification of possible target genes of human-accelerated-elements and disease-associated mutations.

Further analysis revealed connection between genes and expression quantitative trait locus (eQTL) across different cell-types of human brain and along with insights into the target genes of human-accelerated-elements and disease-associated mutations.

Due to availability of low amounts of relevant DNA and stochasticity, "single-cell open-chromatin profiles" have high drop-out rate and noise. To tackle this challenge, we developed a method called forest of imputation trees (FITs) to restore original signals from noisy and sparse single-cell open-chromatin profiles. Our algorithm, FITs is designed in such a way that it avoids bias during the restoration of read-count matrices. For this purpose it build forest of multiple imputation trees. FITs has resolved the challenging issue of recovering single-cell epigenome profiles without compromising the information at genomic sites with cell-type-specific activity. FITs-based imputation has not only improved the accuracy in the detection of enhancers but it has also increased reliability in estimating pathway enrichment score for every single-cell as well as predicting chromatin-interactions.

To utilize the knowledge of chromatin interaction, we propose an approach to study the activity of topologically associating domains (TADs) in cancer cell lines. A TAD is a self-interacting genomic region; DNA sequences within a TAD physically interact more frequently with each other than with sequences outside the TAD. TAD boundaries contribute to complex-trait heritability, especially for immunologic, hematologic, and metabolic traits. We have analyzed the variation in the activity of TADs in different phenotypic conditions across cell-types, creating a resource for understanding the role of chromatin interactions at different phenotypic contexts.

Our proposed methods can help utilize chromatin structure to highlight regulatory elements and genes that influence disease state and drug-response of cells for deciding hypothesis-driven therapeutics.

LIST OF PUBLICATIONS

Publications and Preprints

1. **Neetesh Pandey,** Omkar Chandra, Shreya Mishra, and Vibhor Kumar. "Improving Chromatin-Interaction Prediction Using Single-Cell Open-Chromatin Profiles and Making Insight Into the Cis-Regulatory Landscape of the Human Brain." ***Frontiers in genetics*** 12 (2021)
2. Sharma, Rachesh$^+$, **Neetesh Pandey**$^+$, Aanchal Mongia, Shreya Mishra, Angshul Majumdar, and Vibhor Kumar. "FITs: Forest of imputation trees for recovering true signals in single-cell open-chromatin profiles."***NAR genomics and bioinformatics*** 2, no. 4 (2020): lqaa091. ($^+$**Equal Contribution/co-first**).
3. **Neetesh Pandey**, Madhu Sharma, Arpit Mathur, George Anene Nzelu, Muhammad Hakimullah, Indra Prakash Jha, Omkar Chandra, Shreya Mishra, Ankur Sharma, Roger Foo, Amit Mandoli, Ramanuj DasGupta, Vibhor Kumar. "Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains." ***bioRxiv*** (2023).

Other publications and Preprints

4. Sharma, Madhu, Indra Prakash Jha, Smriti Chawla, **Neetesh Pandey**, Omkar Chandra, Shreya Mishra, and Vibhor Kumar. "Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs." ***Briefings in Bioinformatics*** 23, no. 4 (2022): bbac241.
5. Shreya Mishra, **Neetesh Pandey**, Smriti Chawla, Madhu Sharma, Omkar Chandra, Indra Prakash Jha, Debarka SenGupta, Kedar Nath Natarajan, Vibhor Kumar. "Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis." ***Genome Research*** (2023).
6. Chandra, Omkar, Madhu Sharma, **Neetesh Pandey,** Indra Prakash Jha, Shreya Mishra, Say Li Kong, and Vibhor Kumar. "Inferring functions of coding and non-coding genes using epigenomic patterns and deciphering the effect of combinatorics of transcription factors binding at promoters." ***bioRxiv*** (2022).

INDEX

LIST OF FIGURES

Hi-C profile. (B) The number of predicted interactions within 500kb in different cell-types by cicero and scEChIA [115]

2.4. *Evaluation of scEChIA's ability to predict cis-regulatory interaction with cell-type-specificity. (A) The cell-type-specific FPKM/expression (fold-change above-median across different cell-types) of top 200 genes with the highest relative interactions in the relevant cell-type of the figure panel. For example, in the panel with the label GM12878, the top 200 genes lying within the genomics regions with highest relative number of chromatin-interactions in GM12878 cells (compared to K562 and hESC) were chosen, and their expression levels in the 3 cell-types (GM12878, K562, and hESC) are displayed as box plots. These outcomes suggest that cis-regulatory interactions, which are linked to cell-type-specific expression, can be predicted by scEChIA. (B) The expressed in specific cell-type FPKM level of top 200 genes with the highest relative distal interactions (>500 kbp) in the corresponding cell-type of the figure panel. Numerous ditsal interactions (>500 kbp) predicted via scEChIA are also associated to expression that is specific to particular cell-types. *Significant, **Most Significant. Wilcoxon rank sum test was used to calculate p-value. [115]*

2.5. *Analysis of effect of predicted chromatin-interactions on gene-expression. Here the expression values of top 50 genes with highest and lowest number interactions at their promoter are shown for 3 cell-types (GM12878, K562, hESC) as boxplot. Such as in panel labeled as GM12878, the expression values of top 50 genes which have high or low number chromatin interactions are shown [115].*

2.6 *Inference regarding the known target genes for expression QTL (eQTL) in the human brain and associated cell-type of action. (A) The quantity of seven different types of brain cell-types' predicted chromatin interactions and brain eQTLs with target genes lying more than 500 kbp. (B) Screenshot of the UCSC browser demonstrating the association between an oligodendrocyte precursor's predicted long-range chromatin interaction and a brain eQTL and its target gene SOX10 [115].*

2.7. Findings regarding the target genes of disease related mutations in brain cells. (A) Enhancement of interactions between seven different types of brain cells and GWAS regions associated with mental disorders. P < 0.05 is denoted by the star (*). Using two proportion z-tests, the p-value was calculated. (B) An image from the UCSC browser depicting the promoter of the ALDH1A1 gene and the region with the GWAS SNP (rs3758354) in oligodendrocyte precursors (OPC) [115].

2.8. Snapshots of the UCSC browser for genes related with Human Accelerated Elements (HARS). (A) An estimated interaction between a genomic bin containing a human-accelerated region (HAR) and the SOX2OT promoter in microglia is shown in the UCSC browser snapshot. (B) The image displaying a connection between a HAR-containing area and the gene's promoter for the NRBF2 gene, which is linked to Alzheimer's disease [115].

2.9. Enrichment of motifs of transcription factor at locations with predicted chromatin-interactions in the endothelial cells. (A) The top 3 motifs enriched at chromatin interaction locations (both short and long-range). Also displayed is the p-value for enrichment in two categories of sites (all and only long-range). (B) The top three enriched motifs at locations with distal chromatin-interaction. Also displayed are the p-values for enrichment in two different types of sites. Notice that among top 3 enriched motifs in sites with predicted distal interaction(>500kb) none is enriched at genomic sites with all predicted interactions (which include short and long interactions). [115].

3.1. An explanation of FITs: FITs entails two phases. In phase 1, a large number of imputation trees are constructed to obtain several imputed versions of the original read-count matrix. Inside an imputation tree, each node first does a base method imputation on the non-imputed read-count matrix, then dimension reduction. After that, k-mean clustering is carried out to produce k clusters of cells. One daughter node receives the raw-read count of the cells of each cluster and applies the same imputation and clustering techniques. Every node of the imputation tree drops the sites that have zeros in all cells of its assigned raw read-count matrix. The vectors of the imputed *matrices (shown as red column) are compared to their unimputed original versions using correlation in phase 2 of FITs. For each cell, only the imputed versions that have the strongest connection to its un-imputed read-count vector are selected. (Sharma and Pandey et. al).*

3.2. *FITs capture signals from minor cell-types in an imbalanced data-set of single-cell ATAC-seq profiles from in vivo samples. Here, scATAC-seq read-count matrix for cells in adult mouse liver was used (A) t-SNE based embedding plot for unimputed (raw) and read-counts matrix imputed by different methods. (B) t-SNE based plots made using four other tools (chromVar, cisTopic and SCALE, scOpen) designed for scATAC-seq profiles (Sharma and Pandey et. al.)*

3.3. *Assessment of imputation methods in terms of separability of minor cells in mouse bone marrow scATACseq datasets (A) Visualisation of tSNE results for unimputed and imputed read-counts matrix. The frequency of different cell-types is also mentioned with their names. (B) Scatter plot of results from 4 tools (chromVar, cisTopic and SCALE, scOpen) designed for visualization of scATAC-seq profile. It is to be noticed that dendritic cells and T cells could not cluster together in any of the plots, like FITs-based results (Sharma and Pandey et. al.)*

3.4. *Figure shows how FITs improves analysis utilizing the scATAC-seq profile-calculated gene-set enrichment score for each single cell. (A) The fraction of cells with the correct cell-type terms present in the top five enriched gene-sets, as determined by UniPath calculations. The scATAC-seq profile of cells from Cusonovich et al., was imputed using FITs [130]. (B) Pathway scores generated using read-counts imputed using FITs are used in the t-SNE-based visualisation. (C) A visualization of the t-SNE results for pathway scores derived from UniPath's unimputed read-count (Sharma and Pandey et al.).*

3.5. *The prediction of chromatin-interaction using the scATAC-seq profile is improved by FITs-based imputation. This chart illustrates the fraction of predicted interactions that* *overlap with HiC's chromatin-interaction predictions based on co-accessibility. (A) for GM12878 cells; (B) K562 cells (Sharma and Pandey et al.).*

4.1. *A summary of the chromatin interaction profiles of 4 HNSC cell lines and their profiling and analysis. While the other two are models of primary cancer cells from patients HN137 and HN148, the other two cell lines are from the same patient, HN120. TAD boundaries were obtained and additional downstream analysis was performed using the Hi-C based chromatin interaction profile from HNSC cell lines*

4.2. *A study on the chromatin interaction profiles and the domain activities of cell lines originating from patients with HNSC cancer. (A) The fold change and significance (p-value) of the activity of TAD between 137P and its cisplatin-resistant variant 137PCR are shown in volcano plot. (B) The volcano plot of fold change and significance (p-value) for TAD activity differences between the 137M cell line and its cisplatin counterpart. (C) A visualization of the chromosome 11 interaction matrix. The region inside the middle-positioned rectangle represents the downregulated TADs (in the 11q13.3 band) found in all three HNSC cell lines. Other areas that interact distally with the TAD in the 11q13.3 band are indicated by arrows as R1, R2, R3, and so on. D) The 11q13.3 band and the R1, R2, R3, R4, R5, R6, R7 and R8 area include elevated TAD and the wrenched motifs in H2K27ac. Enriched gene-ontology keywords of genes located in various locations that interact with the 11q13.3 band.*

4.3. *Preservation of TAD boundaries and pattern of drug response-activity association. (A) Visualization of TAD boundaries and chromatin interaction from two cell lines, (GM12878 and HN137P). For two cell lines, the TAD boundaries of the chromatin interaction in chromosome 10 are depicted in a zoomed-in form (green and blue). In the two cell lines, the TAD borders are almost identical. B) Overlapping TAD border areas between several cell types (within 1 KB of the boundary). The proportion of borders overlapping between two cell lines is displayed on the Y axis. (C) From the CCLE database, a heatmap was created to show the relationship between TAD-activity and PIC50 values for drugs in HNSC cell lines. Only TADs (in rows) whose activity has an* *absolute correlation value of at least 0.3 with the PIC50 of at least one drug are displayed. Similar to this, only drugs that have an absolute correlation value of greater than 0.3 with the activity of at least one TAD are displayed. Here, four drug sub-clusters are also highlighted.*

4.4. *Heatmap shows the relationship between drug PIC50 values and TAD-activity for cell lines from different types of cancer in the CCLE database. Only TADs (rows) or medicines (columns) with at least one absolute correlation value greater than 0.3 are displayed. A few drug sub-clusters are also displayed*

4.5. *A. Barplot illustrating the best correlation between the PIC50 values of a few drugs and single gene expression (blue) or TAD-activity (green). B. A bar graph showing the association between the PIC50 values that were actually obtained and those that were predicted using a machine learning model and features that were either TAD activity or gene expression. Even when many genes are utilised to estimate PIC50 values for various drugs, the efficiency (or accuracy) was much lower than the result obtained by TAD-activity score*

4.6. *TAD activity in different cancers TCGA samples and its relationship to survival. Using TCGA data, a heatmap showing the fold increases in the median TAD activity (tumour vs. normal sample) for each of the 16 cancer types was created.*

4.7. *(A) Kaplan Mier (KM) plot for survival rate for head and neck cancer (HNSC) patients. Here KM plot was made for two groups of HNSC patients; high - top 80 patients sample where activity of TAD (hg19:chr22-391000000_39825000) was high and low- 80 patients with lowest activity of TAD. On the right panel, the top enriched ontology terms for genes in the TAD, are shown hand in hand with involved genes. (B) KM-plot for survival rate for HNSC patients based on activity of TAD (hg19:chr11-70150000_71300000) and top enriched terms for involved genes are shown. (C) KM-plot for survival rate for HNSC patients based on activity of TAD (hg19:chr7_51500000_57100000) is shown. The enriched terms for genes inside that TAD are also shown.*

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Overview of genomics and epigenomics

Through whole-genome identification, molecular characterization, and cloning, the multidisciplinary field of genomics seeks to understand the function, structure, and evolution of genes as well as their relationships, with the ultimate goal of understanding organism phenomics. Genomic research includes identifying the overall number of genes, counting transcripts and proteins that an organism encodes, analyzing their interactions, and understanding metabolic pathways. Genomics can be further studied under three subdomains: structural, functional, and comparative. Contrary to the gene-by-gene approach of traditional molecular biology techniques, the area of functional genomics uses genome-wide approaches to make an effort to comprehend the activities and interactions of genes and proteins. It incorporates information gathered from many processes that affect DNA sequence, gene expression, and protein function, involving protein-DNA, protein-RNA, protein-protein interactions, coding, non-coding transcription, and protein translation. These pieces of information are used to create dynamic and interactive networks to understand gene expression and cell differentiation, including cell cycle progression (Bunnik and Le Roch; Skolnick and Fetrow). The aim of using structure to assign numerous levels of function, which is a lengthy process and one that has gained importance with the development of structural genomics, is being achieved with this first stage (Skolnick and Fetrow). In addition, a rapidly growing field of comparative genomics has produced impressive findings. The availability of numerous fully sequenced genomes has made comparative genomic analysis possible. The ability to compare the entire genomes of various organisms enables the development of global perspectives on genome evolution, and the availability of many genomes that have been completely sequenced increases the predictive power for uncovering the latent information of genome structure, function, and evolution. As a result, comparing human genes with genes from other genomes in a genomic landscape may aid in determining novel roles for genes that are not yet annotated (Sivashankari and Shanmughavel).

Epigenetic processes include “DNA-methylation, RNA-methylation, covalent histone modifications, and chromatin assembly states”. These epigenetic processes influence gene-expression without establishing immediate alterations to the DNA sequence (Sakurada; Zhao et al.). Local variations in those epigenetic markers between individuals in a population, known as epigenetic variants or epialleles, might have comparable dynamics to genetic variants. Because epigenetic mechanisms underpin an organism's ability to respond to alternations in the environment, certain “epigenetic-marks” connected to these responses are progressively vulnerable to environmental input. Whereas others “epigenetic-marks” linked to processes like embryonic development, core cellular functions with state changes like differentiation, are more stable (Barrera-Redondo et al.). In a procedure, epigenetic changes are frequently passed down from generation to generation and studied under the topic of trans-generational epigenetic inheritance. The overall significance of “transgenerational epigenetic inheritance” in the evolutionary process of plants and animals is currently under discussion. However, several occurrences in both plants and animals were reported. (Heard and Martienssen).

Many epigenetic techniques have been useful in understanding gene regulation and cell identity better. Different levels of functionality for epigenomic traits can be assessed using different techniques discussed in the subsequent subsection. Data integration allows us to infer functionality from complex data sets, and epigenomic profiling provides a descriptive perspective of the chromatin landscape. A way to directly assess the functionality of epigenomic characteristics is made possible by epigenome editing. Recent studies have demonstrated the universal application of epigenome editing and its ability to regulate gene expression. (Stricker et al.) (Figure 1.1).

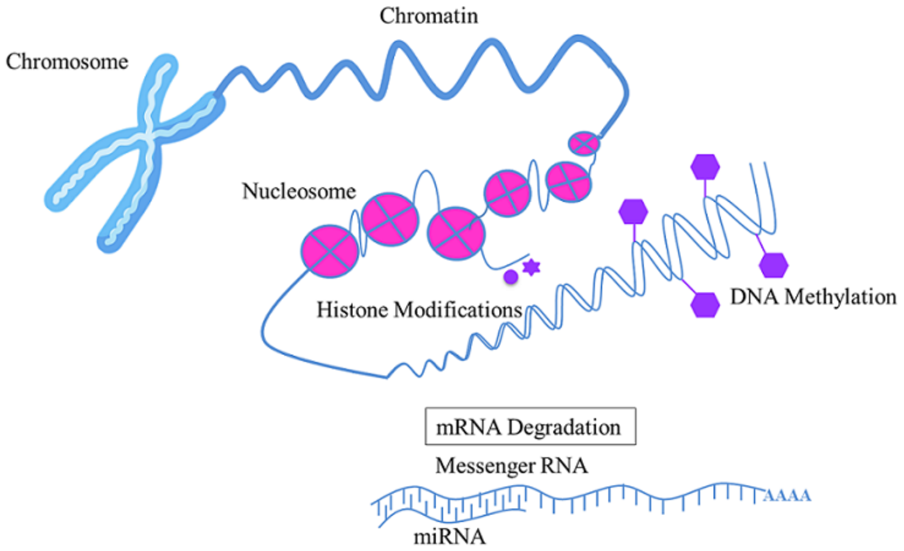


Figure 1.1. Chromatin structure and epigenetic regulation of gene expression

1.1.2 Fundamentals of chromatin profiling

Both structurally and functionally, the chromatin can be labeled as active or repressive. The open-chromatin regions, also known as euchromatin, are known to be associated with active gene regulatory mechanisms, whereas the heterochromatin is nucleosome-dense and crucial for defining transcriptionally-inactive regions. These two main structural groups of chromatin are associated with different histone modifications. A fundamental building block of chromatin is a nucleosome, an octamer of the histone protein that wraps ~147 base pairs of DNA at regular intervals throughout the genome. Chromatin accessibility approaches, such as "MNase-seq (micrococcal nuclease digestion of chromatin followed by sequencing)" and "DNase-seq (DNase I hypersensitive sites sequencing)", can be used to assess the accessibility of chromatin "unwound open-chromatin" "and/or" the arrangement of nucleosomes at genomic locations, which represent their regulatory potential. In order to allow RNA polymerases or transcription factor (TF) binding in modulating gene expression, active regulatory areas are typically assumed to be depleted of nucleosomes. Additionally, "DNA breathing"—a quick unwrapping of the DNA around nucleosomes to permit the binding of regulatory factors—can occur (A. Chawla et al.; G. Li et al.). ATP-dependent chromatin remodeling complexes, like; “Switch/Sucrose non-fermentable and the nucleosome remodeling and deacetylase complex”, increase the stability of nucleosome post-translational changes (A. Chawla et al.). The “N-terminal tails” and core of histone-proteins can undergo covalent modifications more frequently due to histone remodeling enzymes like; “histone-acetyl- or methyl-transferases” (Bannister and Kouzarides). Overall, histone modification pattern/s at regulatory locations, including enhancers and promoters, could affect local chromatin accessibility to transcription factor/s while controlling proximal transcriptional or activity of genes. On the other hand, nucleosome redistributing and extensive histone alterations, either in a direct or indirect way., lead towards remodeling of chromatin-accessibility patterns. Also, they may impact the long-range regulation of gene expression, which can be assessed by assays for open-chromatin and chromatin-interaction. Additionally, those interactions may be reversible (for example, to preserve cellular activities) or irreversible (for example, during neurodevelopment) to define cell lineages (Patrick et al.).

1.1.2.1 Techniques of 3D chromatin interactions

Structure comes before function in biology; for instance, chromosomal territories separate “gene-rich” and “gene-poor” regions and its reorganization was observed under diseased conditions (Borden and Manuelidis). In addition, interference with three-dimensional (3D) chromatin interactions in gene regulatory areas may have functional repercussions. Point mutations, for instance, have been discovered to repel the development of chromatin loops between gene promoter-terminator sequences in the RNA polymerase II-related transcription factors, disabling numerous transcriptional cycles and changing gene expression (Kadauke and

Blobel). Further research is necessary to determine whether such chromatin loops are disturbed spatially as well as temporally when intervening in complicated illnesses.

1.1.2.2 Chromosome conformation capture methods

Chromosome organization in the nucleus is well understood, and this spatial arrangement of the genome is important for both controlling genes and keeping the genome stable. To understand the fundamental mechanism behind genome architecture, a variety of techniques have been created and put to use, including chromosome conformation capture (3C) and methodologies deriving from 3C. High-throughput chromatin architecture experiments at the "genome-scale" can be carried out using 3C and 3C-derived techniques (Kadauke and Blobel; Han et al.).

3C:- one to one mapping

The 3C approach for determining genomic architecture is established on quantification of the frequency of connections among distal DNA segments within cell populations (Louwers et al.). Unlike cytogenetic techniques, the 3C-based genomics method produces unmatched information-dense data. Also, explaining the genome topology of the entire genome, allowing better-structured genome topology analyses at a more heightened resolution and throughput, delivering significant insight into genome organization and dynamics and its influence on genome role. In order to study the organization of the genome from individual loci to the entire genome, 3C technologies are evolving this process (Han et al.).

4C:- one to many mapping

The follow-up to 3C, 4C, is much better than the original. The fundamental 3C protocol is improved upon in various ways by 4C. The ability to detect unknown DNA regions interacting with the region of interest is its most significant benefit. Until the crosslinks are switched around, the 4C technique is identical to the 3C procedure. A second, typically cutting restriction enzyme is then used to treat the DNA, and it has a distinctive recognition sequence than the first one. The known DNA of interest and its interacting DNA are produced as sticky-ended fragments that have the potential to circularize. After that, the unknown DNA is amplified outward around the circle utilising primers developed to bind the known DNA. This library can then be described by microarray or DNA sequencing. The size of 4C over 3C is significantly increased by its compatibility with whole-genome technologies and capacity to amplify unknown interactors. Similar to 3C, 4C features a lot of background noise (Han et al.; Dekker).

5C:- many to many mapping

For in-depth analysis of interactions with specific loci of interest, use 5C. In order to anneal across the ligated junction of the DNA fragments, 5C uses specialized primers. All ligation products can be amplified because the primer tails contain a universal sequence. Using microarray or sequencing, this carbon copy library is described. 5C does not take the place of 4C. First of all, because not all sites allow the design of 5C primers, 4C can have a greater resolution. Additionally, because each primer needed for 5C must be designed, it is not possible to examine the complete genome because millions of primers would be required (Dostie and Dekker).

Hi-C:- all to all mapping

A technique known as “Hi-C”, a variation of the 3C-method, allows scientists to determine the interaction frequency to essentially every locus in a genome in a high throughput way as well as on a genome-wide scale. (Oluwadare et al.). Thus, “Hi-C” allows experimenters to characterize read-pairs interactions based on an “all-versus-all”, that is, to profile interactions for every pair of reads within a genome using next-generation sequencing techniques. The sequenced paired-end reads from the “Hi-C” assay are used to Determine the frequency of chromosomal interactions. For example; the frequency of intra-chromosome interaction, or the frequency of interactions between distinct chromosomes, known as the inter-chromosome interaction frequency. The sections of a chromosome into which it has been fragmented are called fragments, sometimes known as bins or genomic loci. The amount of base pairs (bp) that make up each fragment gives it a specific length or size. The resolution determines the size of the fragment; for example, a 1 MB resolution means that each fragment contains 1,000,000 bp. The acquired IFs are typically displayed as a 2-D matrix called a contact matrix. In that contact matrix; where the rows, as well as columns, correspond to the number of chromosomal or genome fragments. The interaction frequency/s produced by the "Hi-C" approach can be utilized to build 3D chromosomal and genome architectures, making it particularly relevant. These structures in turn aid in the explanation of a number of phenomena, including the regulation of genes, genome-foldings and the relationship with regulatory components and higher-order structural elements in the cell nucleus (Misteli; Cremer and Cremer; de Laat and Grosveld; Ron et al.; Fraser and Bickmore).

1.1.3 Open chromatin accessibility in single cell

The massive analysis of open-chromatin in cancer cohorts was made possible by the assay for "transposase-accessible chromatin utilizing sequencing (ATAC-seq)". ATAC-seq became the standard approach for detecting open-chromatin due to its ease of use and minimal cell quantity requirements. Additionally, a detailed analysis of the digestive processes by the enzyme Tn5 provided information on regulatory elements like nucleosome locations (Buenrostro, Paul G. Giresi, Zaba, Chang, et al.; Corces, Granja, et al.; Schep, Buenrostro, et al.), transcription factor binding sites, and TF activity levels (Z. Li, Schulz, et al.).Determining the status of accessible chromatin in thousands of single cells via both normal and diseased tissues, single-cell sequencing in combination with ATAC-seq (scATAC-seq) substantially broadened the applicability of ATAC-seq (Buenrostro, Wu, Litzenburger, et al.; Buenrostro, M. Ryan Corces, Lareau, Wu, et al.; Domcke et al.; Satpathy et al.; Z. Li, Kuppe, et al.; Thornton et al.). In order to ease library construction through transposition *in situ* and to provide a chromatin accessibility signal, single-cell ATAC-seq requires the isolation and processing of nuclei while maintaining the nuclear scaffold (Thornton et al.).

1.1.4 Noise and bias in the single-cell ATAC estimation

The emerging novel technology known as “single-cell assay of transposase-accessible chromatin followed by sequencing (scATAC-seq)” enables the investigation of gene regulation with single-cell resolution. Bulk sequencing was the standard technique before single-cell sequencing was introduced. It made it possible to investigate the transcriptome (RNA) and genome (DNA), among other omics. Since the advent of single-cell sequencing, scientists have been able to forecast the progression of diseases better, explain the mode of action of specific medications, identify novel cell-types, and do much more. The technology has tremendous potential for treating conditions like cancer and neurodegeneration (Figure 1.2). Even within the same cell-type, the data from “scATAC-seq” are uniquely sparse, binary, and widely varied. The signals in “scATAC-seq” data have low similarity and are sparse and noisy across cells. Additionally, since the majority of the genome is only present in two copies per cell and the transposase can only cleave and add adaptors once per copy, only two sequenceable fragments, or two reads per locus, can be produced for each open-chromatin area after removing PCR duplicates. Because of this, its most binary nature is “scATAC-seq” data, which denotes an open/closed status. (Urrutia et al.).

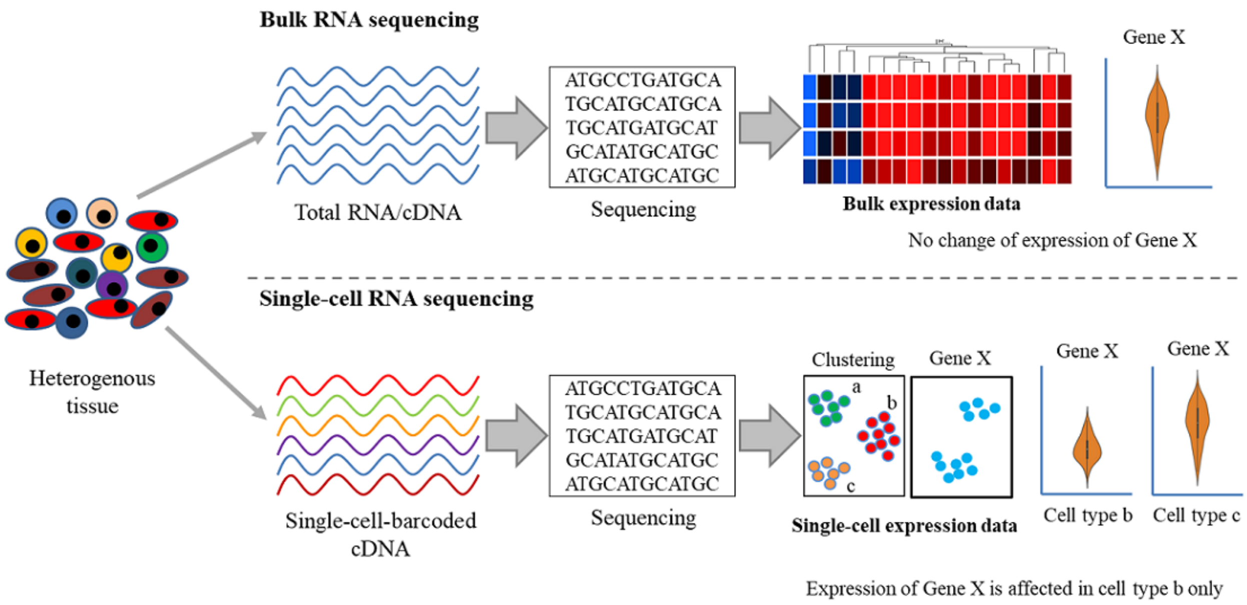


Figure 1.2. Overview of Single-cell Vs. Bulk sequencing

1.1.5 Applications based on cellular phenotypes utilizing chromatin and epigenome profiles

Due to the differential expression of genes, multicellular organisms' cells are genetically homogeneous yet morphologically and functionally varied. Allelic differences brought on by mutations in the DNA sequence provide the basis for the traditional Mendelian inheritance of phenotypic traits. Other genetic events, however, demonstrate non-Mendelian inheritance patterns, including X chromosome inactivation while in initial embryonic growth in female mammals, position-effect striation in flies, along with chromosomal imprinting. The phrase "epigenetic landscape" was used by Conrad Waddington to refer to "the interactions of genes with their environment, which bring about the phenotype" (Wang and Chang; Waddington). The term "epigenetics" was first coined about a century ago, and after that scientists, doctors, and others investigated dark depths of the genome to find clues which suggested that gene function could be affected by multiple other factors also besides sequence variations. There are mulitple evidences which connect diseases, behaviors, and other health indicators with epigenetic changes. Such as researchers have revealed some epigenetic states associated with numerous types of cancer, cognitive dysfunction, and respiratory, autoimmune, cardiovascular, reproductive, and neurobehavioral disorders. Epigenetic processes can also be triggered by many kinds of exposure to cells such as heavy metals, pesticides, polycyclic aromatic hydrocarbons, tobacco smoke, hormones, radiation, viruses, bacteria, and essential nutrients (Weinhold). DNA methylation, histone modification, chromatin remodeling, and non-coding-RNA-associated pathways are a few of the best-described epigenetic changes hypothesized to make and maintain epigenetic adaptations (Egger et al.). Early research established the concept of an "epigenetic code" that controls the state of chromatin and, as a result, expression of genes. These studies demonstrated that “heterochromatin and euchromatin” are linked to distinctive DNA-methylation and histone-modification trends that associate with specific gene activity states. For many diverse processes in mammals, including development, cell differentiation, and proliferation, epigenetic regulation is essential. A vital research objective with significant consequences for human health is the comprehensive knowledge of the networks of regulation and epigenetic processes underlying context-specific gene expression programs as well as cellular phenotypes (Wang and Chang; Y. Wang et al.; Zhu et al.).

1.2 Epigenetics And Chromatin-Based methods as Transcriptional Regulator And Their Cellular Diversity

1.2.1 Spatial structure of the human genome

In vivo, the human genome functions as a complexly folded, 3D chromatin polymer. Therefore, a key component of comprehending how genes are controlled in healthy conditions. Also, gene dysregulation in diseases is comprehending how the human genome is spatially organized and folded within the cell nucleus. Now, set light microscopic-based techniques, along with more contemporary molecular 3C techniques, are working together to provide us with undiscovered unattainable comprehension of this intriguing area of genomics (Egger et al.; Bickmore). Only taking into account the main DNA sequences, as well as linear mappings of post-translational histone alterations that correspond to whether transcription is active or not, is insufficient to comprehend how the human genome functions (Ecker et al.). Instead, research into and comprehension of the spatial organization and 3D folding of chromosomes within the nucleus is needed to fully comprehend how the genome functions in vivo. “Fluorescence in situ hybridization (FISH)” has historically been used to visualize the location and organization of chromosomes, chromatin domains, and specific genes in this field of study. Excitingly, intramolecular DNA ligation and cross-linking are being used in high-throughput molecular tests for determining the spatial relationships of different loci, including complete genomes. These methods are an evolution of the first “chromosome conformation capture (3C)” approach (Bickmore; de Wit and de Laat).

Within a eukaryotic nucleus, the genome is compartmentalized and organized according to two main principles. First, separate compartments are frequently formed within the nucleus by chromosomal areas with comparable biochemical and functional characteristics (placed either on the same or on different chromosomes). Second, TADs which are genomic segments that exhibit extensive self-interactions and are spatially isolated from neighboring segments, are used to partition interphase chromosomes. Within the nucleus, chromatin compartments are discrete genomic areas with different biochemical and functional characteristics. Microscopy has been used extensively to identify and study the majority. Long DNA segments that are frequently referred to as domains are usually found in these compartments. The nucleolus, which is organized around rRNA gene repeats, is the archetypal nuclear compartment. Here, RNA Polymerase I synthesizes rRNA, and ribosomes are put together. Today, it is believed that liquid phase separation contributes to the compartmentalization of nucleoli. Rather than being solid aggregates, nucleoli now seem to be fluid, droplet-like structures that separate from the rest of the nucleoplasm due to their unique physicochemical characteristics (van Steensel and Furlong; Brangwynne et al.).

