COMPUTATIONAL METHODS FOR BULK AND SINGLE-CELL CHROMATIN INTERACTION DATA

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Chanaka Sampath Cooray Bulathsinghalage

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

March 2024

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

COMPUTATIONAL METHODS FOR BULK AND SINGLE-CELL CHROMATIN INTERACTION DATA

By

Chanaka Sampath Cooray Bulathsinghalage

The supervisory committee certifies that this dissertation complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Lu Liu
Chair

Dr. Changhui Yan

Dr. Anne Denton

Dr. Mingao Yuan

Approved:

26th of March 2024
Date

Dr. Simone Ludwig
Department Chair

ABSTRACT

Chromatin interactions occur when the physical regions of chromatin in close proximity interact with each other inside the nucleus. Analyzing chromatin interactions plays a crucial role in deciphering the spatial organization of the genome. Identifying the significant interactions and their functionalities reveals great insights on gene expressions, gene regulations and genetic diseases such as cancer. In addition, single cell chromatin interaction data is important to understand the chromatin structure changes, diversity among individual cells, and the genomics differences between different cell types. In recent years, Hi-C, chromosome conformation capture with high throughput sequencing, has gained widespread popularity for its ability to map genome-wide chromatin interactions in a single experiment and it is capable of extracting both single cell and bulk chromatin interaction data.

With the evolution of experimental methods like Hi-C, computational tools are essential to efficiently and accurately process the vast amount of genomic data. Since the experiment costs are notably higher, optimized computational tools and methods are needed to extract most possible information from the data. Moreover, processing single cell Hi-C data imposes number of challenges due to its sparseness and limited interaction counts. So the development of computational methods and tools to process data from both single cell Hi-C and bulk Hi-C technologies are focused in this work and those are proven to be enhancing the efficiency and accuracy of Hi-C data processing pipelines.

In this dissertation, each chapter consists of a single individual method or a tool to enhance chromatin interaction processing pipelines and the final chapter focuses on the interplay between epigenetic data and chromatin interactions data. The studies that are focused on building computational methods include increasing data read accuracy for bulk Hi-C, identifying statistically significant interactions at single cell Hi-C data, and imputation of single cell Hi-C data to improve quality and quantity of raw reads. It is anticipated that the utilization of the tools and methods outlined in these studies will significantly enhance the workflows of future research on chromatin organization and its correlation with cellular functions and genetic diseases.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Lu Liu for his continued support, help and direction. This would not have been possible without his invaluable guidance. I wish to extend my gratitude to Dr. Changhui Yan, Dr. Anne Denton and Dr. Mingao Yuan for being on my graduate committee and for the invaluable support. I would like to thank Casey Bartlett for assisting me with reviewing the manuscripts. I wish to express my deepest gratitude to my father, Mr. Hemasiri Cooray, my mother, Mrs. Chandra NAA, my beloved wife, Dr. Wathsala Jayawardana, and all my family and friends for their constant support and encouragement throughout this study.

I am deeply grateful for the free education system in my home country, Sri Lanka. Special acknowledgments are due to my primary school, Wickramashila College, my daham pasala, Minioluwa vidyavasa piriwena temple, my secondary school, Bandaranayake College, and my alma mater, University of Moratuwa. Each of these esteemed places has played a crucial role in shaping me into the individual I am today.

I deeply appreciate WSO2 Inc for providing me with invaluable internship and employment opportunities in the field of computer science, both during and after my undergraduate studies. These experiences have not only allowed me to refine my skills but have also ignited my passion for advancing my education through graduate studies. I would like to extend special thanks to my team lead, Chamith Kumarage, all other mentors and colleagues at WSO2 Inc who have offered guidance and support throughout my journey.

I would like to express my sincere gratitude to Meta Platforms Inc. for providing me with a great opportunity to participate in summer PhD internships. This experience allowed me to work on a variety of exciting projects, significantly enhancing the skills crucial for my research related to machine learning and data mining. I am particularly thankful to my manager, Shiyi Tu, for his unwavering support. Additionally, I extend my appreciation to all my teammates and friends whose support was invaluable throughout this journey.

I extend my deepest thanks to North Dakota State University and the Department of Computer Science for granting me the opportunity to pursue my graduate studies. This experience not only enhanced my academic skills but also exposed me to a new cultural environment, significantly enriching my journey far from home.

Finally, my heartfelt thanks goes out to the Bartlett, Carlson, and Hall families for their unwavering support throughout my graduate journey in the USA. Their warm hospitality during family gatherings, holidays, and various fun activities greatly enriched my experience and provided a much-appreciated sense of home away from home. I am truly thankful for their kindness and encouragement.

The work is supported by the National Science Foundation under NSF EPSCoR Track-1 Cooperative Agreement OIA #1946202 and used advanced cyber infrastructure resources provided by Computational Research Center (CRC) at University of North Dakota and the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Bioinformatics involves the storage, retrieval and analysis of large amount of biological data including genomic information, nucleotide, amino acid, protein structures and regulatory information. It enables scientists to understand complex biological processes and diseases, and plays a key role in developing new medical treatments and personalized therapeutic strategies by identifying genetic markers and pathways associated with various conditions across a vast collection of genomic datasets[79]. The primary application of bioinformatics is the analysis of DNA, RNA, and protein sequences, and their functional implications on cellular activities. This includes identifying genes, analyzing regulatory elements, predicting the functionality of proteins, and their relationship with genomic structure. Advanced computational techniques are often required to address complex biological challenges related to sequencing and processing due to the massive volume of data involved. With its strong foundation in algorithms, data structures, and high performance computing, Computer Science principles serve as the backbone for developing computationally effective tools and software that can efficiently store, manage, process, and visualize biological data. Thus the integration of numerous computer science and statistical concepts, such as data mining and machine learning, with the analysis of biological processes has greatly accelerated biological research in many aspects including gene and protein expression analysis, mutations in cancer, and modeling biological systems [137] [91] [169] [95] [103].

Deciphering the three-dimensional organization of the genome still remains a major focus in the field of biology as it can disclose great insights into gene regulation and their correlation with genetic diseases such as cancer [148]. This involves analyzing chromatin interaction data generated using experimental methods such as chromosome conformation capture [143]. More advanced experimental methodologies, such as high-throughput sequencing (the most comprehensive method to analyze genome-wide chromatin interactions) can often result in millions to billions of reads per sample [104]. These need to go through numerous processing pipelines to filter out higher-quality and meaningful reads. Despite the development of numerous tools for curating this information 1.1 1.2, not all challenges associated with processing chromatin interactions data have been resolved. In addition, some of the existing methodologies face various practical challenges due to the

limitations of available datasets and computational resources. To overcome these limitations, the development of more robust and efficient computational tools is necessary. further research should be conducted to utilize various computer science concepts, including different algorithms and data structures. Furthermore, the application of advanced data mining and machine learning techniques to genomics data can facilitate the extraction of meaningful patterns, the inference of relationships, and the generation of insights from large datasets.

The analysis of genome structure data has significantly advanced since the introduction of the first experimental method, chromosome conformation capture, in the early 2000s [160]. More robust and advanced experimental methods to analyze genome wide chromatin organization, such as Hi-C, were developed in 2009 [104]. However, effective computational methods for processing and analyzing this data, particularly for single-cell data, have only recently been introduced [191]. This may be attributed to several factors, including the limited availability of public datasets and the need for high-performance computing resources. Moreover, researchers are continuously making effort to model genomic data into more accurate representations to enhance our understanding of its functional applications. The development of more efficient and effective computational methodologies could streamline this process, bringing us closer to a deeper comprehension of the complex interactions between chromatin structures and their functions. Such advancements could lead the way for novel therapeutic strategies for a variety of genetic diseases, including cancer and other disorders.

All the studies presented in this thesis introduce computational methods specifically designed for analyzing and processing chromatin interaction data, particularly generated using Hi-C methodology. These methods have proven to enhance both single-cell and bulk Hi-C data processing pipelines and are expected to contribute to further genomic research focused on analyzing chromatin organization and its biological functional implications. Section 1.1.1 offers a concise overview of chromatin organization topology and its role in gene regulation. Section 1.1.2 describes the relationship between epigenetics and chromatin interactions, along with their functional implications for gene regulation. Section 1.1.3 discusses how analyzing chromatin organization can provide valuable insights into human health and genetic diseases. Section 1.1.4 introduces the experimental methodologies used to capture data on chromatin interactions. Section 1.1.5 outlines the techniques for capturing epigenetic data. Section 1.1.6 provides a brief overview of the existing com-

putational methods for analyzing chromatin organization and addresses the need for more advanced computational methodologies to process chromatin interaction data, considering the limitations and challenges of existing techniques and methodologies. Section 1.1.7 discusses various representation capabilities of chromatin interaction data and their impact on computational methods. Section 1.2 outlines our motivation for pursuing the studies presented along with the problem definition. Finally, Section 1.3 lists our contributions.

1.1. Background

1.1.1. Chromatin Architecture and Gene Regulation

Chromatin organization within the nucleus is a hierarchical structure ranges from the smallest loops formed by DNA and histone proteins to the higher level organization such as compartmentalization and chromosome organization. This multi-tiered structure plays a critical role in gene regulation, genome stability and cellular function. Sections 1.1.1.1-1.1.1.4 provide brief explanation to the most significant units in the chromatin organization and their functional implications.

1.1.1.1. Formation of Chromatin Interaction

The genetic information of an organism is stored in DNA which is composed of two complementary strands that form a double-helix structure. The DNA is primarily made up of a sugar phosphate backbone and four types of nucleotides: Adenine, Thymine, Cytosine, and Guanine, represented as A, T, C, and G bases. Within the DNA, Adenine pairs with Thymine, and Cytosine pairs with Guanine.

Chromatin interactions play a pivotal role in the regulation of gene expression which influences the cell and genomic functionality. These interactions occur within the nucleus of a cell, where DNA is packaged into a complex structure known as chromatin. This structure undergoes various modifications and reorganizations that enable or restrict access to specific genetic information. Through mechanisms such as looping, chromatin brings distant genes or regulatory elements into close proximity which facilitates or hinders the recruitment of transcription factors and other regulatory proteins. This dynamic interplay is crucial for the orchestration of developmental processes, the maintenance of cellular identity, and the response to external signals.

Chromatin interactions can be further categorized into those occurring within the same chromosome (intra-chromosomal interactions) and those occurring between different chromosomes (inter-chromosomal interactions) (Figure 1.1).

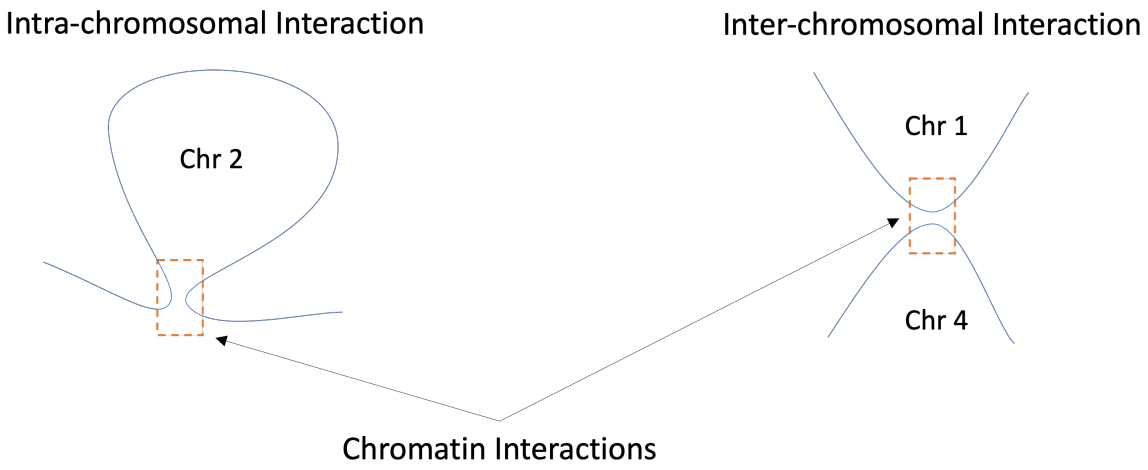

Figure 1.1. Intra vs Inter chromosomal interactions

1.1.1.2. Role of Chromatin Interactions in Gene Regulation

Gene regulation is essential for the proper functioning of an organism. It ensures that specific genes are expressed at the appropriate times, within the correct cells, and in precise quantities. This process ensures that cells are adapted to various environmental conditions, facilitating proper development, functioning, and survival. Disruptions in gene regulation can lead to the development of various diseases, such as cancer, or other abnormalities [82]. Therefore, precise gene regulation is fundamental for proper complex biological processes and the maintenance of healthy functioning cell.

Regulatory elements acts as critical components for controlling the flow of genetic information from DNA to functional proteins and those are key components in gene regulation. Those elements such as promoters, enhancers, silencers and insulators, interact with transcription factors and other associated proteins which influence the transcriptional machinery inside a cell. Promoters are DNA sequences located directly upstream of the corresponding gene and they provide the binding sites for RNA polymerase and transcription initiation. Enhancers are located far away from the gene and are responsible for enhancing the transcription of a gene by increasing the rate of transcription. Silencers can repress gene expression and decrease the rate of gene transcription. Insulators are regulatory elements that restrict the influence of enhancers or silencers on the expression of nearby genes and act as boundaries and ensure that those regulatory elements affect only their target genes. These regulatory elements are associated with each other to control and

maintain the complex and dynamic process of gene expression and essential for proper functioning of cells, normal growth, development and response to environmental changes [23].

Chromatin interactions and loops bring distant genomic regions and their respective regulatory elements into close proximity. Known regulatory elements, such as enhancers and promoters, often interact with each other to initiate gene transcription, despite being physically separated by large genomic distances. Chromatin loops form when two such distant genomic regions, separated by thousands to millions of base pairs, are brought into close proximity within the three-dimensional organization of chromatin. These loops are frequently anchored by protein complexes, including Cohesin and CTCF. Consequently, this looping mechanism plays a crucial role in gene regulation and the control of gene expression. By examining these loops, scientists can understand how alterations in the three-dimensional genome structure are linked to a range of applications, including disease-associated studies [148], epigenetic research [133], and gene regulation across different cell types under various conditions [99].

1.1.1.3. Topologically Associating Domains (TADs)

Topologically associating domains (TADs) are representation of genomic regions where DNA sequences interact more frequently with each other than with sequences outside the region. These regions typically can span to hundereds of kilobases to several megabases and varies among different organisms [10] [32]. TADs act as a structural unit that organize and regulate gene expression in which regulatory elements such as enhancers and promoters are brought into close proximity. Even though the genomic regions in these regions tend to interact more frequently with each other, the cross-TAD interactions were also observed in recent studies [68]. Topologically Associating Domains (TADs) are crucial for identifying regulatory elements and understanding their correlations with gene regulation across different cell types or conditions [74]. Furthermore, analyzing disruptions in TADs can provide insights into the mutations present in cancer genomes and the regulatory mechanisms of oncogenes [173] [64] [44] [78].

1.1.1.4. Compartmentalization of Chromatin(A/B compartments)

Compartmentalization of Chromatin refers to the higher order organization of chromatin in the nucleus into two main distinct types such as A compartments and B compartments. A compartments refer to actively transcribed regions and those are usually located more internally within the nucleus. A compartments associate with the regions of the genome that are involved in

active transcriptional process. In contrast, B compartments correspond to inactive regions. These regions typically has a lower density of genes and low levels of gene expression. B compartments generally located near the nuclear lamina.

The dynamic interplay between A and B compartments within the nuclear architecture reflects not only the current state of cellular function but also plays a crucial role in revealing their relationship with gene regulation [62]. Moreover, this interplay significantly impacts the changes in gene expression associated with cancer progression [140].

1.1.2. Epigenetics and Chromatin Interactions

Epigenetics is a field that focuses on the chemical modifications of DNA and its associated proteins, which can affect gene expression without altering the corresponding genetic sequence [72]. Two common epigenetic modifications are DNA methylation and histone modification. DNA methylation is a biochemical process involving the addition of a methyl group to the DNA molecule which is typically at cytosine bases, leading to changes in gene expression [164]. As the evidence suggests, changes in methylation in the promoter region of a gene can impact gene silencing [116] [180]. Additionally, disruptions in DNA methylation patterns can lead to altered gene functionalities and are implicated in various diseases, including cancer [152] [75] [192].

Histone modifications refer to the chemical changes in histone proteins which serve as the structural framework around which DNA coils [132]. For instance, histone modifications associated with transcriptional activation relax the chromatin structure, making DNA more accessible to the transcription machinery [29]. However, histone modifications can lead to either activation or repression of gene expression, depending on the specific methylated amino acids and methyl groups involved. Histone modification is crucial for cell cycle regulation and development and is also associated with various genetic diseases [8] [151].

Both epigenetic modifications and chromatin interactions are associated with the control of gene expression and gene regulation. Epigenetic modifications can activate or repress gene expression without altering the genetic sequence, using mechanisms such as DNA methylation and histone modifications. Similarly, chromatin interactions control gene regulation by bringing distant genomic regions and regulatory elements into close proximity. Thus, understanding epigenetics and chromatin interactions, as well as their interplay, is crucial for unraveling the dynamics of gene expression across various environments and cell development [120] [24].

1.1.3. Chromatin Organization in Health and Disease

The structural and organizational dynamics of chromatin are fundamental to the regulation of gene expression. Disruptions in this organization are closely linked to a spectrum of health complications and genetic diseases, including cancer and various other pathologies. Higher-order chromatin organizations, such as topologically associating domains (TADs), are essential for orchestrating gene regulation. Alterations in TADs can significantly affect the regulatory landscape over long distances and potentially lead to the emergence of disease-related phenotypes [36] [113].

Instances of human limb malformations have been linked to genomic structural changes, such as deletions, inversions, or duplications within the TAD-spanning locus of WNT6/IHH/EPHA4-/PAX3 [113]. Laboratory studies, involving mice engineered to carry similar genomic rearrangements, have replicated these findings, emphasizing the critical role of TADs in regulating gene expression. Moreover, alterations in the structure and organization of the genome are associated with changes in gene expression levels, which contribute to the analysis of various pathological conditions [58] [159].