A eukaryotic organism's genome comprises a supra-molecular complex comprising three-dimensional (3D) structures and chromatin fibers. Gene expression is influenced by chromosomal interactions and topological modifications based on environmental as well as developmental factors. Genome integrity, gene expression, and DNA replication are all significantly influenced by chromatin architecture. Chromosomal territories, A-compartment /B-compartments, chromatin loops and TADs are higher-order chromatin arrangements. These higher-order chromatin arrangements differ between cells, tissues, and species relying on the developmental stage as well as environmental factors. In the interphase nucleus of most eukaryotes, each chromosome has its own territory and makes up the topmost layer of the hierarchical structure. TADs are the structural building blocks of chromatin, whereas the A-compartment and B-compartments are related to active "euchromatic" and inactive "heterochromatic" chromatin, respectively, with clearly specified genomic and epigenomic properties. The chromatin activity, which changes under environmental stresses as a result of the relocalization of the architectural proteins, is correlated with chromatin architecture such as TADs, in addition to the local interactions along with promoter and regulatory elements (S. Kumar et al.). Heterochromatin and euchromatin, the other two noticeable nuclear compartments, were initially identified based on variations in apparent compaction, as seen under a microscope. Generally, transcribed parts of the genome are euchromatic, whereas transcriptionally inactive or repressed regions are heterochromatic. Based on the linked protein sets and histone modifications, chromatin types are in various flavors. Particularly, trimethylated H3K27 or di- or trimethylated H3K9 often serve as markers for heterochromatin (van Steensel and Furlong; Heitz; Filion et al.; ENCODE Project Consortium).

1.2.2 High throughput chromatin accessibility

Determining the epigenome of phenotypically different states of cells inside complicated primary tissue is a key challenge in systems biology. In order to achieve this, single-cell chromatin accessibility measurements, which record the physical accessibility of alleged functional components throughout the genome, offer a crucial epigenetic perspective of the regulatory environment within individual cells (Mezger et al.; Corces, Buenrostro, et al.; Cusanovich, Reddington, et al.; Buenrostro, Wu, Litzenburger, et al.; Pott; Jin et al.). Combinatorial indexing techniques show great promise for ultra-high throughput accessibility profiling applications, but they catch less accessible fragments in each cell alongside single-cell isolation techniques (Buenrostro, Wu, Litzenburger, et al.) and cannot be integrated by single-cell microscopy or different multi-omic assays that need complete, living cells (Mezger et al.).

Through the use of the following methods, several gene regulatory systems can be examined:

1.2.2.1 DNase I hypersensitive sites sequencing (DNase-seq)

Determining chromatin accessibility and its underlying regulatory lexicon have been extensively studied using “deoxyribonuclease I (DNase I)-hypersensitive site sequencing (DNase-seq)”. An enzyme called deoxyribonuclease I (DNase I) desultorily cuts DNA. Chromosome DNA is packaged in eukaryotes as a chain of nucleosomes that repeats itself on a regular basis. These nucleosomes will prevent DNase-I from efficiently nicking DNA, making the “accessible nucleosome-free regions” more vulnerable to DNase-I cleavage. The nucleosome state of activated genes is greater likely to be changed (Weintraub and Groudine; Elgin). That renders DNase-I digestion an appropriate reference test for identifying genomic regulatory elements. The interest in DNase I investigations quickly peaked in the 1980s but then slowly declined because the term "DNase I-hypersensitive sites" (Elgin) was developed to designate open-chromatin-regions. The largest obstacle to continued development seems to be the absence of high-throughput analysis, as the data produced by conventional approaches was insufficient to reach meaningful findings at a genome-wide level. But with the advent of next-generation sequencing, interest in DHS profiling has lately increased (NGS) “DNase I-hypersensitive site sequencing (DNase-seq)” is a newly developed approach that was made possible by combining high-throughput sequencing with DNase-I digestion (Boyle et al.). DNase-seq enables genome-wide mapping of DNase-I cleavage processes at nucleotide resolution and shows an enhanced “signal-to-noise” ratio comparison to its forerunners (Weintraub and Groudine; Boyle et al.; Y. Liu et al.).

1.2.2.2 Formaldehyde Assisted Isolation of Regulatory Elements with Sequencing (FAIRE-seq)

An approach is describe as “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements)” for separating “nucleosome-depleted DNA” via human chromatin. Chromatin is sheared by sonication, crosslinked along formaldehyde *in-vivo*, and pulled using phenol-chloroform. Fluorescently tagged DNA that was recovered from the aqueous phase is hybridized to a “DNA microarray”. When FAIRE is used in human cells, DNA that is close to “DNase I hypersensitive sites, transcriptional start sites, and active promoters” is substantially enriched. Additionally, there is proof of cell-type-specific FAIRE enrichment patterns given. FAIRE can be used as a positive selection method of genomic areas linked to regulatory activity, as well as those that are typically picked up by assays that measure nuclease hypersensitivity (Giresi et al.).

1.2.2.3 Micrococcal Nuclease Digestion of Chromatin Followed by Sequencing (MNase-seq)

A popular technique for “mapping nucleosome positions and occupancy” affects the “digestion of chromatin with micrococcal nuclease (MNase)”. MNase is an “endo-nuclease” and “exo-nuclease”, which selectively breaks down the naked DNA among nucleosomes, and releases the nucleosomes from chromatin as well as enriched DNA fragments saved by nucleosomes. The resultant undigested DNA undergoes high-throughput sequencing “MNase-seq” and is mapped to a reference genome in order to identify nucleosome locations and occupancy. Because of its sequence biases, “MNase adversely cleaves DNA” around 30-times more quickly “upstream of an A or T than it does 5′ of a G or C” (Chereji, Bryson, et al.; Chereji, Kan, et al.; Chereji, Ocampo, et al.; Mieczkowski et al.). As a result, the level of MNase digestion has a significant impact on the nucleosome occupancy profiles that are produced by counting all mononucleosomal reads covering each genomic location. The regions of the genome which had more sequenced DNA fragments after MNase based digestion are considered to be occupied with nucleosomes. These fragments released from chromatin more quickly, are enriched in the samples that have undergone mild digestion, whereas the regions that are less accessible to MNase are underrepresented in the samples as they undergo mild digestion and get rejected during size selection to allow sequencing of only smaller DNA sequences(Chereji, Bryson, et al.; Chereji, Kan, et al.; Chereji, Ocampo, et al.; Mieczkowski et al.).

1.2.2.4 Assay for transposase-accessible chromatin using sequencing (ATAC-seq)

“ATAC-seq” needs 10 times less nuclei in comparison to other open-chromatin assay like DNase, FAIRE, or MNase-seq but it may capture multi-nucleosome region of accessible chromatin with a higher signal to noise ratio. “Buenrsotro et.al,” introduced “ATAC-seq”, which needs a prokaryotic “Tn5-transposase” modified with point mutations to enhance its activity of enzyme and adaptors to tag open chromatin. The “Tn5-transposase” can be used with single or bulk/multiple isolated nuclei. Particular primer pairs are utilized to amplify the “cut and tagged segments of DNA”. Also, This is observed by high-throughput sequencing. “ATAC-seq” library often shows a trend along 200bp periodicity, comparable to DNA location devoid of one (200 bp) or excess nucleosomes (Buenrostro, Paul G. Giresi, Zaba, Chang, et al.; Buenrostro, Wu, Chang, et al.).

1.2.3 Recognizing the phenotype of the cells through molecular profiling

The use of precision medicine utilizing genetic profiling technology has gradually been coupled with conventional clinico-pathological experiments in order to enhance diagnosis, prognosis, and clinical outcome prediction. Despite the clear accomplishments of the molecular characterisation era, the value of NGS alongside additional omics-based assays is still largely unknown. A future vision for precision medicine includes thorough multi-omic tumour characterisation, dynamic liquid biopsy sample monitoring, annotation automated by "artificial-intelligence" advancements, although guided through clinical input from experts, and patient enrollment in state-of-the-art clinical trials that examine the usefulness of alternative drug-assignment algorithms in addition to testing molecular profile-drug matches, and the use of precision medicine in real-life scenarios (Mittra and Moscow; Malone et al.). The development of personalized medicine will need not only technical advancement in addition to genomics but also involving the education of end-users like; doctors and patients, the modernization of access to genotype-drug matching through adaptive and additional novel clinical trial procedures, and the encouragement of data sharing to achieve maximum knowledge gained (Malone et al.).

When it comes to cancer, molecular profiling uses various methods to uncover cancer biomarkers; the results let doctors know if a tumor is likely to respond well to treatment or be resistant. A cancer biomarker is linked to the body's presence of cancer. A biomarker may be a distinct bodily reaction to the presence of cancer or it may be produced by the tumor itself. Fluorescence in situ hybridization, immunohistochemistry, next-generation sequencing, and quantitative polymerase chain reaction (qPCR) are examples of molecular profiling technologies. In order to prevent, identify, and cure diseases, precision medicine, also known as "personalized medicine," makes use of data about a person's lifestyle, environment, and biology.

1.2.3.1 Molecular profiling technologies

Immunohistochemistry (IHC)

The use of “monoclonal-” and “polyclonal-antibodies” for detecting particular antigens in tissue samples is known as immunohistochemistry (IHC). It is also among the most valuable appliances in the diagnostic surgical pathologist's toolbox. to ascertain the tissue distribution of a desired healthy and health-related antigen, “IHC” is a crucial application of monoclonal and polyclonal antibodies. It is frequently utilised for cancer detection due to particular tumour antigens expressing “de-novo” or being up-regulated in certain tumors. “IHC” is essential to pathology, specifically in the fields of oncologic pathology, neuropathology, and hematopathology (Kaliyappan et al.).

Fluorescence in situ hybridization (FISH)

“Fluorescence in situ hybridization (FISH)” is a cytogenetic approach introduced in the early 1980s. Fluorescent DNA probes are used in FISH to hybridize to complementary DNA at a particular chromosomal region within the nucleus, to generate vibrant signals visible through a fluorescent microscope. FISH requires less cell culture than the traditional cytogenetic (CC) metaphase karyotype investigation and can evaluate interphase nuclei directly using fresh or paraffin-embedded nuclei (Hu et al.). The potential uses of FISH have increased to encompass solid tumours, hematologic malignancies, and genetic disorders recently identified due to the identification of additional disease-related genes. In reference to chronic myeloid leukaemia (Hu et al.; Jiang et al.), breast cancer, and lung adenocarcinoma, correspondingly, the “FISH” detection of “BCR/ABL1 translocation”, HER2 amplification, also ALK rearrangement is crucial for guiding targeted therapy. “FISH” tests are now understood to be essential elements of personalized medicine (Hu et al.).

Next-generation sequencing (NGS)

Through the giving of a diagnosis and important clinical information, tissue examination provides information on patient treatment. A tissue sample's morphological characteristics and, in rare instances, the pattern of expression of a small number of biomarkers are assessed by the diagnostic pathologist. DNA and RNA from tissue samples may now be analyzed in great detail because of recent advances in sequencing technology. “Next-generation sequencing (NGS)”, the aggregate name for these novel technologies, produces vast volumes of data that can be utilized to enhance patient treatment. The interpretation and integration of the molecular data with the morphological data are necessary to optimize the utility of tissue interrogation. The pathologist must, however, be aware of the uses and limitations of NGS data in analyzing the molecular data (Ilyas).

quantitative Polymerase Chain Reaction (qPCR)

A category of methods referred to as quantitative polymerase chain reaction (qPCR) are exploited to determine how many different template DNA sequences there are. In one strategy, the binding of a reporter dye to “double-stranded DNA” is utilized to quantify the progress of the PCR at individual cycles of synthesis. This growth pattern, whose mimics sigmoidal or exponential expansion, is noticed till the fluorescence reaches its plateau situation. Based on the quantification principles, procedures started with rarer DNA copies need more replication cycles to make a product, like; athletes starting a race from a larger distance need to carry more steps to cross the finished bar (Hirakawa et al.).

1.2.4 Cellular heterogeneity

The issue is no longer to show that populations of "apparently similar" cells are heterogeneous after decades of probing, measuring, and studying the actions of single cells. Indeed, at a sufficiently fine level of observation, phenotypic distinctions between cells are always visible.

Instead, the difficult task is to determine whether any of the elements of the observed variability in cells have a biological process or represent important knowledge (Brock et al.; Altschuler and Wu).

During the progression of cancer, cell heterogeneity becomes complex. A single or homogeneous cell's molecular constituents can now be extracted and analysed because of developments in single-cell isolation technology, largely by microdissection. With the use of modern technologies, it is possible to learn more about transcriptomics, genomics, and metabolomics. Data on "methylation, histone, microRNA, and nucleosomal localization" at the single-cell level are now available due to the advances in single-cell epigenomic technologies. Integrating epigenomic data with different omics data can help us understand how cells interact and perform, also how changes to those molecular processes might result in aberrant physiology and the emergence of diseases. The epigenomic approaches, that can be used at the single-cell basis, place emphasis on the possibility of future advancements (Verma and Kumar).

1.2.5 Investigation of disease phenotype using fundamental and high-throughput genomic methods

Numerous associations between complex human phenotypes and genetic variations are known to exist, and the frequency of new findings is increasing rapidly. Because the linked variations are predominantly located in non-coding areas of the genome where there are few established rules for predicting their function, translating these relationships through comprehending of diseases conditions are still a basic problem. It has been discovered as phenotype-related variations are strongly emphasized in potential regulatory-components by combining the compilation of reported genetic connections using mapping of the human genome's regulatory activities. The next stage in this strategy might be to use newly developed high-throughput assays to determine the functionality of likely regulatory elements. Prioritizing variations that most likely have a functional impact on the development of diseases may be done through allele-specific analyses of the regulation of genes. Together, these technologies have made it possible to conduct systematic and empirical tests of hypotheses regarding the role of noncoding variations and “haplotypes” at the scale required for thorough and systematically assessing genetic association studies (Lowe and Reddy).

Genome-wide association studies (GWAS) and other studies have shown that over 15000 SNPs are associated through a complicated disease as of February 2015 (Welter et al.). The processes behind such relationships are, however, still largely unknown. In general, little is known about the underlying structure of complicated diseases and features. Originally, the disease-common variant hypothesis presumed to be widespread variants that are existing among every group of populations are responsible for Phenotype diversity or vulnerability to disease, furthermore; such variants collectively having an “additive or multiplicative” impact on disease risk variants (Lowe and Reddy; Gibson). A percentage of disease associations may be regulated via gene regulation, according to the enrichment of disease-associated single-nucleotide polymorphisms (SNPs) among expression quantitative trait loci (eQTL) SNPs. Researchers would anticipate the eQTL and disease signal to exist at the same causal variant if a disease association is mediated by its impact on gene expression. However, demonstrating that a single SNP is associated with both traits is insufficient to confirm colocalization because causal variants cannot be identified from genetic association data alone, and the same situation could occur if two different causal variants were in linkage disequilibrium (LD) (Hui Guo et al.).

It is generally known that chromatin accessibility to certain enzymes, like; DNase-I, having a sign of genomic-wide regulatory functional activities. This idea is used by “high-throughput sequencing-based assays” like; “DNase-seq and ATAC-seq” to provide detailed mapping of “chromatin-accessibility” throughout the human genome. In a similar fashion, ChIP-seq, a high-throughput-sequencing variant of “chromatin immunoprecipitation”, is currently often utilized to pinpoint transcription factor binding sites and histone changes linked to human genome’s regulatory state. With “DNase-seq and ChIP-seq”, binding events or changed histones can be localized to within 50 bp. That is now substantial proof that genetic variations in potential regulatory regions discovered by ChIP-seq or DNase-seq influences human phenotypes (Hui Guo et al.; Lowe and Reddy).

1.2.6 Role of epigenetics in the regulation of transcription and cellular identity

Despite having almost identical DNA, all cells in an organism have various kinds and activities due to qualitative and controlled differences in gene expression. The controlling of "gene activity/expression is most necessary for proper differentiation and development. As cells divide during mitosis, the gene expression patterns that distinguish differentiated cells are established during development. Thus, in addition to receiving genetic information, cells also inherit information known as epigenetic knowledge that is not encoded in the DNA nucleotide sequence. The study of "mitotically/meiotically" heritable changes in expression of genes those are not induced by modifications in DNA-sequence (Waterland) is just how epigenetics is defined. Cell-type identity and function are known to be determined by epigenetic processes. As a result, rewiring the underlying epigenome is essentially necessary for the conversion of one kind of cell into another. Through cellular reprogramming, somatic cells may be transformed towards "induced pluripotent stem cells", that, subsequently, can be regulated to develop into specific cell-types. On the other hand, trans-differentiation or direct reprogramming entails the direct transformation of one cell-type into another (Basu and Tiwari).

1.2.7 Methods for examining epigenetic alterations

The analysis of DNA methylation and chromatin modification patterns observed throughout numerous biological processes constitutes a large element of current epigenetic research. Recent studies have been focused on a genome-wide analysis, reflecting a progressive shift in epigenetic research over the past few years. The regulation of gene activity must include both DNA methylation and chromatin changes. By adding a "methyl group to a cytosine base's fifth carbon”, DNA methylation substantially reduces gene activity. Less specifically, a variety of methods can modify the chromatin structure, which can either cause an up- or down-regulation of the related gene. The two most popular methods for monitoring changes in DNA methylation and chromatin structure, respectively, are bisulfite modification and chromatin immunoprecipitation (ChIP), among the many experiments used to evaluate the consequences of epigenetic modifications (DeAngelis et al.).

DNA methylation and Histone regulation

“DNA-methylation” in the mammalian genome includes the insertion of a methyl group to the cytosine's C5 position to create 5-methylcytosine. By attracting proteins involved in gene repression or by preventing transcription factor(s) from binding to DNA, DNA methylation often makes chromatin inactive. De novo DNA methylation and demethylation both play active roles in the dynamic process that alters the pattern of gene-regulation during development. Differentiated cells create a strong and distinctive DNA methylation pattern that controls the transcription of genes pertaining to particular tissues (Moore et al.). Important chromatin structure and gene transcription are regulated by histone modifications, which have an impact on a number of significant cellular phenotypes. A growing number of research over the past ten years have suggested that alterations in different histone modifications have a major impact on the ageing process. Furthermore, it has been found that dietary factors, which can influence gene expression and lifespan, can have an impact on the quantity and localisation of histone modifications. This lends credence to the idea that histone alterations might act as the primary cellular platform for signal integration (Molina-Serrano et al.).

1.2.8 Histone modifications as key epigenetic factors

Post-translational histone modifications, like DNA methylation, don't alter the DNA's nucleotide sequence, but they can affect how easily accessed it is by the transcriptional machinery. In general mechanisms of histone modifications in the context of their important function, the epigenetic changes controlling the gene expression, even these histone modifications also play some other roles. For example, histone phosphorylation is renowned for its contribution to DNA repair in response to cell damage. While there are many other kinds of “histone-modifications, acetylation, methylation, phosphorylation, and ubiquitination” are the most well-studied and crucial for controlling chromatin structure and (transcriptional) activity (Alaskhar Alhamwe et al.).

Histone acetylation

The epigenetic alteration of histone acetylation is unmistakably linked to a higher propensity for gene transcription. Increases in histone acetylation generally promote learning and memory and can be seen of as molecular memory aids because gene transcription is a key component of long-lasting types of memories (Gräff and Tsai).

Histone methylation

Histone methylation is the process of adding methyl groups to histone proteins; it is frequently referred to as a "gene silencer." In contrast to histone acetylation, histone methylation has been proposed to be a complicated, yet more long-lasting and stable posttranslational modification. It can take the form of mono-, di-, or trimethylation (me3), with each methylation event having a different impact on gene transcription (Werner et al.).

Histone phosphorylation

Histone phosphorylation primarily takes place during DNA damage, cell division, apoptosis and transcription activation. The enzymes in charge of this PTM are the protein kinases yTel1 as well as yMec1 (ATM and ATR in mammals), which can target tyrosine, serine, and threonine residues (Albini et al.).

Histone ubiquitination

Nearly every DNA-related process, including DNA replication, transcription, and repair, depends on histone ubiquitination. Histone ubiquitination was initially discovered as a transcriptional regulator, but it is now understood that these PTMs serve a variety of roles in the context of "DNA double-strand" breaks repair. With the help of deubiquitinating enzymes (DUBs), dynamic PTM of ubiquitination can be reversed. As implied by their name, the main job of these enzymes is to modify or eliminate ubiquitin chains. The site-specific detection of histone ubiquitination by chromatin-related factors is a key element in the varied outcomes (Aquila and Atanassov).

1.3 Topologically Associating Domains (TADs): "Regulatory interactions between *cis*-regulatory elements and promoters"

1.3.1 A/B compartment dynamics in 3D spatial genome organization

The entire genome is capable of being divided into two components, called the A/B compartments, according to analysis of "Hi-C" data. These compartments relate to open and closed chromatin and are unique to certain cell-types. "Lieberman-Aiden et al.", first described "Hi-C", a technique for measuring long-range interactions in the genome, and Dekker et al., reviewed it (Fortin and Hansen). At a specific resolution set by sequencing depth, a "Hi-C" assay generates a commonly-named genome contact matrix, which quantifies the level of interaction between two location in the genome. The majority of interactions between loci are restricted to those involving loci from the same compartment. It was discovered that the B compartment was linked to closed chromatin and the A compartment to open-chromatin (Fortin and Hansen; Lieberman-Aiden et al.).

Large multi-Mb chromosomal domain clusters, most of which are on the same chromosome but can sometimes be on distinct chromosomes, make up the A and B compartments. The latter kind of linkage most likely takes place in the zone where chromosome territories mix.

These results show an overall pattern for loci with similar genomic properties and chromatin states to be adjacent to one another while remaining distant from locations with the oppositional states. Recent study using "Hi-C" data from human and mouse cells revealed that there is a continuum between the 2-types of opposite chromatin states that preferentially associate in the A-copartment and B-compartments, respectively. Domains of chromatin preferably connect with those that are similarly active. At transcription factory-like locations, colocalization of genes that are actively transcribed has been observed. assemblages of active genes, occasionally found on various chromosomes, are discovered to interact alongwith loci enriched in "RNA polymerase" and someother transcriptional regulators at these locations (Gibcus and Dekker) (Figure 1.3).

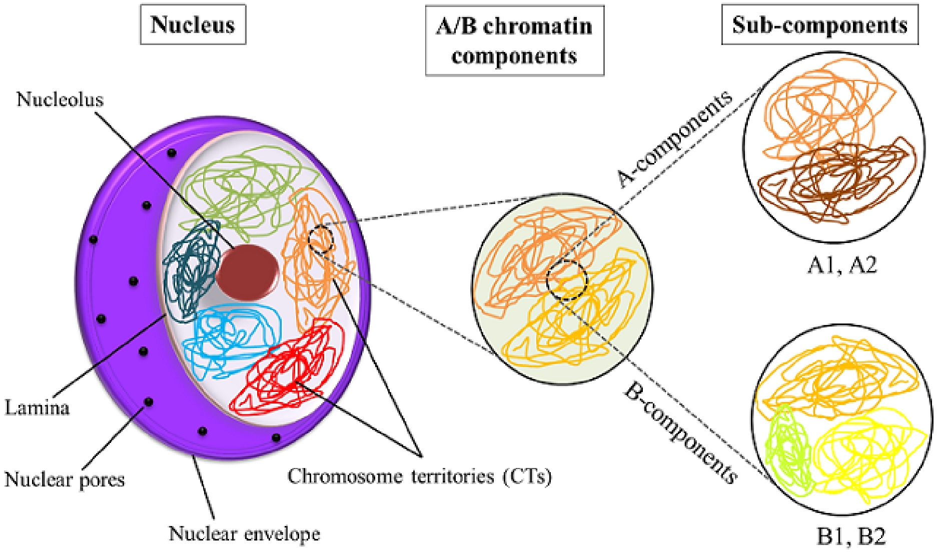


Figure 1.3. 3D genome reorganization indicating dynamics of Compartment A/B

1.3.2 Challenges in TADs analysis

The spatial organization of the genome is essential for controlling how genes are expressed. TADs and sub-domains are the primary building blocks of the 3D-genome, according to recent chromatin interaction mapping research. Comprehending the 3D structure-function association of the genome requires an understanding of such hierarchical systems, which is a crucial first step. For high-resolution “Hi-C” data, existing computational approaches are computationally inefficient and lack statistical domain prediction assessment (Yu et al.). Since precisely matching boundaries might occasionally be difficult, accurate boundary recognition does not always equate to accurate TAD prediction. In general, boundary prediction out-performed than complete TAD detection in terms of tool performance. The TAD boundary predictions produced by TopDom, DomainCaller, and HiCSeg are generally in great concordance and resistant to changes in resolution and sequencing coverage (Dali and Blanchette).

1.3.3 TADs as regulatory regions in development, evolution, and diseases

It has been mentioned by several researchers that TADs isolate a genomic region from the effect of neighboring locations. For cis-regulatory components as well as their target promoters, TADs often feciltate frequent interactions, which would not otherwise happen often to guarantee a sustained gene expression (Galupa and Heard; Tena and Santos-Pereira). TADs have therefore been suggested as a scaffolding of chromatin to control local cis-regulatory landscape, and described as "large genomic areas containing numerous long-range-acting regulatory sequences that affect one or more target genes in a coordinated manner”. Nevertheless, whether TADs epitomize an exclusive functional unit in the chromatin scaffolding has been complicated by the enhancing resolution of Hi-C assays. They have revealed nested structures at the class of subTAD through relative insulation between them. However there is no doubt that TADs can be used to study the relationships between the 3D genome and gene expression, as well as their dynamics in relation to development, disease, and evolution. (Tena and Santos-Pereira; Rao et al.).

1.3.4 Regulation of gene expression in cancer disorders using chromatin-based profiles

Cancer is an effect of the uncontrolled development of abnormal cells, driven by genetic changes that are either attained or inherited from our parents. Each cancer has a unique set of molecular vicissitudes in the cancer cells. Technological advances have made a molecular profiling analysis possible; this allows clinicians to distinguish the molecular alterations amid cancer cells and healthy cells (Martinez-Lage et al.). Molecular profiling utilise different technologies to recognize malignant growth biomarkers; the discoveries illuminate doctors of the possibility that cancers will be sensitive or resistant to treatment. A cancer biomarker is associated with the existence of cancer in the body. A cancer biomarker can be produced by the tumor itself, or it may be an exact response by the body to the occurrence of cancer. Examples of molecular profiling technologies include fluorescence *in situ* hybridization, immunohistochemistry, qPCR, next-generation sequencing. Personalized medicine uses information about a person's way of life, environment, and biology to prevent, diagnose, and treat disease and fragment analysis (FA) to detect changes in RNA or DNA to indicate the presence or absence of genetic marker [102].. Proteins involved in chromatin play a complicated and very context-specific role in cancer. Few chromatin modifiers appear to be capable of initiating the growth of cancer on their own; instead, they are frequently altered in conjunction with crucial tumour suppressors and cell cycle regulators like p53 and CDKN2A. Many mutant chromatin proteins are extremely tissue-specific, despite the possibility that some chromatin regulators, such the MLL3/4-UTX of the COMPASS family, may have a broad tumour suppressor role in a variety of malignancies (Morgan and Shilatifard).

1.4 Scope of thesis work

DNA and proteins combine to form the dynamic structure known as chromatin. Its peculiar features are essential for tightly packing the DNA inside the cell as well as for controlling gene expression and DNA replication. Additionally, it shields DNA from damage. Our understanding of the pathophysiology of numerous chromatin-related diseases is improved by knowing which proteins are involved in the formation of the various complexes inside chromatin. Additionally, it might aid in the development of novel pharmacological targets and more effective treatments. It is difficult to develop a single strategy to investigate chromatin-based study under all physiological conditions since chromatin can exist in a variety of forms under different physiological conditions.

For multiple types of cells in the native state from the *in vivo* samples, the availability of chromatin-interaction profile is rare. It has become important to understand the function of chromatin structures since they are essential in the regulation of the expression of the genes if we are to actually understand the heterogeneity of the cells. However, the data that is currently available on chromatin is derived from only a few cell-types, providing very limited insights into the cellular heterogeneity for developing precision medicine. It could be feasible to determine a cell-type's chromatin-interaction pattern using “single-cell open-chromatin profiles”. A technique known as “single-cell epigenome-based chromatin-interaction analysis” (scEChiA) was created in order to predict chromatin-interaction from single-cell epigenome profiles. “Single-cell open-chromatin profiles” have a high drop-out rate and noise due to stochasticity and the availability of low amounts of relevant DNA. Therefore we also developed a robust technique known as a forest of imputation trees (FITs) to solve this problem (Khan et al.) (Figure 1.4).

1.4.1 Utilizing single-cell open-chromatin profiles and availability of chromatin-interaction

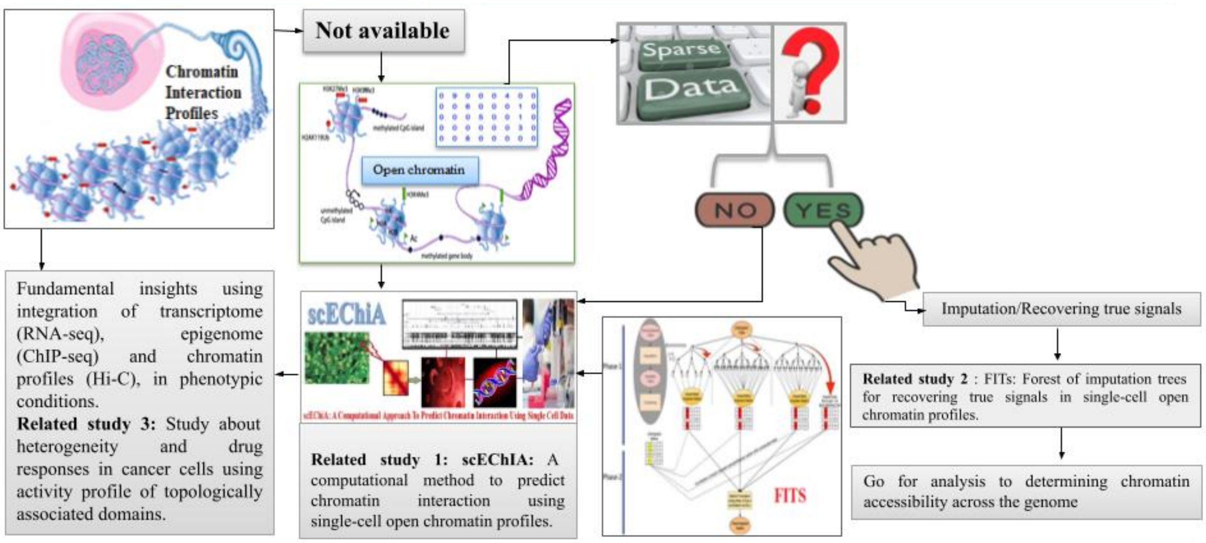
The pattern of chromatin-interaction in a cell-type can be interpreted by utilizing “single-cell open-chromatin profiles”. However, despite the fact that long-distance connections among genomic locations play a substantial role in gene regulation, methods proposed before earlier were biased only for short-range chromatin interaction. Here, we suggest a method that employs “single-cell open-chromatin patterns” to predict both short as well as long-range interactions among genomic regions. Our method called as “single-cell epigenome-based chromatin-interaction analysis" (scEChIA) makes use of signal imputation and refined L1 regularisation. Even in terms of prediction accuracy, scEChIA fared better than other techniques for multiple "single-cell open-chromatin profiles".

1.4.2 Recovery of true signals from single-cell open-chromatin profiles with high noise and sparsity

Here, we present a reliable technique for recovering the original signals from extremely "sparse" and "noisy" "single-cell open-chromatin profiles" called the forest of imputation trees (FITs). To prevent bias at the time of the restoration of read-count matrices, FITs creates multiple imputation trees. It solves the difficult problem of recovering open-chromatin signals at genomic regions with cell-type-specific activity without blurring out information. Along with visualisation and classification, FITs-based imputation increased accuracy in enhancer recognition, pathway enrichment score calculation, and chromatin-interaction prediction. FITs is expanded to provide a broader range of applications, particularly for very sparse read-count matrices. Furthermore, "single-cell open-chromatin profiles" from *in-vivo* samples can benefit greatly from FITs due to their superiority in recovering signals of minority cells.

1.4.3 Using chromatin profile analysis to comprehend cancer-based studies

TADs which are how chromatin are arranged, are defined by preferred contacts between loci that are located inside the same TAD, and it have been associated with the control of the genes they possess by regulating regulatory interactions between "cis-regulatory elements" and promoters. To build regulatory landscapes (RLs), TADs are therefore considered as structural scaffolding. A thorough investigation of the arrangement of the activity of genes with comparable functions in TADs, their colocalization in cancer cells, and their pharmacological responses is limited. To further understand how drugs affect cancer cells, we looked at trends in the activity of TAD gene sets along with chromatin interaction profiles.


Chromatin Interaction Profiles
Not available
Open chromatin
Sparse Data
NO
YES
scEChiA
scEChiA: A Computational Approach To Predict Chromatin Interaction Using Single Cell Data
FITS
Imputation/Recovering true signals

Fundamental insights using integration of transcriptome (RNA-seq), epigenome (ChIP-seq) and chromatin profiles (Hi-C), in phenotypic conditions.

Related study 3: Study about heterogeneity and drug responses in cancer cells using activity profile of topologically associated domains.

Related study 1: scEChIA: A computational method to predict chromatin interaction using single-cell open chromatin profiles.

Related study 2 : FITs: Forest of imputation trees for recovering true signals in single-cell open chromatin profiles.

Go for analysis to determining chromatin accessibility across the genome

Figure 1.4. Coherence among research work

CHAPTER 2

USING SINGLE-CELL OPEN-CHROMATIN PROFILES TO IMPROVE CHROMATIN INTERACTION PREDICTION AND REVEAL INSIGHTS ABOUT THE HUMAN BRAIN'S CIS-REGULATORY LANDSCAPE

2.1 Introduction

For numerous regulatory roles, spatial interactions between diverse genomic sites are essential. To investigate the complex patterns of chromatin architecture involved in controlling gene expression, numerous labs have profiled chromatin-interaction in a variety of cell-types utilizing various experimental high-throughput approaches. The majority of 3C chromosome conformation capture experiments concentrated on local genomic regions (Dekker et al.). The chromatin-interaction analysis via paired-end tag sequencing (ChIA-PET) approach only helps in detecting the binding sites of the protein of interest, but it captures distal interactions (Tang et al.). Across the entire genome chromatin-interaction profile is provided by the “high-throughput chromosome conformation capture Hi-C” experiment, however deep sequencing is needed to provide high resolution (de Wit and de Laat).

2.1.1 Availability of chromatin profile via single-cell open-chromatin profile

Recently, several research teams had already tried to prediction of chromatin interactions utilizing linear, 1D genetics, and epigenetics data (W. Li et al.). The majority of the techniques for interaction prediction rely on epigenetic information using bulk samples, which frequently contain various cell-types (Whalen et al.). Concurrent accessibility of multiple epigenome based profile is presently available for some of the cell-types. Therefore, for many cell-types, predicting chromatin- interactions specific to those cell-types is not an easy task. On the other hand, we might predict chromatin interactions in cell-type if we took use of heterogeneity in the activities of genomic regions in single cells.

The landscape underlying genomic site activity can be provided by single-cell epigenome profiling. The prediction of chromatin interaction, notably for comprehending regulatory mechanisms of minor cell-types in heterogeneous clinical sample for personalised therapies. There wouldn't be simple to regularly profile chromatin-interaction maps for various cell-types for genetically heterogeneous clinical samples of patients utilising experimental assays like 3C and Hi-C or to do computational analyses utilizing bulk epigenome profiles.

2.1.2 Prediction of chromatin interaction using single-cell epigenome for the study of cis-regulatory interaction of human landscape.

For the prediction of local chromatin-interaction employing “single-cell assay for transposase-accessible chromatin utilising sequencing (scATAC-seq)” profiles, A group (Pliner et al.) developed an approachcalled “CICERO”. “CICERO” is build for prediction of interactions between genomic loci which are less than 500-kilobase pairs from each other. A different approach, referred to as “jointly reconstruct cis-regulatory interaction maps” -JRIM (Dong and Zhang), utilizes open-chromatin profiles from multiple cell-types for infering chromatin interactions; as a result, it would not work well for the prediction of a single cell-type. JRIM is intended to predict local chromatin interactions (with in 500Kbp). It has been demonstrated in the past, though, that mutations discovered by “genome-wide association studies” may influence genes which are more than 500 kbp away. In mouse cells, the TAD has a reported median size of 880kbp (Dixon, Selvaraj, et al.). The significance of distal chromatin interactions (>800kbp) between super-enhancers and promoters in poising and activating embryonic stem cells (ESCs) was highlighted by Novo *et al.,*, (Novo et al.). Similar to this, additional research has emphasized the significance of distal chromatin interaction for comprehending gene regulatory patterns and associated epigenetic profiles (Ling and Hoffman). Therefore, a significant outstanding challenge of great utility is the prediction of long-range chromatin interactions utilizing a single-cell epigenome profile. thereby, we've developed a technique termed “single-cell epigenome-based chromatin-interaction analysis” (scEChIA), that uses “single-cell open-chromatin profiles” for predicting interactions across distal sites with good accuracy. We have also demonstrated its value in predicting potential chromatin interactions in brain-cells to generate insightful conclusions.

2.2 Materials and methods

2.2.1 Pre-processing of Data

The genome is initially divided into bins with the appropriate sizes using our method. It utilizes a 25 kbp bin by default. It combines the peaks that are lying in the same bin for a read-count matrix. While combining two peaks, it adds corresponding read counts. After merging these peaks, the resultant read-count matrix is log transformed as follows:

$$\overline{M} = log(M + 1) \tag{1}$$

$$matrix\ M \in \mathfrak{R}^{n*k}$$

$$where\ with\ x_{ij}\ as\ ij^{th}\ element\ of\ matrix$$

2.2.2 Improved Penalizing Parameter in the Gaussian Graphical Model to Integrate Prior Knowledge

Due to the high dimensionality of the single-cell open-chromatin profiles, the counts of peaks (k) is greater than the counts of cells (n) in the read-count matrix. Therefore, it is not simple to estimate a matrix with peak activity covariances. In order to estimate the regularized covariance matrix and its inverse for such issues, a Gaussian graphical model, e.g., the “graphical lasso” technique (Friedman et al.), is helpful. Calculating partial correlations requires the covariance matrix's inverse. After eliminating the impact of confounding factors resulting from other peaks, partial correlation in this case shows a degree of co-accessibility between peaks. For identifying such a direct association between variables, graphical lasso is utilized. Once the strength of their association is weak, the penalty term in the "graphical lasso" leads partial correlations between peak pairs to shrink (Friedman et al.). The graphical lasso technique seeks to maximize:

$$logdet\Theta - tr(U\Theta) - \rho \,||\, \Theta \,||_1 \qquad (2)$$

where r is the penalty term for L1 norm-based regularization, U is the covariance matrix, and Θ is the inverse covariance matrix. The penalty term could be a matrix containing different ρ value for each pair of variables (peaks).

By utilizing the knowledge of the already available chromatin-interaction profile, the proposed approach is using a "penalty matrix" that is uniquely built. The penalty matrix's components are estimated as follows:

$$\rho_{ij} = \frac{\delta}{h_{ij} + \varepsilon} \qquad (3)$$

The published "Hi-C" profiles of numerous cell-types used to estimated $[h_{ij}]$, where $[h_{ij}]$ is the average enrichment level of chromatin interaction between genomic bins $[i\ and\ j]$.

While [δ] is a constant that can be changed to either increase or decrease the number of predicted interactions at the cost of accuracy. The idea behind our technique is also meant to address the following situations:

1. When two interacting sites exhibit high activity across all cell-types and their read-count drop-out is because of stochasticity and reduced sensitivity during "scATAC-seq" profiling, the covariance between them may be underestimated. Although, with interactions data for all the cell-types, it would be easier to retrieve the information by providing a lower penalty or greater prior value.

2. If the noise in read-count matrix of "single-cell open-chromatin profiles" is intense, a priori knowledge of the background could improve the prediction of interaction.

3. Since the penalty isn't exponentially high, it might still be retrieved if two sites have cell-type-specific interactions as well as a decent covariance value. Decent in this context refers to a higher value than the majority of other elements in the covariance matrix.

Therefore, prior information (or heuristics-based values of the penalty matrix) is a vital step. Matrix-factorization is used by ScEChIA to minimize noise in the read-count in an effort to better enhance the prediction. Below is a description of the matrix-factorization that scEChIA employs.

2.2.3 Enhancing Co-occurrence Estimation Through Matrix Factorization

For low-rank matrix completion issues, matrix factorization provides a solution. Y is a matrix of observed read counts, where each row represents a cell and the columns indicate peaks, and is referred to as a sampled representation of the true ideal matrix X in the same dimension (m x n). like that

$$Y = A(X) \tag{4}$$

Here, A is an operator matrix with $0's$ for missing elements of X in Y and $1's$ where it's present. However, if X is known to have a rank $r(< m, n)$, X can be expressed as the product of two matrices $U_{m\,x\,r}$ and $V_{r\,x\,n}$. Consequently, Y may be written as

$$Y = A(X) = A(UV) \tag{5}$$

In order to recover X, we attempt to find matrix U and V via minimizing the Frobenius norm of subsequent cost function

$$\underset{u,v}{min} ||Y - A(UV)||_F^2 \tag{6}$$

majorization-minimization (MM) is a technique that we use to optimize bilinear problems (Y. Sun et al.). For majorization-minimization based optimization, the objective function is majorized by a surrogate function. After that, the surrogate function is minimized as long as a local optimum is obtained. The majorization phase is conducted in order to reduce the overall cost function as described in Eq. 6 so that we can optimize

$$\overset{min}{u,v} ||B - A(UV)||_F^2 \tag{7}$$

$here,\ B\ updates\ incrementally\ with\ iteration\ k\ using\ B_{k+1} = X_k + \frac{1}{a} A^T\ (Y - A(X_k))$

$here,\ \ a\ is\ scalar\ and\ X_k\ is\ matrix$

$The\ matrix\ X_k\ updated\ incrementally\ with\ iteration\ k$

$using\ X_k = U_k V_k$

$Whereas,\ Matrix\ U\ and\ V\ are\ also\ updated\ by\ setting\ one\ matrix\ fixed\ and\ other\ incremented$ $using\ equation$ (8) and (9).

$$U_k = \left|\left|B - U_{k-1} V_{k-1}\right|\right|_F^2 \tag{8}$$

$$V_k = \left|\left|B - U_k V_{k-1}\right|\right|_F^2 \tag{9}$$

We retain a non-negativity constraint on X which truncates the element that has a negative value in $[X_k]$ to zero after each iteration. After performing singular value decomposition (SVD) on X, we initialize factor V as a matrix with the r right singular vector of X. SVD is an overview of eigenvalue decomposition for rectangular matrix; as a result, the representation of the matrix X (size: m x n) is

$$X = L \sum R \tag{10}$$

$Here,$

$\sum -\ \ rectangular\ diagonal\ matrix\ (m \text{ x } n)$

$L -\ \ left\ diagonal\ matrix\ (m \text{ x } n)$

$R -\ \ Right\ diagonal\ matrix\ (n \text{ x } n)$

To create “initial guess of matrix $[V]$", select r vectors from the “right matrix $[R]$"

2.2.4 Analyzing the Accuracy of Chromatin Interaction Prediction

We utilized published “Hi-C” profiles in the relevant cell-type to evaluate the accuracy of the chromatin interaction prediction. Using the juicer tool, we initially extracted chromatin interactions in text layout at a resolution of 25 kbps from “.hic” files (Durand et al.). Three-column juicer tool's output was modified to a seven-column file format. The topmost enriched chromatin interactions from the “Hi-C” profiles were selected for evaluation purposes based on a threshold. Topmost enriched chromatin interactions from the “Hi-C” profile using two different modes. First, we selected the “top 60,000” chromatin interactions on each chromosome. Using the second mode, the count of chosen chromatin contacts via the “Hi-C” profiles was proportional with chromosome size. As a result, the longest chromosome having the largest number of interactions at 60,000. The predicted chromatin interactions was intersected with the “Hi-C” based output using PGLtool (Greenwald et al.).

2.2.5 Chromatin-Interaction Prediction Parameters

CICERO offered a few functions for pre-processing, such as “make_atac_cds, aggregate_nearby_peaks, estimateSizeFactors, detectGenes, reduceDimension, make_cicero_cds, generate_cicero_models, assemble_connections, and estimate_distance_parameter”. The parameters for function “reduceDimension, max_components was 2, num_dim = 3, reduction_method = tSNE, perplexity = 5 and aggregate_nearby_peaks was distance = 25,000”. Use of the hg19 genome version was made for subset function and window size of 500,000 was provided to estimate_distance_parameter function. The default parameters were utilised for the remaining functions.

We used functions with various rho parameters for scEChIA, including rhomatAvg and Interaction Prediction 1. We compute the average of two distinct “Hi-C” files by using rhomatAvg function. chrNo and patternf were provided in accordance with the chromosome number, and a bin size of 25 kbp was selected. Considering the information in the background in terms of the average "Hi-C" matrix and other variables such as "chrinfo, data, rhomatrix, chrNo, startCell, endCell, and chromSize", the function_Interaction_Prediction_1 was utilised to predict chromatin interaction. Function "ucscTrack" was used to make track file of ucsc based on predicted chromatin interaction. We utilised the "Interaction_Prediction_2" function, which was based on a "constant rho 0.01".

2.2.6 Sources of data

With the GEO ID: GSE65360, Buenrostro *et al.,* (Buenrostro, Wu, Litzenburger, et al.) provided the scATAC-seq profile for "K562", "H1ESC", and "GM12878" cells to the public. The "single-cell expression" and "open-chromatin profiles" for brain cells used in this project were published by Lake *et al.,* (Lake et al.) (GEO ID: GSE97942). The GEO ID for the "single-cell open-chromatin" profile for cardiomyocytes (Domcke et al.) is GSE149683. "Hi-C" based chromatin interaction profile, used here for evaluation are accessible at 4D_nucleome database1 through IDs: "Astrocytes, cardiomyocytes, K562, GM12878, hESC". "Hi-C" data from Rao et al., (2014) "GEO ID: GSE63525" was also utilized to verify the findings for "K562", "H1ESC", and "GM12878' cells.

2.3 Single-cell epigenome-based chromatin interaction analysis (scEChIA) results and applications

Tang *et al.,* (2015) have demonstrated that in spite of several cell-type-specific interactions, many chromatin interactions show high similarity among different cell-type (Tang et al.). It is well-known that chromatin-interaction and looping mediated by the CCCTC binding factor (CTCF) are largely conserved and have a massive impact on chromatin architecture. Similarly, the TADs' boundaries are defined by a large number of short tandem repeats, they are often conserved in different cell-types (J. H. Sun et al.). Our computational strategy is based on the well-known characteristic of DNA looping and chromatin conformation conservation.

Consequently, to overcome the drawbacks of earlier approaches, while estimating the Gaussian graphical model, we employed the pre-existing knowledge of chromatin-interactions in several cell-types as a constraint factor, using "single-cell open-chromatin profiles", to predict chromatin-interaction utilizing L1 regularization. To compute the L1 regularization (r) parameter, we utilize the average value of enrichment of established chromatin interactions across a variety of cell-types. To lessen noise in the read-count matrix and enhance the accurateness of chromatin-interaction predictions, scEChIA usese built-in function for matrix factorization along to L1 normalization.

2.3.1 Single-cell epigenome-based chromatin-interaction analysis enhances sensitivity for accurate prediction of distal interactions

We compared the accuracy and sensitivity of our method with the well-known CICERO method. For this, we used the "single-cell open-chromatin profiles" of astrocytes (Lake et al.) and cardiomyocytes (Buenrostro, Wu, Litzenburger, et al.) as well as the "scATAC-seq" dataset of K562, GM12878, and H1ESC (Domcke et al.). In order to predict chromatin-interaction for a cell-type, we estimated the regularisation parameter ρ in the graphical lasso (Glasso) model using the average of known chromatin-interaction in other cell-types.

For instance, we utilized a prior (or regularisation parameter ρ) estimated using the average of the "Hi-C" profiles of GM12878 and H1ESC cells to predict chromatin-interaction in K562 cells (Rao et al.). We used the scATAC-seq profile the created by Buenrostro *et al.,* (Buenrostro, Wu, Litzenburger, et al.) for GM12878 cells and the regularization parameter is calculated using the average of the "Hi-C" profiles of K562 and H1ESC. We conducted an evaluation using enriched chromatin-interaction profiles of relevant cell-types that were based on "Hi-C" (see "Materials and Methods" section).

We observed that Glasso did not perform as well in predicting chromatin-interaction when only a single regularization parameter (constant ρ) value was used (Figure 2.1A). Due to the refined regularization matrix, scEChIA outperformed CICERO for 2 out of the 5 evaluated cell lines (cardiomyocytes and astrocytes) for all chromosomes (see Figure 2.1A, Figure 2.2). ScEChIA and CICERO performed similarly for the other three types, in comparison. By utilising two different thresholding criteria to select meaningful chromatin interactions from “Hi-C” data, we were able to prove our findings. We first used the top 60,000 chromatin interactions in “Hi-C” profile of each chromosomes as a positive set for evaluating predicted interactions, as shown in Figure 2.2. Additionally, we confirmed our findings when the frequency of “Hi-C” based interactions fluctuated according to the chromosome's size (Figure 2.3A). As a result, when it comes to correctly predicting interactions on some data sets of the “single-cell open-chromatin” profile, scEChIA tends to do better than other techniques.

In order to show importance of long range interaction prediction by scEChIA, we needed to determine the size of TADs in human cells. We used published TAD boundaries from the “TADKB-database” and observed that the median TAD size in human cells is also greater than 500 kbp (Figure 2.1B). After confirming the immense sizes of TADs, we calculated the count of long-range interactions predicted through various approaches. Exactly as anticipated, scEChIA predicts a significantly higher frequency (nearly 100-times) of long-range interactions with a gap of more than 500 kbp between interacting sites (Figure 2.1C) while maintaining sensitivity for short-range chromatin contacts (Figure 2.3B). We verified using cell-types “GM12878, K562, H1ESC, astrocytes, and cardiomyocytes” that scEChIA has significantly better sensitivity for long-range interaction (>500 kbp). Overall, the estimation of larger TAD sizes emphasises the significance of detecting long-range interactions to capture inter-TAD interactions, which could be achieved through scEChIA employing “scATAC-seq”.

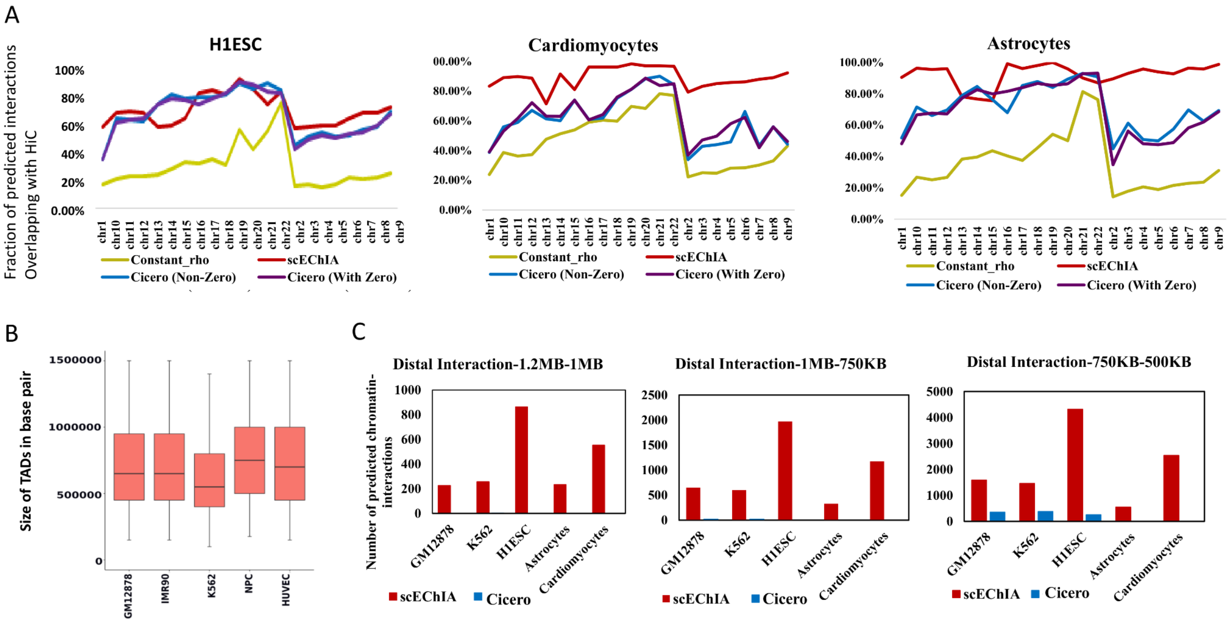


Figure 2.1. Evaluation of the chromatin-interaction prediction's sensitivity and accuracy. (A) Three approaches are used to compare the accuracy of chromatin-interaction prediction: Glasso with a univariate regularization parameter (constant rho), Cicero, and scEChIA. There are two outcomes for Cicero. one does not include predicted interactions and has zero scores, and other includes all interactions. The fraction of predicted chromatin-interactions that overlapped with Hi-C based interactions (top 60,000 in every chromosome) in the respective cell is used as a measure of accuracy. (B) The width of TADs in various human cell-types. The boundaries of TADs are made available at (http://dna.cs.miami.edu/TADKB/) by Liu et al., (2019). The average TAD width is greater than 500 kbp. (C) The sensitivity of prediction of distal chromatin-interaction for various ranges of the distances between interacting loci, is displayed in this figure. In order to accurately predict distal chromatin interactions (>500 kbp), scEChIA overcomes the limitation and it's sensitivity for long range interactions is 1000 times more than other tools. These results were published in our manuscript Pandey et el.(Pandey et al.).

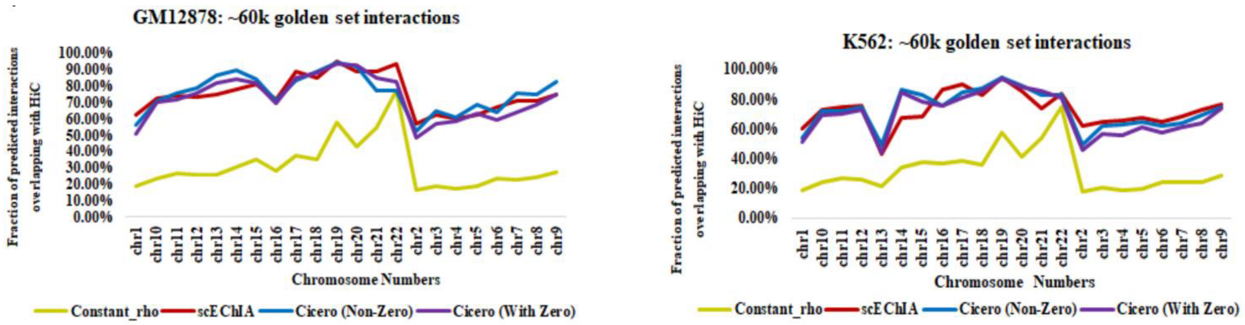


Figure 2.2. Evaluating the accuracy of chromatin interaction predictions using data from two different cell-types' scATAC-seq profiles. The percentage of inferred chromatin interactions that overlap with enriched interactions in the HiC profile is used to measure accuracy. For each chromosome, the number of enriched chromatin interactions from the HiC profile was fixed at 60000(Pandey et al.).

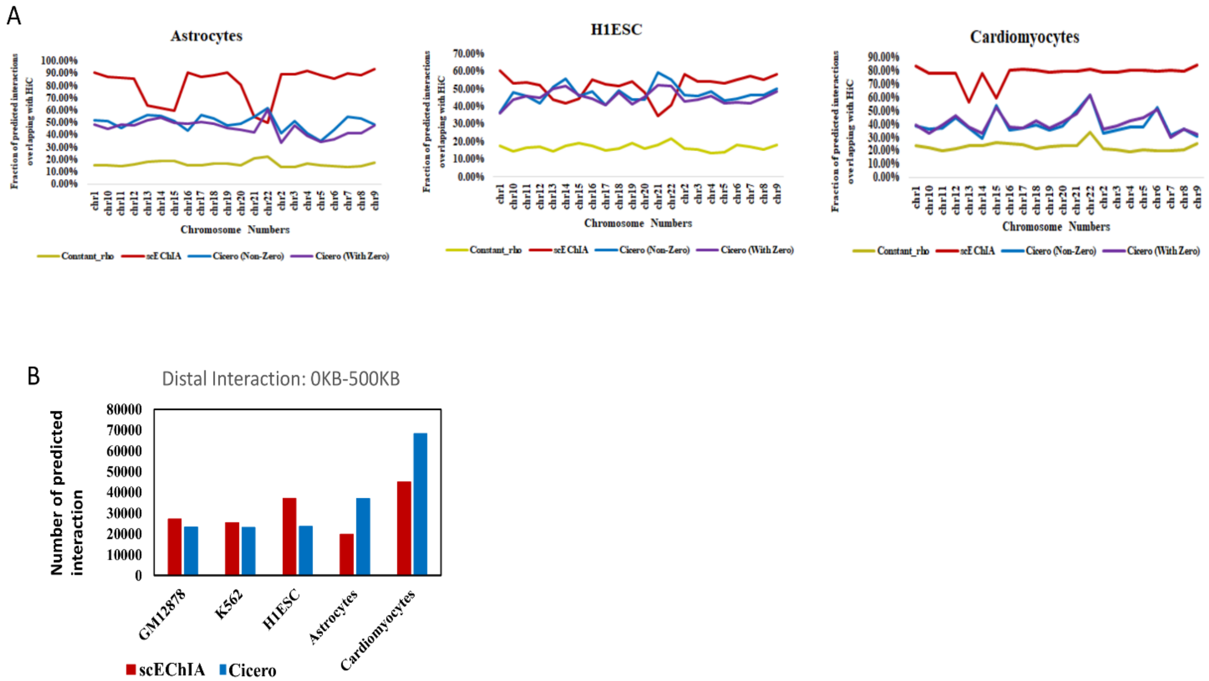


Figure 2.3. Assessment of predicted chromatin interactions. (A) Assessment of prediction using scATAC-seq profiles from 3 cell-types. When the selection of the number of positive sets from HiC is based on the sizes of chromosomes. The accuracy is measured as the fraction of all predicted chromatin-interactions. It overlaps with enriched interactions in Hi-C profile. (B) The number of predicted interactions within 500kb in different cell-types by cicero and scEChIA (Pandey et al.).

2.3.2 Assessing the cell-type specificity of predicted interactions and their influence on gene expression

Although both short-range and long-range genomic interactions could be predicted, there was still some debate over their influence in gene expression and “cell-type-specificity”. Our study showed that the gene’s expression lying within the higher number of genomic interactions was more compared to genes lying at genomic loci with fewer interactions (genes with poor connectivity) (Figure 2.5). Therefore, scEChIA's predicted interactions tend to be coherent with their respective gene-expression profiles. Additionally, we tried emphasizing predicted cell-type-specific interactions and how they affect gene expression as a result. The predicted chromatin interactions in three different cell-types: “K562, GM12878, and H1ESC” were compared, and we showed that several genes had a larger proportion of chromatin interactions at their promoters. These findings demonstrate that the number of predicted chromatin interactions at the promoters of various genes changed based upon the cell-types.

Additionally, genes lying within the genomics regions with higher number of predicted interactions had higher expression in the matching cell-type for which the “scATAC-seq” profile was utilised to make the prediction (Figure 2.4A). Therefore, scEChIA predicts genomic interactions specific to cell-types, which regulate the specificity of genes’ activity in accordance with cell-type. Additionally, we repeated the process using just the long-range interactions that were predicted. Also, the genes with the higher number of predicted long-range chromatin interactions were expressed more in their respective cell-types (Figure 2.4B). Such results confirmed that scEChIA can predict cell-type-specific long-range chromatin interactions, involved in activation of genes by distal sites.

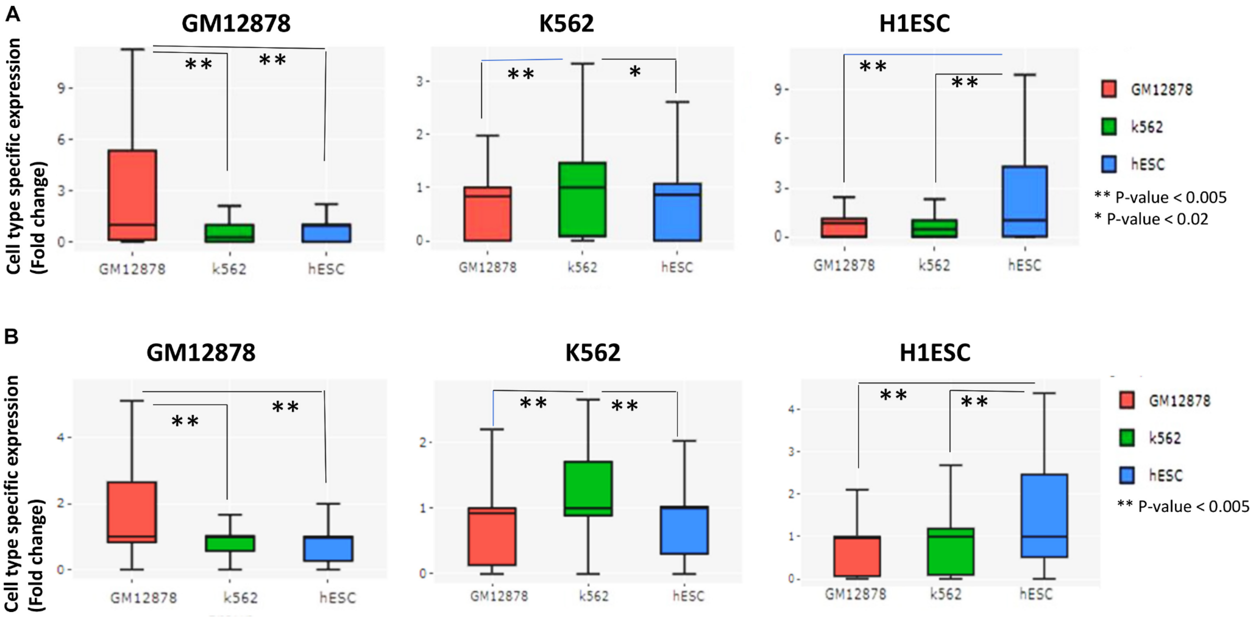


Figure 2.4. Evaluation of scEChIA's ability to predict cis-regulatory interaction with cell-type-specificity. (A) The cell-type-specific FPKM/expression (fold-change above-median across different cell-types) of top 200 genes with the highest relative interactions in the relevant cell-type of the figure panel. For example, in the panel with the label GM12878, the top 200 genes lying within the genomics regions with highest relative number of chromatin-interactions in GM12878 cells (compared to K562 and hESC) were chosen, and their expression levels in the 3 cell-types (GM12878, K562, and hESC) are displayed as box plots. These outcomes suggest that cis-regulatory interactions, which are linked to cell-type-specific expression, can be predicted by scEChIA. (B) The expressed in specific cell-type FPKM level of top 200 genes with the highest relative distal interactions (>500 kbp) in the

corresponding cell-type of the figure panel. Numerous ditsal interactions (>500 kbp) predicted via scEChIA are also associated to expression that is specific to particular cell-types. *Significant, **Most Significant. Wilcoxon rank sum test was used to calculate p-value. (Pandey et al.)

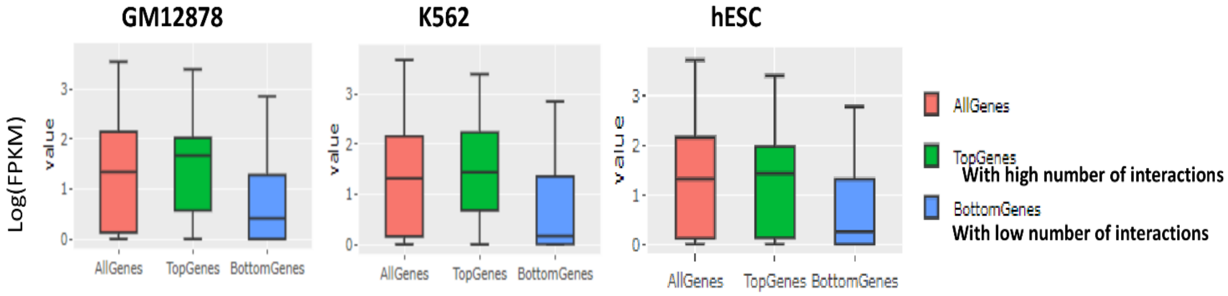


Figure 2.5. Analysis of effect of predicted chromatin-interactions on gene-expression. Here the expression values of top 50 genes with highest and lowest number interactions at their promoter are shown for 3 cell-types (GM12878, K562, hESC) as boxplot. Such as in panel labelled as GM12878, the expression values of top 50 genes which have high or low number chromatin interactions are shown (Pandey et al.).

2.3.3 Human brain's chromatin-interaction landscape

“Single-cell RNA-seq” and “single-cell open-chromatin profiles” of cells originating from the human brain were recently published by Lake *et al.,* (Lake et al.). “Single-cell transposome hypersensitive-site sequencing (scTHS-seq)”, which is more sensitive than ATAC-seq, was utilised by Lake *et al.,* (2018) to profile single-cell open-chromatin patterns. The excellent accuracy in predicting chromatin interactions in astrocyte using the scTHS-seq profile by scEChIA (Figure 2.1A) also suggests good prospects of correct prediction for the other six brain cell-types. Therefore, utilizing the scTHS-seq profile published by Lake *et al.* (Lake et al.). We utilized scEChIA to predict chromatin-interactions in the other six types of brain cells. Astrocytes, inhibitory neurons, oligodendrocytes precursor, oligodendrocyte, microglia, excitatory neurons and endothelial cells are the cell-types for which chromatin interactions were predicted. There were a total of 0.7 million predicted chromatin interactions across all cell-types, ranging from 188857 in microglia to 25838 in oligodendrocytes. Using PGLtool (Greenwald et al.), we intersected our predicted chromatin-interaction with existing expression quantitative trait loci (eQTL) in the brain (Ng et al.) to identify possible cell-types in which the “eQTLs” are linked to their target genes.

It is a challenging task to get information on cell-types for the activity of the published brain eQTLs in the absence of chromatin interactions data of the brain cells. The findings of the intersection of the eQTL dataset and predicted chromatin interaction in seven different types of brain cells are shown. As a result of the intersection, we identified several eQTLs whose target regions genes lying 500 kbps away and corroborated with the predicted long-range chromatin interactions via scEChIA.

Figure 2.6A shows the number of eQTLs with target genes lying at least 500 kbp away and supported by predicted interaction. One illustration of a long-range effect is the brain's eQTL (rs12165519) for SOX10 gene activity. Through a long-range chromatin interaction in oligodendrocyte precursor cells, our investigation showed that the eQTL (rs12165519) might be linked to the target SOX10 promoter and overlaps a peak of open-chromatin profile (ATAC-seq) in the brain (Figure 2.6B).

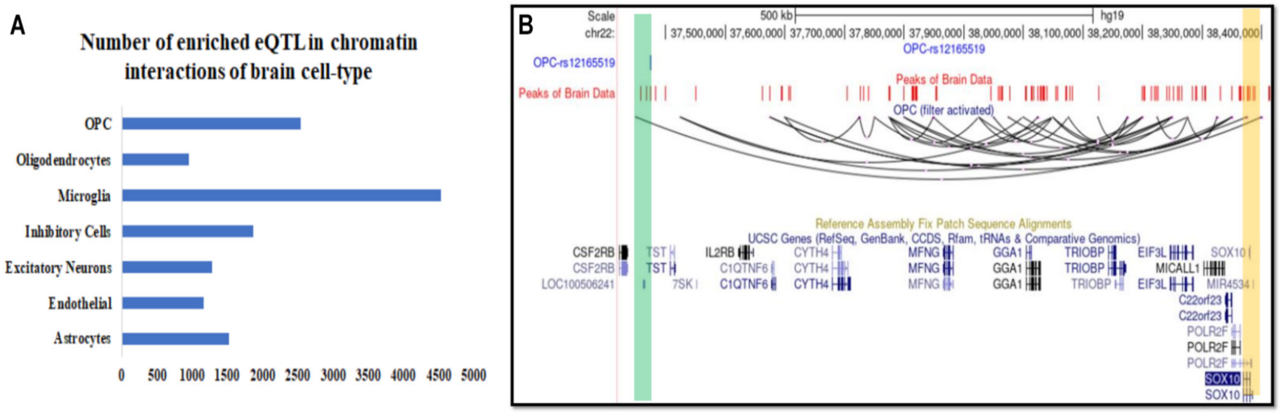


Figure 2.6. Inference regarding the known target genes for expression QTL (eQTL) in the human brain and associated cell-type of action. (A) The quantity of seven different types of brain cell-types' predicted chromatin interactions and brain eQTLs with target genes lying more than 500 kbp. (B) Screenshot of the UCSC browser demonstrating the association between an oligodendrocyte precursor's predicted long-range chromatin interaction and a brain eQTL and its target gene SOX10 (Pandey et al.).