In cancer, alterations and disruptions in the three-dimensional organization of chromatin play a significant role in the progression of the disease. These disruptions can manifest through various mechanisms, including copy-number variation, long-range epigenetic changes, and the activation of atypical gene expression programs, particularly in prostate cancer cells [165]. Despite cancer cells' ability to organize their genomes into TADs, these domains are often smaller with additional cancer-specific domain boundaries. These newly formed boundaries frequently occur with areas of copy-number variation and leads to altered chromatin interactions and regulatory region activities. This results in long-range epigenetically activated or silenced regions with concordant gene activation or repression in prostate cancer. It illustrates the relationship between long-range epigenetic and genomic dysregulation with the changes in higher-order chromatin interactions in cancer [165].

The phenomenon of long-range epigenetic silencing (LRES) affecting neighboring genes has been observed across various cancers [124] [134] [162]. That shows how 3D chromatin architecture influences cancer hallmarks such as sustaining proliferative signalling, evading growth suppressors, resisting cell death, activating invasion and metastasis, enabling replicative immortality, inducing

angiogenesis, reprogramming of energy metabolism, creating the tumour microenvironment, inflammation, evading immune destruction, and genome instability due to mutations [57]. Therapeutic interventions targeting spatial genome organization such as curaxins, have shown promising results in contributing to affect this regulatory level [81] [80].

The relationship between three-dimensional genome organization and active mutational processes influences the observed large-scale variations in mutation rates across human cancers. An analysis of 3,000 tumor-normal paired whole-genome datasets across 42 types of cancer revealed a significant correlation between somatic mutations and topologically associating domain (TAD) boundaries. This finding indicates that somatic mutational load in cancer genomes co-localizes with TAD boundaries, suggesting a significant impact of genome architecture on mutation rates [2].

The molecular mechanisms that underlie transcriptional dysregulation in cancer, including dysregulated enhancers and aberrant enhancer-promoter interactions, offer new insights into cancer development and progression [163] [63] [13]. They suggest potential therapeutic targets, indicating that alterations in chromatin topology can activate oncogenes and contribute to cancer phenotypes [52]. Structural variants such as inversions [52] and translocations [123] can facilitate the expression of oncogenes by positioning enhancers proximal to oncogene promoters. This highlights the complex relationship between chromatin structure and the evolution of cancer [63].

1.1.4. Chromosome Conformation Capture Methodologies

Chromosome conformation capture methods have been developed to map chromatin interactions within cells. Numerous experimental approaches exist to extract these chromatin interactions, as described below. Comprehensive list is available at [117].

1.1.4.1. Chromosome Conformation Capture (3C)

The Chromosome Conformation Capture (3C) technique is the foundational method for identifying locations of chromosomal interactions [34]. It has served as the foundation for many subsequent methodologies and is utilized to analyze the frequency of interactions between specific genomic regions, providing a one-to-one mapping. The 3C technique has the capability of confirming the existence of chromatin loops between proximal chromatin regions.

The 3C procedure involves several steps, beginning with the cross-linking of spatially proximal regions within the nucleus using formaldehyde, which stabilizes the contacts. Subsequently, the DNA is fragmented using a restriction enzyme to isolate these contacts, followed by the ligation of

the DNA fragments. The DNA is then purified, and the genomic sites of interaction are identified using Polymerase Chain Reaction (PCR).

1.1.4.2. Circular Chromosome Conformation Capture (4C)

The 4C (Circular Chromosome Conformation Capture) method, an evolution of the 3C (Chromosome Conformation Capture) technique, is adept at identifying genomic sites across the entire genome that interact with a specific locus of interest (one-to-many mapping) [156]. This method can generate high-resolution contact maps surrounding the target genomic site and requires fewer reads compared to methods such as Hi-C, making it more efficient in specific contexts, such as analyzing interactions related to a particular locus or gene[157].

The 4C protocol includes several initial steps from the 3C process, such as crosslinking at the ligation sites and fragmenting DNA using a primary restriction enzyme. Following in situ ligation of these fragments, the crosslinks are reversed, and the DNA is purified. Then the purified fragments are cut using a secondary restriction enzyme and ligated once more to create circularized DNA molecules. These circularized molecules are then processed through inverse PCR, which cleaves the ligations and attaches primers specific to the region of interest. Finally, the fragments are sequenced using next-generation sequencing techniques. The contact frequencies are determined by analyzing the proportion of reads mapped to particular genomic sites.

1.1.4.3. Chromosome Conformation Capture Carbon Copy or 3C-Carbon Copy(5C)

The 5C technique represents an extension of the 3C method, involving high-throughput and comprehensive analysis of many interactions concurrently [38]. It involves the simultaneous examination of interactions between multiple loci. Similar to the 3C method, the 5C approach starts with the cross-linking of ligation sites, followed by fragmentation using a restriction enzyme. After that, 5C utilizes ligation-mediated amplification to investigate interactions between multiple loci. The amplified products are then subjected to sequencing or microarray analysis to generate chromatin interactions.

1.1.4.4. Chromosome Conformation Capture with High Throughput Sequencing (Hi-C) and Variants

The Hi-C method is capable of identifying genome-wide chromatin interactions, and it has become increasingly popular due to its ability to generate a vast number of genome-wide chromatin interactions compared to earlier methods [174] [11]. There are two primary types of Hi-C methods:

single-cell Hi-C and bulk Hi-C. Single-cell Hi-C captures chromatin interactions within individual cells, whereas bulk Hi-C captures chromatin interactions from a mixture of cells.

The Hi-C method expands upon the 3C process by labeling the ends of DNA fragments with biotin, assisting in the identification of ligation sites. This method involves ligating the fragments, shearing the DNA to remove cross-links, and finally analyzing the chimeric reads using high-throughput paired-end sequencing.

Due to its widespread adoption, several variants of the Hi-C method have been introduced to address different research needs. Diploid Chromosome Conformation Capture (Dip-C) is a variant designed for analyzing chromatin interactions at the single-cell level, thus providing insights into cell-to-cell heterogeneity and the dynamics of chromosome organization. In situ Hi-C improves upon the original protocol by performing the proximity ligation step within intact nuclei, thereby reducing DNA loss during the process and enhancing the resolution and efficiency of interaction detection. Micro-C utilizes micrococcal nuclease (MNase) for chromatin digestion, in contrast to the restriction enzymes used in traditional Hi-C, resulting in finer resolution maps of chromatin interactions [87]. This method is particularly effective in mapping nucleosome-nucleosome interactions and revealing detailed chromatin organization. Lastly, HiChIP modifies the Hi-C protocol by incorporating a chromatin immunoprecipitation (ChIP) step, making it valuable for studying chromatin interactions mediated by specific proteins of interest, similar to the ChIA-PET method.

1.1.4.5. Chromatin Interaction Analysis by Paired-End Tag Sequencing(ChIA-PET)

Compared to the Hi-C method, which provides a comprehensive overview of all chromatin interactions within the nucleus, ChIA-PET specifically targets interactions mediated by particular proteins. It combines chromatin immunoprecipitation (ChIP) with DNA sequencing to identify interactions between DNA regions bound by a specific protein. We will discuss another variation of the Chromatin Immunoprecipitation method, called ChIP-Seq, which focuses on the interactions between DNA and proteins, in later sections on epigenetic analysis.

This method involves several steps, including cross-linking to stabilize protein-DNA interactions, immunoprecipitation to enrich DNA segments bound by specific proteins, and sequencing to identify the interacting DNA regions. This technique is often utilized to analyze the role of transcription factors in the formation of interactions between DNA elements and their relationship to gene regulation. Consequently, ChIA-PET is particularly useful for revealing the role of pro-

teins such as transcription factors [195], estrogen receptors [59], CTCF binding factors, and histone proteins in the organization of the genome into functional domains.

1.1.5. Techniques in Epigenetic Analysis

Epigenetic analysis techniques are essential for understanding gene regulation beyond mere DNA sequence analysis. These methods focus on studying DNA methylation, histone modifications, and DNA-binding proteins, and etc. In this dissertation, we analyzed data generated using techniques specifically aimed at processing histone modifications, DNA methylation, and transcription factor binding sites. Chromatin Immunoprecipitation Sequencing (ChIP-Seq) stands out as a robust and powerful technique that merges chromatin immunoprecipitation with high-throughput DNA sequencing to investigate protein-DNA interactions within the genome. The ChIP-Seq process begins by crosslinking proteins to DNA, fragmenting the DNA, and then selectively isolating specific DNA-protein complexes. Then the DNA is purified, sequenced, and mapped to a reference genome for enrichment analysis. The enrichment of DNA sequences in corresponding genomic regions signifies the locations of those specific protein binding sites.

H3K4me3 and H3K27ac are examples of histone modifications, each representing a distinct epigenetic mark that plays a crucial role in the regulation of gene expression. H3K4me3 involves the addition of three methyl groups to the lysine 4 residue of histone H3 and is strongly associated with actively transcribed genes, primarily located near the promoter regions. This modification serves as an indicator of active gene promoters, facilitating transcription initiation [66]. Conversely, H3K27ac, which involves the addition of an acetyl group to the lysine 27 residue of histone H3, is associated with chromatin relaxation and active gene transcription. Typically found near enhancer regions, H3K27ac serves as an indicator of active enhancers [26]. The presence and patterns of these epigenetic markers are essential for understanding gene activity and the various aspects of active gene regions.

1.1.6. Existing Computational Methods for Analyzing Chromatin Interactions

In this thesis, we propose advanced computational methodologies for processing chromatin interaction data. This section analyzes existing computational techniques and organizes them based on their application to various aspects of chromatin interaction analysis. These include methods for single-cell interactions, bulk interactions, the analysis of raw interaction reads, and integrative analytical approaches. The underlying principles and methodologies of these existing techniques

have provided a solid foundation, enabling us to introduce novel computational methods. These existing methods are crucial for identifying appropriate reference datasets, conducting thorough benchmarking of results, and perform comprehensive comparisons. Additionally, we do not cover the most common genomic tools such as FastQC, BWA, and Bowtie, as they are widely recognized for quality control and alignment tasks in various sequencing analyses. Instead, our focus is on more specialized computational methods specifically designed for chromatin interaction processing workflows.

Tables 1.1 and Table 1.2 list the majority of the Hi-C tools and methodologies utilized in our proposed computational methodologies for various purposes, including preprocessing, visualization, and benchmarking, as well as comparing results.

1.1.6.1. Computational Methods for Analyzing Bulk Chromatin Interactions

Bulk chromatin interaction analysis enables researchers to examine the three-dimensional structure of the genome and its functional implications. The Bulk Hi-C method is widely regarded as the most effective for analyzing chromatin interactions, due to its ability to generate large volumes of experimental data. However, the raw data from Hi-C experiments include noise, biases, and artifacts resulting from experimental procedures and sequencing technologies. Advanced computational methods are required to correct these biases, normalize the data, and transform the raw interaction frequencies into meaningful biological insights. These methods should be capable of identifying chromatin interactions with higher confidence along with the comparison of chromatin structures across different cell types or conditions, and assisting to uncover the underlying principles of genome organization. Without such computational preprocessing, our ability to explore the complexities of genomic architecture and its impact on cellular functions would be significantly limited.

Numerous bulk Hi-C datasets are available across various organisms, including different human tissues, disease cells, various animal species, and even plants [31]. This extensive collection enables the exploration of chromatin interactions between different species, cell cycles, disease phases, and between normal and disease cells. Comprehensive end-to-end pipelines exist, ranging from the processing of raw interaction data to the generation of processed contact matrices and visualizations. HiC-Pro [150], hiclib [71], Hicup [182] and Juicer [40] are among the most commonly used tools in the research community.

Table 1.1. Existing Hi-C tools and methodologies-1

Tool	Purpose	Source	Reference
HiC-Pro	Pipeline to process data from raw reads to normalized contact maps	https://github.com/nservant/HiC-Pro	[150]
FitHiC and FitHiC2	Identify significant inter-chromosomal and intra-chromosomal interactions from given contacts	https://github.com/ay-lab/fithic	[6], [85]
HiCCUPS	Identify significant inter-chromosomal and intra-chromosomal interactions from given contacts	https://github.com/aidenlab/juicer	[136]
hiclib	Generate contact maps from raw reads	https://github.com/mirnylab/hiclib-legacy	[71]
scHiC-Explorer	Set of tools to analyze, process and visualize hi-C and single cell Hi-C data	https://github.com/joachimwolff/scHiCExplorer	[184]
cooltools	A toolset to analyze Hi-C data for various tasks including normalization, compartment and TADs analysis	https://github.com/open2c/cooltools?tab=readme-ov-file	[126]
mHi-C	Recover multimapping reads when aligning hi-c raw data	https://github.com/keleslab/mHiC	[200]
HiCrep	Measure reproducibility of Hi-C contact matrices	https://github.com/TaoYang-dev/hicrep	[189]
HiC-DC	Identify Significant interactions	https://bitbucket.org/leslielab/hic-dc/src/master/	[21]
diffHic	Detect differential genomic interactions in Hi-C data	https://www.bioconductor.org/packages/release/bioc/html/diffHic.html	[111]
SnapHiC	Identify significant interactions from single cell Hi-C data	https://github.com/HuMingLab/SnapHiC	[191]
Hicup	Provide a pipeline to process raw fastq reads including steps: Truncating, mapping, filtering and deduplicating	https://github.com/StevenWingett/HiCUP	[182]
Hic-inspector	Provide a suite of tools designed for Hi-C data processing tasks including aligning, counting, filtering, and generating contact maps	https://github.com/HiC-inspector/HiC-inspector	[22]

Table 1.2. Existing Hi-C tools and methodologies-2

Tool	Purpose	Source	Reference
Hippie	A pipeline to extract intra and inter-chromosomal enhancer–target gene relationships	http://wanglab.pcbi.upenn.edu/hippie/	[70]
Hicdat	Provide a graphical interface to perform hic processing tasks along with other data types including Chip-seq and RNA-seq	https://github.com/MWSchmid/HiCdat	[145]
Hifive	A set of tools for processing HiC and 5C data	https://github.com/bxlab/hifive	[144]
Hic-bench	A set of pipelines for Hi-C and ChIP-Seq analysis	https://github.com/NYU-BFX/hic-bench	[93]
Hic-spector	A matrix library for spectral and reproducibility analysis of Hi-C contact maps	https://github.com/gersteinlab/HiC-spector	[188]
Hibrowse	A locally deployable browser designed for the visualization and analysis of Hi-C data, along with its genetic and epigenetic annotations	https://github.com/lyotvincent/HiBrowser	[101]
Juicebox	visualization software for Hi-C contact maps	https://github.com/aidenlab/Juicebox	[40]
HiCPlus	Enhance the resolution of Hi-C contact maps utilizing convolutional neural networks	https://github.com/zhangyan32/HiCPlus	[198]
HiCNN	HiCPlus iteration using deep convolutional neural networks	http://dna.cs.miami.edu/HiCNN/	[109]
HicGAN	Improve resolution of Hi-C maps using generative adversarial networks (GANs)	https://github.com/kimmo1019/hicGAN	[108]
DeepHiC	Enhance the resolution of Hi-C contact maps through Generative Adversarial Network	https://github.com/omegahh/DeepHiC	[65]

HiC-Pro offers an integrated pipeline for processing Hi-C data, starting with the alignment of reads to the reference genome. These reads are then mapped to restriction fragments, followed by the classification and removal of invalid interaction pairs. The pipeline concludes by providing raw contact matrices alongside ICE-normalized contacts. Additionally, HiC-Pro supports allele-specific analysis when relevant data are supplied [150]. Similarily, hiclib is a Python library that provides a flexible framework for Hi-C data analysis. It supports preprocessing, mapping, and filtering of Hi-C data, allowing users to interact with the data throughout each step [71]. However, hiclib requires a more involved setup and poses a steeper learning curve for individuals unfamiliar with programming. Juicer is another platform that specializes in generating Hi-C maps from raw reads. It offers a suite of command-line tools for various annotations and analyses. Typically, these pipelines focus solely on uniquely mapping reads and overlooked multi-mapping reads and other read types. To overcome this limitation, multi-mapping reads recovery algorithms similar to mHiC have been developed [200]. Drawing inspiration from these, this thesis proposes a heuristic strategy-based method to recover multi-mapping reads.

However, the above mentioned tools lack the capability to identify statistically significant and more meaningful interactions. To address this, tools such as FitHiC , followed by FitHiC2 [6] [85] and HiCCUPS [136], were introduced to filter out statistically significant interactions. FitHiC applies a statistical approach to assign confidence scores on interactions based on the frequency of contact between genomic loci and utilizes a spline regression model to represent distance-dependent interaction frequencies. FitHiC2 was later introduced to enhance the effectiveness of distinguishing between random noise and biologically meaningful interactions. HiCCUPS, similar to FitHiC, is a peak-calling algorithm and it identifies areas where interactions between parts of the genome are unusually high considering the surrounded local neighborhood. It compares the frequencies of pixels in the contact matrices to those of surrounding areas and identifies statistically significant peaks according to the predefined four neighborhoods around the corresponding pixel.

Furthermore, various computational tools have been implemented to serve different purposes in Hi-C processing pipelines. HOMER is a comprehensive suite designed to provide functionalities such as annotation, normalization, integration with other genomic data, and visualization [60]. However, HOMER does not include read mapping functionality. GOTHiC provides a probabilistic model to identify genuine interactions using a binomial test while correcting for biases [115]. As

the need to compare Hi-C data increases, tools such as diffHiC [111] have been introduced to detect differential genomic interactions in Hi-C data. HiCExplorer [184] is another comprehensive suite that includes the functionalities for processing, normalization, analysis, and visualization of Hi-C data. It includes additional capabilities such as the identification of topologically associating domains (TADs) and A/B compartments. These tools and frameworks have laid a strong foundation for Hi-C data analysis and opened up numerous possibilities for continuing to enhance and introduce more advanced computational methodologies.

1.1.6.2. Computational Methods for Analyzing Single Cell Chromatin Interactions

Single-cell Hi-C data provide an opportunity for researchers to analyze the heterogeneity and dynamic nature of genome architecture across different cell types, developmental stages, and disease states [121] [17]. Unlike bulk Hi-C data, which aggregates chromatin interactions from millions of cells, single-cell Hi-C captures the chromatin interactions within individual cells. Consequently, single-cell Hi-C data present unique challenges, including sparse contact matrices and increased noise and variability between cells due to lower number of reads generated per experiment. Furthermore, there are fewer single-cell Hi-C datasets available compared to bulk Hi-C datasets. For example, a dataset comprising 10,696 mouse and human single cells, introduced by Ramani et al. [135] as part of the single-cell combinatorial indexed Hi-C (sciHi-C) method, contains an average of 25,632 contact pairs per cell. Followed by that, Kim et al. [86] generated data from over 19,000 cells across five human cell lines (GM12878, H1Esc, HFF, IMR90, and HAP1) using the sci-Hi-C method, averaging 8,167 contacts per cell. Consequently, most available single-cell experimental data introduce significant sparsity into the contact matrices. The single-cell dataset of the human brain, generated by Lee et al. [94] using the single-nucleus methyl-3C sequencing method, consists of 398,726 contacts per cell. Similarly, the cell cycle dataset of mouse embryonic cells produced by Nagano et al. [121] shows comparable numbers of contacts per cell. However, these figures still represent a relatively low sequencing depth for revealing insightful patterns.