2.3.4 Coverage of genome-wide association studies, cell-type specificity, and mutations

Lake *et al.,* (2018) analyzed the enrichment of open-chromatin signals within 100 kbp around single nucleotide polymorphisms (SNPs) using GWAS data to assess the cell-type specificity associated with a mental disorder (Lake et al.). They did not, however, attempt to identify the GWAS SNP's target genes. We identified the target genes of GWAS SNPs linked to mental disorders employing our dataset of the predicted chromatin-interaction in seven brain cell-types. We consider a gene as a target only if its promoter lying within the 25 kbp genomic bin interacts with the bin that contains the GWAS mutation. We also compared the enrichment of mental diseases at GWAS regions that overlapped with sites interacting directly with a gene. Normalizing the fraction of non-brain disorder GWAS SNP overlap with sites interacting with promoters (promoter-connected) allowed for the calculation of enrichment. To get relative enrichment, we built null-model comprising of GWAS mutations associated with non-brain disorders like–breast cancer, bladder cancer, hepatitis A, ulcerative colitis, lung cancer, hepatitis C, waist-to-hip, lung adenocarcinoma, and lung disease severity in cystic fibrosis, platelet count, bone mineral density. We observed a higher enrichment of risk variants for some mental diseases compared to the null model demonstrating cell-type specificity in relation to promoters, supporting earlier reports (Figure 2.7A). For instance, promoter-connected regions in microglia were more enriched in Alzheimer's disease risk variants (Figure 2.7A). According to reports, the development of late-onset Alzheimer's disease is correlated with increased activity of microglia hallmark genes in the cortex (Zhang et al.). Through our investigation we identified a few novel associations between genes and mental diseasase. For some diseseas, we identified the genes' possible role in the cell-type that might contribute to disease development. For instance, the promoter of the gene ALDHA1 in oligodendrocyte precursors cells is interacting with a region containing a mutation (SNP id: rs3758354) linked to schizophrenia, depression and bipolar disorder (Figure 2.7B). It's interesting to note that ALDHA1 is also known to be essential for appropriate oligodendrocyte precursor differentiation through activating the retinoic acid receptor (RXR) (Huang et al.). Its relationship to the SNP rs3758354, particularly in oligodendrocyte precursor cells, is unknown.

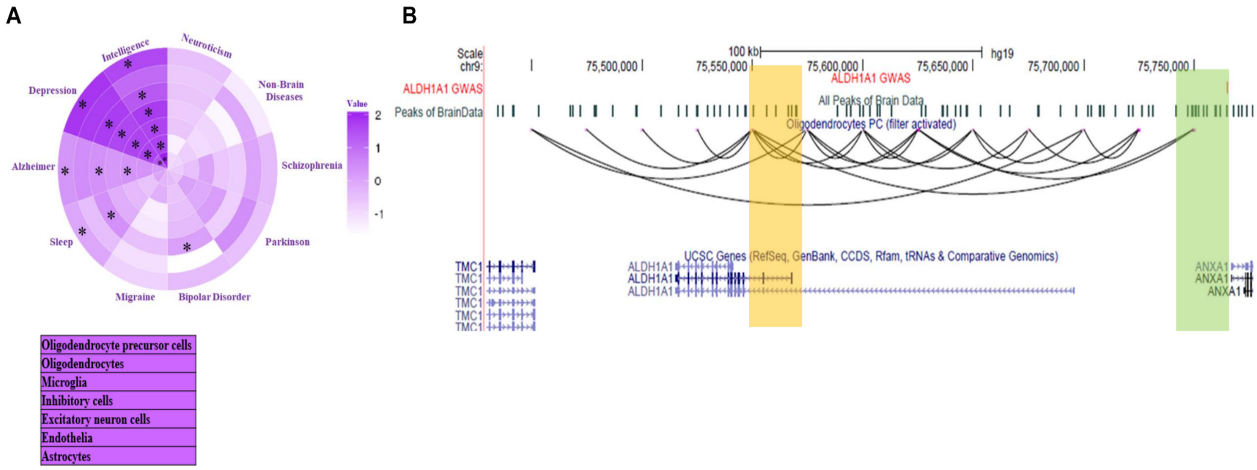


Figure 2.7. Findings regarding the target genes of disease related mutations in brain cells. (A) Enhancement of interactions between seven different types of brain cells and GWAS regions associated with mental disorders. P < 0.05 is denoted by the star (*). Using two proportional z-tests, the p-value was calculated. (B) An image from the UCSC browser depicting the promoter of the ALDH1A1 gene and the region with the GWAS SNP (rs3758354) in oligodendrocyte precursors (OPC) (Pandey et al.).

2.3.5 Human accelerated regions' targets across brain cell-types

Numerous human accelerated regions (HARs) have been found, however, only a small number of HARS's mechanisms of action and effects are understood (Hubisz and Pollard). Given that human brain structures are more complicated compared to other species, our prediction may be a valuable tool to unravel the HARs' target genes in brain cells. As a result, we intersected genomic regions lying within predicted chromatin interactions against known HARs (Hubisz and Pollard). Several HARs target genes were identified by our study (see Figure 2.8). For example, the scEChIA predicted interaction between the promoter of the SOX2OT gene and the HAR ANC980 (Figure 2.8A). According to Amaral *et al.*, (2009), SOX2OT is involved in the regulation of expression of SOX2 and neurogenesis and has numerous transcription start sites (Amaral et al.). Another intriguing interaction between HAR (ANC518) and the promoter region of a NRBF2 gene in astrocytes was also identified by our findings. The promoter of NRBF2, which is more than 500 kbp away, seems to be interacting with the HAR ANC518, which is found in the intron of the gene ZNF365 (Figure 2.8B). Therefore, employing current techniques with a “single-cell open-chromatin profile” would not have been capable of predicting such distal interactions (distance > 500 kbp).

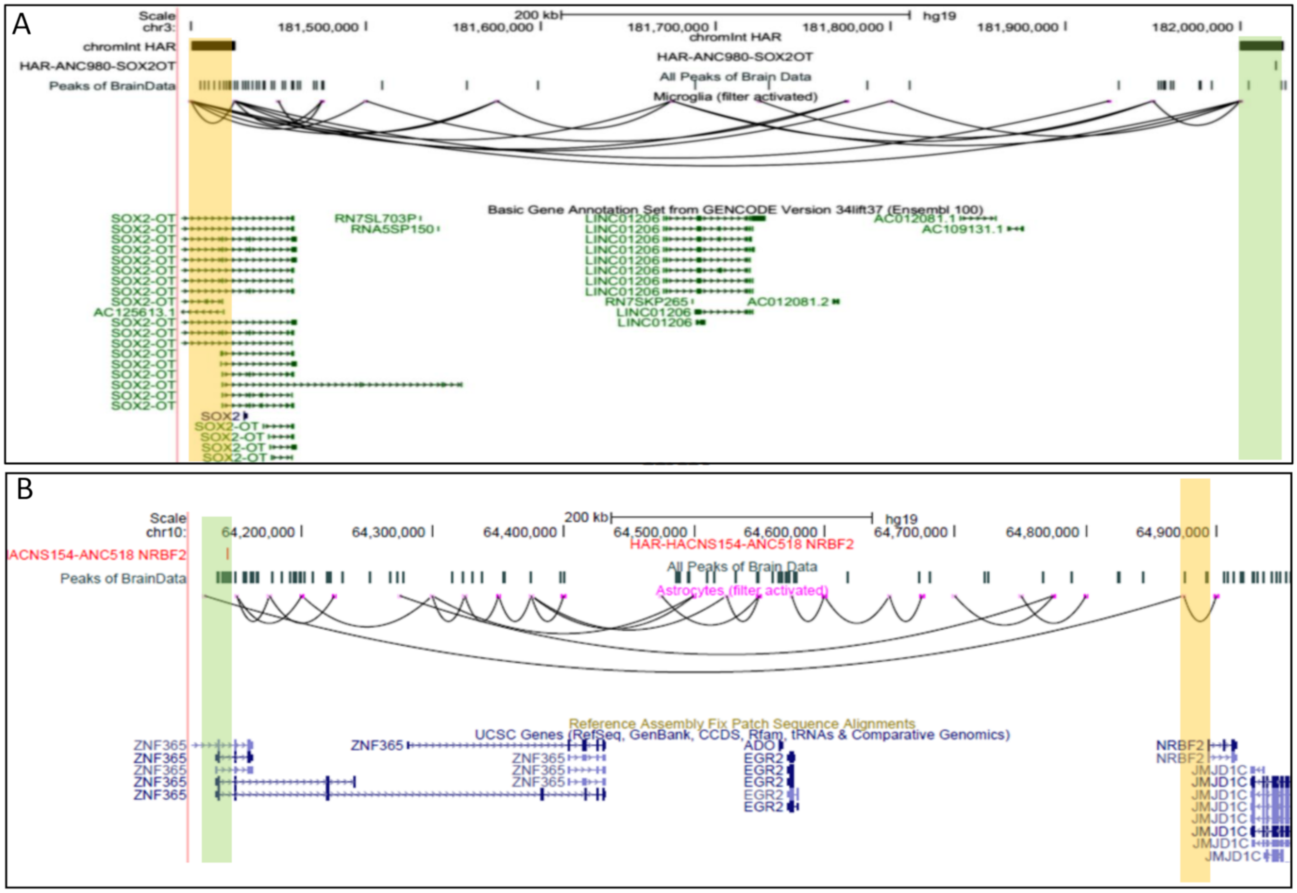


Figure 2.8. Snapshots of the UCSC browser for genes related with Human Accelerated Elements (HARS). (A) An estimated interaction between a genomic bin containing a human-accelerated region (HAR) and the SOX2OT promoter in microglia is shown in the UCSC browser snapshot. (B) The image displaying a connection between a HAR-containing area and the gene's promoter for the NRBF2 gene, which is linked to Alzheimer's disease (Pandey et al.).

2.3.6 Predicted distal chromatin interactions provide insights regarding regulatory transcription factors

To clarify the significance of identifying long-distance chromatin interactions to infer regulatory networks in brain cells, we carried out TF motif enrichment at non-promoter sites with predicted chromatin interactions. Primarily, we used HOMER (Heinz et al.) to perform motif enrichment analysis for non-promoter genomic regions with chromatin interactions for endothelial cells.

Then, in endothelial cells, we chose non-promoter genomic locations with long-distance chromatin interactions (>500 kbp). We discovered that the majority of the TF motifs significantly enriched in all interacting sites were also enriched within genomic loci with long-range interactions. Nevertheless, interferon regulatory factors (IRF) did not appear to be as enriched in locations with long-range interactions (>500 kbp) in endothelial cells, among the top three enriched motifs in all the genomic loci with predicted chromatin interaction (Figure 2.9A). Other TF motifs that were significantly enriched in genomic loci with predicted long-range contact did not seem to be significantly enriched in endothelial cell sites with all chromatin contacts. For instance, the top 3 TF motifs [STAT6, histone nuclear factor P (HINFP), and EBNA1] were not significantly enriched in all chromatin interaction sites in endothelial cells, despite being enriched in sites with long-range interactions (Figure 2.9B). The Epstein-Barr virus is connected to the viral protein known as EBNA1. Further research is required to understand HINFP's (or MIZF's) function in endothelial cells. Among the most intriguing and enriched motifs is TF STAT6, which, according to a few studies, is activated in endothelial cells from the brain in response to external stimuli (Tozawa et al.; Dozio and Sanchez; Fasler-Kan et al.). This finding implies that STAT6 may be poising or regulating the gene expression of endothelial cells via long-range chromatin interactions. It also emphasises how our approach can make it possible to gain such knowledge about how TFs regulate cellular processes utilizing their “scATAC-seq” profiles.

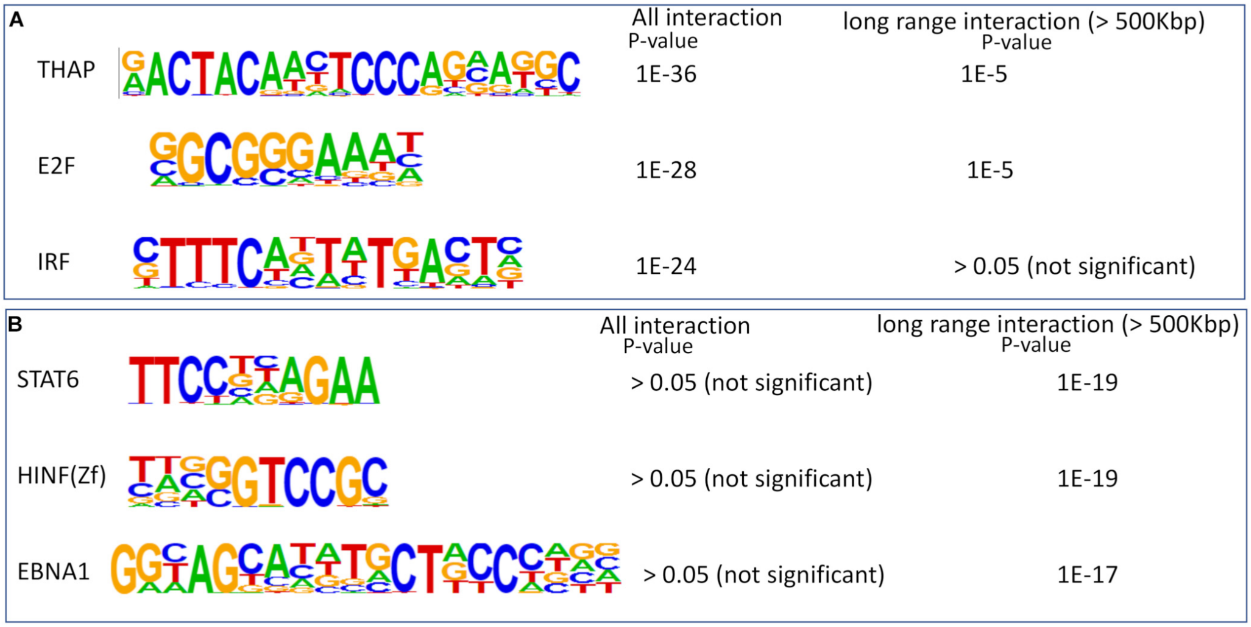


Figure 2.9. Enrichment of motifs of transcription factor at locations with predicted chromatin-interactions in the endothelial cells. (A) The top 3 motifs enriched at chromatin interaction locations (both short and long-range). Also displayed is the p-value for enrichment in two categories of sites (all and only long-range). (B) The top three enriched motifs at locations with distal chromatin-interaction. Also displayed are the p-values for enrichment in two different types of sites. Notice that among top 3 enriched motifs in sites with predicted distal interaction(>500kb) none is enriched at genomic sites with all predicted interactions (which include short and long interactions). (Pandey et al.).

2.3.7 Discussion

Assessment of the co-accessibility of the genomic sites enables the prediction of chromatin interactions but co-accessibility between genomic sites could be because of various reasons, therefore previous methods' predictions were restricted within 500 kbp. Our method overcomes these obstacles by employing prior information as a prior for calculating a constrained estimate of chromatin-interaction. In comparison to other methods, scEChIA's feature like adaptive L1 normalization teachnique for the estimation of "Gaussian graphical model" and noise reduction using "matrix factorization" allows for the prediction for a greater number of distal interactions (distance >500 kbp) utilizing "single-cell open-chromatin profiles". Additionally, we have demonstrated that, even in terms of accuracy, our method may be superior to currently used comparable methods for few sparse "single-cell open-chromatin profiles". There are numerous applications for chromatin interaction prediction utilising a “single-cell open-chromatin profile”. Researchers may find it helpful to use the predicted chromatin-interaction in seven different types of brain cells in this work help better understand regulation in the human brain. The availability of the chromatin interaction profile is rare, particularly for cells from the *in vivo* brain sample that are in their natural state. The usefulness of predicted long-range chromatin-interactions through scEChIA is reflected by a large number of overlapping brain eQTLs and its target gene contacts (1,000–4,500 eQTLs). With the use of our methodology and our predictions, number of conclusions can be drawn, including the cell-type specificity of the GWAS loci's target, novel associations between genes as well as alternative promoters and disorders, targets of HARS, and alternative splicing due to cis-regulation. For example, our data shows that a SOX2OT gene's promoters may be controlled via a HAR and may have a human-specific mechanism for regulating brain structure and function. The relationship between a HAR and NRBF2 gene that is located more than 500 kbp aside was identified as a very pertinent example by our prediction of chromatin interaction in astrocytes. It is well known that Alzheimer's disease, the 7th leading cause of mortality worldwide, causes the autophagy-associated gene NRBF2 to express less in human brain with Alzheimer's disease (Lachance et al.). We can therefore gain insight into therapeutically significant regulatory mechanisms by using our technique to identify chromatin interactions.

Previous research has highlighted a few cases of how long-range chromatin interactions allow TFs to regulate gene expression. The priming of ESCs by NANOG is a very important example (Novo et al.). Long-range promoter-SE interactions are more frequent in ESCs compared to ESCs with a deficiency in Nanog (Novo et al.). Our investigation shows that differential enrichment for transcription factor motifs in sites with all predicted interactions and just long-range interactions in endothelial cells also indicates an important regulatory pattern. Both STAT6 as well as IRF are implicated in endothelial cells' inflammatory response (Tozawa et al.; Yan et al.). We observed that the IRF motif was absent at genomic loci with distal contacts but enriched at sites with chromatin interaction. However, in brain endothelial cells, the STAT6 motif were only enriched at locations with distal interactions. Our findings led us to hypothesize that STAT6 would preferably activate genes in brain endothelial cells by long-range chromatin interactions, while IRF might do so via short-range chromatin interaction. These illustrations show how our approach can be used to infer gene-regulatory networks from scATAC-seq patterns.

CHAPTER 3

RECOVERY OF TRUE SIGNALS FROM SINGLE-CELL OPEN-CHROMATIN PROFILES USING A FOREST OF IMPUTATION TREES

3.1 Introduction

Epigenome profiles can now be applied more widely for investigating biological and clinical samples by using high-throughput sequencing. Numerous epigenome profiles, including DNA-methylation patterns, chromatin-accessibility, and histone modifications (Ernst et al.), are being used for looking into active, poised, and repressed regulatory components with in genomes (Rivera and Ren). Epigenome profiles have proven to be particularly helpful for describing non-coding regulatory regions such as enhancers (V. Kumar, Rayan, et al.). Epigenome profiling was mostly carried out in the preceding decade utilizing bulk samples made up of millions of cells. In samples of tumors or early embryonic stages, bulk sample epigenome profiles do not aid in the identification of poorly described cell populations and rare cell-types. Heterogeneity is exhibited among single cells of the same tissue in response to environmental stimuli, including among *in vitro* culture of cells. Bulk epigenome profiles generally fails to capture this heterogeneity. Furthermore, the transcriptome and epigenome patterns of cells can also exhibit variability among cells. Single-cell RNA-seq (scRNA-seq) profiles may not accurately depict phenomena like chromatin poising or bivalency at numerous genes. Researchers have created methods to characterize genome-wide epigenome patterns in single cells to help explain these problems. “Single-cell open-chromatin” detection technology has recently been used to create large-scale single-cell epigenome profiles (Cusanovich, Hill, et al.; Cao et al.), besides the fact that profiling of histone modification and DNA methylation (Hongshan Guo et al.) for single-cells is possible (Rotem et al.). Different types of procedures, such as DNase-seq (Dnase I hypersensitive sites sequencing), MNase-seq (Micrococcalnuclease-based hypersensitive sites sequencing), and ATAC-seq (Transposase-Accessible Chromatin utilising sequencing), can be used for “single-cell open-chromatin profiling” (Buenrostro, Wu, Litzenburger, et al.). A “single-cell open-chromatin profile” may be able to identify a genome's active and poised regulatory regions. Most significantly, it has recently aided in the understanding of how transcription factors (TFs) regulate the behavior of cells that are transitioning (Jia et al.). “Single-cell open-chromatin profiles” have shown to be helpful for identifying chromatin-interaction patterns in addition to offering a view of heterogeneity among cell states (Pliner et al.). After merging reads from many cells or using bulk samples that match, peak-calling is the first step in the analysis of single-cell open-chromatin profiles. The total number of reads lying on peaks is then calculated for each cell. In order to capture the signals at cell-type-specific regulatory elements in heterogeneous cell-types, researchers typically use a greater number of peaks, sometimes exceeding more than 100000 in number (Cusanovich, Hill, et al.).

3.1.1 Enhancing the single-cell epigenome signal to examine cellular heterogeneity

The read-count matrices for single-cell epigenome profiles are often very sparse owing to poor sequencing depth and the limited volume of genetic material obtained from single-cells, which calls for imputation approaches. It may be possible to emphasize only universally open sites, such as insulators and the promoters of housekeeping genes, by utilizing a minute number of hyperactive peaks to minimize sparsity. On the other hand, “single-cell open-chromatin profiles” with a lot of peaks have a better probability of incorporating cell-type-specific sites, but this comes at the cost of an excessive amount of noise and sparsity. Mostly, there are two causes for the sparsity in a read-count matrix of “single-cell open-chromatin profiles”. The first cause can be high drop-out rate, that limits the detection of many active genomic regions (false zeros). The existence of several repressed genomic sites due to their cell-type-specific activity is a real biological phenomenon, is the second cause. As a result, read-count matrix of the "single-cell open chromatin-profile" has higher fractions of true and false zeros than the scRNA-seq data. Given these restrictions with "single-cell open-chromatin profiles", it is challenging to classify and divide cells into subgroups, which is a requirement for many imputation techniques. Most imputation techniques designed for single-cell RNA-seq (scRNAseq) profiles may not work satisfactorily on "single-cell open-chromatin datasets" for the reasons stated above. Therefore, a second-generation imputation approach is required for an accurate assessment of DNA accessibility using "single-cell open-chromatin profiles". This method must be able to surpass the inability of various other similar methods to manage high degree of noise and sparsity. The optimal signal-recovery approach should be able to identify these sites, which operate as enhancers because there are many non-coding sites with activity specific to particular cell-types. The issue of imputation becomes increasingly pressing and difficult, especially with the current droplet-based single-cell ATAC-seq (scATAC-seq) methodology (Lareau et al.), which provides profiles of many cells with low sequencing depth. Even while "scATAC-seq" profile imputation has received less attention, various imputation strategies for scRNA-seq datasets have been developed. The pioneer approach for imputing scRNA-seq profiles is called MAGIC (van Dijk et al.). Utilizing the method of heat diffusion, MAGIC detects missing expression values by disseminating information among related cells. The method used by MAGIC involves standardizing the similarity ratings among single-cells to create a Markov transition matrix (van Dijk et al.). The transition matrix calculates the weights for additional cells while imputed expression for a single cell is computed. In contrast to traditional KNN-based imputation methods (Troyanskaya et al.), MAGIC uses a variable value for K in its K nearest neighbour (KNN) approach. Because they treat nearly all zero read-counts as missing values, methods like MAGIC have tendency to introduce artifacts and blur out real biological variability. Another attempt at imputation on drop-out genes was made by the scImpute approach (Li and Li). scImpute first uses a mixture model for the distribution of read counts and estimate the chance of drop-out for each gene in each cell. Utilizing the information derived from the same gene in similar cells with comparable characteristics, scImpute estimates the missing values at false zeros. Due to the irregular distribution of tag-counts with a very high drop-out rate and noise, techniques like scImpute, which rely on parametric method for estimating drop-out rate, may not be successful for single cell epigenome profiles. The likelihood of locating the appropriate neighborhood of cells is a vital step for the majority of imputation methods like scImpute and MAGIC. However step of estimating correct neighborhood in not efficient with the "scATAC-seq" profiles due to the high amount of sparsity and noise. Deep count autoencoder (DCA) (Eraslan et al.) is a deep learning-based method that has most recently been suggested for denoising and imputing single-cell expression profiles. DCA models the distribution of the genes using an auto-encoder and makes predictions about it using a prior with zero-inflated negative binomial distribution. The distribution's mean parameter is used with DCA to find representative denoised reconstruction of missing values. It would seem logical to apply DCA to single-cell open-chromatin profiles. Nevertheless, single-cell chromatin profiles are significantly more sparse than single-cell transcriptome profiles, hence it may not always be successful to simulate the distribution of their read-counts. Several techniques for visualizing and grouping "scATAC-seq" profiles have been developed (H. Chen et al.). In order to recover missing read-count values in the "scATAC-seq" profile, SCALE (Xiong et al.) also employs auto-encoder, while scOpen relies on positive-unlabeled (PU) learning strategy for imputation (Z. Li, Kuppe, et al.). In a similar manner, signal amplification and extraction processes are carried out by SCATE (Ji et al.), but this method uses previously published bulk open-chromatin profiles that contain known peaks.

3.1.2 Identification of minor cell-types underlying cellular heterogeneity

Every technique has its own advantages and disadvantages. For instance, autoencoder-based learning is frequently impacted by the majority group, which increases the possibility of losing knowledge about minor cell-types. Therefore, in addition to clustering and visualization, recovering drop-out read-count in the “scATAC-seq” profiles has remained an unresolved issue that needs to be addressed for numerous applications of “scATAC-seq”. We were aware of the limitations for “single-cell open-chromatin profiles” caused by incorrect classification and modeling of the distribution of tag counts to estimate the drop-out rate. Therefore, using an ensemble of imputing trees, we created a strategy that can get around these limitations by avoiding suboptimal solutions. Our ensemble-based methodology is called the Forest of imputation trees (FITs). Using criteria that are helpful for single-cell open-chromatin analysis, we benchmarked FITs using multiple “scATAC-seq” profiles from several cell-types. We demonstrate that FITs accurately recover the chromatin accessibility of locations such as enhancers with cell-type-specific activity without performing over-imputation using the “scATAC-seq” profiles. We also demonstrate that FITs improves dimension reduction and classification accuracy for “scATAC-seq” profiles more effectively than other approaches. Furthermore, we demonstrate that, in contrast to existing imputation techniques, FITs can manage unbalanced scATACseq datasets, minimise false heterogeneity detection, and enhance minor cell-type detection. Next, we demonstrate that “scATAC-seq” profile prediction of chromatin interaction is similarly improved by FITs-based imputation of the read-count matrix.

3.2 Materials and methods

3.2.1 Data preprocessing

We initially evaluate data's quality and eliminate any peaks which have non-zero read-count within all cells. We take log transform of the data and normalise the read-counts from scATAC-seq. As a result, the read-count x_{ij} on a site g_j in cell i is shown as:

$$\overline{x_{ij}} = log(x_{ij} / \mu_i + 1.01) \qquad (1)$$

where μ_i is mean read-count in the cell i. FITs receive as input the log of the read-normalized count's matrix. In order to prevent the possibility of infinite values during optimization, we have utilised pseudocount of 1.01 here rather than 1, similar to scImpute (Li and Li).

3.2.2 Improving imputation using clustering with randomized features

Finding the relevant subclasses is a difficult endeavor because of the sparsity and noise in the "single-cell open-chromatin profile". So, we combine clustering with imputing in a semi-randomized manner.

Our technique involves two stages: The repeated clustering in the form of a hierarchical tree and imputation at each node constitute the first phase, which involves creating numerous imputed versions of the same raw matrix. In contrast to other suggested techniques, we do not perform clustering using all of the features at once, neither we fully rely on the classification using the raw read-count matrix. We employ an iterative methodology for each tree, with preliminary imputation at each parent node and cell categorization at the subsequent level. Classification is performed at all nodes, not just the lowest nodes (at the third layer here). Using results from several trees in the first phase, a final imputed matrix is put together in the second step.

Phase-1: The first phase is implemented as described below:

Step 1: The transpose of the read-count matrix X is given, where cells are depicted by columns and peaks by rows, use a base technique to implement a preliminary imputation over matrix X utilizing all of the cells in one class.

Step 2: Using t-SNE or SVD (singular value decomposition) to reduce the dimensions of the imputed data, choose n sites (peaks) randomly. The range of 50 to 100 percent of all peaks is used to choose the number of selected peaks, n, in this case. K-mean clustering is applied to categorize the cells after lowering dimension. The number of clusters k is randomly selected in range of 2-8.

Step 3: The raw read-count of the cells in every class is assembled after classes are identified using k-means clustering. Some peaks appear to have 0 read-count (minimum) in all cells of that class after assembling the raw read-count matrix for that class. The peaks with all zero signals are therefore treated as true zeros and deleted when imputation is performed on the raw read count matrix of cells belonging to a class.

Step 4: For the different classes of cells, the base method of imputation is applied separately.

Step 5: Using the method described above in steps 2 and 3, the imputed matrix of each class is then utilized to find sub-classes. Once more, the peaks (features) and values for k for the k-mean clustering are chosen at random.

Step 6: A separate matrix is constructed using the non-imputed raw read-count vectors of the cells in a given subclass. Once more, peaks with 0 read-count in all cells of a sub-class are eliminated, and imputation is carried out separately for the matrix of every sub-class.

Step 7: A full matrix is created by compiling the imputed read-count matrix from each sub-class. As a result, the value zero is assigned to the sites dropped in cells that belong to a subclass. Notice that a full matrix version is also created using imputation for several classes at the first level. We thus get two different versions of the imputed matrix from each tree. To obtain an ensemble of imputation trees, above steps 1 to 7 are repeatedly performed.

In phase 2, using the procedures outlined below, the results of many trees from phase 1 are further processed.

Phase-2: Following steps are taken in phase 2

Correlations between a cell's unimputed read count vector and the imputed versions from Phase-1 are calculated for each cell. The final imputed vector is the average of the m most correlated imputed versions for each cell. One only needs to take the highest correlated imputed version when m = 1. The amount of m is determined by the user and may range from 1 to the number of trees created in phase 1. For benchmarking FITs on various datasets, we have utilized the default value of m = 3 in this instance. If there are less than three imputed versions and the user does not specify a choice (m = 1), all of them are used to create the final vector.

Below is an explanation of the reason and logic behind some phase 1 and phase 2 steps:

i. To improve the likelihood of obtaining the correct cluster, initial imputation is performed at every node of the tree before dimension reduction and clustering.

ii. In addition, sub-classification is carried out to assign cells belonging to a minor cell-type or cell-state the chance to come together and enable more precise imputation.

iii. In order to avoid forcing the imputation of cells that belong to a majority class using smaller groups of cells, the imputed version of the read-count at level-1 of a tree is also collected.

iv. To select the best k imputed version in phase 2, we apply spearman correlation. When comparing unimputed and imputed read-count vectors using various distance measures, we found that spearman correlation-based selection of the most appropriate imputed version produced the greatest results.

The phase of selecting imputed vectors with the strongest correlation with unimputed read-count is inspired by the minimization criteria used by practically all imputation algorithms. The difference between imputed and non-imputed matrices is reduced at observable features in classical imputation approaches focused on discovering lower rank matrices and preventing under and over imputation. In other words, we can state that FITs uses the reduction criteria twice, once during imputation at each tree node in phase 1 and once during phase 2.

3.2.3 FITs base imputation techniques

Even though FITs is made to be resilient to manage imputation-related errors, it is worthwhile to go through the base imputation technique that underlies FITs. As will be discussed below, the base imputation method employs the nuclear norm minimization with singular value soft-thresholding strategy.

Given a read-count matrix (Y) of a set of cells, where the rows correspond to individual cells and the columns to peaks. One may refer to the observed read-count matrix Y as a sampled representation of the true ideal matrix X. It can be represented as:

$$Y = A(X) \tag{2}$$

$here,$
$X \; - \; True\ ideal\ matrix\ of\ shape\ (n, m)$
$Y \; - \; read\ count\ matrix\ of\ shape\ (n, m)$
$A \; - operator\ matrix$
$n \; - \; number\ of\ individual\ cells$
$m - \; number\ of\ peaks$

Here, A is an operator matrix that subsamples and has 0's for elements of X that are not observed and 1's for those that are observed. Given the read-counts in Y and the sub-sampling mask A, the imputation challenge at hand is to recover the entire matrix X.

The approximate rank of the matrix X is frequently unknown, making it difficult to find a solution to equation 2. Researchers employ a different approach to address these problems. Researchers attempt to solve equation (2) with the restriction that the solution is low-rank for this reason. It is possible to write this mathematical illustration as:

$$minrank(X)\ such\ that\ Y\ =\ A(X) \quad (3)$$

The issue itself is NP-Hard. Nevertheless, numerous research (Candès and Recht; Candes and Plan) use nuclear norm minimization, which is its closest convex surrogate, for matrix completion. The term for the nuclear norm minimization is:

$$\overset{min}{X} ||\ X\ ||_{*}\ such\ that\ Y\ =\ A(X) \quad (4)$$

The nuclear-norm here, represented by $||.||_{*}$, is the total of the singular values in the data matrix X. The l_1 norm of the vector of singular values of X can be used as a rigorous and convex replacement for this minimum rank constraint. Consequently, the following modified form of the aforementioned equation is suggested as a solution (Candes and Plan):

$$\overset{min}{X} ||\ Y\ -\ A(X)||_{F}^{2}\ +\ \lambda||X||_{*} \quad (5)$$

The Lagrange multiplier λ is shown here. The issue in the equation does not have a closed-form solution (Rotem et al.) . As a result, it is solved repeatedly. At iteration k, as shown below, we employ the majorization-minimization strategy to tackle this problem.

$$\overset{min}{X} ||\ B\ -\ X\ ||_{F}^{2}\ +\ \lambda||X||_{*} \quad (6)$$

$$Where\ B_{k+1} = X_{k} + \frac{1}{\alpha}A^{T}\ (Y\ -\ A(X_{k})) \quad (7)$$

$Utilizing\ the\ equation\ \ \overset{min}{X}||M1\ -\ M2||\ >\ ||s1\ -\ s2||$

We can formulate the minimization problem as where s1 and s2 are singular values of matrices M1 and M2.

$$\overset{min}{X} ||s_{B} - s_{X}||_{2}^{2}\ +\ \lambda||s_{X}||_{*} \quad (8)$$

Where $||s_{x}||$ is the absolute sum of the singular values of X, and s_{B} and s_{x} represent the singular values of B and X, respectively (Cai et al.). Thus, soft thresholding (Cai et al.) is frequently used to solve the minimization problem in equation (Cao et al.) in the manner described below.

$$s_X = sign(s_B)\, max(0, |s_B| - \lambda/2\ ) \tag{9}$$

It's been observed that the method is resistant to the value as long as it is relatively small (Li and Zhou) .

3.2.4 Analyzing chromatin interaction predictions and calculating co-accessibility among sites

To determine genomic site co-accessibility, a covariance matrix was computed using a read-count matrix representing a single cell's open-chromatin profile. However, estimating the real covariance matrix is not trivial because the number of elements to be estimated in the covariance matrix is typically more than the amount of data-points in the read-count matrix. The covariance matrix might also not always accurately reflect direct interactions between genomic locations. As a result, Graphical Lasso (Friedman et al.) is a good solution for this kind of issue. Calculating partial correlations between variables (genomic site) can be done with the help of the Graphical Lasso approach, which also aids in estimating regularized covariance and the inverse of the covariance matrix (Friedman et al.). When the effects of other factors are taken into account, the partial correlation serves as a measure of the degree of relationship between two variables. Graphical Lasso (GLasso) seeks to identify a limited proportion of real partial correlations between variables given the noise and small quantity of the data. GLasso uses a penalty term which shrinks partial correlations between many pairs to value zero, if there is not enough strength in the estimate of their association. Our aims is to maximize the objective function of GLasso:

$$log\ det\, \Theta\ - tr\,(U\Theta) - \rho||\Theta\,||_1 \tag{10}$$

$where$

$\Theta:\ inverse\ covariance\ matrix\ having\ the\ dependence\ structure\ of\ variables\ and$

$U:\ their\ covariance\ matrix\ and$

$\rho : \; penalty\ term\ for\ L1\ norm\ based\ regularization$

Worth mentioning that, we did not adopt Cicero's (Pliner et al.) method of including a penalty term that varied with the distance between genomic regions because we did not want to neglect distal interaction. Additionally, the purpose of this analysis was solely to assess the improvement achieved in the prediction of co-accessibility by imputation. In this case, we set ρ equal to 0.001. We combined peaks that were within 25 kbp of one another and also added their read counts prior to determining co-accessibility. In other words, read-counts were calculated using bins of 25 kbp. Both imputed and non-imputed read-count matrices were subjected to this task, which was followed by the computation of their covariance matrices and the use of Graphical Lasso. Using the inverse covariance matrix obtained by Graphical Lasso, we determined partial correlation values i.e. co-accessibility scores, between each pair of the genomic region (merged peaks or bins). We obtained the HiC data and processed chromatin interaction files for K562 and GM12878 from Rao et al., (Rao et al.) in HiC data format. We used Juicer (Durand et al.) to extract the interaction from files in .hic format, and we then transformed the result to a six-column bed format with scores. We selected high confidence interactions with p-values <1E-9 out of all interactions. To compare high-confidence HiC-based chromatin interactions with co-accessibility-predicted interacting peak-pairs, we utilized PGLtool (Greenwald et al.) .

3.2.5 Clustering and separability evaluation metrics

After dimension reduction of the imputed and non-imputed read-count matrices using t-SNE (t-distributed Stochastic Neighbor Embedding) (Walldén et al.) , we carried out k-means clustering. To evaluate various characteristics of clustering and imputation, we used two evaluation metrics.

The first method, known as the adjusted Rand index (ARI), has a cost for false positives and false negatives, where "positive" denotes the assignment of two identical cells to the same cluster and "negative" denotes the assignment of two similar cells to different clusters. Let $V = [v_1, v_2, .., v_k]$ be the clustering result with k clusters having n_i number of observations in cluster v_j and $T = [t_1, t_2, .., t_p]$ represent the true p classes consisting of $n_i$number of data in class t_i. The ARI is computed as:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}\binom{n_{ij}}{2} - \left[\sum_{j=1}^{p}\binom{n_i}{2}\sum_{j=1}^{k}\binom{n_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{j=1}^{p}\binom{n_i}{2} + \sum_{j=1}^{k}\binom{n_i}{2}\right] - \left[\sum_{j=1}^{p}\binom{n_i}{2} + \sum_{j=1}^{k}\binom{n_i}{2}\right]/\binom{n}{2}} \qquad (11)$$

$where$

$$n = \sum_{j=1}^{k} n_j = \sum_{j=1}^{p} n_i$$

Further, the second method which we used was cell-type separability (CTS). We first look for the spearman correlation between the read counts of each potential cell pair, and then calculate CTS. The intra-cell-type median correlation is then determined by calculating the median correlation for pairs of cells that belong to the same type. Then, only those pairs where one cell belongs to each of the two kinds are used to determine the inter-cell-type median correlation. The difference between intra-cell-type and inter-cell-type median correlations is termed as CTS. Therefore, the CTS value is always calculated between two cell-types.

3.2.6 Data sources

The datasets used for this manuscript can be found in open access databases. Downloaded in SRA format were the reads for a "scATAC-seq" profile by Buenrostro *et al.,* (Buenrostro, Wu, Litzenburger, et al.) (SRA ID: SRP052977). In order to match the reads to the human genome's hg19 version, Bowtie was used (Langmead and Salzberg) . The same study by Buenrosto *et al.,* at GEO database (GSE65360) provides the peaks of the ATAC-seq human cell-types used here.

Bedtools were used to merge the peaks. Next, for each peak in the combined peak list, the read-count for each cell was estimated. The GEO database's (ID: GSE96772) scATAC-seq read-count for immune cells was downloaded (Buenrostro, M. Ryan Corces, Lareau, Wu, et al.). GEO ID: GSE111586 makes single-cell ATAC-seq read-counts for mouse adult liver and bone marrow cells available. We also employed pathway enrichment score-based analysis for the dataset with the same GEO ID: GSE111586. About 39 primary cell-types, including "collision" and "Unknown," got annotations. We used the peak-list of three cell-type ATAC-seq profiles from another GEO database: BJ, GM12878, and H1ESC for evaluation (GEO ID: GSE65360). Other bulk ATAC-seq profiles used in this study included those with the GEO IDs BJ: GSE113414, GM12878: GSM1155958, H1ESC: GSM2083754, HL60: GSM2083754, and K562: GSM1782764. The K562 and GM12878 cell lines' chromatin interaction files were obtained from Rao *et al.* (GEO ID: GSE63525) (Rao et al.) .

3.3 Findings and applications of Forest of imputation trees for detecting true signals in single-cell open-chromatin profiles (FITs)

The idea that biologically similar cells would have comparable levels of activity at a regulatory site can be utilized to infer the missing values. Therefore, an imputation approach has a high likelihood of producing accurate results if we cluster similar cells into a subcluster. Nevertheless, establishing the correct sub-cluster is not an easy process given the noise, sparsity, and imbalance in the “single-cell open-chromatin dataset”. Consequently, our approach combines randomization with multiple hierarchical tree-based grouping, imputations, and a base technique (Figure 3.1). Every node in the tree uses the base imputation approach, which employs a well-known method of soft thresholding of singular values for matrix completion (see the "Materials and Methods" section). Our method uses a hierarchical tree-based approach, and as a result, we first do an initial imputation for all the cells, grouping them together. We divide the cells into K classes using the initial imputed read-count matrix (nodes). We do imputation using each cell's raw read count for each class, disregarding the previously imputed matrix. Multiple peaks (genomic sites), however, have zero read counts in all the cells that comprise a class when we construct the read count of those cells. This is just what we expected, and we make use of it to improve imputation. When performing class-wise imputation, we treat those peaks that have 0 read counts in every cell in a class as true-zeroes and remove them. For additional classification, we once more employ the imputed read-count of cells with non-dropped peaks that belong to a class. As a result, we obtain cell subclasses, and we once more aggregate the raw read count of cells that belong to a subclass. One crucial detail to be aware of is the fact that we randomly select between 50% - 100% of the non-dropped peaks for classification for each level. At each phase, we also generate a random number for the number of classes k in the k-mean clustering. As a result, we carry out numerous instances of this hierarchical tree-based clustering and imputation while selecting the features and k (number of classes) at random. The best jth column of the a cell j the multiple-imputed read-count matrix is chosen based on correlation of the vector raw read-count and final imputed matrices from numerous such trees. It is based on our finding that the correlation between cells with similar types has a higher spearman correlation than cells with different types. Therefore, in comparison to accurate imputation, the imputed vector for a cell will have a poorer correlation with the unimputed version if imputation is performed by clubbing it with incorrect neighbors. A critical filtering step that helps FITs prevent over-imputation is the final step of selecting the best m vectors from the many imputed versions (Figure 3.1).

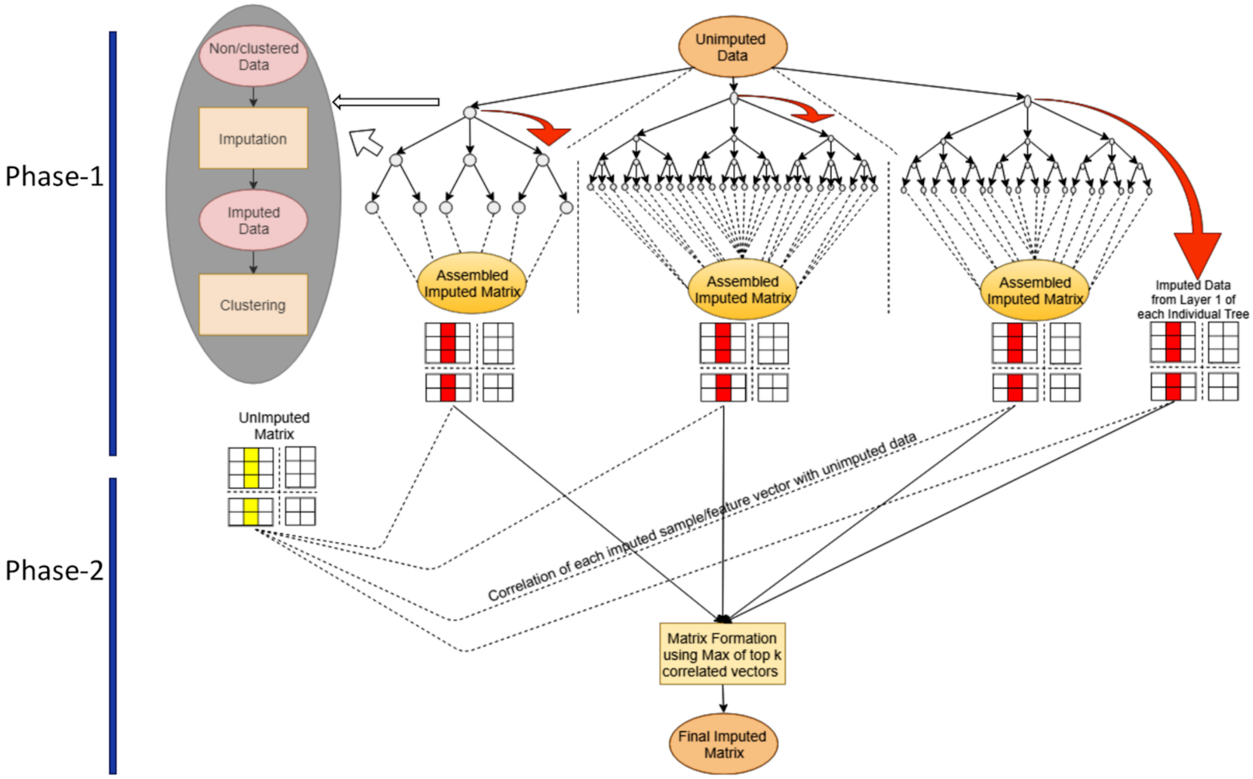


Figure 3.1. An explanation of FITs: FITs entails two phases. In phase 1, a large number of imputation trees are constructed to obtain several imputed versions of the original read-count matrix. Inside an imputation tree, each node first does a base method imputation on the non-imputed read-count matrix, then dimension reduction. After that, k-mean clustering is carried out to produce k clusters of cells. One daughter node receives the raw-read count of the cells of each cluster and applies the same imputation and clustering techniques. Every node of the imputation tree drops the sites that have zeros in all cells of its assigned raw read-count matrix. The vectors of the imputed matrices (shown as red column) are compared to their unimputed original versions using correlation in phase 2 of FITs. For each cell, only the imputed versions that have the strongest connection to its un-imputed read-count vector are selected. (Sharma and Pandey *et. al.*).

3.3.1 FITs can manage unbalanced read count matrices and recover the signal of minor cell populations

Even in the presence of dominant signals from major cell-types, the method of hierarchical steps of imputation and clustering combined with randomization in FITs has the potential to reveal minor population clusters. In order to assess how well imputation techniques performed on the unbalanced “scATAC-seq” profile dataset from adult mouse bone marrow and liver published by Cusanovich *et al.*, (Cusanovich, Hill, et al.), we performed imputation. Cusanovish *et al.*, performed annotation and assigned cell-type for the majority of the cells in the adult mouse “scATAC-seq” dataset. Bone marrow data contained 4033 cells, and liver data had 6167 cells in the combined “scATAC-seq” dataset. After imputation, we used the eight most common cell-types and cells with the label "unknown" to perform t-SNE-based dimension reduction and visualization for both datasets. There was an imbalance in numbers even among cell-types that had been preserved.

There was an even greater imbalance in the number of different cell-types in the liver “scATAC-seq” data. Hepatocytes made up 91.4% of the retained cells in the liver “scATAC-seq” data, while 6 other cell-types had frequencies of less than 1.6%. Results for imputation of liver “scATAC-seq” profiles also showed that other approaches (KNNimpute and DCA) entirely failed and were unable to assist in segregating minor cells in t-SNE-based visualization (Figure 3.2A). However, FITs-based imputation resulted in the division of minor cell-types into different groupings. Few cells labeled as "unknown" in the FITs-based imputed version of the liver dataset co-localized with hepatocytes in the t-SNE-based scatter plot (Figure 3.2A). To highlight enhancers in "unknown" cells overlapping hepatocytes, we adjusted the raw read-counts. As a result, annotation of cells marked as "unknown" was also made possible via FITs-based imputation. None of the four evaluated tools (chromVar, cisTopic, SCALE, and scOpen) could display minority cell-types independently like FITs-based data for the liver “scATAC-seq” profile (Figure 3.2B).

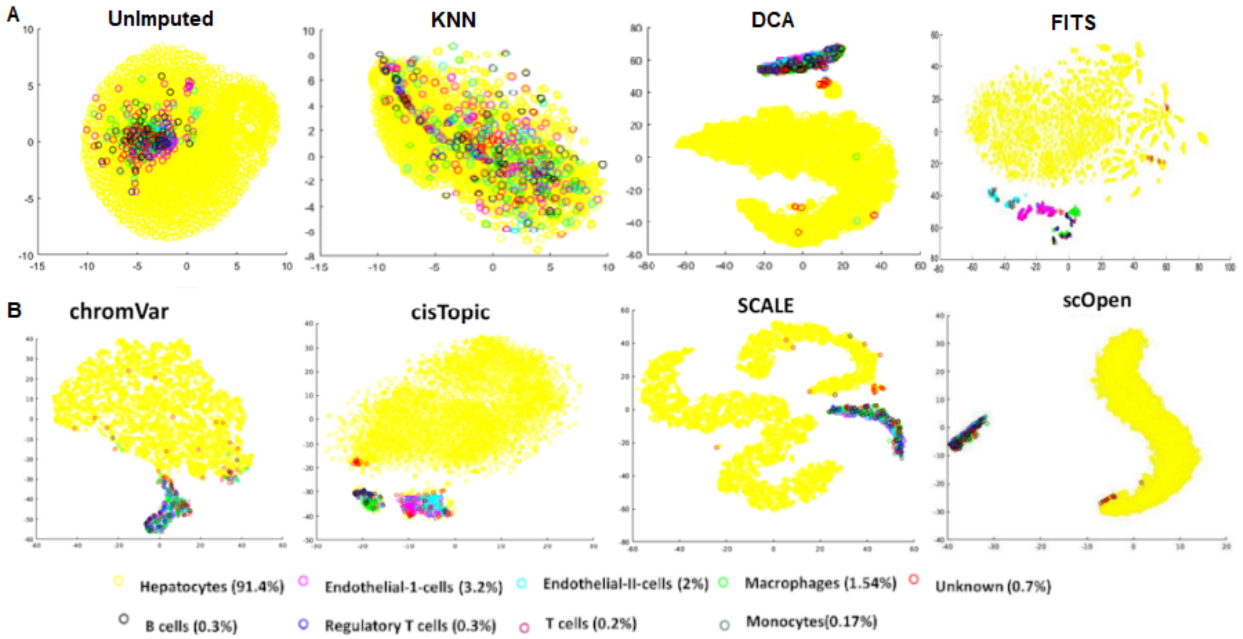


Figure 3.2. FITs capture signals from minor cell-types in an imbalanced data-set of single-cell ATAC-seq profiles from *in vivo* samples. Here, scATAC-seq read-count matrix for cells in adult mouse liver was used (A) t-SNE based embedding plot for unimputed (raw) and read-counts matrix imputed by different methods. (B) t-SNE based plots made using four other tools (chromVar, cisTopic and SCALE, scOpen) designed for scATAC-seq profiles (Sharma and Pandey et. al.).

3.3.2 Assessment of imputation techniques in mouse BoneMarrow scATACseq datasets with regard to separability of minor cell-types

Four cell-types (macrophages, B cells, dendritic cells, and T cells) were found in the bone marrow "scATAC-seq" dataset with a frequency of less than 2%, while two cell-types (hematopoietic progenitor + erythroblast) accounted for more than 70% of the total cells.

The visualization of t-SNE results for the standard imputation methods such as KNNimpute and DCA, were shown to be more ineffective at recovering signals of minor cell-types from bone marrow datasets. In t-SNE plots, minor cell-type sites and major cell-type locations were messed up for KNNimpute. However, FITs was effective in recovering signals even for minor cell-types for bone marrow data. Whereas, the t-SNE plots for the read-count matrix imputed by FITs demonstrated the separability of tiny groups of cells. For example, macrophages formed a group that was very distinct from other cell-types. We also evaluated how well FITs performed in comparison to four additional tools, chromVar (Schep, Wu, et al.), cisTopic (Bravo González-Blas et al.), SCALE (Xiong et al.), and scOpen (Z. Li, Kuppe, et al.), which have been demonstrated to be effective for visualizing and grouping "scATAC-seq" profiles.

We discovered that the other four methods (chromVar, cisTopic, SCALE, and scOpen) could not recover the signal of dendritic and T cells, which had frequency 1.3% (Figure 3.3B), using the boneMarrow "scATAC-seq" profile.

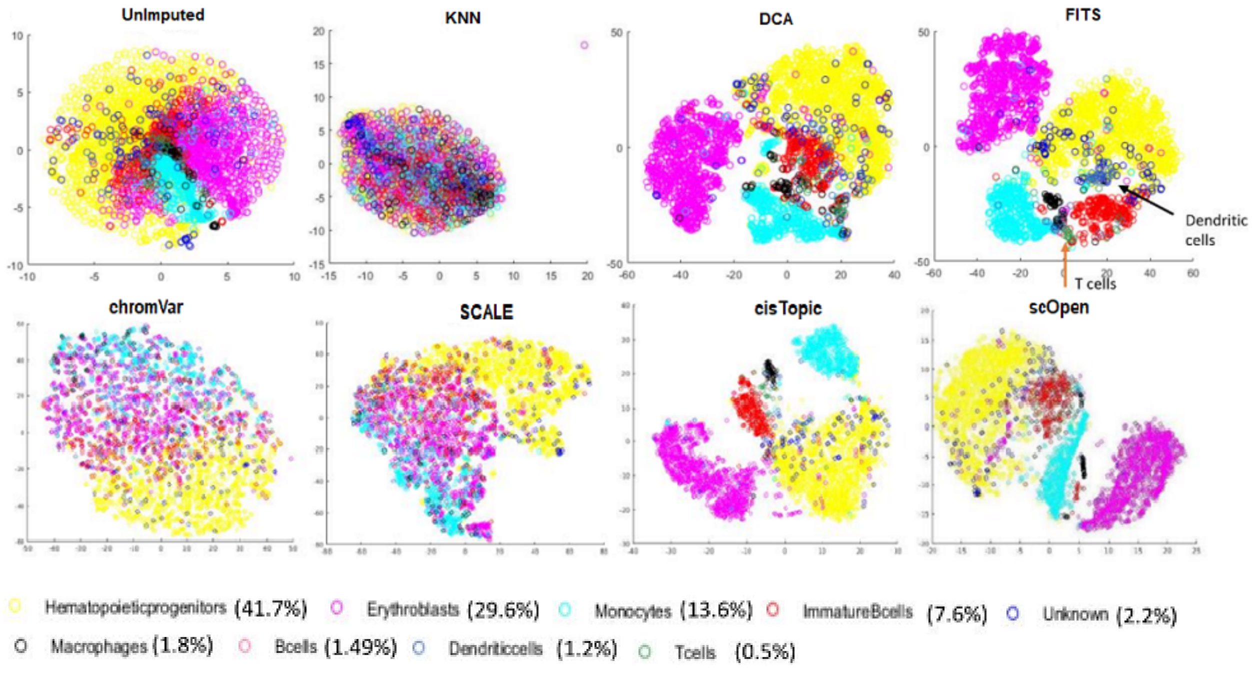


Figure 3.3. Assessment of imputation methods in terms of separability of minor cells in mouse bone marrow scATACseq datasets (A) Visualisation of tSNE results for unimputed and imputed read-counts matrix. The frequency of different cell-types is also mentioned with their names. (B) Scatter plot of results from 4 tools (chromVar, cisTopic and SCALE, scOpen) designed for visualization of scATAC-seq profile. It is to be noticed that dendritic cells and T cells could not cluster together in any of the plots, like FITs-based results (Sharma and Pandey et. al.).

3.3.3 FITs increases the precision of analyzing atlas size scATAC-seq profiles and corresponding calculations of cell-wise gene-set enrichment

Other than visualization and clustering, “scATAC-seq” profiles have several other applications also. For example, Chawla *et al.*, (S. Chawla et al.) recently created a technique called UniPath to transform “scATAC-seq” read-count to gene-set enrichment score for every single-cell. Gene-set enrichment can be used to infer regulatory patterns in a cell. It is possible to use UniPath to annotate the cells using their “scATAC-seq” profile when gene-sets of cell-type markers are utilized (S. Chawla et al.). By dividing the read count by a previously created global accessibility score, UniPath first highlights cell-type specific peaks (potentially enhancers) for the purpose of estimating gene-set enrichment. Then, it computes gene-set scores using genes close to peaks with strong cell-type specificity as a foreground. We initially assessed whether imputation using FITs can improve the output of UniPath using gene-sets of known cell-type markers.

FITs-based imputation significantly enhanced UniPath's performance when estimating the geneset enrichment score. The fraction of cells with the correct cell-types in the top five terms is significantly greater with imputed read-count than with their unimputed version, as shown in Figure 3.4A for three sets of cells. According to Chawla *et al.*, (S. Chawla et al.), the conversion of read-counts to pathway scores offers a further method for addressing large-scale “scATAC-seq” profiles while maintaining consistency and horizontal scalability. Chawla *et al.*, only visualized atlas scale data for the scRNA-seq profile; they did not do so for the “scATAC-seq” profile. Because of this, we visualized the atlas scale “scATAC-seq” profile created by Cusanovich et al., utilizing the read-count to pathway scores transformation for more than 68,000 cells. Using pathway scores computed for unimputed and FITs-based imputed read-counts, we performed t-SNE-based visualization. Cells of the same kind co-localized in the t-SNE plot were created using the pathway enrichment score for the imputed read-count matrix, as shown in Figure 3.4B. The t-SNE plot, on the other hand, revealed significant overlap across various cell-types using pathway enrichment scores derived from unimputed read-count. Our findings thus demonstrate that the analysis of atlas-scale “scATAC-seq” profiles using gene-set enrichment scores can be significantly enhanced by FITs-based imputation.

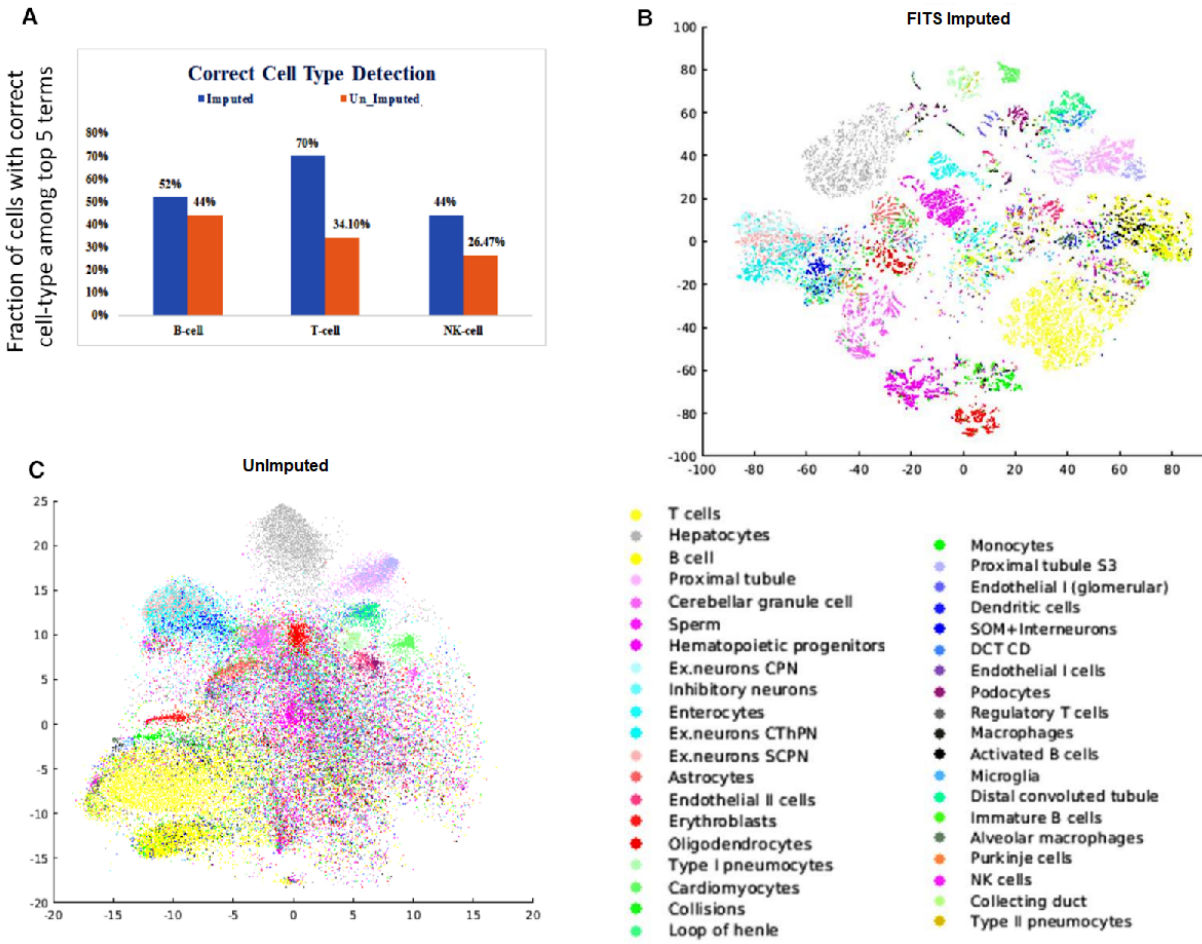


Figure 3.4. Figure shows how FITs improves analysis utilising the scATAC-seq profile-calculated gene-set enrichment score for each single cell. (A) The fraction of cells with the correct cell-type terms present in the top five enriched gene-sets, as determined by UniPath calculations. The scATAC-seq profile of cells from Cusonovich *et al.*, was imputed using FITs (Cusanovich, Hill, et al.). (B) Pathway scores generated using read-counts imputed using FITs are used in the t-SNE-based visualisation. (C) A visualization of the t-SNE results for pathway scores derived from UniPath's unimputed read-count (Sharma and Pandey et al.).

3.3.4 FITs enhance single-cell open-chromatin profile's ability to detect chromatin interaction

After assessing FITs for their potential to aid in the calculation of cell similarity, we looked into whether imputation might be used to estimate co-accessibility between sites. According to a recent theory put forth by Pliner (Pliner et al.), areas with significant co-accessibility in "single-cell open-chromatin profile" are very likely to be interacting. The comprehension of gene regulation and the identification of the target genes of non-coding genomic regions highlighted by GWAS (Zhang and Lupski) are two examples of how chromatin interaction maps are useful in many processes. To forecast interactions between genomic locations, Pliner *et al.*, used a Graphical Lassov-based (Friedman et al.) technique. The performance of the graphical Lasso approach (Friedman et al.), which is used to lessen the impact of noise and calculate direct interactions, could be enhanced by giving less sparse data. We used a graphical Lasso-based approach to assess imputation-based advancement in the "scATAC-seq" based prediction of chromatin interaction. For evaluation, we used the Rao *et al.*, published HiC-based chromatin interaction profile for K562 and GM1287 (Rao et al.). For both the K562 and GM12878 cell lines, we were able to recover high-confidence chromatin interactions at a resolution of 25 Kbp using hic files. We combined the peaks within 25 Kbp in the "scATAC-seq" read-counts matrix for both K562 and GM12878 cells and used Graphical Lasso to find intrachromosomal chromatin interactions. The overlap between predicted and truly high confidence chromatin interactions for the two cell-types K562 and GM12878 was indeed enhanced by 10–30% for various chromosomes using FITs-based imputation (Figure 3.5). One crucial point to note is that, in contrast to CICERO, we did not concentrate on predicting interaction solely within a specific distance range. In contrast, we also predicted intrachromosomal interaction among sites lying far apart. The examination of single-cell open-chromatin profiles can be done in a variety of ways using FITs, including chromatin interaction prediction.

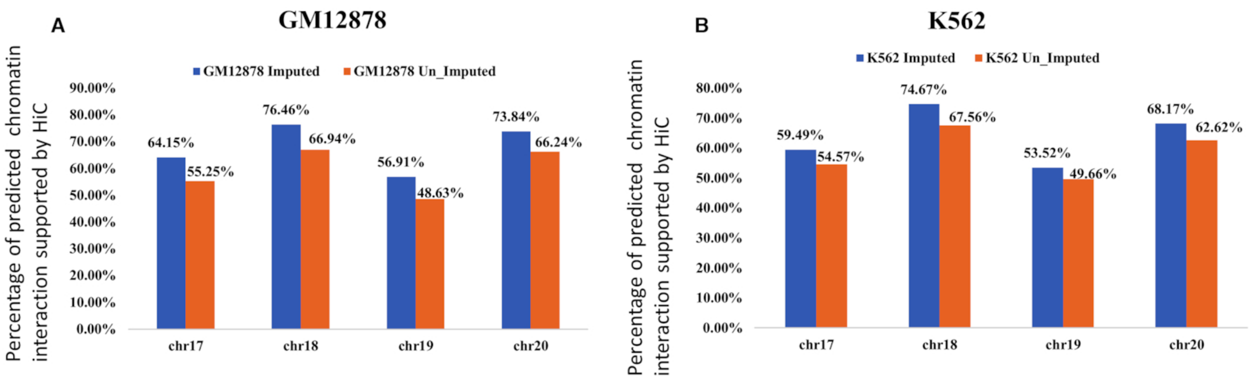


Figure 3.5. The prediction of chromatin-interaction using the scATAC-seq profile is improved by FITs-based imputation. This chart illustrates the fraction of predicted interactions that overlap with HiC's chromatin-interaction predictions based on co-accessibility. (A) for GM12878 cells; (B) K562 cells (Sharma and Pandey et al.).

Discussion

Compared to RNA-seq and DNA methylation profiles, the pattern of signal and sparsity in "single-cell open-chromatin" are different. Because of this, the imputation of "scATAC-seq" profiles requires attention and cannot be handled similarly to "scRNA-seq dataset". Furthermore, it is not entirely acceptable to perform binarization instead of imputation for "scATAC-seq" read-count. Since every single strand for two homologous chromosomes might contribute to the pool of DNA fragments, we can have a read-count value of four reads despite having a narrow peak of 200 bp. For "single-cell open-chromatin profiles", it is necessary to investigate the effects of imputation in various downstream analysis steps. In order to improve the downstream analysis needed for "single-cell open-chromatin profiles", we first assessed whether imputation of "scATAC-seq" might be helpful. Second, we created a technique for imputing "single-cell open-chromatin profiles" that can manage significant levels of noise, sparsity, and cell-composition bias. The three features that make FITs effective are randomised sub-clustering, imputation within multiple trees to prevent suboptimal solutions, drop-out decision after clustering, and selection of imputed vectors depend on correlation with unimputed version to prevent false-imputation.

Here, we also, demonstrated how approaches that rely on parameters for a single clustering phase increase the likelihood of artefacts resulting from classification errors. Completely relying on a small number of non-randomized classification steps further increases the chance of becoming stuck in local minima and missing actual heterogeneity. In contrast to other imputation techniques, we have demonstrated here that FITs executes imputation in such a way for the “scATAC-seq” profile that there is reduced risk of identifying false heterogeneity. Particularly when the dataset is unbalanced, classification frequently fails to recognize the minority population as a separate class, leading to artefacts when other imputation techniques are used. The detection of unusual cell states utilizing “scRNA-seq profiles” has been the subject of numerous research; however, using “scATAC-seq”, such analysis is rare because of overwhelming noise and imbalances in datasets. Our analysis of FITs suggests that it may be possible to identify rare cell states using “scATAC-seq” read counts, and that this may open up new possibilities for the study of clinical *in vivo* samples. FITs outperformed four other “scATAC-seq” profile-specific approaches (chromVar, cisTopic, SCALE, and scOpen) for the “scATAC-seq” profiles of cells from *in vivo* samples. We demonstrated the application of FITs for analysis steps specific to “scATAC-seq” profiles, chromatin interaction prediction, and calculation of gene-set enrichment score for single cells, in addition to the recovery of minor-cell signals. In order to improve inference and provide new applications for “scATAC-seq", FITs can collaborate with a variety of existing tools.

A benefit of FITs is that, due to horizontal scalability, they can accommodate massive read-count matrices. The massive read-count matrix can be divided into numerous smaller matrices with randomly selected cells to execute Phase-1 of FITs. The same cell may appear in two smaller matrices, but the union of all the smaller matrices should correspond to every cell in the original dataset. Before the final matrix compilation occurs in Phase 2, the Phase-1 of FITs can be executed for a number of small matrices on several computers. Large read-count matrices are rarely supported by other imputation techniques. Thus, the issue of imputing larger read-count matrices is also solved by FITs.