To mitigate these issues, imputation methods have been introduced to enhance single-cell data by predicting missing interactions and reducing data sparsity. These imputation algorithms utilizes probabilistic approaches, such as the random walk with restart [191], to impute and filter out significant interactions. Despite their benefits, these algorithms also have limitations, which are discussed in subsequent chapters. Beyond identifying significant interactions, clustering within

single cells according to the cell type or phase has emerged as a major application. Approaches based on random walks and linear convolution have been adapted to implement a single-cell clustering algorithm[201], facilitating the analysis of TADs (Topologically Associating Domains) across single cells and enabling visualizations. Additionally, methods based on nearest neighbors and unsupervised embedding have been utilized in clustering single cells [183] [106].

Single-cell Hi-C analysis tools and methodologies are still in their early stages, and researchers often resort to applying bulk Hi-C methodologies to single-cell data to avoid limitations. Further research is needed to experiment advanced computational algorithms and data structures on single cell hi-c data to uncover hidden patterns and variations. As a result, in this work, we introduce two novel computational tools designed to identify statistically significant interactions from single-cell Hi-C data.

1.1.6.3. Computational Methods for Integrative Analysis

Hi-C data serves as a valuable resource for deciphering chromatin topology and functional regulatory elements. However, Hi-C data alone may not provide a complete picture of the relationship between gene regulation and chromatin structure. The capabilities of Hi-C experiments are primarily limited to reflecting genomic regions in close proximity without necessarily representing the functional relationship between them. Additionally, various noises and biases associated with Hi-C data complicate the differentiation between specific interactions occurring due to random noise or actual functional relationships.

By integrating Hi-C data with other omics datasets, such as gene expression, DNA methylation, histone modifications, transcription factor binding, and chromatin accessibility, researchers can gain a comprehensive overview of the regulatory mechanisms correlating gene expression with cellular function. Moreover, one-dimensional (1D) chromatin data, generated using experimental approaches like ChIP-Seq, tend to produce genomic signals at a much higher resolution than Hi-C data and it offers a more fine-grained analysis of genomic structure. Numerous computational methods have been introduced to understand the interplay between different types of data related to chromatin structure [56] [76] [5]. These methodologies enhance our understanding of the complex interactions within the genome and provides a more holistic view of cellular function and regulation.

Higashi is a computational strategy for integrating single-cell Hi-C data with methylation data [196]. It utilizes a hypergraph neural network to model the relationships between different

chromatin regions and generate low dimensional embedding which allows for the characterization of genome structures into compartments and TADs along with cell type classification. The method offers a refined perspective on chromatin organization, illustrating how the integration of Hi-C data with other omics datasets at the single-cell level can support the development of an embedding model that reflects diverse cell types and cellular states.

Graph embedding techniques have successfully been used to identify genomic subcompartments from Hi-C data along with integration of other omics data for evaluation [4]. By converting high-dimensional chromatin interaction data into a more manageable lower-dimensional space, these methods unveil patterns and structures that remain obscured by standard Hi-C workflows and pipelines. Unsupervised learning algorithms can identify clusters within this data, corresponding to genomic regions with similar interaction profiles. Similarly, graph embeddings of both 1D genomic signals and interactions have also been utilized to annotate chromatin domains [155].

Approaches based on Hidden Markov Models have also been applied in the annotation of chromatin states using epigenomic signals. DeepChIA-PET, a deep learning framework, is designed to predict ChIA-PET interactions from Hi-C and ChIP-seq data through a convolutional neural network. This reveals that integrating ChIP-seq data enhances model performance compared to using the Hi-C network alone which implies that the combination of different omics datasets leads to a more nuanced understanding of chromatin complexity [110].

Resources such as LungCancer3D [185] offer comprehensive databases for merging lung cancer chromatin architecture information with multi-omics data. These databases are crucial in understanding disease-specific alterations in chromatin organization and their implications for cancer biology. Moreover, the integration of Hi-C data with gene expression data can reveal how chromatin contacts affect gene regulation and co-expression, as demonstrated by studies from [141] [100] [146].

1.1.7. Computational Concepts and Data Representations

Experiments such as Hi-C, which generate genome-wide chromatin interactions, consists of a high noise-to-signal ratio in the output. This issue arises from various factors, including the inherent nature of the biological elements, limitations of experimental techniques, and the complexity of the data being collected. Additionally, the costs associated with expanding experimental data coverage are substantial which makes it crucial to implement sophisticated computational strategies to derive meaningful biological insights from the data. To manage these types of data, advanced data mining

and statistical techniques are often applied due to their ability to filter out noise and identify patterns hidden within vast datasets. Moreover, robust data structures are required for efficient analysis and processing of this information. The unique characteristics of chromatin interaction data offer numerous opportunities to apply advanced computational concepts tailored to their specific nature and representations. Chromatin interaction data are deposited as raw FASTQ reads in repositories. After alignment to the reference genome, these data are typically represented as interaction pairs, where two genomic regions are shown to interact with each other. These paired reads enable representation through various data structures, including matrices, graphs, and contact maps.

Representing interactions as matrices is the most common approach, as it is capable to handle the application of various statistical and matrix operations efficiently. When converting paired reads into matrices, genomic bins are often used to generate more meaningful interaction pairs based on the resolution and nature of the experimental data. A genomic bin is created by dividing the entire genome, or a chromosome, into regions of equal size. The resolution of bins in Hi-C experiments can range from kilobases to millions of bases, determined by the study's nature and functional implications. Kilobase resolutions are utilized to analyze regulatory elements, loops, and other functional elements associated with gene regulation. For analyzing higher-order chromatin topologies, such as topologically associating domains (TADs) and compartments, megabase resolutions are often used. After grouping the raw reads into genomic bin pairs, 2D matrices are created, representing the linear genome divided into genomic bins as dimensions. This results in a symmetric matrix in which the entries correspond to the number of raw reads associated with each pair of genomic bins.

Another representation of chromatin interactions is through graphs or networks. An interaction graph can be constructed with nodes representing genomic bins or regions, and edges indicating whether a corresponding pair of bins contains an interaction. This graph may also be weighted, where the weight corresponds to the raw read count or a normalized value that represents the significance of the edge. Various graph learning techniques and network related information such as centrality and connected components can be utilized to decipher underlying hidden representations and patterns within these graphs [127]. Additionally, graph embedding algorithms can be applied to learn the latent representations of these graphs [4].

Chromatin interaction data are represented using various visualization techniques to extract information from different perspectives, including heatmaps, arc diagrams, 3D genome models, and browser-based tools. Heatmaps are particularly useful in analyzing global interaction patterns across the entire genome or within specific chromosomal regions. They aid in identifying regions of high interaction (hotspots) that could be crucial for gene regulation and chromatin domain formation including Topologically Associating Domains (TADs) and compartments. In arc diagrams, such as Circos Plots [193], the genome is typically arranged in a circular layout, facilitating the visualization of interactions between distant genomic regions. These diagrams are especially effective for visualizing inter-chromosomal interactions. Furthermore, 3D genome models have been developed recently due to advances in technology and computational methods, to model the organization of chromatin within the nucleus in three-dimensional space. These models are designed to study the spatial context of genomic interactions and get accurate representations of physical interactions in chromatin. Browser-based tools, such as the UCSC Genome Browser [83] or WashU epigenome browser [98], allow users to visualize chromatin interaction data alongside a wide array of other genomic signals, including epigenetic and regulatory element information. These tools assist in studies that are focused on integrative genomic analysis and hypothesis generation.

1.2. Motivation and Problem Definition

Within the broad domain of genomics, understanding the three-dimensional organization of the genome is essential for deciphering the complexities of cellular function and gene regulation. Chromatin interaction data, particularly generated by Hi-C, capture genome-wide interactions, which can be used to explain functional relationships and regulatory mechanisms in cellular growth, cancer development, and other pathologies. However, the interpretation of Hi-C data remains a significant challenge due to various computational challenges in analysis, processing, and integration with other genomic datasets.

The main obstacle in interpreting Hi-C data lies in the limitations of current computational methodologies, which are often constrained by data, resolution, and accuracy. These issues need to be addressed at various stages of the Hi-C processing pipelines. Firstly, it is necessary to recover as many reads as possible from the raw genomic data during alignment to the genome and to assess the possibility of recovering multi-mapping reads in addition to unique mapping reads to enhance the output. Secondly, after generating read pairs, it is crucial to filter out significant interactions to

accurately distinguish biologically meaningful interactions. Currently, very few methodologies have been developed for identifying significant interactions from single-cell Hi-C data. Lastly, integrating Hi-C data with other genomic datasets, such as epigenomic data, is essential for uncovering hidden representations and identifying the role of different data in genome organization.

In response to these challenges, our research is aimed at developing advanced computational strategies to overcome the limitations present in various stages of the Hi-C processing pipeline. By leveraging advancements in data mining, machine learning, and high-performance computing, we propose the construction of diverse computational tools and methodologies specifically designed to process, analyze, and interpret Hi-C data with greater accuracy and efficiency. We anticipate that the studies detailed in this thesis will significantly enhance future research efforts to decode the complexities of chromatin organization and provide new insights into a deeper understanding of human health and disease, ultimately contributing to the development of novel therapeutic strategies.

1.3. Contributions

We introduced a novel methodology for identifying statistically frequent inter-chromosomal interactions using single-cell Hi-C data. To the best of our knowledge, this is the first implementation of a tool for this purpose that is publicly available. In the proposed method, inter-chromosomal interactions are represented as a network. This is followed by the application of a Binomial distribution measurement for filtering, to identify loci with statistically significant interactions. The results were evaluated both statistically and biologically, in comparison with existing literature [18].

We proposed a methodology for recovering multi-mapping reads using a heuristic strategy to enhance Hi-C data. The method involves the recovery of reads based on their distance from the restriction enzyme cutting sites. The performance was compared with that of mHi-C, the only other existing tool of its kind. Additionally, the results were evaluated through biological interpretation [19].

We proposed a computational method to filter statistically significant intra-chromosomal interactions from single-cell Hi-C data. This proposed method comprises three key steps: imputation based on the nearest neighbor, normalization, and identification of significant interactions. SnapHiC, the only existing method, was utilized to benchmark the results, and ChIP-seq data, along with promoter data, were used to evaluate the biological interpretation of the output [20].

We conducted an integrative analysis of chromatin interaction data and epigenetic signals using a graph embedding model. In this study, we focused on identifying the roles of interaction and epigenetic data in constructing chromatin organization. The results were evaluated using a statistical approach, ensuring the global preservation of the chromatin network.

1.4. Dissertation Overview

The dissertation introduces advanced computational strategies designed to enhance both the efficiency and accuracy throughout various phases of the Hi-C processing pipeline. Typically, Hi-C processing pipelines encompass several critical stages, including the mapping and filtering of reads, the paring of valid reads, and the identification of significant chromatin interactions, as illustrated in Figure 1.2. These identified valid pairs and significant chromatin interactions are pivotal for downstream analytical tasks. Such tasks include visualizations, analysis of higher-order chromatin organization, and integrative analysis, also depicted in Figure 1.2.

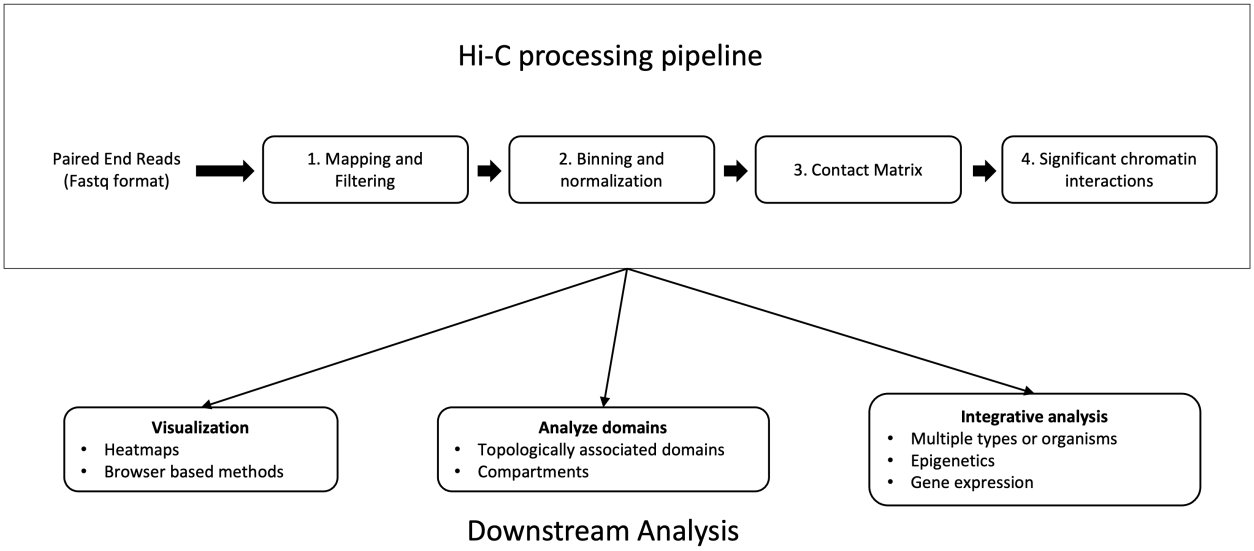

Figure 1.2. Hi-C processing pipeline

Each chapter introduces a distinct computational methodology or tool tailored for specific stages of the process. Specifically, the methodologies discussed in Chapters 2 and 4 concentrate on detecting significant chromatin interactions at the single-cell level, correlating with Step 4 of the Hi-C processing pipeline, as illustrated in Figure 1.2. Chapter 3 is dedicated to improving the mapping and filtering stages of Hi-C reads, aligning with Step 1 in Figure 1.2. Chapter 5 delves into

an integrative analysis that combines chromatin interaction data with epigenetic markers, correlated to the downstream analysis stage.

2. NETWORK-BASED METHOD FOR REGIONS WITH STATISTICALLY FREQUENT INTERCHROMOSOMAL INTERACTIONS AT SINGLE-CELL RESOLUTION

2.1. Abstract

2.1.1. Background

Chromosome conformation capture-based methods, especially Hi-C, enable scientists to detect genome-wide chromatin interactions and study the spatial organization of chromatin, which plays important roles in gene expression regulation, DNA replication and repair etc. Thus, developing computational methods to unravel patterns behind the data becomes critical. Existing computational methods focus on intrachromosomal interactions and ignore interchromosomal interactions partly because there is no prior knowledge for interchromosomal interactions and the frequency of interchromosomal interactions is much lower while the search space is much larger. With the development of single-cell technologies, the advent of single-cell Hi-C makes interrogating the spatial structure of chromatin at single-cell resolution possible. It also brings a new type of frequency information, the number of single cells with chromatin interactions between two disjoint chromosome regions.

2.1.2. Results

Considering the lack of computational methods on interchromosomal interactions and the unsurprisingly frequent intrachromosomal interactions along the diagonal of a chromatin contact map, we propose a computational method dedicated to analyzing interchromosomal interactions of single-cell Hi-C with this new frequency information. To the best of our knowledge, our proposed tool is the first to identify regions with statistically frequent interchromosomal interactions at single-cell resolution. We demonstrate that the tool utilizing networks and binomial statistical tests can identify interesting structural regions through visualization, comparison and enrichment analysis and it also supports different configurations to provide users with flexibility.

2.2. Introduction

Stretching the DNA in a human cell, it would be about two meters long, but how can it fit into a tiny space of about 6 microns across? DNA of cells of different tissues (e.g. neural cells and heart cells) are essentially the same, but why do these cells function disparately and what factors turn the genes' on and off and result in the disparities? To gain insights into these questions, advances in chromosome conformation capture-based technologies have provided researchers a great opportunity to study the higher-order spatial organization of chromatin. A popular method is chromosome conformation capture with high-throughput sequencing (Hi-C), in which genomes are cross-linked with formaldehyde, fragmented with enzymes, randomly ligated in proximity and finally sequenced by next-generation sequencing platforms. After raw reads are processed by bioinformatics pipelines, a genome-wide contact map of a collection of cells is generated and it reveals intrachromosomal interactions and interchromosomal interactions. Intrachromosomal interactions refer to the valid ligations between DNA fragments of the same chromosome and interchromosomal interactions refer to the valid ligations between DNA fragments of different chromosomes. Intrachromosomal interactions are the majority of chromatin interactions in Hi-C experiments and their interaction frequencies are genomic distance dependent [90]. Interchromosomal interactions are two orders of magnitude weaker than intrachromosomal interactions [142] and interchromosomal interactions contain a higher proportion of noise than intrachromosomal interactions [105].