Fewer groups have attempted to use single-cell profiles in multiomics investigations to provide a global perspective on development and disease (Lake et al.). Due to the reliability, it offers during analysis, the main benefit of FITs is that it would motivate more researchers to investigate single-cell open-chromatin profiles for multiomics studies. Few studies have also used other single-cell epigenome profiles, including histone modifications, MNAse-seq, and DNAse-seq. Because of FITs' universality, it is suited for other types of single-cell epigenome datasets as well. FITs may therefore be further modified for other types of single-cell epigenome profiles in the future.

CHAPTER 4

ANALYSIS OF ACTIVITY AND INTERACTIONS OF CHROMATIN DOMAINS IN CANCER SAMPLES AND THEIR RELEVANCE IN DRUG-RESPONSE OF CANCER CELLS

4.1 Introduction

4.1.1 Use of chromatin profiles in the regulation of gene expression in cancer

The organization of the chromatin and the regulation of the activity of numerous genome segments perform a important function in cellular state transitions and cell responsiveness. The relationship between chromatin organization and various characteristics of cancer cells (L. Li et al.) and variable rates of mutation in their genomes (Polak et al.) has been researched by a number of groups. According to Jones *et al.*, mutations in genes of enzymes involved in chromatin organization are present in more than 50% of human malignancies (Jones et al.). Even if genetic mutation and evasion of apoptosis perform a function in cellular the establishment of a cancer state, regulatory modifications caused by perturbations in the process of distinct chromatin segments are every time crucial to such transitions. Previous research indicates that cancer therapies are likely to prevent certain cancer cells, perhaps if they don't have any resistant mutations (Pisco and Huang). Several previous research studies have attempted to identify the causes of cancer cell drug resistance. Using mutation and expression profile, the “cancer cell line encyclopaedia (CCLE)” team attempted to predict responsiveness to various drugs (Barretina et al.). Many teams have also attempted for providing non-genetic explanations utilizing reprogramming as well as stemness, that are related to enhanced xenobiotic resistance, specifically because of the expression of “efflux pumps” (Kenyon and Gerson; DeNicola et al.; Nakasone et al.; Rosenzweig; Challen and Little; Zhou et al.), greater effective DNA repair (Kenyon and Gerson), resilience and stress response (Dean et al.; Donnenberg and Donnenberg; Medema). Other research teams have tried to articulate it in the context of the heterogeneity of tumours and the ability of a small number of cancer cells to switch states. The progression of cancer, including associated drug-resistant various forms is largely influenced by the organisation of chromatin in each case.

4.1.2 Regulatory role of TADs in the context of disease development

It is defined that chromatin is organised into a variety of domains, which are sometimes referred to as TAD. TADs have naturally found segregated genomic regions. It has been found that TADs remain mostly stable at the time of differentiation (Dixon, Jung, et al.) as well as across a variety of cell types (Dixon, Selvaraj, et al.; Rao et al.). Non-mammalian genomes are likewise organized in these domains in addition to mammalian genomes. In reality, several studies have shown that a small number of TAD boundaries have been conserved over evolutionary history, with about 54 percent of TAD boundaries throughout the human genome likewise being preserved at homologous regions of mouse genomes. (Dixon, Selvaraj, et al.). In addition to the presence of TADs in the genome, the co-localisation of genes also show a pattern influenced by evolution and cellular processes. The greatest known cluster of tumour suppressor genes, for instance, is known to be located at the gene-dense region “3p21.31” on chromosome 3. Previously there have been study of multiple deletions in location “3p21.31” on chromosome 3 in several tumors (Angeloni). There could be a number of causes for the synteny and co-localization of genes with same or related functions. Pressure from selection to prevent the recombination of desirable alleles inside a locus is one of the causes. Paralogs are more frequently found within the same TAD as well as random pair of genes, according to research by Salem *et al.,* (Ibn-Salem et al.). TADs specify gene-sets that eukaryotic cells activate at various stages of development and in various cellular states in well orchestrated manner. The rate of mutations over TADs and its boundaries are being examined in contexts of cancer progression and survival studies (L. Li et al.). TAD-boundaries were shown to co-localize with modifications related to the somatic mutational impact of the cancer genome, according to a study by Akedmir *et al.,(Akdemir et al.)*.

TAD-boundary changes in the progression of cancer development has been examined in several studies. Gene and enhancer activity is insulated by TAD boundaries from the influence of neighbouring domains (Gong et al.). Therefore, aberrant gene activity within a TAD may be a increasingly obvious sign of the efficient breakdown of this TAD boundary. In order to explore the impact of cancer and its relationship to drug resistance, we employed a method to quantify the activities of a gene-set composed of genes within individual TADs.

“HNSC (Head and neck cancer)” cell lines' chromatin interaction patterns were the starting point for our initial analysis. In order to study how cancer cells respond to drugs, we used “patient-derived primary cultures (PDCs) of head and neck squamous cell carcinomas” developed from Chia *et al.,* (Chia et al.) and studied in other research (Suphavilai et al.). In order to comprehend the patterns of activation of genomic domains in drug resistance, we characterized the chromatin interaction landscapes for 4-HNSC cell lines. By analysing the transcriptome for “819 cell lines contained in the CCLE database” and the responsea to 544 drugs listed in the CTRPv2 portal, we broadened our investigation. Additionally, we applied our method to interpret the transcriptome-profiles of 9014 patients with 20 different cancer-types that were made accessible via TCGA consortium.

4.2 Material And Methods

4.2.1 Experimental method

At the Genome Institute of Singapore, HNSC patient-derived cell lines "HN137P, HN148P, HN120P, and HN120M" were cultured, along with subsequent ChiP-seq as well as “Hi-C” based profiling in addition sequencing was carried out. The generation of PDC model for the HNSC cell lines is explained by Chia et al. (Chia et al.). *In-situ “*Hi-C” was performed as previously described (Lee et al.). “NEBNext ChIP-seq library Prep kit” was used to create multiplexed ChIP-seq libraries (NEB) and sequencing was carried out utilising Illumina HiSeq-2500, as described by Sharma *et al.*, (Sharma et al.) .

4.2.2 Computational analysis

4.2.2.1 Analysis of Hi-C profile

Figure 4.1 highlights our method for analyzing the profiles of chromatin interactions for several HNSC cell lines through TAD based profiling. To develop maps that cover the entire genome comprising millions of interactions among genomic loci pairs, “Hi-C” protocol generates hundreds millions of read-pairs. Recent advances in bioinformatics have enabled efficient sequence read filtering, alignment, and preprocessing as well as normalization of contact matrices to eliminate biases and estimate chromatin architecture (Forcato et al.). We used the HiCUP tool to handle “Hi-C” data for our investigation. HiCUP is a pipeline for analyzing sequence data from investigations that look at the 3D architecture of the genome through using “Hi-C” and Capture “Hi-C” (CHi-C) technologies. The pipeline removes artifacts that could otherwise hinder downstream analysis and maps data to a specific reference genome (Wingett et al.; Aljogol et al.).

4.2.2.2 HiCUP pipeline control script

Many programs, “hicup_truncater, hicup_mapper, hicup_filter, and hicup_deduplicator (hicup_digester creates the genome digest file used by hicup filter)”, are included in the HiCUP pipeline. “Hi-C di-tag paired reads” were produced from paired FASTQ files and matched with the hg19 reference genome. This HiCUP scripts regulates the pipeline's operations, which passed output from one phase of the programme to the next, and it executed each script in the proper order. The configuration document pertaining to the HiCUP programme also contains the parameters for the complete workflow. The configuration file specified the filenames of the FASTQ file pairs that need to be analyzed (Wingett et al.).

4.2.2.3 Implementation

We utilized the DpnII digest profiles of the hg19 reference genome produced by “HiCUP-digester” for computational analyses. The following data set has been utilized for recognizing "Hi-C" artefacts, represents all potential DpnII fragments inside the genome. HiCUP-v0.5.10 through Bowtie2 (hg19) was utilized for the alignment procedure, and the minimum and maximum di-tag ranges had been chosen to 150 and 1000, respectively, in the filtering stage. Whereas, for other parameters, default values were chosen. With an acceptable unique alignment score >65%, the finalized BAM output was generated.

4.2.2.4 Generation of HiC contact Matrix

To generate an HiC contact matrix, we initially converted binary alignment mapping file to an arrow-head input form by exploiting SAMtools’ ‘View’ command. The awk command was also employed to acquire the required information via a bam file. By using pre-defined command from Juicer Tools and the arrow-head input file, the normalised “Hi-C” interaction matrix were generated. With balanced KR normalization (Knight-Ruiz normalization), The captured interaction frequency for every region from the "Hi-C" map was exported using Juicer Tools' ‘dump’ command. We were able to accomplish this and get the “Hi-C” matrix at 25kb of resolution for each and every chromosome in .txt format.

4.2.2.5 Detection of TADs

To find TADs at 25kb resolution across all chromosomes, we applied a Domaincaller method by Dixon et al., (Dixon, Selvaraj, et al.; Dali and Blanchette). We began using a square interaction matrix of 25kb resolution “Hi-C” interaction map, following we used DI from matrix.pl to generate directionality indices with bin sizes of 25,000 and windows of 125,000. The input for the “HMM_calls.m” program was the vector containing directionality indices. Further, we processed HMM using “file_ends_cleaner.pl, converter_7col.pl, hmm_probability_correcter.pl, hmm-state_caller.pl, and hmm-state_domains.pl” to extract TAD calls. The following input parameters were used to execute the HMM probability correction script: “min = 2, prob = 0,99, and bin_size = 25,000” (Dali and Blanchette). Additionally, we obtained TADs using the TADKB database, It provides TADs at 50-kb and 10-kb resolutions from a range of cell types, includes human “HMEC, IMR90, K562, NHEK, GM12878, and KBM7” (Rao et al.; T. Liu et al.). Directionality Index (DI) was used to determine the TAD sites for each cell type (T. Liu et al.).

4.2.2.6 Analysis of ChIP-seq profiles and Calculation of TAD activity score

Using DFilter (V. Kumar, Muratani, et al.) the ChIP-seq profile was analyzed to call peaks and create custom tracks. The DFilter help page's guidelines were followed while setting the parameters for calling peaks. Big data sets across different assays having a variety of complicated phenotypes are widely available, which motivated our own selves to research the fundamental mechanism underlying TADs from the perspective of various biological questions.

Gene set enrichment (GSE) was carried out to make it easier to organize and evaluate TAD-genesets in our investigations. Gene expression profiles can be used to summarize a signature or a pattern using a method named gene-set enrichment analysis (GSEA). This method is preferable to single gene analysis in terms of biological interpretability, noise and also dimensionality reduction (Hung et al.). Gene-set enrichment variability inside the sample population was assessed using the GSVA programme (Hänzelmann et al.). GSVA is a non-parametric, unsupervised method that provides an unconventional approach of explicitly modelling phenotypes with gene-set enrichment scores. The most often used models for studying cancer biology, finding cancer targets, and evaluating medication efficacy includes cancer cell lines. The gene expression data for the drug-based study of numerous cell lines was acquired by browsing the “Cancer Cell Line Encyclopedia (CCLE)” (Nusinow et al.). Additionally, RNASeq information was obtained via TCGA portal (https://portal.gdc.cancer.gov/) for TAD-based survival analysis of PAN-cancer. Using data from the CCLE and TCGA databases, the TAD based gene set enrichment scores was generated for various cancer types. Heatmap was used to visualize the distinct TAD gene-set enrichment.

4.2.2.7 Calculation of survival p-values

R libraries' ‘survival’ and ‘survminer’ were utilized to determine the survival p-value (Tm and Grambsch). GSVA score and clinical data from Xena-browser, including days-to-death, gender, etc., have been the input files. Clinical data's "Status" column has two values: 0 for alive and 1 for died. The median of every row of the GSVA has been determined in order to determine the survival p-value for every TAD. A new column called "median-group" was then created, denoting 1 for GSVA score larger-than median along with 0 for GSVA score less-than median. This was followed by the merging of clinical data (days_to_death, status ‘0/1’, patient ID) with GSVA score and median group columns. In order to fit the data inside the "survfit" function of the R package's “survival and survminer”, we prepared a file for each TAD. The output of Survfit provides the survival p-value for every TAD in question. Additionally, we used the coxph tool to calculate the hazard ratio for every TAD. Additionally to the MaxStat package (Lausen and Schumacher), several cutoff combinations, such as; “51-49, 60-40, and 70-30 percentiles”, were used to describe the patient group depending on the activities of a TAD to investigate TADs importance in survival (high-low). Afterwards, TADs that were common of those combinations were chosen for further study. Maxstat allows both the cutpoint estimate, and the test technique above with various p-value approximation along with graphing of the empirical processes of standardised statistics (Lausen and Schumacher).

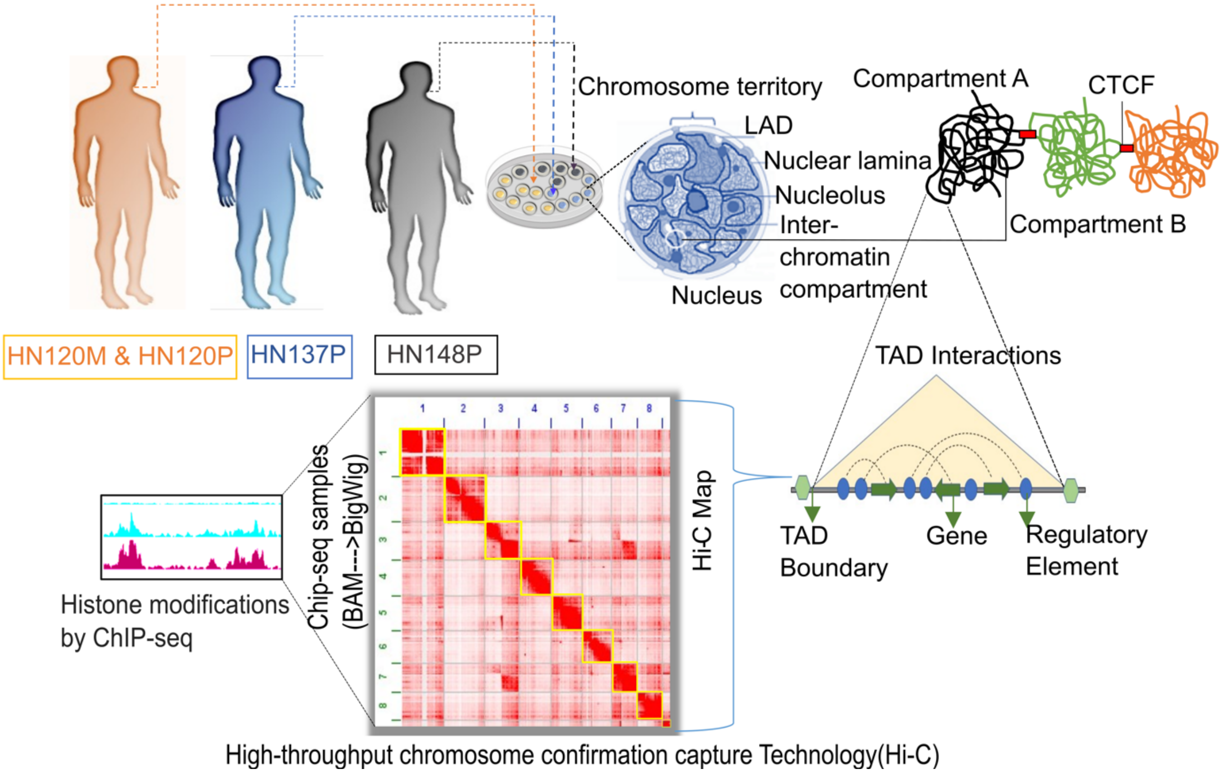


Figure 4.1. A summary of the chromatin interaction profiles of 4 HNSC cell lines and their profiling and analysis. While the other two are models of primary cancer cells from patients HN137 and HN148, the other two cell lines are from the same patient, HN120. TAD boundaries were obtained and additional downstream analysis was performed using the Hi-C based chromatin interaction profile from HNSC cell lines.

4.3 Results

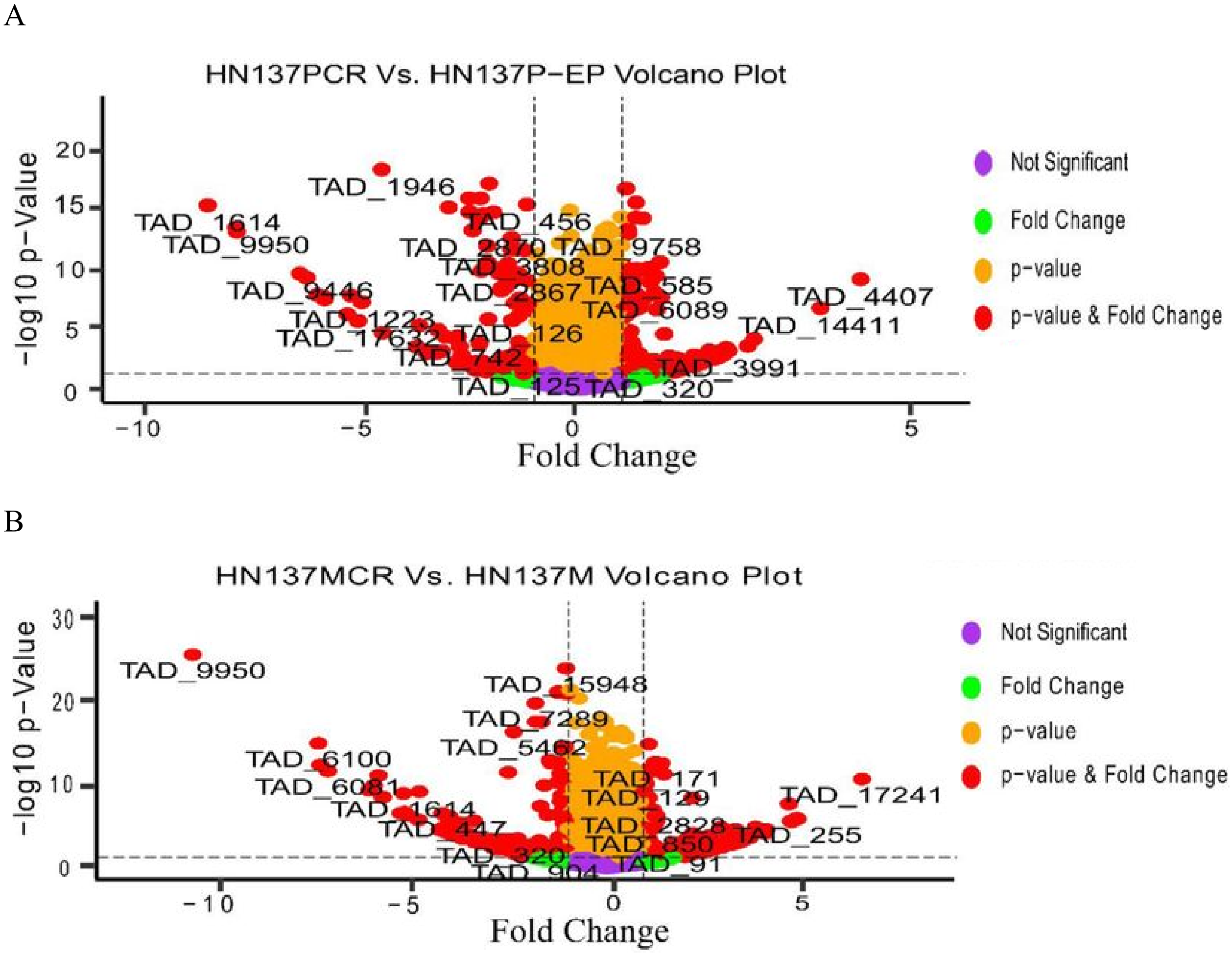
4.3.1 Oral cancer HiC profiling, patterns of TADs and association with drug resistance

We did “Hi-C” in 4 HNSC cell lines to examine their chromatin organization. Primary oral cancer cells from the primary tumours of three patients with HN120, HN137, and HN148 (HN120P, HN137P, HN148P), as well as metastatic cancer cells from patient “HN120 (HN120M)”, make up the four HNSC cell lines that we utilized for profiling chromatin interaction. In HN137P, HN120M, and HN148P, we also conducted ChIP-seq for the histone modifications H3K4me3 and H3K27ac. We used the method of assessing a TAD's activity by calculating the enrichment of the genes-set that it included. We utilised UniPath to estimate TAD gene-set enrichment with “single-cell transcriptomics” and used GSVA to compute TAD activity utilizing a bulk expression profile. (Hänzelmann et al.). We also utilized mean histone acetylation (H3K27ac) ChIP-seq read-count to confirm that TAD gene-set enrichment was indicative of its activity. Mean read-counts in TADs for H3K27ac showed positive correlation with the TAD gene-set enrichment score (p-value < 0.001). TAD gene-set enrichment scores are referred to here as TAD activity.

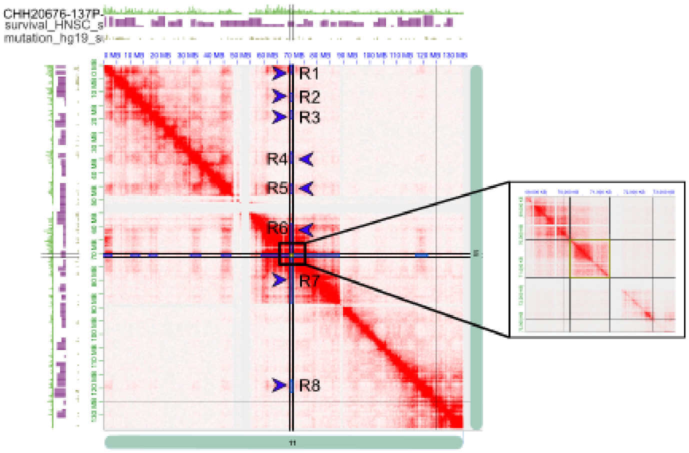
4.3.2 Change in activity of TADs with development of cisplatin resistance

We compared the TAD activity scores within oral cancer cell lines as well as those of its cisplatin-resistant versions. The “single-cell expression profiles” of the HNSC cell lines “HN137P (paired-end), HN137M, HN137PCR, and HN137MCR”, which were published by Sharma et al. (Sharma et al.) and proven to have the same genotype, were utilized for this purpose (Ramazzotti et al.). Since it was confirmed that the cell lines HN120M and HN120MCR had the exact genotype, we additionally utilised single-cell expression profiles from these lines. We discovered that some of the TADs were significantly downregulated in cisplatin-resistant versions of many HNSC cell lines “HN120MCR, HN137PCR, and HN137MCR” via the wilcoxon rank sum test with fold-change computation (Figure 4.2 A-B). There exist three crucial TADs - “chr11_70150000_71300000, chr11_70200000_71300000, chr11_70225000_71350000” are positioned in the 11q13 region of the chromosome, namely in the “11q13.3 and 11q13.4” bands, and they possessed a hierarchical structure where one TAD was a subset of the others. Five available keratin-associated protein genes; “KRTAP5-7, KRTAP5-8, KRTAP5-9, KRTAP5-10, and KRTAP5-11”, 1 actin-binding protein (CTNN) plus some additional genes with clear and unclear functions were part of gene-set of TADs found in the “11q13.3 and 11q13.4” bands and upregulated within cisplatin-resistant versions of multiple HNSC cell lines. We observed that the genomic region at 11q13.3 is closely related to different regions when we visualized chromatin interaction landscapes for chr11 (refer to Figure 4.2 C, regions “R1, R2 R3, R4, R5, R6, R7, R8”). A broader regulatory condensate may include TADs in “11q13.3 and 11q13.4”, as shown by these findings. We therefore select H3K27ac peaks (active-enhancers and promoters) found in frequently upregulated TADs and areas (R1-8, figure 4.2C) connected to them. Top motifs for the ETS family-related TFs “EWS:ERG, Fli1, and Etv2” were revealed by the motif enrichment study for chosen peaks. (Figure 4.2 D). There have been reports of “ERG and Fli1” involvement in several cancers (Martens et al.), the Fli1 itself is recognised to play a function in cell cycle regulation (Fuchs) and cisplatin resistance (Shen et al.) in cancer cells.

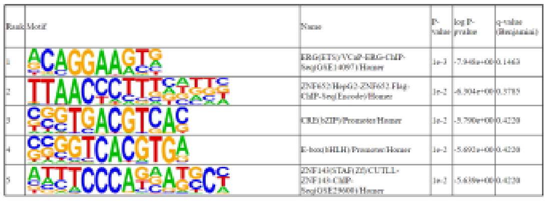
We further carried out gene-ontology enrichment for H3K27ac peaks that are found in the regions of chromosome 11 that interact with the 11q13.3 and 11q13.4 band. Further research found that, whereas other regions connecting to region 11q13.3 and region 11q13.4 lacked copy number variation, these two regions did. The region R1 interacting to the cisplatin-induced upregulated TAD in 11q13.3, also contains a large number of genes that produce keratin related protein “KRTAP5-1, KRTAP5-2, KRTAP5-3, KRTAP5-4 and KRTAP5-5”. Therefore, the HNSC chromatin profile showed these genes from the KRTAP5 family exist close together in 3D chromatin configuration and are potentially part of the exact regulatory unit, which activity is stimulated in response to cisplatin. In addition, different interacting areas R6 as well as R7 include genes related to the stress response (Figure 4.2 D). Similar to this, other accompanying areas include genes with a functional enrichment for stress response. Thus a hub of chromatin interactions may be implicated in stress response as well as cell survival in the existence of cisplatin. That was discovered by comparing TAD activity between HNSC cell lines and its cisplatin-resistant versions.



C



D



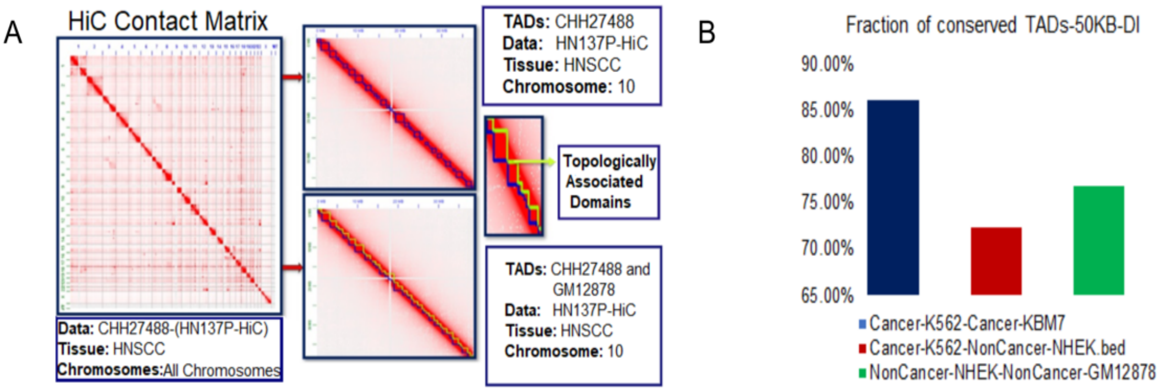
Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)
1	ACAGGAAGTG	ERG(ETS)/VCaP-ERG-ChIP-Seq(GSE14097)/Homer	1e-3	-7.948e+00	0.1463
2	TTAACCCTTTCATTTC	ZNF652/HepG2-ZNF652.Flag-ChIP-Seq(Encode)/Homer	1e-2	-6.304e+00	0.3785
3	CGGTGACGTCAC	CRE(bZIP)/Promoter/Homer	1e-2	-5.790e+00	0.4220
4	CCGGTCACGTGA	E-box(bHLH)/Promoter/Homer	1e-2	-5.692e+00	0.4220
5	ATTTCCCAGAATGCC	ZNF143|STAF(Zf)/CUTLL-ZNF143-ChIP-Seq(GSE29600)/Homer	1e-2	-5.639e+00	0.4220

R1	High sulphur keratin-associated protein
R2	positive regulation of cell growth
R3	Serum amyloid A protein
R4	Ets domain
R5	positive regulation of collagen biosynthetic process
R6	Stress response
R7	response to superoxide
R8	T Cell Receptor and CD3 Complex

Figure 4.2. A study on the chromatin interaction profiles and the domain activities of cell lines originating from patients with HNSC cancer. (A) The fold change and significance (p-value) of the activity of TAD between 137P and its cisplatin-resistant variant 137PCR are shown in volcano plot. (B) The volcano plot of fold change and significance (p-value) for TAD activity differences between the 137M cell line and its cisplatin counterpart. (C) A visualization of the chromosome 11 interaction matrix. The region inside the middle-positioned rectangle represents the downregulated TADs (in the 11q13.3 band) found in all three HNSC cell lines. Other areas that interact distally with the TAD in the 11q13.3 band are indicated by arrows as R1, R2, R3, and so on. D) The 11q13.3 band and the R1, R2, R3, R4, R5, R6, R7 and R8 area include elevated TAD and the wrenched motifs in H2K27ac. Enriched gene-ontology keywords of genes located in various locations that interact with the 11q13.3 band.

4.3.3 Generalization of our insight on drug-resistance and TAD association through analysis of CCLE data

The conservation of TAD boundaries across cell types and species has been documented by many studies, and we have further verified this pattern in cancer cells. Using the TAD boundaries from the TADKB database, we found that TAD boundaries are also preserved in a variety of cancer cell types (Figure 4.3 A-B). These findings motivated us to determine their TAD-activity amongst 819 CCLE cell lines by using a common TAD-list. In order to create a commonly available TAD list, we created a union list of TADs discovered in HNSC cell lines and TADs present in the TADKB database. We estimated the correlation among the TAD activity of TADs from the union group and the stated IC50 value for drugs in CCLE repository. When correlating data from HNSC cell lines, an intriguing pattern emerged. There was a strong association amongst the IC50 values of several different drugs. For instance, the TAD’s activity of TAD at the positions “hg19:chr13_58180000_59940000” exhibited a strong correlation “R=0.79” with the drug Importazole's IC50 value. A different TAD's activity was found to be negatively correlated to the IC50 (and positively correlated with PIC50) values of numerous drugs, namely; "Bemcentinib, Dactolisib, and LRRK2-IN-1” at the chromosomal region “hg19:chr1_12310000_13770000”. We further conducted clustering of correlation score between TAD activity and drug PIC50 values in order to investigate a systematic analysis (see Figure 4.3 C). We splitted the TAD into eight primary clusters and identified a category of TADs whose activities correlated positively with PIC50 values for one set of drugs while negatively towards PIC values for another class of drugs. The TAD activity of cluster-3 exhibited a negative correlation towards the PIC50 values of the drugs in the D1 class, as illustrated in Figure 4.3 C. The PIC50 values of the drug class D2 were positively correlated with TADs in the same cluster-3. In contrast, a drug from class D4 had a PIC50 value that was positively correlated to a TAD from cluster-2 and negatively correlated to the activities of TADs in cluster-5.



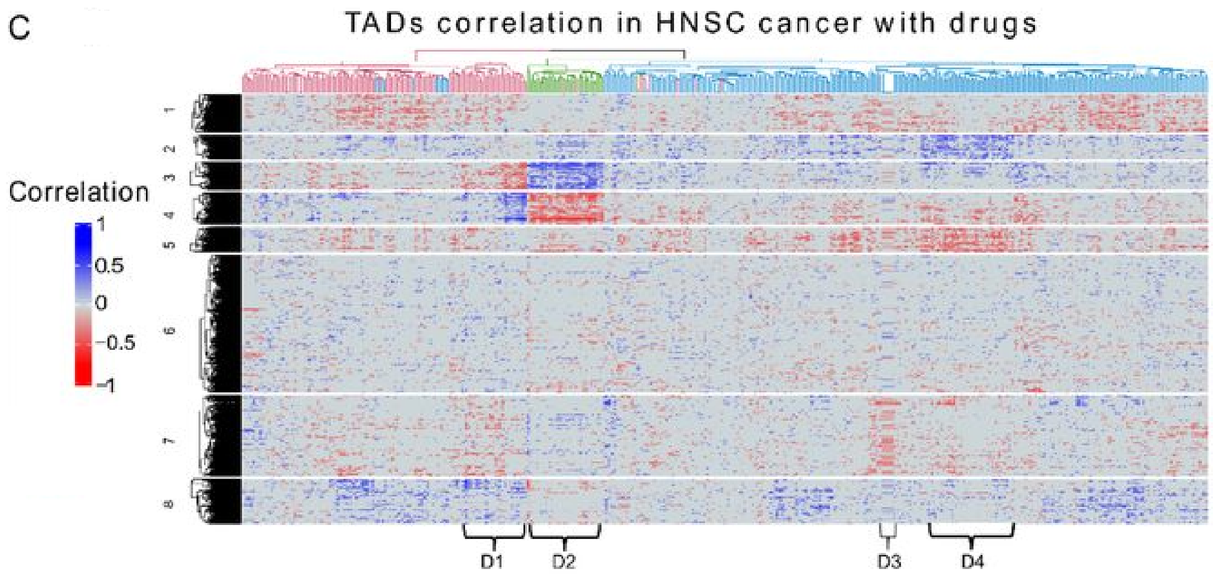


Figure 4.3. Preservation of TAD boundaries and pattern of drug response-activity association. (A) Visualization of TAD boundaries and chromatin interaction from two cell lines, (GM12878 and HN137P). For two cell lines, the TAD boundaries of the chromatin interaction in chromosome 10 are depicted in a zoomed-in form (green and blue). In the two cell lines, the TAD borders are almost identical. B) Overlapping TAD border areas between several cell types (within 1 KB of the boundary). The proportion of borders overlapping between two cell lines is displayed on the Y axis. (C) From the CCLE database, a heatmap was created to show the relationship between TAD-activity and PIC50 values for drugs in HNSC cell lines. Only TADs (in rows) whose activity has an absolute correlation value of at least 0.3 with the PIC50 of at least one drug are displayed. Similar to this, only drugs that have an absolute correlation value of greater than 0.3 with the activity of at least one TAD are displayed. Here, four drug sub-clusters are also highlighted.