As the popularity of the Hi-C approach grows, large amounts of data have been generated and significant endeavors are devoted to developing computational methods and tools. These computational methods and tools can be coarsely divided into two categories, Hi-C data processing and downstream analysis. For the first category, there are some existing tools used to generate valid chromatin interactions from raw sequencing reads [182][71][22][69][145][150][144][41][93]. They follow similar processing steps and may adopt different sequence alignment strategies (pre-truncation, iterative and trimming), filtering criteria (read-level, read-pair level, strand and distance) and normalization methods (explicit-factor correction, matrix balancing and joint correction). Besides, there are some computational tools to exam the quality of Hi-C data by measuring the reproducibility of Hi-C replicates [144][189][172][188]. For the second category, there are several major analysis tasks to gain insights into the spatial structure and function of chromatin. A/B compartments which

correspond to open and closed chromatin can be identified by using Principle Component Analysis on transformed chromatin contact maps [104]. Megabase-sized Topologically Associating Domains (TADs) can be discovered by using a Hidden Markov Model with a directionality index [36]. There are other methods available to detect TADs [43][136][97][149][181][154]. As TADs are defined as continuous chromosomal loci, these methods only take intrachromosomal interactions into consideration. Statistically significant long-range chromatin interactions are extracted from Hi-C data. As there is no prior knowledge about interchromosomal interactions, computational methods focus on intrachromosomal interactions because the frequency of interactions between two intrachromosomal loci heavily depends on the genomic distance between the loci. Some methods identify statistically significant chromatin interactions by fitting the frequencies of intrachromosomal interactions with certain distributions, such as power-law [104], double-exponential [168] and negative binomial [74]. Instead of assuming a certain distribution, a nonparametric method [6] identifies statistically significant chromatin interactions by estimating the genomic distance-dependence relationship with splines. Furthermore, there is a method [136] extracting significant chromatin interactions as calling peaks in a chromatin contact map within the surrounding two-dimensional region. Hi-C data are also used to construct three-dimensional models of chromatin structure. Some methods [168][39][7][175][199][12][9][96] try to learn a consensus chromatin structure of a collection of cells. Some methods [138][49][67][179][130][171] are intended to learn a set of chromatin structures representative of the observed chromatin interaction data. Besides the above downstream analysis tasks, there are some computational methods to carry out differential analysis on Hi-C data [111][107] and multiple two-dimensional visualization tools exist [202][129][40]. For a comprehensive list of computational tools on Hi-C data, please check out the Omictools website [61] on high-throughput chromosome conformation capture data analysis software tools.

There are substantial computational methods and tools for downstream analysis of Hi-C data, however, most of them focus on intrachromosomal interactions and little attention is paid to interchromosomal interactions. Partly because there is no prior knowledge such as the strong genomic distance-dependence relationship between frequency of intrachromosomal interactions and the genomic distance. In addition, the frequency of the interchromosomal interactions is much lower than intrachromosomal interactions while their search space is much larger (bin pairs across chromosomes VS bin pairs within chromosomes). To the best of our knowledge, there are few

computational studies that are dedicated to bulk Hi-C interchromosomal interactions. One study presents an investigation on human and mouse interchromosomal contacts and provides insights into mammalian chromatin organization [36]. A recent work develops a computational method based on an autoencoder and a multilayer perceptron classifier to impute high-resolution interchromosomal interactions [187]. Another paper presents two computational methods to estimate the transcription factors enriched in the interchromosomal interactions in yeast [27].

With the development of single-cell technologies, some single-cell Hi-C (scHi-C) approaches [121][45][135] are invented and therefore we can examine chromatin interactions at single-cell resolution. They also bring a new type of frequency information, the number of single cells with chromatin interactions between two disjoint chromosome regions. Generally these chromosome regions are defined by dividing chromosomes into equal-sized bins according to a resolution specified by users. Considering the lack of computational methods on interchromosomal interactions and the obvious pattern of intrachromosomal interactions along the diagonal of a chromatin contact map, we propose a computational method dedicated to analyzing interchromosomal interactions of single-cell Hi-C with this new frequency information. The fundamental difference between our research and previous research on interchromosomal interactions is our research is based on the new frequency information observed from each cell among all cells profiled. Since a bulk Hi-C experiment pools cells together at the very beginning so it can't discern whether a chromosomal interaction is shared by single cells or not. Therefore, computational methods on bulk Hi-C experiments don't consider the new frequency information at single-cell level, which is not available in bulk Hi-C experiments. In addition, when dealing with frequent interchromosomal interactions our method takes multiple contact maps as its inputs while computational methods on bulk Hi-C take one contact map as their inputs. What is more, to the best of our knowledge there is no tool available for frequent interchromosomal interactions. Specifically, we develop a computational tool to identify regions with statistically frequent interchromosomal interactions and make it accessible to the public. We believe that the regions associated with statistically frequent interchromosomal interactions under the single-cell context may be helpful for new hypotheses and functionally important therefore deserve more attention. Finally, frequent pattern mining is a longstanding topic in data mining research [55].

Our contributions may be stated as follows:

- We propose a computational method to identify regions associated with statistically frequent interchromosomal interactions at single-cell resolution.
- To the best of our knowledge, we are the first to implement a tool to serve the purpose and make it open to the public. To accommodate different scHi-C experiments, the tool is flexible on configurations.
- We demonstrate that using our proposed tool on two real scHi-C data sets, it can identify interesting structural regions.

The rest of chapter is organized as follows. The "Method" delineates our proposed method in detail. The "Data" introduces two scHi-C data sets as our inputs. The "Results and discussion" demonstrates that our proposed tool's usability on identifying interesting regions and flexibility of configurations. The "Conclusion" sections concludes that the tool will be useful for analyzing scHi-C interchromosomal interactions.

2.3. Method

In Fig. 2.1, the workflow of our proposed tool is illustrated and it includes three steps, network construction, statistical measurement calculation and region selection. The inputs of our tool are chromatin interactions of single cells, which are represented in heatmaps and can be easily generated with scHi-C processing pipelines such as NueProcess [161]. The outputs of our tool are identified regions, whose interchromosomal interactions are statistically frequent, along with frequencies and p-values. They are provided to help users refine identified regions with some frequency or p-value cutoff.

First, we construct a network by using interchromosomal interactions for each cell respectively. Due to low read coverages of scHi-C experiments and the more complex chromosomal structures of larger mammalian genomes, i.e. homo sapiens and mus musculus, chromosomes are divided into equal-sized bins to accumulate sufficient signals. Each bin is represented as a node with an index, and if there is an interchromosomal interaction whose two ends fall within two bins then the corresponding two nodes are connected with an edge. Instead of counting the number of interchromosomal interactions between bins, we are more concerned about their presence or absence because of the scarcity and variability of interchromosomal interactions in single cells. Therefore, an unweighted network is constructed for each cell.

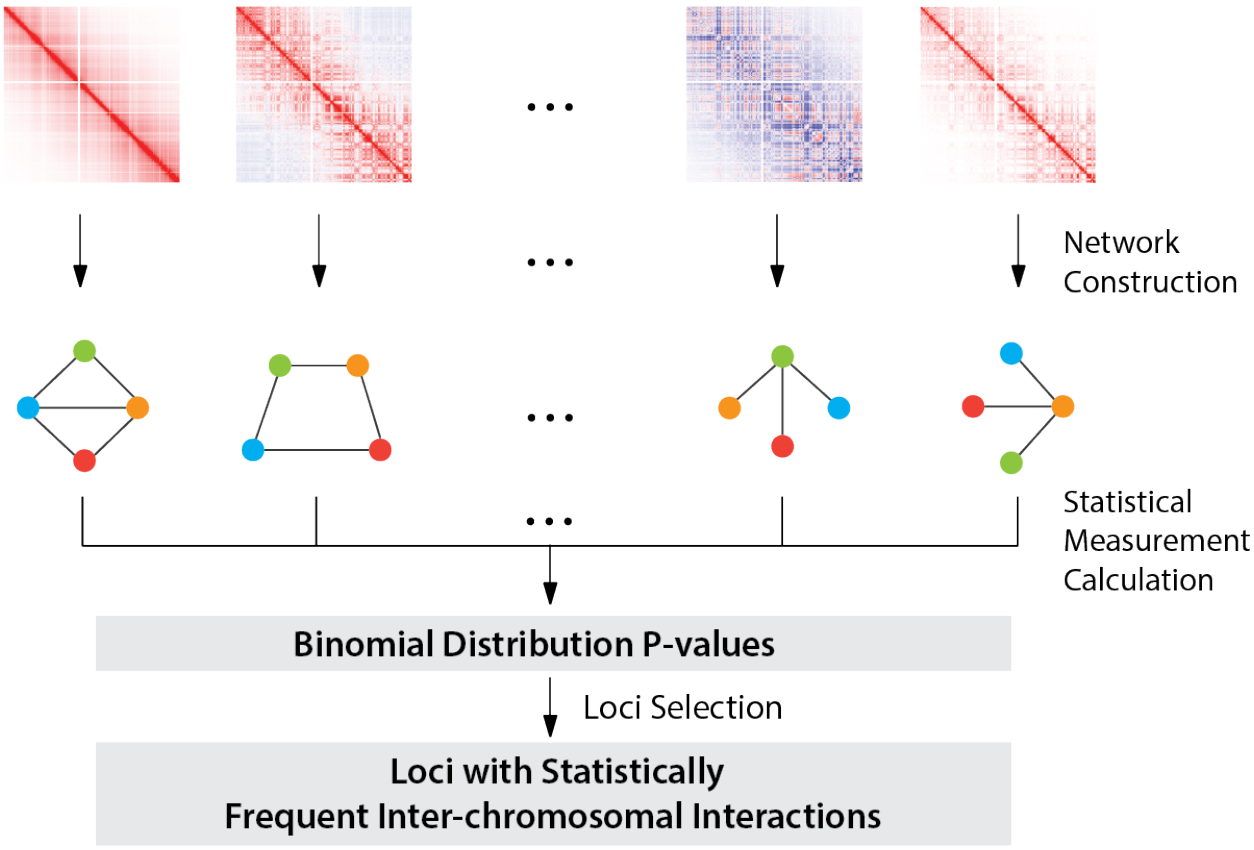

Figure 2.1. Workflow of the proposed method based on networks and statistical tests.

Second, we develop a measurement to quantify how statistically frequent for an edge to be detected among single cells. To avoid an overestimation of this measurement and therefore reduce false positives, we first remove nodes without any intrachromosomal and interchromosomal interactions among all cells to narrow down the search space of edges, which originally is all node pairs of different chromosomes. Assume the number of edges in the edge search space is M, the number of single cells is N, and the number of interchromosomal interactions for cell i is represented as n_i. Then $\frac{n_i}{M}$ represents the probability for cell i to have an edge between two nodes of different chromosomes. If a given edge is observed in t cells, we can use the following equations(2.1, 2.2, 2.3) to calculate its p-value.

$$p - value = \sum_{i=t}^{N} \binom{N}{i} p^i (1-p)^{N-i} \tag{2.1}$$

$$p = func(\frac{n_1}{M}, \frac{n_2}{M}, ..., \frac{n_{N-1}}{M}, \frac{n_N}{M}) \tag{2.2}$$

$$func \in \{max, mean, min\} \tag{2.3}$$

Similar to previous research [39][88][84], in Eq. 2.1 the binomial distribution is applied to estimate the p-value that reflects how likely it is for an edge to be observed in at least a given number of cells among all single cells. The rationality behind the selection of the binomial distribution is assuming whether there is an edge between two nodes of different chromosomes is a Bernoulli trial, the binomial distribution can capture edges that appear so frequent in multiple single cells that they reach statistical significance among all single cells. These frequent edges can only be detected in scHi-C experiments instead of bulk Hi-C experiments because subtle single-cell level information is pooled in bulk Hi-C experiments. Equation 2.2 is used to quantify the probability of an edge with all cells considered, which is determined by a function in Eq. 2.3. Users can configure the selection of these functions through a parameter. For scHi-C experiments with larger genomes or low sequencing depths, it is recommended to use max to select regions with highly statistically frequent interchromosomal interactions; therefore fewer regions would be selected. To the contrary, min is applied to select more regions. For scHi-C experiments with smaller genomes or high sequencing depths, min increases the odds for some regions to be selected while max may find nothing. mean is a balance between max and min, so the number of identified regions falls between them.

At last, p-values are adjusted by the Bonferroni correction and a user provided p-value cutoff, e.g. 0.05, is applied to select regions associated with statistically frequent interchromosomal interactions.

2.4. Data

To demonstrate that our proposed tool can be used to identify interesting structural regions, we use data from two existing scHi-C studies as our input data sets.

The first study [122] investigated the cell-cycle dynamics of chromosomal organization at single-cell resolution. The authors processed single F1 hybrid 129 $\times$ Castaneus mouse embryonic stem cells (mESCs) grown in 2i media using 1.5 million reads per cell on average. They analyzed 1,171 cells with fluorescence-activiated cell sorting, which labeled these cells to different cell-cycle phases based on levels of the DNA replication marker geminin and DNA content. Among them, 280 cells with a prefix of 1CDX1 were labeled as G1 phase; 303 cells with a prefix of 1CDX2 were labeled as Early-S phase; 262 cells with a prefix of 1CDX3 were labeled as Mid-S phase; 326 cells with a prefix of 1CDX4 were labeled as Late-S phase. We treat cells of different cell-cycle phases separately and feed them as inputs of our tool respectively. Therefore we identify regions with statistically frequent interchromosomal interactions for different cell-cycle phases.

The second one [45] developed a single-nucleus Hi-C protocol which provides >10-fold more contacts per cell than the previous method [121] to investigate chromatin organization at oocyte-to-zygote transition in mice. There are 40 transcriptionally active oocytes labeled as non-surrounded nucleolus (NSN), 76 transcriptionally inactive oocytes labeled as surrounded nucleolus (SN), 30 maternal nuclei from zygotes and 24 paternal nuclei from zygotes. Maternal and paternal nuclei are extracted from predominantly G1 phase zygotes.

2.5. Results and Discussion

Both data sets have single cells/nuclei of four conditions, therefore we run the proposed tool on single cells/nuclei of each condition respectively. Since the genomes used in the two experiments are large and sequencing read coverages are low, to accumulate sufficient interchromosomal interactions in a bin, we set the bin size to 500 kilobases (kb), which is also used in other existing studies [84][106]. We first show that our tool can identify regions with statistically frequent interchromosomal interactions, then demonstrate that our tool is flexible to different configurations, which support sliding windows for region diversity, different functions to estimate the probability

of having an edge between two nodes thereby providing adaptability of identified regions, and a configuration of different bin sizes e.g. 500kb VS 1 megabases (Mb).

2.5.1. Usability of Identifying Interesting Regions

To demonstrate the usability of our proposed method, we first display identified regions in visualization, then compare the identified regions and at last carry out enrichment analysis with other genomics features such as CTCF binding sites and enhancers etc.

2.5.1.1. Identification of Statistically Frequent Regions

In Fig. 2.2, identified regions associated with statistically frequent interchromosomal interactions among single cells of the cell-cycle data set are visualized in Circos [89]. The max function is configured for our method. The banded ideograms are mouse chromosomes (1-19, X and Y) and the black lines between them are interchromosomal interactions and the ends of these lines correspond to identified regions in chromosomes. Figure 2.2a shows the results of single cells of G1 phase; Fig. 2.2b shows the results of single cells of Early-S phase; Fig. 2.2c shows the results of single cells of Mid-S phase; and Fig. 2.2d shows the results of single cells of Late-S phase.

Among all four Circos plots, there is an apparent common hub in chromosome 11 (between 3Mb and 3.5Mb) whose interchromosomal interactions are highly enriched. The finding of this hub is corroborated by previous research with bulk Hi-C experiments to study interchromosomal contact networks in mammalian genomes [84]. They also discovered this hub in the mouse genome. Our finding confirms the hub's existence at single-cell level and rules out the possibility that its existence is solely contributed by very few cells with a large amount of interchromosomal interactions in the region. In addition, these four Circos plots are similar but not exactly the same, which means single cells of different cell phases share some interchromosomal interactions but also have some variabilities on interchromosomal interactions.

In Fig. 2.3, identified regions associated with statistically frequent interchromosomal interactions among single cells/nuclei of the oocyte-to-zygote data set are visualized. Figure 2.3a shows the results of single oocytes labeled as NSN; Fig. 2.3b shows the results of single oocytes labeled as SN; Fig. 2.3c shows the results of single maternal nuclei from zygotes; and Fig. 2.3d shows the results of single paternal nuclei from zygotes. Our tool reports much fewer regions on this data set and there is no hub. The absence of the hub may be partly because of cell discrepancies on cell types and cell cycles. To be more specific, in the second research, oocytes and maternal/paternal nuclei

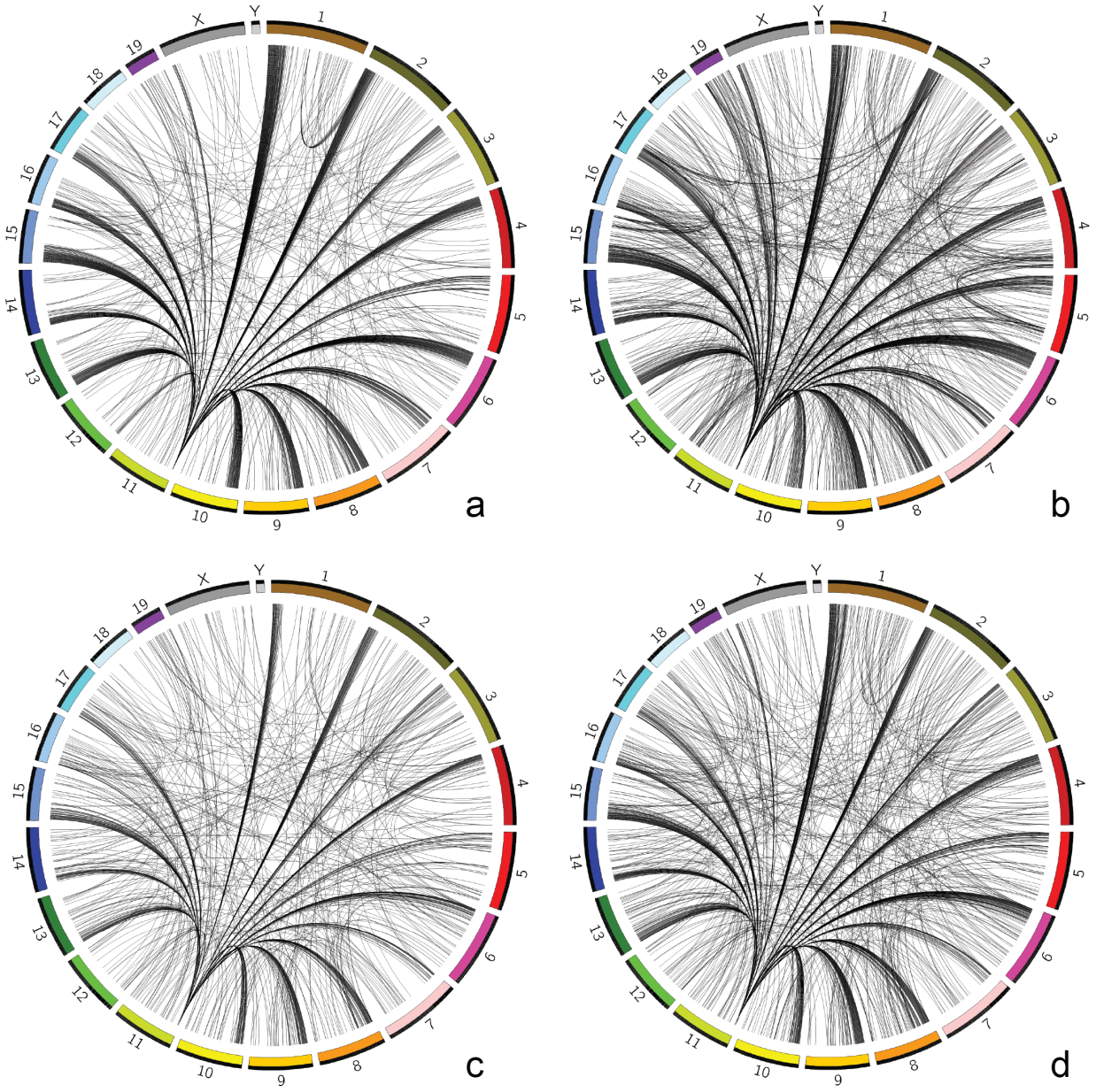

Figure 2.2. Identified regions of the cell-cycle data set. Visualizing genome-wide identified regions and their interchromosomal interactions of the cell-cycle data set with an adjusted p-value cutoff of 0.05 in Circos plots. a single cells of G1 phase; b single cells of Early-S phase; c single cells of Mid-S phase; d single cells of Late-S phase

Table 2.1. Pairwise comparisons of the cell-cycle data set

Comparison	Common	Unique in former	Unique in latter
G1 VS Early-S	757	219	569
G1 VS Mid-S	526	450	198
G1 VS Late-S	708	268	335
Early-S VS Mid-S	595	731	129
Early-S VS Late-S	767	559	276
Mid-S VS Late-S	597	127	446

from zygotes only contain a single set of chromosomes. However, for the chromosome 11 from 3Mb to 3.5Mb, there are comparatively more interchromosomal interactions among all four Circos plots. Additionally, a similar interchromosomal interaction pattern is observed: there are some shared interchromosomal interactions but there are also some variabilities at single-cell resolution.

2.5.1.2. Pairwise Comparisons of Identified Regions

For the cell-cycle data set, we compare the identified regions from single cells of different phases and examine the similarity and dissimilarity. In Table 2.1, single cells of different phases share a lot of common regions. There are some unique regions in each phased single cells. All pairs have more common regions than unique regions except the comparison between Early-S and Mid-S. Because the number of common regions is limited by the identified regions from single cells at Mid-S phase and single cells at Early-S phase report the most identified regions.

We also compare the identified regions from single cells of the oocyte-to-zygote data set. In Table 2.2, single cells of different conditions share some regions and there are more unique regions than common regions. This phenomenon seems inconsistent with what we have observed in the cell-cycle data set. But it does make sense and reflects the different types of single cells/nuclei used in their experiments. When identified regions from oocytes labeled NSN are compared with the ones from other cells/nuclei, the oocytes labeled SN share the most common regions because both of them are the same type of cells and their common regions are limited by the identified regions from oocytes labeled NSN; single maternal nuclei share more regions than single paternal nuclei because oocytes and single maternal nuclei are both from females while single paternal nuclei are from males. The same reason can also be applied to explain why oocytes labeled SN share more common regions with single maternal nuclei than single paternal nuclei. At last, single maternal

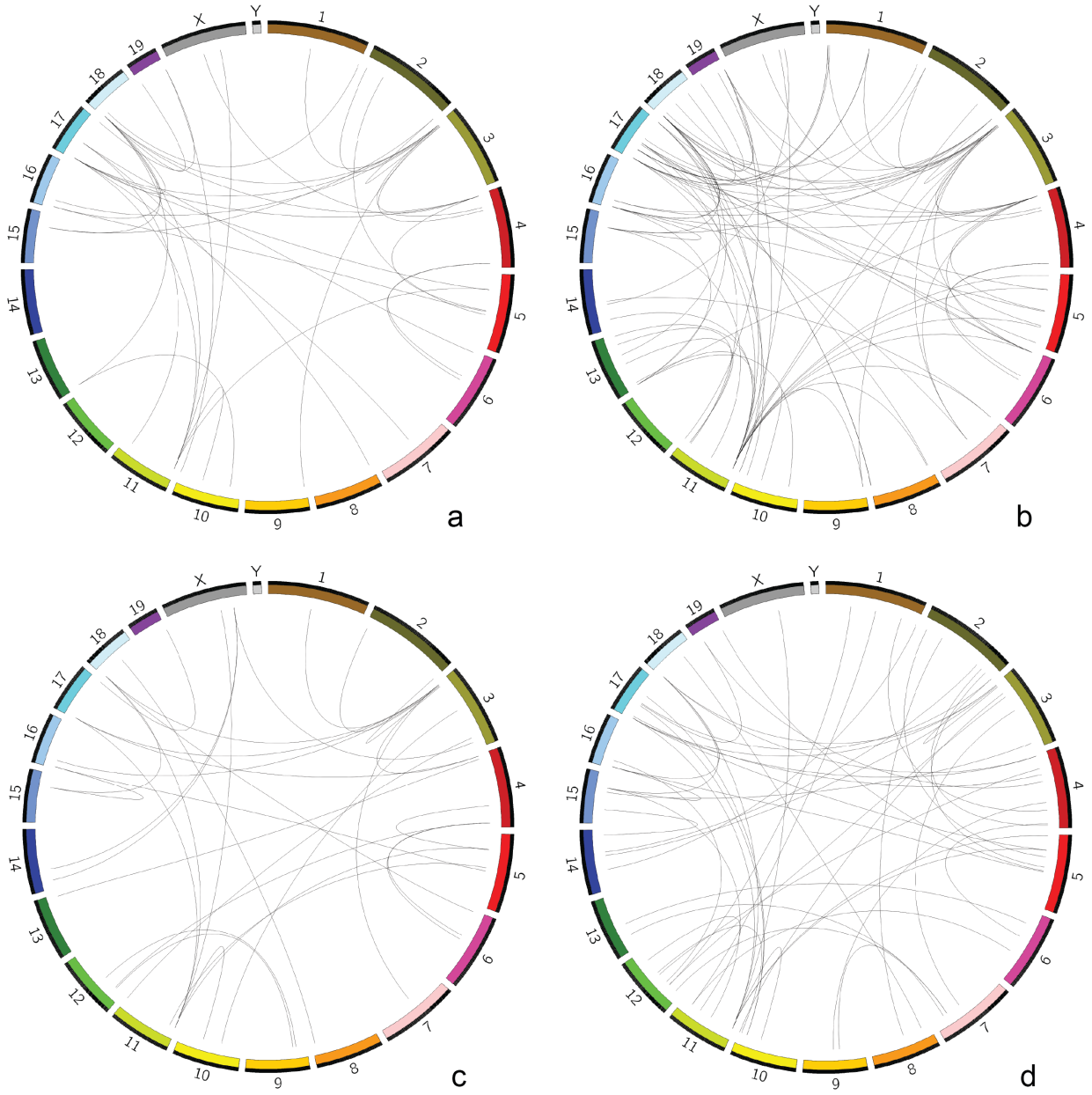

Figure 2.3. Identified regions of the oocyte-to-zygote data set. Visualizing genome-wide identified regions and their interchromosomal interactions of the oocyte-to-zygote data set with an adjusted p-value cutoff of 0.05 in Circos plots. a single oocytes labeled as NSN; b single oocytes labeled as SN; c maternal nuclei from zygotes; d paternal nuclei from zygotes

Table 2.2. Pairwise comparisons of the oocyte-to-zygote data set

Comparison	Common	Unique in former	Unique in latter
NSN VS SN	**35**	2	49
NSN VS maternal	18	19	15
NSN VS paternal	15	22	36
SN VS maternal	**21**	63	12
SN VS paternal	19	65	32
maternal VS paternal	**13**	20	38

nuclei and single paternal nuclei share the fewest common regions because some are from females and the others are from males.

2.5.1.3. Enrichment Analysis of Identified Regions

To improve the interpretation of identified regions, we carry out enrichment analysis of identified regions with genomic features, which are available in the cell-cycle data set. As there are too many identified regions in the data set, we select top ranked regions/nodes according to the numbers of statistically frequent unweighted edges with a cutoff (≥ 3 except ≥ 4 for single cells at Early-S phase because there are too many top regions). Therefore we obtain 16 regions for single cells at G1 phase, 37 regions for single cells at Early-S phase, 34 regions for single cells at Mid-S phase and 47 regions for single cells at Late-S phase. Genomic features of mESC cell line are downloaded from this paper [153] and they are CTCF binding sites, enhancer sites, H3K4me3 peaks, H3K27ac peaks and Pol II peaks.

For the above selected regions of each phase, the numbers of genomic features are counted respectively. Then we ranomly select the same number of regions and count the numbers of genomic features falling into these randomly selected regions respectively. We carry out this randomization strategy 50,000 times and therefore we obtain empirical background samples for each genomic feature. We calculate the z-score for each genomic feature. In Table 2.3, most of genomic features are enriched (≥ 1.97, which corresponds to 0.05 in p-value) except enhancer. What is more important, for single cells at Early-S phase, all the genomic features are highly enriched. (When ≥ 3 is used as the cutoff, the results become more enriched.) H3K4me3 and H3K27ac are active gene transcirption marks. Pol II plays very important roles in gene transcription. An enhancer increases the likelihood of gene transcription. CTCF plays important rols in chromatin structure and insolates the spread

Table 2.3. Identified Regions' Enrichment Analysis of the cell-cycle data set

Input	CTCF	enhancer	H3K4me3	H3K27ac	Pol II
G1	2.82	1.05	1.75	2.63	2.48
Early-S	**10.86**	**9.81**	**12.48**	**12.05**	**12.58**
Mid-S	2.81	1.48	3.08	2.74	3.64
Late-S	3.37	1.74	4.33	4.36	5.05

of heterochromatin. Early-S phase corresponds to the commencement of DNA replication. These genomic features seems working coordinately to facilitate the initialization of DNA replication.

2.5.2. Flexibility of Configurations

To make our tool flexible to accommodate different scHi-C experiments, we support different configurations, which include sliding windows for region diversity, edge probability functions for adjustability of identified regions and different bin sizes.

2.5.2.1. Configuration of Sliding Windows

By default, our tool divides chromosomes into bins from the first bases of chromosomes to the last ones, which limits the starting and ending positions of regions. To overcome this limitation, our tool supports a sliding window strategy by moving bins toward the last bases certain bases (e.g. 100kb). It lets users decide where their regions' starting and ending positions through a parameter. In Table 4, we adopt four sliding windows of sizes of 100kb, 200kb, 300kb and 400kb and compare the identified regions with the ones by default (no sliding window). If identified regions mediated by some interchromosomal interactions from the no sliding window condition overlap with identified regions from a sliding window condition at both ends, we treat these regions as common identified regions; otherwise they are different. Therefore, we can calculate the common identified regions between no sliding window and sliding windows. In Table 2.4, we conclude that most identifed regions between no sliding window and sliding windows are common because some shared interchromosomal interactions fall into these regions. But as these common regions' starting and ending positions are different, our tool diversifies the identified regions to users. What is more interesting is the single cells at Early-S phase share the fewest identified regions between no sliding window and sliding windows of different sizes. As DNA synthesis commences at Early-S phase, interchromosomal interactions may vary or involve in DNA synthesis initialization activites more at

Table 2.4. Overlapping identified regions of the cell-cycle data set with no sliding window and sliding windows of different sizes

Input Data	100kb	200kb	300kb	400kb
G1	92.11%	92.01%	92.01%	95.49%
Early-S	85.52%	86.05%	86.73%	89.22%
Mid-S	90.33%	89.92%	91.16%	93.65%
Late-S	93.19%	91.08%	91.66%	94.44%

Table 2.5. Overlapping identified regions of the oocyte-to-zygote data set with no sliding window and sliding windows of different sizes

Input Data	100kb	200kb	300kb	400kb
oocyte NSN	100%	92.01%	92.01%	95.49%
oocyte SN	86.90%	89.29%	89.29%	92.86%
pronucleus maternal	93.94%	93.94%	90.91%	100%
pronucleus paternal	94.12%	90.20%	90.20%	92.16%

this phase than other phases. In Table 2.5 of the oocyte-to-zygote data set, we can reach the same conclusion that most identified regions are common between no sliding window and sliding windows of different sizes and meanwhile there are some different regions.

2.5.2.2. Configuration of Edge Probability Functions

Our proposed tool supports three functions, max, mean and min, to estimate the probability of an edge between two nodes of different chromosomes, therefore improving adjustability of identified regions. In Table 2.6 of the cell-cycle data set and Table 2.7 of the oocyte-to-zygote data set, our tool configured with the max function identifies the fewest regions; our tool configured with the min function identifies the most regions and our tool configured with the mean funciton falls between them. This is because if we fix other variables except p in Eq. 2.1, a large p entails a large *p-value* and a small p entails a small *p-value*. As we have explained in the second to last paragrpah of Method, users can select these functions according to the sizes of genomes and sequencing depths used in their experiments. Therefore, our proposed tool provides adaptability of identified regions.

2.5.2.3. Configuration of Bin Sizes

Finally, our tool also supports different bin sizes. As scHi-C experiments have low read coverages and scarce interchromosomal interactions, we need to use large bin sizes to accumulate

Table 2.6. Number of identified regions of the cell-cycle data set with edge probability functions

Input Data	max	mean	min
G1	976	1651	2133
Early-S	1326	2579	7714
Mid-S	724	1833	2991
Late-S	1043	1999	6058

Table 2.7. Number of identified regions of the oocyte-to-zygote data set with edge probability functions

Input Data	max	mean	min
oocyte NSN	37	79	199
occyte SN	84	229	1846
pronucleus maternal	33	50	268
pronucleus paternal	51	51	274

sufficient interchromosomal interactions in a bin. We run our tool with bin_size=1Mb on the two data sets and compare the identified regions with the ones of bin_size=500kb. We find that the identified regions of bin_size=500kb and bin_size=1Mb are quite similar for most single cells except the Early-S phased single cells in the cell-cycle data set. In Fig. 2.4b of bin_size=1Mb, the hub of the chromosome 11 at 3Mb becomes less obvious as it is overshadowed by enrichment of other interchromosomal interactions because of the increased bin size and single cells of this particular cell phase. Therefore, different bin sizes may affect the identified regions.

2.6. Conclusion

In this paper, we introduce a computational method to identify regions associated with statistically frequent interchromosomal interactions at single-cell resolution and implement it as an open source tool, which is the first serving the purpose to the best of our knowledge. Its workflow includes network construction, binomial statistical measurement calculation and region selection. We demonstrate its usability on two existing scHi-C data. On the cell-cycle data set, the tool discovers a hub in the mouse chromosome 11 from 3Mb to 3.5Mb, which is endorsed by a previous study on interchromosomal contact networks with bulk Hi-C experiments. On the oocyte-to-zygote data set, there is no apparent hub at the region, but comparatively interchromosomal interactions are enriched. Identified regions' pairwise comparisons show that our method identifies common

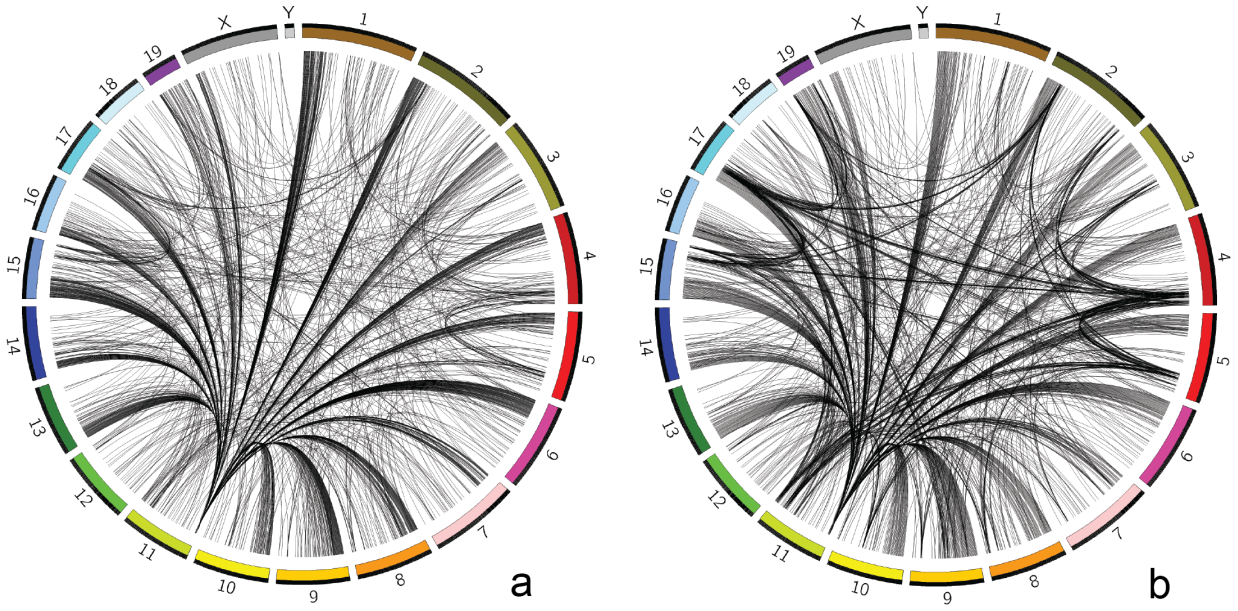

Figure 2.4. Comparing identified regions of Early-S phased single cells with different bin sizes. a bin_size=500kb b bin_size=1Mb

regions between different data sets and also reflects the true dissimilarity such as different cell types. Identified regions' enrichment analysis helps improve the interpretation of top ranked identified regions and these genomic features are highly enriched for single cells at Early-S phase, which implies our top ranked regions may be functionally important. We also exhibit our proposed tool's flexibility on configurations, which support sliding windows for diverse regions, edge probability functions for adjustable regions and different bin sizes. Overall, it will be a useful tool for analyzing scHi-C interchromosomal interactions.

Due to low sequencing depths of scHi-C experiments and the paucity of interchromosomal interactions, identifying high resolution regions of several kilobases (e.g. 8kb) is extremely difficult. Our tool can run with this resolution but due to the limitation of scHi-C data, it can't identify any regions passing the statistical tests. We will try to mitigate this problem by imputing high-resolution interchromosomal interactions with data of other experiments such as interchromosomal interactions from bulk Hi-C experiments. In addition, further research is needed to improve the signal-to-noise ratio for scHi-C experiments.