The correlation among TAD activities with the drug PIC50 value was reduced when the study was performed using cell lines from various types of cancer in the CCLE database (Figure 4.4). Such a reduction in correlation values suggests that the context of the cancer type affects the relationship between TAD activity and cancer cell responsiveness to medication. We discovered various connections between the exact set of TAD and drug covering multiple cancer types. Even with the usage of cell-lines from all different forms of cancer in the CCLE dataset, however, apparently remained, some TADs of which activity was strongly associated to drug responses. For example; TADs in cluster-3-6 and cluster-8 demonstrated activity with a significant and positive correlation to its IC50 value "i.e., a negative correlation with the PIC50" of a number of drugs (Figure 4.4). A biological processes with the most strongly enriched gene sets in TADs of cluster-2 include keratinization, cornified envelope, and laminin interactions. To stop premature cell death, cornification has anti-apoptotic and anti-necroptotic properties (Eckhart et al.). But excessive cornification is also thought to be a form of cell death. The expression profile reported by CCLE was based on cancer cell lines that had been grown without drug stress; as a result, it suggests that cancer cell lines (or cells) with higher keratinization and cornification activities are more likely to have higher IC50 values.

On the other hand, the activity of cluster-1 TADs was correlated negatively to IC50 values (positively correlated to PIC50). Cluster-1 gene-enrichment revealed certain essential biological mechanisms involved in drug response. These biological mechanisms include NF-kappaB signalling, Rac protein signal transduction, negative control of interleukin-6 synthesis, etc. Interleukin (IL)-6 levels have been found to be greater in persons with various cancer types. In addition, RAC1 hyperactivation is commonly occuring in human cancer and could have been caused via overexpression, uncontrolled degradation, abnormal upstream inputs, as well as changed intracellular location (Olson et al.; Z. Wang et al.; Wu et al.; Mizukawa et al.). The overall relationship between TAD-activity and drug IC50 values for several cancer cell types indicated common drug response pathways.

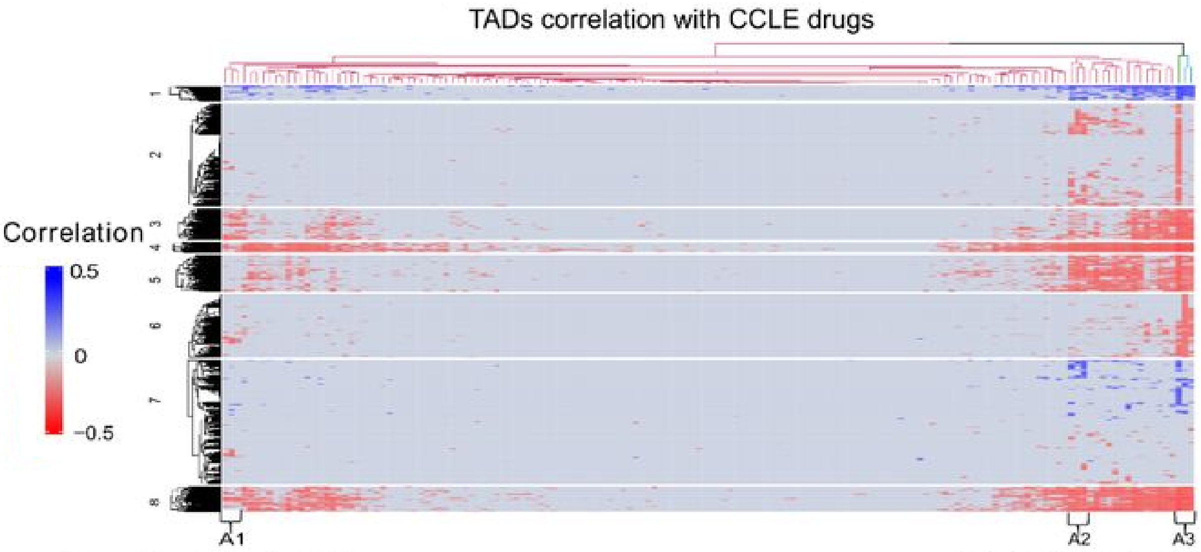


Figure 4.4. Heatmap shows the relationship between drug PIC50 values and TAD-activity for cell lines from different types of cancer in the CCLE database. Only TADs (rows) or medicines (columns) with at least one absolute correlation value greater than 0.3 are displayed. A few drug sub-clusters are also displayed.

The best correlation values between drug TAD-activities and IC50 values were equivalent to single-gene values when we solely utilised data from HNSC cell lines available within CCLE repository. However, the associations among activities of certain TADs with IC50 values of specific drugs were significantly stronger while we utilised various kind of cancer cell lines available in CCLE database than it was when we used any single-gene estimate (Figure 4.5 A). This outcome demonstrates the possibility of our method of leveraging the TAD gene-set to identify shared drug-response mechanisms in multiple cancer types. Therefore, even if drug response genes varied between cancer types, the underlying process would still have been the same. Our findings can be easily interpreted given that several paralogous genes with equivalent functions frequently co-localize. Additionally, we attempted to estimate drug PIC50 values using either gene expression profiles or TAD-activity scores (Figure 4.5 B). We discovered that for many drugs, the accuracy achieved by gene based prediction model was substantially less than the values obtained via utilising TAD-activity score, even when many genes were used to predict the PIC50 values. Such findings suggest that, despite the heterogeneity of cancer cells, TAD-activity studies have the ability to uncover a shared drug response mechanism. For drugs including ko-143, fortinib, CIL70, lenvatinib, and others, TAD activity-based prediction was substantially more accurate than gene-based prediction.

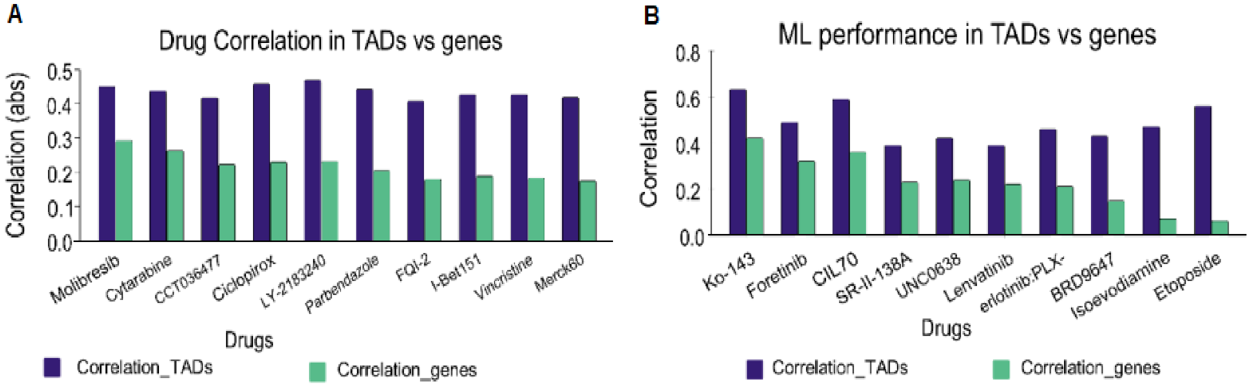


Figure 4.5. A. Barplot illustrating the best correlation between the PIC50 values of a few drugs and single gene expression (blue) or TAD-activity (green). B. A bar graph showing the association between the PIC50 values that were actually obtained and those that were predicted using a machine learning model and features that were either TAD activity or gene expression. Even when many genes are utilised to estimate PIC50 values for various drugs, the efficiency (or accuracy) was much lower than the result obtained by TAD-activity score.

4.3.4 Pattern of TAD activity in TCGA pan cancer (studying variation in TAD activity)

There might be a wide range of reasons for patient-to-patient variations in TAD activity: 1) Changes in gene expression levels brought on by variations in the cellular stage of cancer cells, mutations on TAD boundaries, as well as genomic structural variations that may generate genomic region-shuffles and the development of mini TADs. 2) CNV of TAD-containing genomic region. Nevertheless, each of these occurrences has an impact on the projected gene activity inside defined TADs. Therefore, utilising the transcriptome of cancer samples provided through TCGA group, we investigated changes in TAD activity. To do this, we created a union-list of TADs reported in the TADKB database (T. Liu et al.) and TADs discovered in HiC profiles in HNSC cells. First, we calculated the fold change of the median TAD activity between tumour versus normal samples from the same tissue type. According to the type of cancer, we discovered many TADs with increased activity. However, a small number of TADs did exhibit increased activity in multiple cancer types as compared to normal tissues. TADs from cluster-12 was highly active in various kind of cancer, as indicated in figure 4.6. The cellular glucuronidation enrichment term having P-value: 0.0000014, harbouring UGT genes that play an important role in cancer progression belonged to the TADs in class 11 (Allain et al.). The deregulation of UGT expression and activity has been associated with the growth of various malignancies (Stingl et al.). Although the execution of immune evasion by cancer cells is well established, our TAD-based analysis revealed a common region and related mechanism that may be implicated in many cancers.

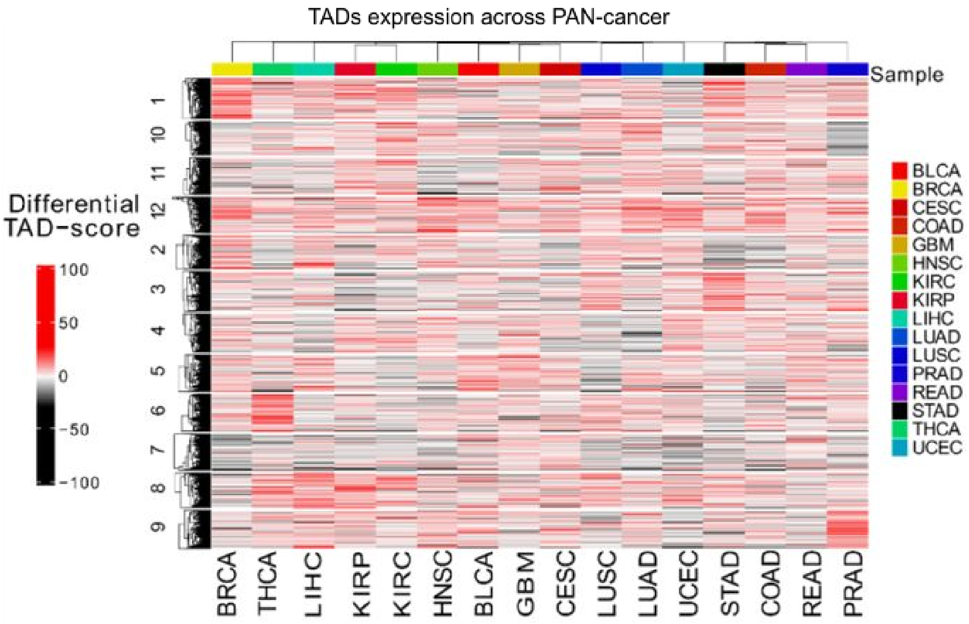


Figure 4.6. TAD activity in different cancers TCGA samples and its relationship to survival. Using TCGA data, a heatmap showing the fold increases in the median TAD activity (tumour vs. normal sample) for each of the 16 cancer types was created.

The correlation among TAD activity followed by survival throughout a variety of cancer types indicated an intriguing pattern. Six paralogous APOBEC3 family genes were present in the top TADs for HNSC whose activity showed the most significant correlation with survival (Figure 4.7 A). Innate and adaptive immunity are triggered by APOBEC3 family genes like APOBEC3A, which are known to be involved in single-stranded DNA deamination and anti-retrotransposition based anti-viral actions. Additionally, RNA deamination activity for RNA editing to protect cells from viruses and cancer is known to be regulated by APOBEC3 genes. Our findings demonstrate that survival in HNSC was higher when TAD with cluster of APOBEC genes had higher activity, which is consistent with a previous report (T. W. Chen et al.). It should be emphasised that HPV frequently causes HNSC, and that APOBEC3 is known to defend against HPV through an RNA editing mechanism (Cytidine to Uracil). None of the APOBEC3 family genes showed a significant correlation during survival analysis utilising only individual genes. It has therefore become clear that single-gene based study could not replicate TAD-activity based conclusions. In light of this, the TAD activity based method highlighted TAD with APOBEC3 family genes' cumulative influence upon survival, even though HNSC cells from various patients may express APOBEC3 genes in a heterogeneous manner. The TAD "hg19:chr11-70150000_71300000", located on "11q13.4" region contains, KRTAP5 family genes; "KRTAP5-7, KRTAP5-10, and KRTAP5-11". These KRTAP5 family genes are involved in keratinization. This TAD (hg19:chr11-70150000_71300000) was also included among other top TADs whose activities were related to a higher survival rate in HNSC (Figure 4.7 B). A TAD "hg19:chr7-51500000_57100000" harbouring genes (particularly EGFR-gene) relevant in EGFR-related pathways were also found amongst the other important TADs whose activity were significantly linked to HNSC survival (Figure 4.7 C). EGFR activity is connected with tumour development (Sasaki et al.).

Based on our prior study in terms of survival p-value calculation, we used a cutoff of 51-49 to classify the patient group based on TAD activity to explore TAD's significance in survival (high-low). TAD (hg19:chr22-391000000_39825000) was on top and picked as the best candidate for further research. Further, we have expanded using other types of threshold also and some literature support. As discussed in the method section, we have determined the survival p-value on the basis of several cutoff combinations such as 51-49, 60-40, and 70-30 percentiles, and we have also used the MaxStat package, which was used to describe the patient group established on the activity of TAD to investigate TADs importance in survival ‘high-low’. TADs (chr11-70150000_71300000 and chr7 51500000 57100000) that were prominent in those combinations were then chosen for further investigation. TAD (hg19:chr22-391000000_39825000) was also found in cutoff combinations 51-49 and the MaxStat approach.

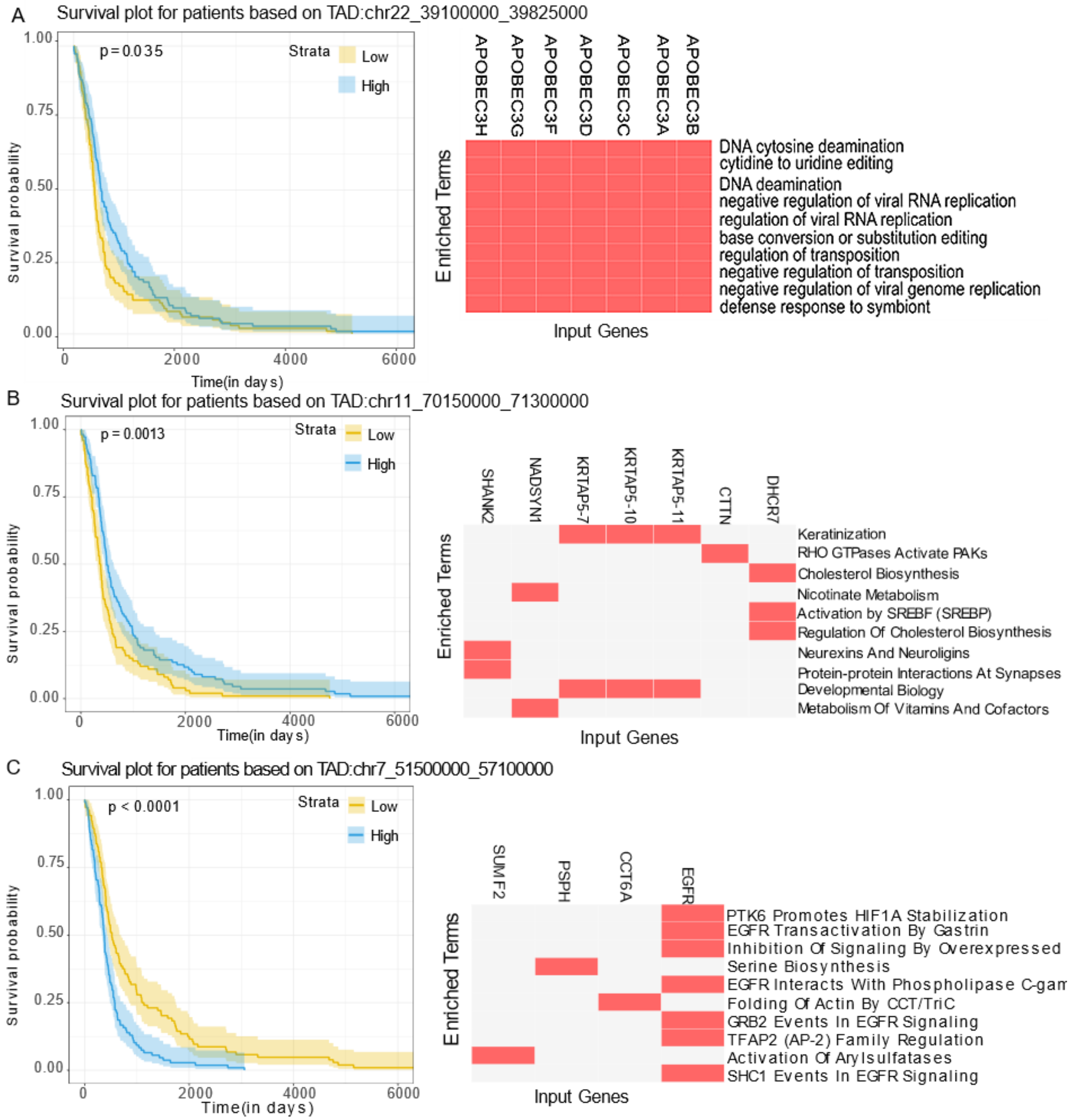


Figure 4.7. (A) Kaplan Mier (KM) plot for survival rate for head and neck cancer (HNSC) patients. Here KM plot was made for two groups of HNSC patients; high - top 80 patients sample where activity of TAD (hg19:chr22-391000000_39825000) was high and low- 80 patients with lowest activity of TAD. On the right panel, the top enriched ontology terms for genes in the TAD, are shown hand in hand with involved genes. (B) KM-plot for survival rate for HNSC patients based on activity of TAD (hg19:chr11-70150000_71300000) and top enriched terms for involved genes are shown. (C) KM-plot for survival rate for HNSC patients based on activity of TAD (hg19:chr7_51500000_57100000) is shown. The enriched terms for genes inside that TAD are also shown.

4.3.5 Discussion

We applied a unique strategy to calculate activity of TAD utilizing gene-expression profiles. TAD activity also indicates the activity of its enhancers and other regulatory components. The numerous interactions within TADs further support the notion that they are regulatory units. Studying the function of these regulatory units and interactions with one another in relation to how cancer cells respond to drugs can therefore reveal shared mechanisms of action. We identified a region at chromosome 11 in the "11q13.3–11q13.4" bands, which may be assisting survival during cisplatin-based stress, for example, through study of TAD-activity alteration in response to cisplatin in HNSC cell lines. The relationship of TADs in the 11q13.3–11q13.4 band with cisplatin resistance, however, has only sometimes been reported. The contribution of TADs in the "11q13.3–11q13.4" bands in giving resilience to cancer cells against cisplatin-based stress has also been expanded by our investigation to include chromatin structure. According to our findings, TADS in the "11q13.3 and 11q13.4" bands may contain enhancers or genes that assist, activate or regulate additional connected regions that contain stress-response genes. Much significantly, a region R1 that interact with TADS at "11q13.3 and 11q13.4" band also has a large number of genes for keratin associated proteins. With the exception of a few case studies, keratin related proteins have not been frequently connected with characteristics of cancer cells. (Berens et al.). However, it's probable that during drug-induced stress, clusters of genes for keratin related proteins as well as other response genes co-localize to co-express. Based on our study using TCGA HNSC samples, we found that several of the TADs related to survival involve genes for enriched pathway terms. These pathways include keratinization, expression and translocation for olfactory receptors, "PI3K-AKT activation", and "EGFR-PTK6-based HIF1A stabilisation". Clusters of several paralogous genes related to HNSC survival were discovered, which hints at the significance of these genes in both physiological and pathological functions as a result of their evolutionary conservation.

Our study of the TCGA HNSC samples showed that the top TADs linked to survival, included clusters of several APOBEC genes associated with RNA editing and virus protection. Regarding the relationship between RNA editing and high or low survival, there have been opposing opinions (Paz-Yaacov et al.). However, there are primarily two-types of RNA editing: “adenosine-to-inosine (A-to-I) editing”, which is mediated by ADARs, and “cytosine-to-uracil (C-to-U) editing”, which is mediated through APOBEC family of genes (Kurkowiak et al.). Our work has suggested that C-to-U RNA editing, which is mediated by APOBEC, may be offering protection and prolong life, particularly in case of HNSC .

Our data unequivocally indicates TAD function could also be employed as a biomarker for estimating survival or predicting drug response to deal with tumor heterogeneity. Also, common mechanism revealed through TAD-based analysis can serve as a guide in developing the rules for finding effective drug combinations. These efficient drug combinations serve to minimise redundancy in the present paradigm of cancer chemotherapy, which involves several types of drugs and their combinations being explored by various consortiums such as CCLE. Our unbiased investigation of the relationship between TAD based regulatory units and cancer survival revealed a few potential targets for chemotherapy. For example, APOBEC3 could be used to benefit from innate immunity depending on the C-to-U RNA-editing mechanism to treat cancer in HNSC patients. Another advantage of TAD-based drug-resistance and survival research is that it may be used to identify super-enhancers and genes that should be the focus of cancer therapies.

CHAPTER 5

CONCLUSION

The analysis of chromatin-based data using various computational biology and bioinformatics methodologies is the main emphasis of this thesis. Integrative analysis of genetic data from many studies is a crucial component of our work. We also have the findings of my thesis work, which generated thorough epigenome, and 3D genome organization for the essential cellular phenotypic developmental stages. We have investigated how the 3D genome and epigenome regulate gene expression throughout the onset of disease using computational methods. In addition to offering a unique method for predicting distal chromatin interactions from “single-cell open-chromatin profiles”, we constructed a strategy to deal with the problem of “single-cell open-chromatin profiles” having sparse data. In order to comprehend the mechanisms behind drug response in cancer cells, we also examined trends in the activity of gene sets with TADs.

5.1 Overview of the contribution

In this section, we provide a concise synopsis of the chapters that present the thesis in its entirety.

5.2 Using single-cell open chromatin profiles to improve chromatin interaction prediction and reveal insights of the human brain's cis-regulatory landscape

“Single-cell open-chromatin profiles” (scATAC-seq) can potentially determine patterns of chromatin interactions across cell types. Current cis-regulatory network predictions approaches employing “single-cell open-chromatin profiles” concentrate mainly on nearby chromatin interactions; nevertheless, it is clear that interactions across distant genomic regions play an important role in controlling gene expression. We suggested a method for predicting short-term and long-term interactions between genomic regions utilizing scATACs-seq profiles. The methodology we've adopted, known as “single-cell epigenome based chromatin interaction analysis (scEChIA)”, makes use of the imputation of signals and enhanced L1-regularization. “scEChIA” outperformed other techniques in context of prediction accuracy for a few “single-cell open-chromatin profiles”. We predicted about 0.7-million interactions between genomic locations in the human brain using scEChIA. Further investigation showed the cell type for the association among genes and “expression quantitative trait locus (eQTL)” in the human brain, as well as new information regarding the target genes of human-accelerated elements and disease associated mutations. Our “scEChIA-enabled” study also points at the possible action of a few transcription factors (TFs), particularly via long-range interaction in brain endothelial cells. These findings demonstrate the merit of our strategy for developing gene-regulatory networks from “single-cell open-chromatin profiles”. Particularly for cells originating from *in-vivo* samples that are less abundant.

5.3 Recovery of true signals from single-cell open-chromatin profiles using a forest of imputation trees

Single-cell resolution investigation of the heterogeneity of regulatory area activity has been made possible by the development of “single-cell open-chromatin profiling” technologies. Even so, stochasticity and the scarcity of relevant DNA result in a high drop-out rate with noise in the data “single-cell open-chromatin profiles”. To recovering original signals through severely sparse and noise “single-cell open-chromatin profiles”, we propose here a reliable approach named as “forest of imputation trees (FITs)”. To eliminate bias when restoring read-count matrices, FITs build multiple imputation trees. The difficult problem of retrieving “open chromatin signals” at genomic locations with cell-type specific activity without obscuring information has been resolved. Along with classification as well as visualization, FITs-based imputation increased accuracy in enhancer recognition, pathway enrichment score calculation, and chromatin-interaction prediction. FITs is enhanced to provide a broader range of applications, particularly for very sparse read-count matrices. Furthermore, “single-cell open-chromatin profiles” from *in-vivo* samples can benefit greatly from FITs due to their superiority in restoring signals of minority cells.

5.4 Analysis of activity and interactions of chromatin domains in cancer samples and their relevance in drug-response of cancer cells

Chromatin is most likely the cell's most complicated molecular ensemble. It is made up of genomic DNA as well as multiple directly or indirectly related protein and RNA molecules. It comprises histones, DNA-binding factors and its nascent transcripts, replication and repair machines that replicate and preserve DNA, and many additional molecules that interact with any of these components. TADs, which are how chromatin is organized, are defined by preferable contacts between loci within the same genomic region, and linked to gene regulation through tight regulatory interactions between cis-regulatory elements and promoters. TADs are thus regarded as structural scaffolding in the construction of regulatory landscapes (RLs). There hasn't been much research done on how similar-functioning genes are organized in TAD, where they colocalize in cancer cells, and how they react to drugs. To figure out the connection between drugs and cancer cells, we examined chromatin interaction profiles and patterns in the activity of TAD gene sets. We discovered multiple patterns in co-localisation of paralogue genes in 3 dimensional chromatin scaffold. We also discovered the pattern of activation of clusters of paralogous genes in response to drugs in cancer cells. Our proposed approach of analysis using TAD-activity resolves the problem of heterogeneous activation of paralogous or co-localised genes by different types of cancer cells in response to drug. Thus our computational method enables finding common mechanism of drug-response across multiple patients and possibly many cancer types even if classical method fails due to heterogeneity among cancer cells.

5.5 Future directions

Future research into the molecular mechanisms can benefit greatly from our chromatin-based analysis. This technological study may be used in the future to address issues about how chromatin dynamics and structure affect gene regulation, DNA repair, DNA replication, and genome organisation. In relation to transcription, this approach will be able to examine if chromatin undergoes structural reorganisation while changing through inactive to active transcribing state and how epigenetic alterations may affect this process. Additionally, by utilising chromatin structure to highlight regulatory areas that influence disease state and cellular drug response, our presented approaches can help in the selection of hypothesis-driven therapeutics.

References

Akdemir, Kadir C., et al. “Somatic Mutation Distributions in Cancer Genomes Vary with Three-Dimensional Chromatin Structure.” *Nature Genetics*, vol. 52, no. 11, Nov. 2020, pp. 1178–88.

Alaskhar Alhamwe, Bilal, et al. “Histone Modifications and Their Role in Epigenetics of Atopy and Allergic Diseases.” *Allergy, Asthma, and Clinical Immunology: Official Journal of the Canadian Society of Allergy and Clinical Immunology*, vol. 14, May 2018, p. 39.

Albini, Sonia, et al. “Histone Modifications.” *Epigenetics and Regeneration*, 2019, pp. 47–72, https://doi.org/10.1016/b978-0-12-814879-2.00003-0.

Aljogol, Dina, et al. “Comparison of Capture Hi-C Analytical Pipelines.” *Frontiers in Genetics*, vol. 0, 2022, https://doi.org/10.3389/fgene.2022.786501.

Allain, Eric P., et al. “Emerging Roles for UDP-Glucuronosyltransferases in Drug Resistance and Cancer Progression.” *British Journal of Cancer*, vol. 122, no. 9, Apr. 2020, pp. 1277–87.

Altschuler, Steven J., and Lani F. Wu. “Cellular Heterogeneity: Do Differences Make a Difference?” *Cell*, vol. 141, no. 4, 2010, pp. 559–63, https://doi.org/10.1016/j.cell.2010.04.033.

Amaral, Paulo P., et al. “Complex Architecture and Regulated Expression of the Sox2ot Locus during Vertebrate Development.” *RNA* , vol. 15, no. 11, Nov. 2009, pp. 2013–27.

Angeloni, D. “Molecular Analysis of Deletions in Human Chromosome 3p21 and the Role of Resident Cancer Genes in Disease.” *Briefings in Functional Genomics & Proteomics*, vol. 6, no. 1, Mar. 2007, https://doi.org/10.1093/bfgp/elm007.

Aquila, Lanni, and Boyko S. Atanassov. “Regulation of Histone Ubiquitination in Response to DNA Double Strand Breaks.” *Cells* , vol. 9, no. 7, July 2020, https://doi.org/10.3390/cells9071699.

Bannister, Andrew J., and Tony Kouzarides. "Regulation of Chromatin by Histone Modifications." *Cell Research*, vol. 21, no. 3, 2011, pp. 381–95, https://doi.org/10.1038/cr.2011.22.

Barrera-Redondo, Josué, et al. "Genomic, Transcriptomic and Epigenomic Tools to Study the Domestication of Plants and Animals: A Field Guide for Beginners." *Frontiers in Genetics*, vol. 11, 2020, https://doi.org/10.3389/fgene.2020.00742.

Barretina, Jordi, et al. "The Cancer Cell Line Encyclopedia Enables Predictive Modeling of Anticancer Drug Sensitivity." *Nature*, vol. 483, no. 7391, Mar. 2012, p. 603.

Basu, Amitava, and Vijay K. Tiwari. "Epigenetic Reprogramming of Cell Identity: Lessons from Development for Regenerative Medicine." *Clinical Epigenetics*, vol. 13, no. 1, July 2021, p. 144.

Berens, E. B., et al. "Keratin-Associated Protein 5-5 Controls Cytoskeletal Function and Cancer Cell Vascular Invasion." *Oncogene*, vol. 36, no. 5, Feb. 2017, pp. 593–605.

Bickmore, Wendy A. "The Spatial Organization of the Human Genome." *Annual Review of Genomics and Human Genetics*, vol. 14, no. 1, 2013, pp. 67–84, https://doi.org/10.1146/annurev-genom-091212-153515.

Borden, J., and L. Manuelidis. "Movement of the X Chromosome in Epilepsy." *Science*, vol. 242, no. 4886, Dec. 1988, pp. 1687–91.

Boyle, Alan P., et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." *Cell*, vol. 132, no. 2, Jan. 2008, pp. 311–22.

Brangwynne, Clifford P., et al. "Active Liquid-like Behavior of Nucleoli Determines Their Size and Shape in Xenopus Laevis Oocytes." *Proceedings of the National Academy of Sciences* *of the United States of America*, vol. 108, no. 11, Mar. 2011, pp. 4334–39.

Bravo González-Blas, Carmen, et al. "cisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data." *Nature Methods*, vol. 16, no. 5, May 2019, pp. 397–400.

Brock, Amy, et al. "Non-Genetic Heterogeneity--a Mutation-Independent Driving Force for the Somatic Evolution of Tumours." *Nature Reviews. Genetics*, vol. 10, no. 5, May 2009, pp. 336–42.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, et al. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, vol. 109, Jan. 2015, pp. 21.29.1–21.29.9.

Buenrostro, Jason D., M. Ryan Corces, Caleb A. Lareau, Beijing Wu, et al. "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation." *Cell*, vol. 173, no. 6, May 2018, pp. 1535–48.e16.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, et al. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." *Nature*, vol. 523, no. 7561, July 2015, pp. 486–90.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, et al. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods*, vol. 10, no. 12, Dec. 2013, pp. 1213–18.

Bunnik, Evelien M., and Karine G. Le Roch. "An Introduction to Functional Genomics and Systems Biology." *Advances in Wound Care: The Journal for Prevention and Healing*, vol. 2, no. 9, Nov. 2013, pp. 490–98.

Cai, Jian-Feng, et al. "A Singular Value Thresholding Algorithm for Matrix Completion." *SIAM*

Journal on Optimization, vol. 20, no. 4, 2010, pp. 1956–82, https://doi.org/10.1137/080738970.

Candes, Emmanuel J., and Yaniv Plan. “Matrix Completion With Noise.” *Proceedings of the IEEE*, vol. 98, no. 6, 2010, pp. 925–36, https://doi.org/10.1109/jproc.2009.2035722.

Candès, Emmanuel, and Benjamin Recht. “Exact Matrix Completion via Convex Optimization.” *Communications of the ACM*, vol. 55, no. 6, 2012, pp. 111–19, https://doi.org/10.1145/2184319.2184343.

Cao, Junyue, et al. “Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells.” *Science*, vol. 361, no. 6409, Sept. 2018, pp. 1380–85.

Challen, Grant A., and Melissa H. Little. “A Side Order of Stem Cells: The SP Phenotype.” *Stem Cells*, vol. 24, no. 1, Jan. 2006, pp. 3–12.

Chawla, Anjali, et al. “Chromatin Profiling Techniques: Exploring the Chromatin Environment and Its Contributions to Complex Traits.” *International Journal of Molecular Sciences*, vol. 22, no. 14, July 2021, https://doi.org/10.3390/ijms22147612.

Chawla, Smriti, et al. “UniPath: A Uniform Approach for Pathway and Gene-Set Based Analysis of Heterogeneity in Single-Cell Epigenome and Transcriptome Profiles.” *Nucleic Acids Research*, vol. 49, no. 3, Feb. 2021, p. e13.

Chen, Huidong, et al. “Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data.” *Genome Biology*, vol. 20, no. 1, Nov. 2019, p. 241.

Chen, T. W., et al. “APOBEC3A Is an Oral Cancer Prognostic Biomarker in Taiwanese Carriers of an APOBEC Deletion Polymorphism.” *Nature Communications*, vol. 8, no. 1, Sept. 2017, pp. 465–465.

Chereji, Răzvan V., Tsung-Wai Kan, et al. “Genome-Wide Profiling of Nucleosome Sensitivity and Chromatin Accessibility in Drosophila Melanogaster." *Nucleic Acids Research*, vol. 44, no. 3, Feb. 2016, pp. 1036–51.

Chereji, Răzvan V., Josefina Ocampo, et al. "MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-Histone Barriers." *Molecular Cell*, vol. 65, no. 3, Feb. 2017, pp. 565–77.e3.

Chereji, Răzvan V., Terri D. Bryson, et al. "Quantitative MNase-Seq Accurately Maps Nucleosome Occupancy Levels." *Genome Biology*, vol. 20, no. 1, Sept. 2019, p. 198.

Chia, S., et al. "Phenotype-Driven Precision Oncology as a Guide for Clinical Decisions One Patient at a Time." *Nature Communications*, vol. 8, no. 1, Sept. 2017, https://doi.org/10.1038/s41467-017-00451-5.

Corces, M. Ryan, Jason D. Buenrostro, et al. "Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution." *Nature Genetics*, vol. 48, no. 10, Oct. 2016, pp. 1193–203.