2.7. Availability of Data and Materials

For the implementaion details of our tool, please check out it at GitHub. Currently it supports the following genomes, mm9, mm10, hg18 and hg19. It can be easily extended to other organisms.

3. A HEURISTIC STRATEGY FOR MULTI-MAPPING READS TO ENHANCE HI-C DATA

3.1. Abstract

Current Hi-C analysis approaches focus on uniquely mapped reads and little research has been carried out to include multi-mapping reads, which leads to a lack of biological signals from DNA repetitive regions. We propose a heuristic strategy to assign multi-mapping reads to loci according to the distance to their closest restriction enzyme cutting sites. We demonstrate that the heuristic strategy can rescue multi-mapping reads thus enhance the quality of Hi-C data. Compared with mHi-C, it not only improves replicate reproducibility in the same cell type, but also maintains the difference between replicates of different cell types. Moreover, the strategy identifies much more common statistically significant chromatin interactions between Hi-C experiments of different restriction enzymes, improves performance on chromatin state annotation analysis, especially on two repetitive annotations, and has a huge advantage on computing resources. Therefore, the heuristic strategy can be used to enhance Hi-C data by utilizing multi-mapping reads.

3.2. Introduction

Three-dimensional genome organization plays important roles in many biological processes, which include long-range gene regulation [33], DNA replication and repair [114, 42]. The alteration of three-dimensional genome architecture leads to human diseases, such as cancer [48, 3]. As the development of chromosome conformation capture-based technologies, high-throughput chromosome conformation capture (Hi-C) [104] emerges as a popular method to detect genome-wide chromatin interactions. In Hi-C experiments, crosslinked DNA is fragmented with restriction enzymes. Then DNA fragments are ligated, selected, sheared and finally sequenced as paired-end reads. After these paired-end reads are processed by Hi-C analysis pipelines, chromatin contact maps are generated for downstream analysis and exploration. Recent studies have discovered some multi-scale spatial genomic structures, which include A/B compartment [104], topologically associating domains (TADs) [36], chromatin loops [136] and frequently interacting regions (FIREs) [147].

Owing to the sequencing cost, few studies generate high-resolution data sets. To enable high-resolution structure discovery on low-resolution data sets, some computational methods are proposed to enhance Hi-C data with machine learning algorithms. HiCPlus [198] and HiCNN [109] both use deep convolutional neural networks. HicGAN [108] and DeepHiC [65] infers high-resolution Hi-C data with generative adversarial networks. However, all of these methods depend on one high-resolution data set as their training sets and ignore heterogeneity among cell types.

Though machine learning algorithms are popular, they are not the only method to enhance Hi-C data. In fact, for each Hi-C data set, a large number of reads are discarded at the very beginning. Because most Hi-C pipelines only consider uniquely mapped reads (unique reads) and ignore multi-mapping reads, which are mapped to multiple genomic loci. To the best of our knowledge, there is only one study, mHiC [200], accounting for multi-mapping reads. mHi-C assigns multi-mapping reads according to the interacting patterns learned from unique reads, therefore the multi-mapping read assignment depends on unique reads. Here we propose a heuristic strategy which doesn't depend on unique reads to utilize multi-mapping reads. The heuristic strategy not only enhances Hi-C data, but also enables exploration of new interacting patterns.

Our contributions may be stated as follows:

- We propose a heuristic strategy to utilize multi-mapping reads for Hi-C data processing.
- We demonstrate that using our proposed strategy on Hi-C data sets can enhance Hi-C data in quantity and reproducibility and recover more common statistically significant chromatin interactions between experiments of different restriction enzymes.
- Through chromatin state annotation analysis, we show that our proposed strategy can recover more signals at DNA repetitive regions.

The rest of paper is organized as follows. The second section delineates the heuristic strategy to use multi-mapping reads. The third section introduces two human cell lines and two Arabidopsis data sets as our test data. The fourth section evaluates the heuristic strategy by comparing it with mHi-C and a method that only considers unique reads. The last one concludes that the heuristic strategy complements multi-mapping reads in Hi-C analysis.

3.3. Method

We propose a heuristic strategy to utilize multi-mapping reads in Hi-C experiments to strengthen chromatin interaction data. As shown in Figure 3.1A, for Hi-C read ends, there are three possible outcomes, unaligned, unique and multi-mapping reads. Compared with unaligned reads, multi-mapping reads are reads with high quality alignment scores, but their alignment loci cannot be uniquely determined. To avoid the abuse of utilizing multi-mapping reads, we only rescue multi-mapping reads with less than a specific number of alignments. For example, mHi-C by default utilizes multi-mapping reads with less than 100 alignments. In order to assign a multi-mapping read to a unique locus among its alignments, we hypothesize that the locus closer to restriction enzyme cutting sites has a higher probability to be the origin as shown in Figure 3.1B. The hypothesis is based on the Hi-C processing of unique reads. In Hi-C processing pipelines, the closest restriction enzyme cutting sites are picked to filter unique reads. Second according to our empirical experience, an object's breakage because of outside forces is most likely to happen at the object's periphery with defects. In Hi-C experiments at the shearing step, shearing may happen preferentially close to the restriction enzyme cutting sites, which can be viewed as defects as these sites are cut by restriction enzymes before. Therefore, we select the loci for multi-mapping reads according to the distance to the closest restriction enzyme cutting sites. What is more important, as our multi-mapping read assignment is carried out at the sequence alignment step, there is no impact on following Hi-C data processing and the same filtering criteria (such as distance to restriction enzyme cutting sites) can be applied to unique and multi-mapping reads to remove invalid chromatin interactions.

3.4. Data

To demonstrate that the heuristic strategy can rescue multi-mapping reads in Hi-C experiments, thus increasing detected chromatin interactions and expanding the breadth of genome coverage, we test the strategy on Hi-C experiments of two cell lines from a study [36] on revealing topological domains in mammalian genomes and Hi-C experiments of Arabidopsis thaliana seedling tissues from two studies [194, 139] with different restriction enzymes. The first cell line is human embryonic stem cell (hESC) and the second cell line is derived from human fetal lung (IMR90). For each cell line, Hi-C experiments were conducted independently with two biological replicates (r1 and r2) using HindIII as the restriction enzyme to cut crosslinked DNA into fragments. Thereafter,

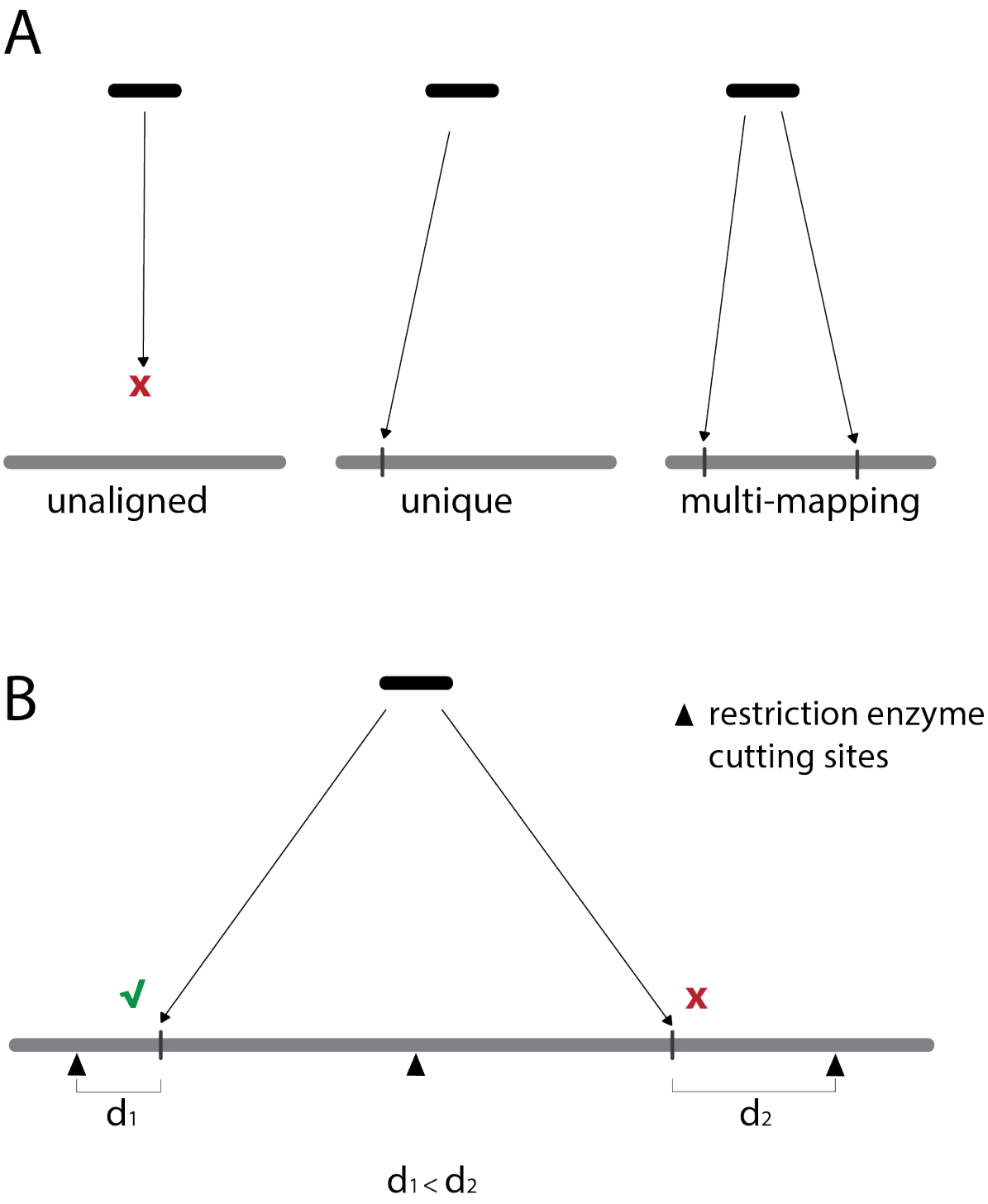

Figure 3.1. Hi-C read alignment outcomes and the heuristic strategy for multi-mapping reads. A: three types of reads, unaligned, unique and multi-mapping reads, B: a multi-mapping read is assigned to a locus closest to restriction enzyme cutting sites.

DNA fragments in close proximity were ligated in a diluted environment and the resulting ligation products were sonicated, filtered and finally sequenced by paired-end sequencing. Therefore, two paired read files were generated for each replicate, e.g. hESC_r1_1 and hESC_r1_2. For Arabidopsis thaliana seedling tissues, the first study [194] carried out the Hi-C experiments using HindIII with two biological replicates (r1 and r2), which are named HindIII_r1 and HindIII_r2. The second study [139] carried out the Hi-C experiments using DpnII with three biological replicates (r1, r2 and r3), which are named DpnII_r1, DpnII_r2 and DpnII_r3.

3.5. Results

3.5.1. Sequence Alignment Statistics Necessitate Utilizing Multi-Mapping Reads

We adopt Hi-C processing pipelines consisting of a sequence of processing functions or commands, for example, Hiclib [71], to process paired reads of hESC and IMR90's replicates. Because it is convenient to incorporate the heuristic strategy into these pipelines and understanding the inner complex logic of a holistic tool is not this study's research focus. As Hi-C processing pipelines ignore multi-mapping reads at the sequence alignment step, we need to carry out our own sequence alignment to keep multi-mapping reads. A sequence alignment tool, for example, Bowtie 1 [92], is applied to align two ends of Hi-C reads independently with its default settings and the statistics of sequence alignment for each replicate are listed in Table 3.1. For each replicate, multi-mapping reads are more than unaligned reads at both ends. This means there are more multi-mapping reads than unaligned reads to be rescued. This phenomenon can be explained by the fact that these reads are short reads which are more likely to be aligned to multiple loci than nowhere. In addition, prevalent short-read sequencing in Hi-C experiments necessitates the need of utilizing multi-mapping reads to enhance chromatin interaction data.

Table 3.1. hESC and IMR90 paired-end sequence alignment statistics.Two ends of Hi-C paired-end reads are mapped independently because distance constraint of paired-end reads doesn't apply to Hi-C reads.

replicate	hESC_r1		hESC_r2		IMR90_r1		IMR90_r2	
#reads	237,662,270		496,522,946		397,194,480		259,123,992	
unique reads(%)	69.77	68.74	72.31	70.96	71.65	69.04	70.44	70.26
unaligned reads(%)	11.99	13.16	9.79	11.45	10.82	13.87	11.74	11.70
multi-mapping reads(%)	**18.24**	**18.10**	**17.9**	**17.59**	**17.53**	**17.09**	**17.82**	**18.04**

3.5.2. The Heuristic Strategy Increases Detected Chromatin Interactions

To demonstrate that the heuristic strategy can strengthen chromatin interaction data, we test the strategy on each replicate with hiclib and mHi-C respectively. hiclib only considers unique reads and incorporating our strategy takes both unique and multi-mapping reads into count. mHi-C leverages multi-mapping reads in a sequence of commands and it is convenient to replace its multi-mapping read assignment method with our strategy. The numbers of detected chromatin interactions for each replicate are shown in Table 3.2. Compared with unique reads, the heuristic strategy increases millions of chromatin interactions because it also accounts for multi-mapping reads. Compared with mHi-C, the heuristic strategy gains chromatin interactions marginally because they both leverage unique and multi-mapping reads.

Table 3.2. hESC and IMR90 chromatin interactions with hiclib and mHi-C under different configurations. hiclib+ represents incorporating hiclib with the heuristic strategy. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	hiclib	hiclib+	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	16,156,824	**21,528,337**	17,043,308	20,325,529	**20,819,070**
hESC_r2	117,150,577	**139,527,552**	105,617,771	124,622,391	**124,955,453**
IMR90_r1	81,524,268	**97,985,497**	83,161,703	97,444,579	**98,380,530**
IMR90_r2	89,322,274	**104,647,014**	83,381,123	96,099,798	**98,325,832**

3.5.3. The Heuristic Strategy Enhances the Reproducibility of Chromatin Interaction Data

Replicate reproducibility is an important measurement used to assess the quality of chromatin interaction data. We calculate the reproducibility scores among hESC and IMR90's replicates by chromosome (from chromosome 1 to chromosome 22) with HiCRep [189]. As shown in Figure 3.2, for each configuration [mHi-C (unique), mHi-C and mHi-C+], there are two types of replicate reproducibility scores. The first type (at the top) represents the average of replicate reproducibility scores in the same cell line (hESC_r1 VS hESC_r2 and IMR90_r1 VS IMR90_r2). The second type (at the bottom) represents the difference between the average of replicate reproducibility scores in the same cell line and the average of replicate reproducibility scores between different cell lines

(hESC_r1 VS IMR90_r1, hESC_r1 VS IMR90_r2, hESC_r2 VS IMR90_r1 and hESC_r2 VS IMR90_r2). For the first type of replicate reproducibility scores, mHi-C and mHi-C+ are better than mHi-C(unique). This means compared with the configuration only utilizing unique reads, configurations utilizing both unique and multi-mapping reads improve the reproducibility between replicates in the same cell line. In addition, mHiC's multi-mapping read assignment method (mHi-C) is slightly better than our strategy (mHi-C+) on improving the reproducibility between replicates in the same cell line. But for the second type of replicate reproducibility scores, our strategy performs better than mHi-C. Among the 22 chromosomes, our strategy has noticeably larger differences on 7 chromosomes, while mHi-C's multi-mapping read assignment method has 2 noticeably larger differences on 2 chromosomes. What is more important, our strategy achieves similar performance with the method only utilizing unique reads. Taking these two types of replicate reproducibility scores into consideration, we conclude that our strategy not only improves the replicate reproducibility in the same cell line, but also maintains the difference between different cell lines.

3.5.4. The Heuristic Strategy Improves Statistically Significant Chromatin Interactions

Enhanced chromatin interaction data enable downstream analysis and exploration for new discoveries. Therefore, we apply Fit-Hi-C [6] to normalized chromatin interactions to identify statistically significant chromatin interactions with respect to a false discovery rate of 0.05. In Table 3.3, both configurations utilizing unique and multi-mapping reads report more statistically significant chromatin interactions than the configuration utilizing only unique reads. In addition, mHi-C's multi-mapping read assignment method seems identifying more statistically significant chromatin interactions than our strategy. It can be explained if we further examine detected chromatin interactions and keep only unique chromatin interactions. As shown in Table 3.4, incorporating our strategy gains much more unique chromatin interactions because mHi-C assigns multi-mapping reads according to the interacting patterns in the unique reads. Therefore, interacting patterns in the unique reads would be enriched to be statistically significant. The heuristic strategy doesn't assign multi-mapping reads according to unique reads and consequently it can explore more interacting patterns. However, these dispersed interacting patterns may become less statistically significant.

To further investigate two approaches utilizing multi-mapping reads on identifying statistically significant chromatin interactions, we apply them on Hi-C experiments of Arabidopsis thaliana seedling tissues from two studies [194, 139] with different restriction enzymes, HindIII and DpnII.

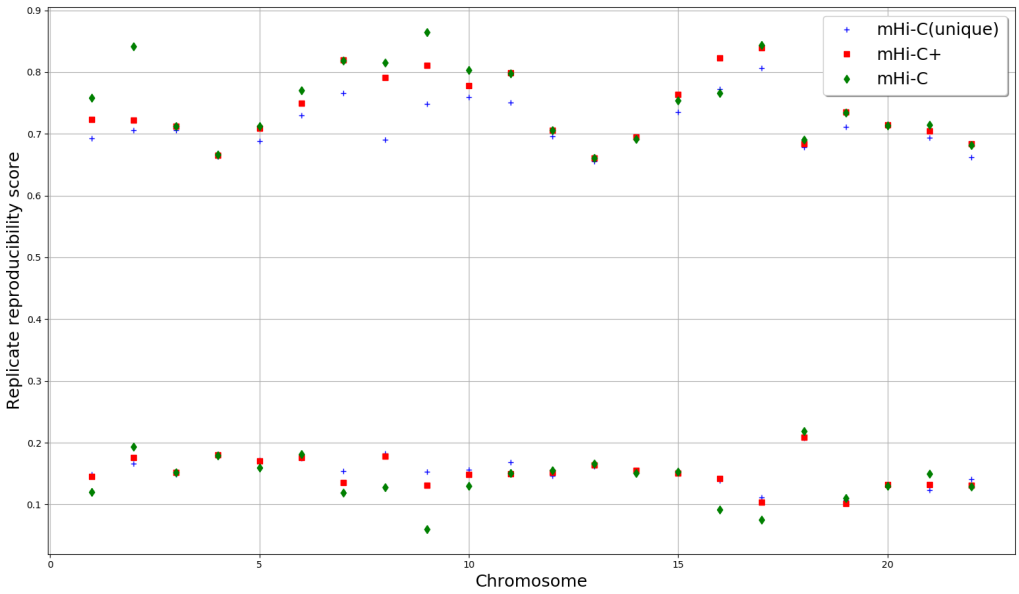

Figure 3.2. Replicate reproducibility scores for human chromosome 1-22. HiCRep is used to calculate reproducibility scores among hESC and IMR90's replicates. For each configuration [mHi-C(unique), mHi-C and mHi-C+], there are two types of replicate reproducibility scores. The first type (at the top) represents the average of replicate reproducibility scores in the same cell line. The second type (at the bottom) represents the difference between the average of replicate reproducibility scores in the same cell line and the average of replicate reproducibility scores between different cell lines.

Table 3.3. Statistically significant chromatin interactions identified by Fit-Hi-C. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	4,206	8,412	7,226
hESC_r2	34,630	54,642	53,236
IMR90_r1	49,500	78,574	69,476
IMR90_r2	55,124	85,160	74,396

Table 3.4. hESC and IMR90's unique chromatin interactions with mHi-C under different configurations. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	11,589,365	12,696,565	**14,656,936**
hESC_r2	48,065,862	51,792,951	**61,564,215**
IMR90_r1	54,974,139	58,975,514	**66,763,164**
IMR90_r2	63,548,605	67,705,423	**76,033,914**

Fit-Hi-C is used to identify statistically significant chromatin interactions with respect to a false discovery rate of 0.05 for each replicate respectively. Pairwise comparison is carried out between replicates of different restriction enzymes and the common statistically significant chromatin interactions are counted as shown in Table 3.5. Our strategy identifies much more common statistically significant chromatin interactions than mHi-C (>32%) because when assigning multi-mapping reads, our strategy does not depend on unique reads and therefore improving the identification of common statistically significant chromatin interactions.

Table 3.5. Common statistically significant chromatin interactions on Arabidopsis thaliana Hi-C experiment. HindIII and DpnII were used on Arabidopsis thaliana seedling tissues. Pairwise comparision between replicates of different restriction enzymes is carried out.

mHiC VS mHiC+	DpnII_r1	DpnII_r2	DpnII_r3
HindIII_r1	1561, **2064**	2079, **2838**	2067, **2877**
HindIII_r2	2020, **3250**	2817, **4083**	2757, **4084**

3.5.5. The Heuristic Strategy Improves Performance on Chromatin State Annotation Analysis

To further investigate the statistically significant chromatin interactions, we download 15 chromatin state annotations of hESC cell line at this website and study how these annotations overlap with statistically significant chromatin interactions. To make a fair comparison, we select the same number of statistically significant chromatin interactions. For each chromatin state annotation, we calculate the average of number of statically significant chromatin interactions overlapping with chromatin regions associated with the annotation.

In Table 3.6, mHi-C's multi-mapping read assignment method achieves similar performance on the first 13 chromatin state annotations with our strategy. But for the two repetitive annotations highlighted in red, our strategy outperforms mHi-C's multi-mapping read assignment method in very large margins. Multi-mapping reads are mostly located at repetitive genome regions because multi-mapping reads are reads that can be mapped to multiple loci. Both strategies utilize multi-mapping reads. Our strategy reports higher overlapping with two repetitive annotations, this means our strategy can recover more signals at repetitive genome regions, which helps exploring these uncharted regions.

Table 3.6. Chromatin state annotations overlapping with hESC statistically significant chromatin interactions. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heurstic strategy.

method	mHi-C	mHi-C+
1_Active_Promoter	1.00	1.00
2_Weak_Promoter	1.09	1.07
3_Poised_Promoter	0.46	0.49
4_Strong_Enhancer	0.35	0.36
5_Strong_Enhancer	0.51	0.49
6_Weak_Enhancer	0.87	0.85
7_Weak_Enhancer	0.46	0.46
8_Insulator	0.77	0.77
9_Txn_Transition	0.16	0.16
10_Txn_Elongation	0.49	0.48
11_Weak_Txn	0.45	0.44
12_Repressed	0.49	0.51
13_Heterochrom/lo	2.00	1.99
14_Repetitive/CNV	**1.08**	**1.19**
15_Repetitive/CNV	**1.52**	**1.68**

3.5.6. The Heuristic Strategy has a huge Advantage on Computing Resources

Computing resources are essential to bioinformatics research, especially for researchers and students with a limited budget. We compare the running time and memory usage on the same computing resource. As some commands (such as sequence alignment) in the pipeline are shared under different configurations, we only summarize the computing resources pertaining to the multi-mapping read assignment in Figure 3.3. mHi-C's multi-mapping read assignment method takes at least five-fold running time and ten-fold RAM than our strategy. When two configurations are applied to high resolution Hi-C data sets, the difference on computing resources becomes more glaring. Therefore, the heuristic strategy has a huge advantage on computing resources than mHi-C's multi-mapping read assignment method.

3.6. Conclusion

In this paper, we introduce a heuristic strategy to include multi-mapping reads into Hi-C analysis by assigning these reads according to the distance to their closest restriction enzyme cutting sites. Through the evaluation of Hi-C human data, we display that there are more multi-mapping reads than unaligned reads to be rescued. Compared with methods only considering unique reads, the strategy improves the quantity and reproducibility of Hi-C data, which enables new discoveries

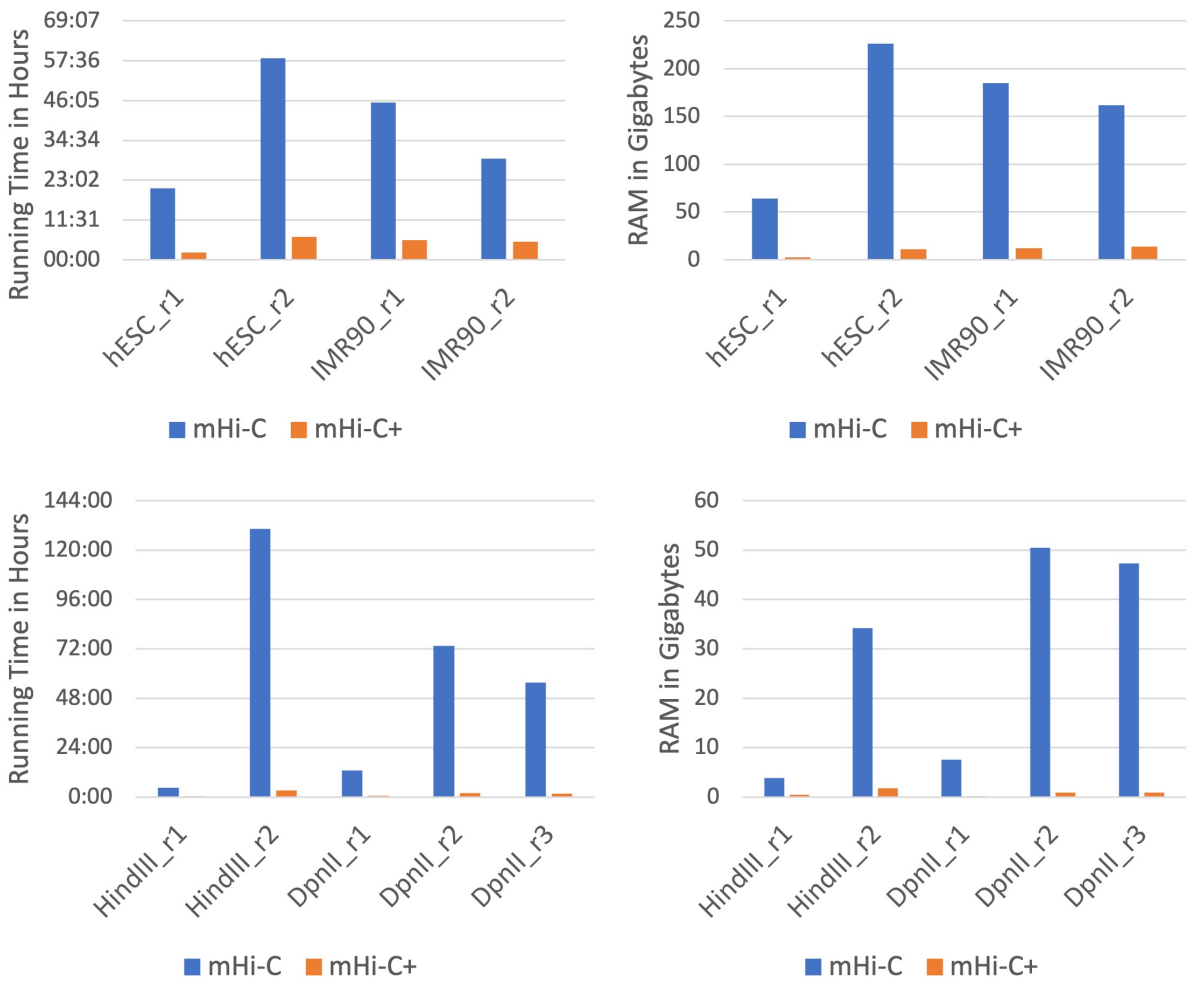

Figure 3.3. Comparison of computing resources (running time in hours and RAM in gigabytes) with mHi-C under different configurations.

of statistically significant chromatin interactions. Compared with mHi-C, the strategy maintains the difference between replicates of different cell lines, reports more common statistically significant chromatin interactions (>32%) between experiments with different restriction enzymes, improves performance on chromatin state annotation analysis, especially on two repetitive annotations and shows a huge advantage on computing resources (at least 5-fold in running time and 10-fold in RAM). Therefore, our strategy is an important complement to incorporating Hi-C multi-mapping reads.

Due to most Hi-C reads used in this paper are short reads (36 base pairs), we didn't rescue unaligned reads. For longer sequence reads, more efforts can be extended to study whether Hi-C data can be further enhanced by rescuing both unaligned reads with recursive mapping and multi-mapping reads with our proposed strategy. We also plan to combine our proposed strategy and machine learning algorithms to achieve high-resolution and high coverage Hi-C data.

4. SCHI-CNN: A COMPUTATIONAL METHOD FOR STATISTICALLY SIGNIFICANT SINGLE-CELL HI-C CHROMATIN INTERACTIONS WITH NEAREST NEIGHBORS

4.1. Abstract

The intricate interplay of regulatory elements, spatial arrangements, and transcription factors shapes the complex chromatin architecture within individual cells, offering valuable insights into cellular diversity and heterogeneity in the realm of chromatin biology. Nevertheless, the analysis of single-cell Hi-C data presents notable challenges due to its sparse nature and limited interaction counts. In this study, we introduce a novel algorithm, scHi-CNN, designed for the detection of statistically significant single-cell Hi-C chromatin interactions. Our method comprises three key steps: imputation of single-cell matrices, normalization, and identification of statistically significant interactions. To assess the robustness and scalability of scHi-CNN across various conditions, we evaluate its performance using three distinct datasets: human cortex cells, mouse embryonic stem cells, and a mouse cell cycle dataset. Moreover, we delve into the biological relevance of the derived significant interactions by examining CTCF binding sites, known promoter-related interactions, and the overlap between different datasets of the same cell type. The results underscore the ability of scHi-CNN to identify more biologically meaningful interactions from single-cell data, facilitating a deeper comprehension of regulatory elements and spatial arrangements within individual cells and across diverse cell types.

Code and sample data for this paper are available on the GitHub repository at `https://github.com/bignetworks2019/scHi-CNN`

4.2. Introduction

Single-cell chromatin interaction data plays a crucial role in unraveling the intricacies of three-dimensional chromatin structure, capturing cellular heterogeneity, and elucidating genomic variations across diverse cell types. Identifying significant interactions from raw interaction data is imperative for examining regulatory elements, spatial arrangements, transcription factor functions,

and other functional elements in individual cells. However, processing single-cell Hi-C data presents several challenges due to its inherent sparseness and limited interaction counts.

Despite the availability of single-cell chromatin interaction datasets to the public, the analysis of significant intra-chromosomal interactions within individual cells is still in its nascent stage. Existing tools primarily focus on imputing and modeling chromatin interactions in single-cell contact matrices, utilizing approaches such as analyzing topologically associating domains, embeddings, and cluster domains[201, 102, 186, 196]. Furthermore, a computational tool has been developed for identifying frequent inter-chromosomal interactions from single cells using a network-based method[18]. However, none of these tools effectively address the identification of significant intra-chromosomal interactions at the single-cell level. In many cases, researchers resort to employing bulk Hi-C technologies like HiCCUPS[136] and FitHiC[6] to derive significant interactions by aggregating individual cell interactions. Unfortunately, these methods typically yield suboptimal results as they are not tailored to identify significant chromatin interactions specifically within single cells.

Recently, SnapHiC, a random walk algorithm-based method, has been introduced as a pioneering computational approach for identifying significant intra-chromosomal interactions from single-cell Hi-C data[191]. The method has shown promise in enabling the analysis of very high-resolution chromatin interactions (e.g., 10kb) from single-cell Hi-C data. However, the high-resolution nature of these chromatin interactions imposes stringent requirements on the raw single-cell Hi-C data. It is recommended that each single cell possesses a minimum of 150,000 raw chromatin contacts, a threshold that most existing unfiltered single-cell Hi-C data fails to meet. Moreover, SnapHiC treats chromatin interactions in each cell as independent entities, disregarding the local similarities of chromatin interactions between different cells. Notably, leveraging local similarities has proven effective in enhancing the analysis of single-cell Hi-C data[196] and single-cell Hi-C data clustering [183]. Furthermore, the majority of single-cell studies[122, 135, 45, 161, 86, 121, 119] have been conducted at resolutions of hundreds of kilobases or several megabases. Consequently, there is a need for new computational methods that can accommodate a wider range of single-cell Hi-C data while considering the local similarities of chromatin interactions between different cells, particularly at a comparatively relaxed resolution (e.g., 100kb).

In this study, we propose a novel algorithm for statistically significant **s**ingle-**c**ell **Hi-C** chromatin interactions with **N**earest **N**eighbors, named **scHi-CNN**. The algorithm comprises three

main steps: imputation of single-cell matrices utilizing a k-nearest-neighbor-based approach, normalization, and identification of statistically significant chromatin interactions. To evaluate the performance of our proposed method, we primarily compared it with the SnapHiC algorithm. We utilized three distinct types of single-cell datasets and compared the counts of significant interactions as well as the overlapping interactions between different datasets of the same cell type. Additionally, we assessed the relevance of the derived significant interactions by analyzing CTCF binding sites considering the fact that CTCF plays an important role in three-dimensional genome organization and presumely contributes to the formation of higher-order chromatin structure [47]. To provide a comprehensive comparison, we utilized bulk Hi-C data and contrasted the outcomes obtained from the different methods. Furthermore, we conducted an analysis of chromatin loops generated using varying numbers of cells, focusing on known regulatory elements. The results demonstrated that our proposed algorithm is capable of identifying more biologically meaningful interactions from single-cell data, even when utilizing a smaller number of cells compared to SnapHiC. We firmly believe that our method serves as a valuable tool for identifying significant chromatin interactions in single-cell data, thereby contributing to the analysis of three-dimensional chromatin organization.

4.3. Background

SnapHiC[191] is a computational pipeline which is designed to identify significant intra-chromosomal chromatin loops from single cell Hi-C data and it is the closest work related to this study. It utilizes the random walk with restart(RWR) algorithm to impute the contact probability between the intra-chromosomal interactions. The primary steps of the SnapHiC method include estimating contact probabilities using the RWR algorithm, normalizing based on genomic distances, identifying loop candidates through statistical measurements, and clustering loop candidates to identify the summits. They offers a comparative analysis of the results between existing bulk Hi-C techniques such as HICCUPS, FastHiC, FitHiC2, and HiC-ACT and provided a tool for public use.

HiCCUPS[136] is another computational tool to capture significant long range chromatin loops using Bulk Hi-C data and does not work with single cell Hi-C data. It analyze the local enrichment patterns comparing to the existing local background. The algorithm checks for significant enrichment relative to four different neighborhoods around the pixel in the contact matrix and identify peaks using a statistical measurement.

FitHiC[6] is a computational tool which is capable of identifying mid range chromatin interactions from Bulk Hi-C data. It uses a spline to map observed contact counts versus their genomic distances and provides a statistical measurement value(corrected p-value) for chromatin interactions using a binomial distribution approach and hypothesis testing correction.

ScHiCluster[201] is a single-cell clustering algorithm for Hi-C contact matrices, which relies on imputations using linear convolution and random walk. scHiCluster demonstrates improved clustering accuracy in low coverage datasets compared to existing methods. After imputation with scHiCluster, topologically associating domain (TAD)-like structures can be identified within single cells, and their consensus boundaries are enriched at TAD boundaries observed in bulk cell Hi-C samples which enables visualization and comparison of single-cell 3D genomes.

4.4. Method

4.4.1. Proposed Algorithm

Our proposed algorithm consists of three key steps: imputation, normalization, and identification of significant chromatin interactions. Our algorithm workflow is visually represented in Fig 4.1.

4.4.1.1. Imputation of Single Cell Contact Matrices

The initial step involved partitioning each chromosome into equal-sized bins for each individual cell. One of our aim is to handle datasets with fewer chromatin interactions, so we've chosen to broaden the bin size resolution of SnapHiC from high resolution bins (like 10Kb, 25Kb) to 100Kb. We then assigned chromatin interactions to specific bin pairs and tallied these interactions to generate contact matrices. For contact matrices that contained empty pixels (i.e., zero contact count), we implemented a strategy to impute these empty pixels. Specifically, we extracted a surrounding region measuring *(2d+1)x(2d+1)* (e.g $d=5$ bin pair differences in each direction from the empty pixel) to identify the closest neighbors. To perform imputation, we only considered pixels that had at least one chromatin interaction within their surrounding region.

Subsequently, we retrieved the surrounding matrices corresponding to the same position in the other cells for the same chromosome. From these matrices, we selected the top k (e.g $k=4$, which is also used in [201]) neighbors based on the Pearson correlation coefficient. The mean of these top k closest neighbors was then used to impute the empty pixel. Note that after the imputation, the empty pixel can still be zero if the same entries are zeros for all top k neighbors.