Corces, M. Ryan, Jeffrey M. Granja, et al. "The Chromatin Accessibility Landscape of Primary Human Cancers." *Science*, vol. 362, no. 6413, Oct. 2018, https://doi.org/10.1126/science.aav1898.

Cremer, T., and C. Cremer. "Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells." *Nature Reviews Genetics*, vol. 2, no. 4, 2001, pp. 292–301, https://doi.org/10.1038/35066075.

Cusanovich, Darren A., Andrew J. Hill, et al. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility." *Cell*, vol. 174, no. 5, Aug. 2018, pp. 1309–24.e18.

Cusanovich, Darren A., James P. Reddington, et al. "The Cis-Regulatory Dynamics of Embryonic Development at Single-Cell Resolution." *Nature*, vol. 555, no. 7697, Mar. 2018, pp. 538–42.

Dali, Rola, and Mathieu Blanchette. “A Critical Assessment of Topologically Associating Domain Prediction Tools.” *Nucleic Acids Research*, vol. 45, no. 6, Apr. 2017, pp. 2994–3005.

DeAngelis, J. Tyson, et al. “An Overview of Epigenetic Assays.” *Molecular Biotechnology*, vol. 38, no. 2, 2008, pp. 179–83, https://doi.org/10.1007/s12033-007-9010-y.

Dean, Michael, et al. “Tumour Stem Cells and Drug Resistance.” *Nature Reviews. Cancer*, vol. 5, no. 4, Apr. 2005, pp. 275–84.

Dekker, Job, et al. “Capturing Chromosome Conformation.” *Science*, vol. 295, no. 5558, Feb. 2002, pp. 1306–11.

---. “The Three ‘C’ S of Chromosome Conformation Capture: Controls, Controls, Controls.” *Nature Methods*, vol. 3, no. 1, 2006, pp. 17–21, https://doi.org/10.1038/nmeth823.

de Laat, Wouter, and Frank Grosveld. “Spatial Organization of Gene Expression: The Active Chromatin Hub.” *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, vol. 11, no. 5, 2003, pp. 447–59.

DeNicola, Gina M., et al. “Oncogene-Induced Nrf2 Transcription Promotes ROS Detoxification and Tumorigenesis.” *Nature*, vol. 475, no. 7354, July 2011, pp. 106–09.

de Wit, Elzo, and Wouter de Laat. “A Decade of 3C Technologies: Insights into Nuclear Organization.” *Genes & Development*, vol. 26, no. 1, Jan. 2012, pp. 11–24.

Dixon, Jesse R., Inkyung Jung, et al. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature*, vol. 518, no. 7539, Feb. 2015, pp. 331–36.

Dixon, Jesse R., Siddarth Selvaraj, et al. “Topological Domains in Mammalian Genomes

Identified by Analysis of Chromatin Interactions." *Nature*, vol. 485, no. 7398, Apr. 2012, pp. 376–80.

Domcke, Silvia, et al. "A Human Cell Atlas of Fetal Chromatin Accessibility." *Science*, vol. 370, no. 6518, Nov. 2020, https://doi.org/10.1126/science.aba7612.

Dong, Kangning, and Shihua Zhang. "Joint Reconstruction of Cis-Regulatory Interaction Networks across Multiple Tissues Using Single-Cell Chromatin Accessibility Data." *Briefings in Bioinformatics*, vol. 22, no. 3, May 2021, https://doi.org/10.1093/bib/bbaa120.

Donnenberg, Vera S., and Albert D. Donnenberg. "Multiple Drug Resistance in Cancer Revisited: The Cancer Stem Cell Hypothesis." *Journal of Clinical Pharmacology*, vol. 45, no. 8, Aug. 2005, pp. 872–77.

Dostie, Josée, and Job Dekker. "Mapping Networks of Physical Interactions between Genomic Elements Using 5C Technology." *Nature Protocols*, vol. 2, no. 4, 2007, pp. 988–1002.

Dozio, Vito, and Jean-Charles Sanchez. "Characterisation of Extracellular Vesicle-Subsets Derived from Brain Endothelial Cells and Analysis of Their Protein Cargo Modulation after TNF Exposure." *Journal of Extracellular Vesicles*, vol. 6, no. 1, Apr. 2017, p. 1302705.

Durand, Neva C., et al. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems*, vol. 3, no. 1, July 2016, pp. 95–98.

Ecker, Joseph R., et al. "ENCODE Explained." *Nature*, vol. 489, no. 7414, 2012, pp. 52–54, https://doi.org/10.1038/489052a.

Eckhart, Leopold, et al. "Cell Death by Cornification." *Biochimica et Biophysica Acta*, vol. 1833, no. 12, Dec. 2013, pp. 3471–80.

Egger, Gerda, et al. "Epigenetics in Human Disease and Prospects for Epigenetic Therapy." *Nature*, vol. 429, no. 6990, 2004, pp. 457–63, https://doi.org/10.1038/nature02625.

Elgin, Sarah C. R. "DNAase I-Hypersensitive Sites of Chromatin." *Cell*, vol. 27, no. 3, 1981, pp. 413–15, https://doi.org/10.1016/0092-8674(81)90381-0.

ENCODE Project Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature*, vol. 489, no. 7414, Sept. 2012, pp. 57–74.

Eraslan, Gökcen, et al. "Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder." *Nature Communications*, vol. 10, no. 1, Jan. 2019, p. 390.

Ernst, Jason, et al. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types." *Nature*, vol. 473, no. 7345, May 2011, pp. 43–49.

Fasler-Kan, Elizaveta, et al. "Cytokine Signaling in the Human Brain Capillary Endothelial Cell Line hCMEC/D3." *Brain Research*, vol. 1354, Oct. 2010, pp. 15–22.

Filion, Guillaume J., et al. "Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells." *Cell*, vol. 143, no. 2, Oct. 2010, pp. 212–24.

Forcato, Mattia, et al. "Comparison of Computational Methods for Hi-C Data Analysis." *Nature Methods*, vol. 14, no. 7, July 2017, pp. 679–85.

Fortin, Jean-Philippe, and Kasper D. Hansen. "Reconstructing A/B Compartments as Revealed by Hi-C Using Long-Range Correlations in Epigenetic Data." *Genome Biology*, vol. 16, Aug. 2015, p. 180.

Fraser, Peter, and Wendy Bickmore. "Nuclear Organization of the Genome and the Potential for Gene Regulation." *Nature*, vol. 447, no. 7143, 2007, pp. 413–17, https://doi.org/10.1038/nature05916.

Friedman, Jerome, et al. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* , vol. 9, no. 3, July 2008, pp. 432–41.

Fuchs, Ota. "Importance and Presentation of Transcription Factor Fli-1 in Hematopoiesis and in

Hematological and Other Malignancies." *ARC Journal of Hematology*, vol. 2, no. 2, 2017, pp. 23–37.

Galupa, Rafael, and Edith Heard. "Topologically Associating Domains in Chromosome Architecture and Gene Regulatory Landscapes during Development, Disease, and Evolution." *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 82, 2017, pp. 267–78.

Gibcus, Johan H., and Job Dekker. "The Hierarchy of the 3D Genome." *Molecular Cell*, vol. 49, no. 5, 2013, pp. 773–82, https://doi.org/10.1016/j.molcel.2013.02.011.

Gibson, Greg. "Rare and Common Variants: Twenty Arguments." *Nature Reviews. Genetics*, vol. 13, no. 2, Jan. 2012, pp. 135–45.

Giresi, Paul G., et al. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin." *Genome Research*, vol. 17, no. 6, June 2007, pp. 877–85.

Gong, Yixiao, et al. "Stratification of TAD Boundaries Reveals Preferential Insulation of Super-Enhancers by Strong Boundaries." *Nature Communications*, vol. 9, no. 1, Feb. 2018, p. 542.

Gräff, Johannes, and Li-Huei Tsai. "Histone Acetylation: Molecular Mnemonics on the Chromatin." *Nature Reviews. Neuroscience*, vol. 14, no. 2, Feb. 2013, pp. 97–111.

Greenwald, William W., et al. "Pgltools: A Genomic Arithmetic Tool Suite for Manipulation of Hi-C Peak and Other Chromatin Interaction Data." *BMC Bioinformatics*, vol. 18, no. 1, Apr. 2017, p. 207.

Guo, Hongshan, et al. "Single-Cell Methylome Landscapes of Mouse Embryonic Stem Cells and Early Embryos Analyzed Using Reduced Representation Bisulfite Sequencing." *Genome*

Research, vol. 23, no. 12, Dec. 2013, pp. 2126–35.

Guo, Hui, et al. “Integration of Disease Association and eQTL Data Using a Bayesian Colocalisation Approach Highlights Six Candidate Causal Genes in Immune-Mediated Diseases.” *Human Molecular Genetics*, vol. 24, no. 12, June 2015, pp. 3305–13.

Han, Jinlei, et al. “3C and 3C-Based Techniques: The Powerful Tools for Spatial Genome Organization Deciphering.” *Molecular Cytogenetics*, vol. 11, Mar. 2018, p. 21.

Hänzelmann, Sonja, et al. “GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data.” *BMC Bioinformatics*, vol. 14, Jan. 2013, p. 7.

Heard, Edith, and Robert A. Martienssen. “Transgenerational Epigenetic Inheritance: Myths and Mechanisms.” *Cell*, vol. 157, no. 1, Mar. 2014, pp. 95–109.

Heinz, Sven, et al. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell*, vol. 38, no. 4, 2010, pp. 576–89, https://doi.org/10.1016/j.molcel.2010.05.004.

Heitz, Emil. *Das Heterochromatin der Moose*. 1928.

Hirakawa, Yasuko, et al. “Quantitative Polymerase Chain Reaction Analysis by Deconvolution of Internal Standard.” *BMC Molecular Biology*, vol. 11, Apr. 2010, p. 30.

Huang, Jeffrey K., et al. “Retinoid X Receptor Gamma Signaling Accelerates CNS Remyelination.” *Nature Neuroscience*, vol. 14, no. 1, Jan. 2011, pp. 45–53.

Hubisz, Melissa J., and Katherine S. Pollard. “Exploring the Genesis and Functions of Human Accelerated Regions Sheds Light on Their Role in Human Evolution.” *Current Opinion in Genetics & Development*, vol. 29, Dec. 2014, pp. 15–21.

Hu, Linping, et al. “Fluorescence in Situ Hybridization (FISH): An Increasingly Demanded Tool for Biomarker Research and Personalized Medicine.” *Biomarker Research*, vol. 2, no. 1, Feb. 2014, p. 3.

Hung, Jui-Hung, et al. “Gene Set Enrichment Analysis: Performance Evaluation and Usage Guidelines.” *Briefings in Bioinformatics*, vol. 13, no. 3, May 2012, pp. 281–91.

Ibn-Salem, Jonas, et al. “Co-Regulation of Paralog Genes in the Three-Dimensional Chromatin Architecture.” *Nucleic Acids Research*, vol. 45, no. 1, Jan. 2017, pp. 81–91.

Ilyas, Mohammad. “Next-Generation Sequencing in Diagnostic Pathology.” *Pathobiology*, vol. 84, no. 6, 2017, pp. 292–305, https://doi.org/10.1159/000480089.

Jia, Guangshuai, et al. “Single Cell RNA-Seq and ATAC-Seq Analysis of Cardiac Progenitor Cell Transition States and Lineage Settlement.” *Nature Communications*, vol. 9, no. 1, Nov. 2018, p. 4877.

Jiang, Hui, et al. “The Utility of Fluorescence in Situ Hybridization Analysis in Diagnosing Myelodysplastic Syndromes Is Limited to Cases with Karyotype Failure.” *Leukemia Research*, vol. 36, no. 4, 2012, pp. 448–52, https://doi.org/10.1016/j.leukres.2011.10.014.

Jin, Wenfei, et al. “Genome-Wide Detection of DNase I Hypersensitive Sites in Single Cells and FFPE Tissue Samples.” *Nature*, vol. 528, no. 7580, Dec. 2015, pp. 142–46.

Ji, Zhicheng, et al. “Single-Cell ATAC-Seq Signal Extraction and Enhancement with SCATE.” *Genome Biology*, vol. 21, no. 1, July 2020, p. 161.

Jones, Peter A., et al. “Targeting the Cancer Epigenome for Therapy.” *Nature Reviews. Genetics*, vol. 17, no. 10, Sept. 2016, pp. 630–41.

Kadauke, Stephan, and Gerd A. Blobel. “Chromatin Loops in Gene Regulation.” *Biochimica et Biophysica Acta*, vol. 1789, no. 1, Jan. 2009, pp. 17–25.

Kaliyappan, Karunakaran, et al. “Applications of Immunohistochemistry.” *Journal of Pharmacy and Bioallied Sciences*, vol. 4, no. 6, 2012, p. 307, https://doi.org/10.4103/0975-7406.100281.

Kenyon, Jonathan, and Stanton L. Gerson. "The Role of DNA Damage Repair in Aging of Adult Stem Cells." *Nucleic Acids Research*, vol. 35, no. 22, Dec. 2007, pp. 7557–65.

Khan, Niamat, et al. "Current Analytical Strategies in Studying Chromatin-Associated-Proteome (Chromatome)." *Molecules* , vol. 26, no. 21, Nov. 2021, https://doi.org/10.3390/molecules26216694.

Kumar, Suresh, et al. "Understanding 3D Genome Organization and Its Effect on Transcriptional Gene Regulation Under Environmental Stress in Plant: A Chromatin Perspective." *Frontiers in Cell and Developmental Biology*, vol. 9, 2021, https://doi.org/10.3389/fcell.2021.774719.

Kumar, Vibhor, Nirmala Arul Rayan, et al. "Comprehensive Benchmarking Reveals H2BK20 Acetylation as a Distinctive Signature of Cell-State-Specific Enhancers and Promoters." *Genome Research*, vol. 26, no. 5, May 2016, pp. 612–23.

Kumar, Vibhor, Masafumi Muratani, et al. "Uniform, Optimal Signal Processing of Mapped Deep-Sequencing Data." *Nature Biotechnology*, vol. 31, no. 7, July 2013, pp. 615–22.

Kurkowiak, Małgorzata, et al. "The Effects of RNA Editing in Cancer Tissue at Different Stages in Carcinogenesis." *RNA Biology*, vol. 18, no. 11, Nov. 2021, pp. 1524–39.

Lachance, Véronik, et al. "Autophagy Protein NRBF2 Has Reduced Expression in Alzheimer's Brains and Modulates Memory and Amyloid-Beta Homeostasis in Mice." *Molecular Neurodegeneration*, vol. 14, no. 1, Nov. 2019, p. 43.

Lake, Blue B., et al. "Integrative Single-Cell Analysis of Transcriptional and Epigenetic States in the Human Adult Brain." *Nature Biotechnology*, vol. 36, no. 1, Jan. 2018, pp. 70–80.

Langmead, Ben, and Steven L. Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods*, vol. 9, no. 4, Mar. 2012, pp. 357–59.

Lareau, Caleb A., et al. “Droplet-Based Combinatorial Indexing for Massive-Scale Single-Cell Chromatin Accessibility.” *Nature Biotechnology*, vol. 37, no. 8, Aug. 2019, pp. 916–24.

Lausen, Berthold, and Martin Schumacher. “Maximally Selected Rank Statistics.” *Biometrics*, vol. 48, no. 1, 1992, p. 73, https://doi.org/10.2307/2532740.

Lee, Dominic Paul, et al. “Robust CTCF-Based Chromatin Architecture Underpins Epigenetic Changes in the Heart Failure Stress-Gene Response.” *Circulation*, vol. 139, no. 16, Apr. 2019, pp. 1937–56.

Li, Cai, and Hua Zhou. “Svt: Singular Value Thresholding in MATLAB.” *Journal of Statistical Software*, vol. 81, no. 2, Nov. 2017, https://doi.org/10.18637/jss.v081.c02.

Lieberman-Aiden, Erez, et al. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science*, vol. 326, no. 5950, Oct. 2009, pp. 289–93.

Li, Gu, et al. “Rapid Spontaneous Accessibility of Nucleosomal DNA.” *Nature Structural & Molecular Biology*, vol. 12, no. 1, Jan. 2005, pp. 46–53.

Li, Lifei, et al. “Cancer Is Associated with Alterations in the Three-Dimensional Organization of the Genome.” *Cancers*, vol. 11, no. 12, Nov. 2019, https://doi.org/10.3390/cancers11121886.

Ling, Jian Qun, and Andrew R. Hoffman. “Epigenetics of Long-Range Chromatin Interactions.” *Pediatric Research*, vol. 61, no. 5 Part 2, 2007, p. 11R – 16R, https://doi.org/10.1203/pdr.0b013e31804575db.

Liu, Tong, et al. “TADKB: Family Classification and a Knowledge Base of Topologically Associating Domains.” *BMC Genomics*, vol. 20, no. 1, Mar. 2019, pp. 1–17.

Liu, Yongjing, et al. “A Practical Guide for DNase-Seq Data Analysis: From Data Management to Common Applications." *Briefings in Bioinformatics*, vol. 20, no. 5, Sept. 2019, pp. 1865–77.

Li, Wei Vivian, and Jingyi Jessica Li. "An Accurate and Robust Imputation Method scImpute for Single-Cell RNA-Seq Data." *Nature Communications*, vol. 9, no. 1, Mar. 2018, p. 997.

Li, Wenran, et al. "DeepTACT: Predicting 3D Chromatin Contacts via Bootstrapping Deep Learning." *Nucleic Acids Research*, vol. 47, no. 10, June 2019, p. e60.

Li, Zhijian, Christoph Kuppe, et al. "Chromatin-Accessibility Estimation from Single-Cell ATAC-Seq Data with scOpen." *Nature Communications*, vol. 12, no. 1, Nov. 2021, p. 6386.

Li, Zhijian, Marcel H. Schulz, et al. "Identification of Transcription Factor Binding Sites Using ATAC-Seq." *Genome Biology*, vol. 20, no. 1, 2019, https://doi.org/10.1186/s13059-019-1642-2.

Louwers, Marieke, et al. "Studying Physical Chromatin Interactions in Plants Using Chromosome Conformation Capture (3C)." *Nature Protocols*, vol. 4, no. 8, July 2009, pp. 1216–29.

Lowe, William L., and Timothy E. Reddy. "Genomic Approaches for Understanding the Genetics of Complex Disease." *Genome Research*, vol. 25, no. 10, 2015, pp. 1432–41, https://doi.org/10.1101/gr.190603.115.

Malone, Eoghan R., et al. "Molecular Profiling for Precision Cancer Therapies." *Genome Medicine*, vol. 12, no. 1, 2020, https://doi.org/10.1186/s13073-019-0703-1.

Martens, Joost H. A., et al. "ERG and FLI1 Binding Sites Demarcate Targets for Aberrant Epigenetic Regulation by AML1-ETO in Acute Myeloid Leukemia." *Blood*, vol. 120, no. 19, Nov. 2012, pp. 4038–48.

Martinez-Lage, Marta, et al. "CRISPR/Cas9 for Cancer Therapy: Hopes and Challenges."

Biomedicines, vol. 6, no. 4, Nov. 2018, https://doi.org/10.3390/biomedicines6040105.

Medema, Jan Paul. "Cancer Stem Cells: The Challenges Ahead." *Nature Cell Biology*, vol. 15, no. 4, Apr. 2013, pp. 338–44.

Mezger, Anja, et al. "High-Throughput Chromatin Accessibility Profiling at Single-Cell Resolution." *Nature Communications*, vol. 9, no. 1, Sept. 2018, p. 3647.

Mieczkowski, Jakub, et al. "MNase Titration Reveals Differences between Nucleosome Occupancy and Chromatin Accessibility." *Nature Communications*, vol. 7, May 2016, p. 11485.

Misteli, Tom. "Beyond the Sequence: Cellular Organization of Genome Function." *Cell*, vol. 128, no. 4, 2007, pp. 787–800, https://doi.org/10.1016/j.cell.2007.01.028.

Mittra, Arjun, and Jeffrey A. Moscow. "Future Approaches to Precision Oncology-Based Clinical Trials." *Cancer Journal* , vol. 25, no. 4, 2019, pp. 300–04.

Mizukawa, Benjamin, et al. "Inhibition of Rac GTPase Signaling and Downstream Prosurvival Bcl-2 Proteins as Combination Targeted Therapy in MLL-AF9 Leukemia." *Blood*, vol. 118, no. 19, 2011, pp. 5235–45, https://doi.org/10.1182/blood-2011-04-351817.

Molina-Serrano, Diego, et al. "Histone Modifications as an Intersection Between Diet and Longevity." *Frontiers in Genetics*, vol. 10, Mar. 2019, p. 192.

Moore, Lisa D., et al. "DNA Methylation and Its Basic Function." *Neuropsychopharmacology*, vol. 38, no. 1, 2013, pp. 23–38, https://doi.org/10.1038/npp.2012.112.

Morgan, Marc A., and Ali Shilatifard. "Chromatin Signatures of Cancer." *Genes & Development*, vol. 29, no. 3, 2015, pp. 238–49, https://doi.org/10.1101/gad.255182.114.

Nakasone, Elizabeth S., et al. "Imaging Tumor-Stroma Interactions during Chemotherapy Reveals Contributions of the Microenvironment to Resistance." *Cancer Cell*, vol. 21, no. 4, Apr. 2012, pp. 488–503.

Ng, B., et al. *Brain xQTL Map: Integrating the Genetic Architecture of the Human Brain Transcriptome and Epigenome*. https://doi.org/10.1101/142927.

Novo, Clara Lopes, et al. “Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition.” *Cell Reports*, vol. 22, no. 10, Mar. 2018, pp. 2615–27.

Nusinow, David P., et al. “Quantitative Proteomics of the Cancer Cell Line Encyclopedia.” *Cell*, vol. 180, no. 2, Jan. 2020, pp. 387–402.e16.

Olson, Michael F., et al. “An Essential Role for Rho, Rac, and Cdc42 GTPases in Cell Cycle Progression Through G $_1$.” *Science*, vol. 269, no. 5228, 1995, pp. 1270–72, https://doi.org/10.1126/science.7652575.

Oluwadare, Oluwatosin, et al. “An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data.” *Biological Procedures Online*, vol. 21, Apr. 2019, p. 7.

Pandey, Neetesh, et al. “Improving Chromatin-Interaction Prediction Using Single-Cell Open-Chromatin Profiles and Making Insight Into the -Regulatory Landscape of the Human Brain.” *Frontiers in Genetics*, vol. 12, Oct. 2021, p. 738194.

Patrick, Ellis, et al. “Deconvolving the Contributions of Cell-Type Heterogeneity on Cortical Gene Expression.” *PLoS Computational Biology*, vol. 16, no. 8, Aug. 2020, p. e1008120.

Paz-Yaacov, Nurit, et al. “Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors.” *Cell Reports*, vol. 13, no. 2, Oct. 2015, pp. 267–76.

Pisco, A. O., and S. Huang. “Non-Genetic Cancer Cell Plasticity and Therapy-Induced Stemness in Tumour Relapse: ‘What Does Not Kill Me Strengthens Me.’” *British Journal of Cancer*, vol. 112, no. 11, May 2015, pp. 1725–32.

Pliner, Hannah A., et al. “Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data.” *Molecular Cell*, vol. 71, no. 5, Sept. 2018, pp. 858–71.e8.

Polak, Paz, et al. “Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer.” *Nature*, vol. 518, no. 7539, Feb. 2015, pp. 360–64.

Pott, Sebastian. “Simultaneous Measurement of Chromatin Accessibility, DNA Methylation, and Nucleosome Phasing in Single Cells.” *eLife*, vol. 6, June 2017, https://doi.org/10.7554/eLife.23203.

Ramazzotti, Daniele, et al. “Variant Calling from scRNA-Seq Data Allows the Assessment of Cellular Identity in Patient-Derived Cell Lines.” *Nature Communications*, vol. 13, no. 1, 12 May 2022, p. 2718.

Rao, Suhas S. P., et al. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell*, vol. 159, no. 7, Dec. 2014, pp. 1665–80.

Rivera, Chloe M., and Bing Ren. “Mapping Human Epigenomes.” *Cell*, vol. 155, no. 1, Sept. 2013, pp. 39–55.

Ron, Gil, et al. “Promoter-Enhancer Interactions Identified from Hi-C Data Using Probabilistic Models and Hierarchical Topological Domains.” *Nature Communications*, vol. 8, no. 1, Dec. 2017, p. 2237.

Rosenzweig, Steven A. “Acquired Resistance to Drugs Targeting Receptor Tyrosine Kinases.” *Biochemical Pharmacology*, vol. 83, no. 8, Apr. 2012, pp. 1041–48.

Rotem, Assaf, et al. “Single-Cell ChIP-Seq Reveals Cell Subpopulations Defined by Chromatin State.” *Nature Biotechnology*, vol. 33, no. 11, Nov. 2015, pp. 1165–72.

Sakurada, Kazuhiro. “Environmental Epigenetic Modifications and Reprogramming-Recalcitrant

Genes." *Stem Cell Research*, vol. 4, no. 3, May 2010, pp. 157–64.

Sasaki, Takamitsu, et al. "The Role of Epidermal Growth Factor Receptor in Cancer Metastasis and Microenvironment." *BioMed Research International*, vol. 2013, Aug. 2013, p. 546318.

Satpathy, Ansuman T., et al. "Transcript-Indexed ATAC-Seq for Precision Immune Profiling." *Nature Medicine*, vol. 24, no. 5, May 2018, pp. 580–90.

Schep, Alicia N., Beijing Wu, et al. "chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods*, vol. 14, no. 10, Oct. 2017, pp. 975–78.

Schep, Alicia N., Jason D. Buenrostro, et al. "Structured Nucleosome Fingerprints Enable High-Resolution Mapping of Chromatin Architecture within Regulatory Regions." *Genome Research*, vol. 25, no. 11, Nov. 2015, pp. 1757–70.

Sharma, Ankur, et al. "Longitudinal Single-Cell RNA Sequencing of Patient-Derived Primary Cells Reveals Drug-Induced Infidelity in Stem Cell Hierarchy." *Nature Communications*, vol. 9, no. 1, Nov. 2018, p. 4931.

Shen, Q., et al. "Super Enhancer-LncRNA SENCR Promoted Cisplatin Resistance and Growth of NSCLC through Upregulating FLI1." *Journal of Clinical Laboratory Analysis*, vol. 36, no. 6, June 2022, https://doi.org/10.1002/jcla.24460.

Sivashankari, Selvarajan, and Piramanayagam Shanmughavel. "Comparative Genomics - a Perspective." *Bioinformation*, vol. 1, no. 9, Mar. 2007, pp. 376–78.

Skolnick, J., and J. S. Fetrow. "From Genes to Protein Structure and Function: Novel Applications of Computational Approaches in the Genomic Era." *Trends in Biotechnology*, vol. 18, no. 1, Jan. 2000, pp. 34–39.

Stingl, J. C., et al. "Relevance of UDP-Glucuronosyltransferase Polymorphisms for Drug

Dosing: A Quantitative Systematic Review." *Pharmacology & Therapeutics*, vol. 141, no. 1, Jan. 2014, pp. 92–116.

Stricker, Stefan H., et al. "From Profiles to Function in Epigenomics." *Nature Reviews Genetics*, vol. 18, no. 1, 2017, pp. 51–66, https://doi.org/10.1038/nrg.2016.138.

Sun, James H., et al. "Disease-Associated Short Tandem Repeats Co-Localize with Chromatin Domain Boundaries." *Cell*, vol. 175, no. 1, Sept. 2018, pp. 224–38.e15.

Sun, Ying, et al. "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning." *IEEE Transactions on Signal Processing*, vol. 65, no. 3, 2017, pp. 794–816, https://doi.org/10.1109/tsp.2016.2601299.

Suphavilai, Chayaporn, et al. "Predicting Heterogeneity in Clone-Specific Therapeutic Vulnerabilities Using Single-Cell Transcriptomic Signatures." *Genome Medicine*, vol. 13, 2021, https://doi.org/10.1186/s13073-021-01000-y.

Tang, Zhonghui, et al. "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription." *Cell*, vol. 163, no. 7, Dec. 2015, pp. 1611–27.

Tena, Juan J., and José M. Santos-Pereira. "Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease." *Frontiers in Cell and Developmental Biology*, vol. 9, July 2021, p. 702787.

Thornton, Casey A., et al. *Spatially Mapped Single-Cell Chromatin Accessibility*. https://doi.org/10.1101/815720.

Tm, Therneau, and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York Springer, 2000.

Tozawa, Hideto, et al. "Genome-Wide Approaches Reveal Functional Interleukin-4-Inducible STAT6 Binding to the Vascular Cell Adhesion Molecule 1 Promoter." *Molecular and*

Cellular Biology, vol. 31, no. 11, 2011, pp. 2196–209, https://doi.org/10.1128/mcb.01430-10.

Troyanskaya, O., et al. “Missing Value Estimation Methods for DNA Microarrays.” *Bioinformatics*, vol. 17, no. 6, 2001, pp. 520–25, https://doi.org/10.1093/bioinformatics/17.6.520.

Urrutia, Eugene, et al. “Destin: Toolkit for Single-Cell Analysis of Chromatin Accessibility.” *Bioinformatics* , vol. 35, no. 19, Oct. 2019, pp. 3818–20.

van Dijk, David, et al. “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.” *Cell*, vol. 174, no. 3, July 2018, pp. 716–29.e27.

van Steensel, Bas, and Eileen E. M. Furlong. “The Role of Transcription in Shaping the Spatial Organization of the Genome.” *Nature Reviews. Molecular Cell Biology*, vol. 20, no. 6, June 2019, pp. 327–37.

Verma, Mukesh, and Vineet Kumar. “Single-Cell Epigenomics: Technology and Applications.” *Single-Cell Omics*, 2019, pp. 215–29, https://doi.org/10.1016/b978-0-12-814919-5.00011-7.

Waddington, C. H. “The Epigenotype.” *International Journal of Epidemiology*, vol. 41, no. 1, 2012, pp. 10–13, https://doi.org/10.1093/ije/dyr184.

Walldén, Marcus, et al. “Accelerating In-Transit Co-Processing for Scientific Simulations Using Region-Based Data-Driven Analysis.” *Algorithms*, vol. 14, no. 5, 2021, p. 154, https://doi.org/10.3390/a14050154.

Wang, Kevin C., and Howard Y. Chang. “Epigenomics.” *Circulation Research*, vol. 122, no. 9, 2018, pp. 1191–99, https://doi.org/10.1161/circresaha.118.310998.

Wang, Yanming, et al. “Beyond the Double Helix: Writing and Reading the Histone Code.” *Novartis Foundation Symposium*, vol. 259, 2004, pp. 3–17; discussion 17–21, 163–69.

Wang, Z., et al. "Rac1 Is Crucial for Ras-Dependent Skin Tumor Formation by Controlling Pak1-Mek-Erk Hyperactivation and Hyperproliferation in Vivo." *Oncogene*, vol. 29, no. 23, June 2010, pp. 3362–73.

Waterland, Robert A. "Epigenetic Mechanisms and Gastrointestinal Development." *The Journal of Pediatrics*, vol. 149, no. 5, 2006, pp. S137–42, https://doi.org/10.1016/j.jpeds.2006.06.064.

Weinhold, Bob. "Epigenetics: The Science of Change." *Environmental Health Perspectives*, vol. 114, no. 3, Mar. 2006, pp. A160–67.

Weintraub, H., and M. Groudine. "Chromosomal Subunits in Active Genes Have an Altered Conformation." *Science*, vol. 193, no. 4256, Sept. 1976, pp. 848–56.

Welter, Danielle, et al. "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations." *Nucleic Acids Research*, vol. 42, no. Database issue, Jan. 2014, pp. D1001–06.

Werner, Craig T., et al. "Mechanisms Regulating Compulsive Drug Behaviors." *Neural Mechanisms of Addiction*, 2019, pp. 137–55, https://doi.org/10.1016/b978-0-12-812202-0.00010-5.

Whalen, Sean, et al. "Enhancer-Promoter Interactions Are Encoded by Complex Genomic Signatures on Looping Chromatin." *Nature Genetics*, vol. 48, no. 5, May 2016, pp. 488–96.

Wingett, Steven, et al. "HiCUP: Pipeline for Mapping and Processing Hi-C Data." *F1000Research*, vol. 4, Nov. 2015, p. 1310.

Wu, Chia-Yen C., et al. "PI3K Regulation of RAC1 Is Required for KRAS-Induced Pancreatic Tumorigenesis in Mice." *Gastroenterology*, vol. 147, no. 6, 2014, pp. 1405–16.e7, https://doi.org/10.1053/j.gastro.2014.08.032.

Xiong, Lei, et al. “SCALE Method for Single-Cell ATAC-Seq Analysis via Latent Feature Extraction.” *Nature Communications*, vol. 10, no. 1, Oct. 2019, p. 4576.

Yan, Rui, et al. “Endothelial Interferon Regulatory Factor 1 Regulates Lipopolysaccharide-Induced VCAM-1 Expression Independent of NFκB.” *Journal of Innate Immunity*, vol. 9, no. 6, June 2017, pp. 546–60.

Yu, Wenbao, et al. “Identifying Topologically Associating Domains and Subdomains by Gaussian Mixture Model And Proportion Test.” *Nature Communications*, vol. 8, no. 1, Sept. 2017, p. 535.

Zhang, Bin, et al. “Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer’s Disease.” *Cell*, vol. 153, no. 3, 2013, pp. 707–20, https://doi.org/10.1016/j.cell.2013.03.030.

Zhang, Feng, and James R. Lupski. “Non-Coding Genetic Variants in Human Disease: Figure 1.” *Human Molecular Genetics*, vol. 24, no. R1, 2015, pp. R102–10, https://doi.org/10.1093/hmg/ddv259.

Zhao, Boxuan Simen, et al. “Post-Transcriptional Gene Regulation by mRNA Modifications.” *Nature Reviews Molecular Cell Biology*, vol. 18, no. 1, 2017, pp. 31–42, https://doi.org/10.1038/nrm.2016.132.

Zhou, S., et al. “The ABC Transporter Bcrp1/ABCG2 Is Expressed in a Wide Variety of Stem Cells and Is a Molecular Determinant of the Side-Population Phenotype.” *Nature Medicine*, vol. 7, no. 9, Sept. 2001, pp. 1028–34.

Zhu, Jiang, et al. “Genome-Wide Chromatin State Transitions Associated with Developmental and Environmental Cues.” *Cell*, vol. 152, no. 3, Jan. 2013, pp. 642–54.