To maintain the integrity of the analysis, we imposed a maximum distance threshold (e.g. 1 million base pairs) for the imputation of interactions. This ensures that the imputed values are derived from nearby genomic regions that are more likely to exhibit chromatin interactions. Also, given the symmetry of a Hi-C matrix, our procedure involved only the imputation of the upper half matrix within the specified distance.

4.4.1.2. Normalization

To standardize our contact matrices, we employed a normalization approach that involved grouping interactions with the same genomic distance, effectively normalizing them diagonally with the same parameters used in SnapHiC [191] for a fair comparison. For each diagonal segment within a contact matrix, we started by filtering the top 1% of the interactions with the highest contact values. Subsequently, we computed the mean and standard deviation using the remaining values and calculated corresponding z-scores. Diagonals with a standard deviation lower than 10^{-6} were disregarded, and those segments were filled with zeros to account for their negligible variability.

4.4.1.3. Identification of Significant Chromatin Interactions

To identify significant chromatin interactions , we implemented similar criteria used in SnapHiC to determine if a interaction bin pair qualified as a peak compared to its surrounding region. For an interaction pair to be considered, its mean normalized contact counts across all cells needed to exceed zero. Additionally, we require that at least 10% of single cells exhibited a normalized contact count greater than 1.96 (corresponding to a $pvalue \leq 0.05$). For interactions that meet these criteria, we conducted a paired t-test with the local neighborhood to assess significance. The local neighborhood was defined as the surrounding regions within a 2-bin genomic distance, excluding the closest neighbors (i.e., bin pairs within a 1-bin genomic difference). Using the mean of the local neighborhood values, we performed the paired t-test and obtained t-statistics and p-values. Subsequently, we grouped the p-values based on genomic distance and converted them into false discovery rates (FDRs) using the Benjamini-Hochberg procedure. Finally, we identified the significant chromatin interactions based on a t-statistic greater than 3 and an FDR value less than 0.1.

4.4.2. Processing Single-Cell Hi-C Data

In this study, we utilized several publicly available single-cell Hi-C datasets. Firstly, for the cell cycle dataset [122], we obtained contact matrices for single cells categorized into four distinct

cell cycle phases. The labels G1 phase, Early-S phase, Mid-S phase, and G2 phase correspond to the datasets 1CDX1, 1CDX2, 1CDX3, and 1CDX4 respectively. Each phase included interaction data for a total of 390 individual cells. Secondly, we acquired contact matrices for Mouse ES cells [94] comprising a total of 475 cells. Lastly, we obtained contact matrices for human frontal cortex single cells [94] that comprise a total of 4,238 cells. To process the single-cell Hi-C data, we applied both the proposed algorithm and the SnapHiC algorithm, allowing for a comparison of the results obtained from each method.

4.4.3. Processing Bulk Hi-C Data

In the study, we obtained the Fastq files for the bulk Hi-C data [16] corresponding to the cell cycle dataset. These files were then processed using HiC-Pro to generate contact matrices [150]. For the bulk Hi-C data related to Mouse ES cells [94], we directly downloaded the contact matrices from the NCBI database. To identify significant chromatin interactions within these contact matrices, we applied the HiCCUPS [136] and FitHiC2[6] algorithms. In order to compare these findings with the single-cell Hi-C data, we focused on the common interactions identified by both HiCCUPS and FitHiC2 algorithms.

4.4.4. Processing CTCF ChIP-Seq Data

The Mouse ES cells CTCF ChIP-seq narrow peak data were obtained from the ENCODE project (ENCSR362VNF) [25]. Similarly, for Homo sapiens neural cells derived from H1, the CTCF ChIP-seq data were downloaded from ENCODE (ENCSR822CEA). To analyze the single-cell Hi-C datasets, we performed a counting of CTCF-enriched interactions. An interaction was classified as CTCF-enriched if both ends of the interaction overlapped with at least one CTCF binding site. This criterion allowed us to identify and examine interactions that exhibited a potential association with CTCF binding events.

4.4.5. Processing Promoter Related Interactions

In this study, we utilized a previously reported set of promoter-related interactions, including promoter-promoter and promoter-other interactions, as a reference dataset [77]. These interactions were derived from a study conducted on human cortex cells. To evaluate the performance of our proposed methodology, we compared our results with those obtained from SnapHiC with varying numbers of cells. We then examined the overlap between these interactions and the reference promoter-related interactions. This analysis allowed us to assess the accuracy and effectiveness

of our methodology in capturing relevant chromatin interactions within the context of promoter activity in human cortex cells.

4.5. Results

4.5.1. Quantity of Significant Chromatin Interactions

For the analysis of human cortex cells, we used both methodologies across varying cell numbers, namely 10, 25, 50, 100, 200, and 500 cells. To ensure unbiased and representative results, we performed multiple random selections of cell numbers, as depicted by the error bars in Figure 4.2C. The outcomes consistently demonstrate that scHi-CNN identifies a significantly higher number of chromatin loops, even when applied to a small cell population. In contrast, SnapHiC's performance appears to be less effective, particularly in detecting interactions among smaller cell groups. To extend our evaluation, we applied both methodologies to the whole and each cell phase in the cell cycle dataset. The performance remains consistent across the cell phases, as illustrated in Figure 4.2E. Also scHi-CNN is capable of identifying the increasing trend of significant chromatin interactions in cells at varying stages of the cell cycle, a phenomenon attributed to DNA replication during the S phase. In contrast, SnapHiC is unable to capture these inherent biological states of the cell. Furthermore, we quantified the raw interactions corresponding to the identified significant interactions (Figure 4.2A and B). Notably, when analyzing smaller cell groups (around 10 cells), scHi-CNN identifies interactions that were present in approximately 60% of the cells, whereas the interactions identified by the SnapHiC method are present in a much smaller fraction of cells. This showcases the superiority of scHi-CNN in identifying frequently occurring chromatin interactions among cells, thereby highlighting its potential to derive more relevant chromatin loops.

4.5.2. CTCF Enriched Interactions

CTCF plays an important role in three-dimensional genome organization and presumely contributes to the formation of higher-order chromatin structure [47]. We assessed the CTCF enrichment of the significant interactions obtained from scHi-CNN and SnapHiC by leveraging previously collected CTCF methylation data (Figure 4.2D,F). Our analysis reveals that the percentage of CTCF-enriched interactions derived from scHi-CNN remains consistent across different cell quantities, whereas SnapHiC struggles to generate CTCF-enriched interactions, especially when dealing with smaller cell populations. SnapHiC requires a minimum of 50-100 cells to produce 50% of the CTCF-enriched interactions. In contrast, scHi-CNN consistently identifies more than 60% of CTCF-

enriched interactions in both human cortex cell and cell cycle datasets, regardless of the number of cells used. These findings suggest that the results obtained using scHi-CNN encompass a higher proportion of biologically meaningful data, indicating an improvement over existing methodologies in terms of data quality and relevance.

4.5.3. Common Interactions between Different Datasets from the Same Cell Type

To investigate the overlap of interactions between different datasets, we analyzed the cell cycle dataset and mouse embryonic stem cell (mESC) dataset, which used the same cell type. We specifically examined the common interactions between each phase of the cell cycle and mESC cells for both scHi-CNN and SnapHiC (Figure 4.3). Additionally, we determined the common interactions across each cell phase within the cell cycle dataset (Figure 4.4). Interestingly, scHi-CNN consistently identifies a significantly higher percentage of common interactions in both cases. In addition, the Figure 4.3 illustrates that scHi-CNN outperforms SnapHiC in terms of stability, as evidenced by a lower maximum variation in the common percentage (6.36% for scHi-CNN compared to 7.86% for SnapHiC). This observation suggests that scHi-CNN excels in deriving meaningful interactions by effectively identifying a greater number of common interactions within the same cell type.

4.5.4. Identified Promoter Centered Interactions

To gain further insights into the identified interactions, we conducted an evaluation using Layer 2/3 (L2/3) type cells from human cortex cells, considering different quantities of cells. In order to facilitate more comparison, we also employed SnapHiC at 10kb resolution with 100 L2/3 cells. We specifically focused on four known promoters and genes associated with cortex and neural cells, as highlighted in previous studies[77, 158, 112], to assess the identified chromatin interactions. Figure 4.5A showcases the identified interactions for each cell quantity using scHi-CNN, while highlighting the promoters of interest. Remarkably, scHi-CNN successfully identifies these promoter-related interactions even with a very low cell count, whereas SnapHiC fails to detect most of these interactions even with a higher cell count at 100kb resolution. Although SnapHiC manages to identify a few promoter-related interactions at 10kb resolution, its performance fells short compared to scHi-CNN. Furthermore, Figure 4.5C,D,E,F illustrates the overlapping promoter-centered interactions identified using 100 cortex single cells, in comparison with the promoter-centered interactions reported in a previous study[77]. Though scHi-CNN identifies less number of significant interactions (Figure 4.5B) than SnapHiC, our method, scHi-CNN, reports a significantly higher percentage of

promoter-centered interactions compared to SnapHiC (Figure 4.5C,D,E,F). These findings further highlight the superior performance of scHi-CNN in identifying a greater proportion of biologically meaningful interactions.

4.6. Conclusion

In conclusion, this study presents a novel and robust methodology for identifying significant intra-chromosomal chromatin loops from single-cell Hi-C data, addressing the limitations of existing tools and expanding our understanding of chromatin architecture in individual cells. Our method consists of three primary steps: 1) imputing contact matrices using a K-nearest-neighbour-based approach, 2) normalization, and 3) identifying significant chromatin interactions using a statistical test. We evaluated the performance of our proposed approach using three distinct datasets, including human cortex cells, mouse embryonic stem (ES) cells, and a mouse cell cycle dataset, with varying numbers of cells to assess the robustness and scalability of our method across different conditions.

To validate the biological relevance of the interactions identified by our approach, we utilized several criteria, including CTCF binding sites, analysis of known promoter-related interactions, and quantification of common interactions between different datasets of the same cell type. Our method shows a greater ability to generate a significantly higher number of biologically meaningful interactions compared to SnapHiC. The capabilities were demonstrated through a higher percentage of CTCF-enriched interactions, greater overlap with known promoter-centered interactions, and increased common interactions between the same cell types, thus highlighting the potential of our method in deciphering complex regulatory networks in single cells.

Future research could focus on refining and optimizing the methodology to further enhance its performance, sensitivity, and generalizability across diverse cell types and conditions. Additionally, integrating our method with other single-cell genomics data modalities, such as single-cell RNA-seq, ATAC-seq, or ChIP-seq, could provide a more comprehensive view of the molecular mechanisms associating with chromatin architecture and gene regulation in single cells. This multi-modal integration would enable researchers to better understand the complex interplay between chromatin structure and function, ultimately leading to novel therapeutic strategies for various diseases, including cancer and developmental disorders, which are often characterized by aberrant chromatin organization and gene expression patterns.

4.7. Acknowledgment

The work is supported by the National Science Foundation under NSF EPSCoR Track-1 Cooperative Agreement OIA #1946202 and used advanced cyber infrastructure resources provided by the University of North Dakota Computational Research Center and the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University.

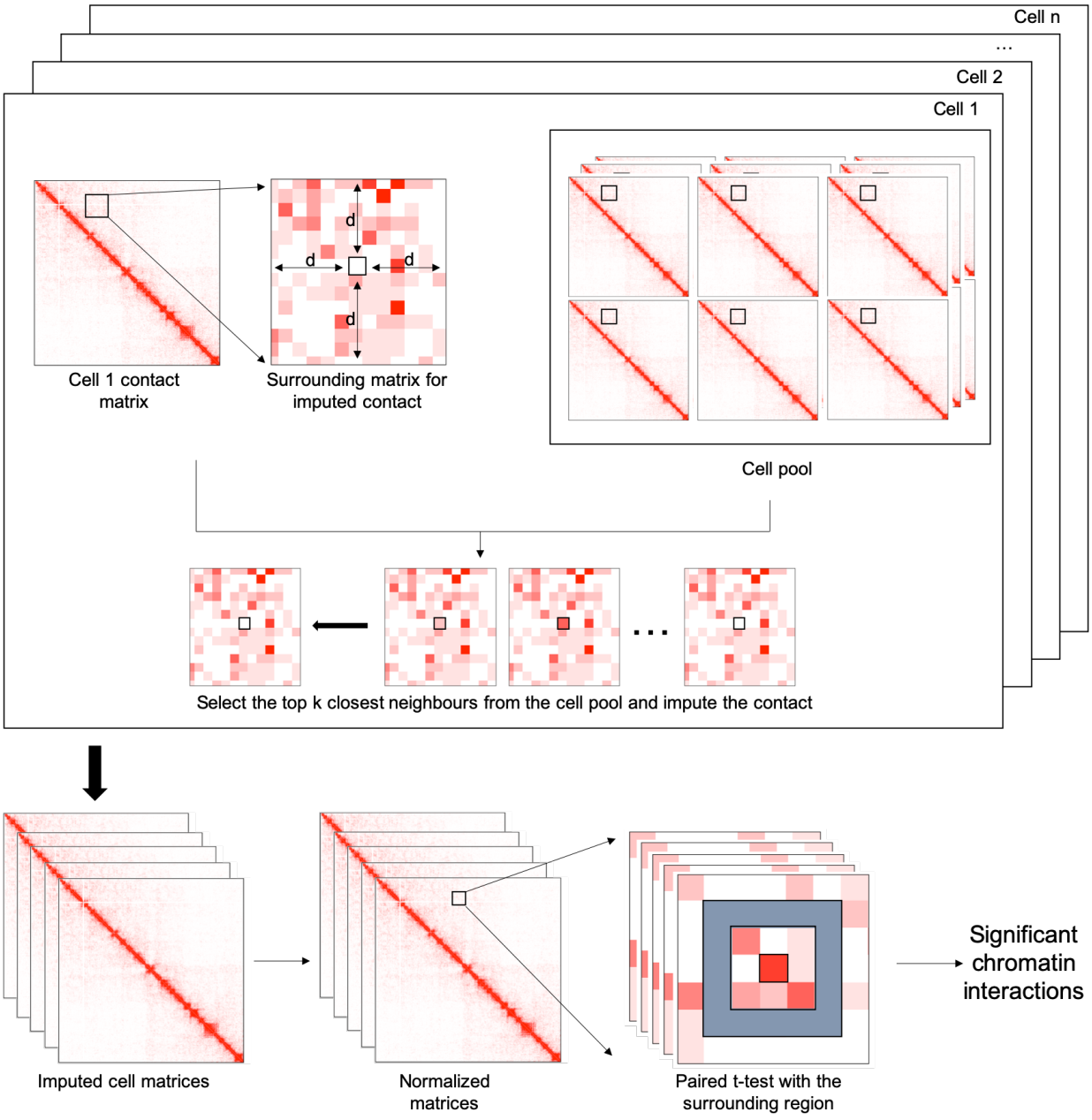

Figure 4.1. Workflow of the Method - 1. Single-cell contact matrix imputation, 2. Normalization process, 3. Identification of significant chromatin interactions

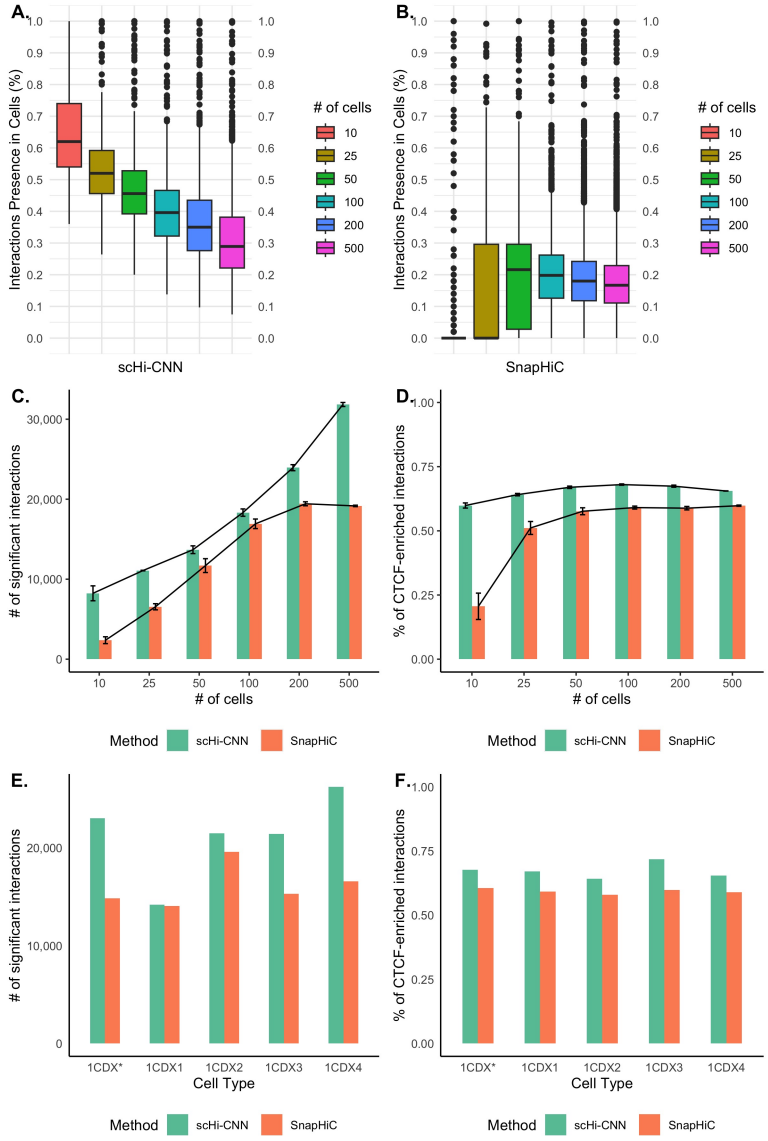

Figure 4.2. A. Distribution of the percentages of the presence of raw interactions corresponding to the identified significant interactions across cells in prefrontal cortex for scHi-CNN (e.g., 0.5 means 50% of the cells contain the identified significant interaction) B. Same as 'A' for the SnapHiC method C. Significant interactions derived using scHi-CNN and SnapHiC for cells in prefrontal cortex. D. Percentage of CTCF enriched interactions identified using the two methods for cells in prefrontal cortex. In A,B,C, and D five random samples for each number of cells were gathered and represented in the figure with the error bars. E. Significant interactions derived using two methods for cell cycle data organized in each cell cycle. F. Percentage of CTCF enriched interactions identified using the two methods for cell cycle data.

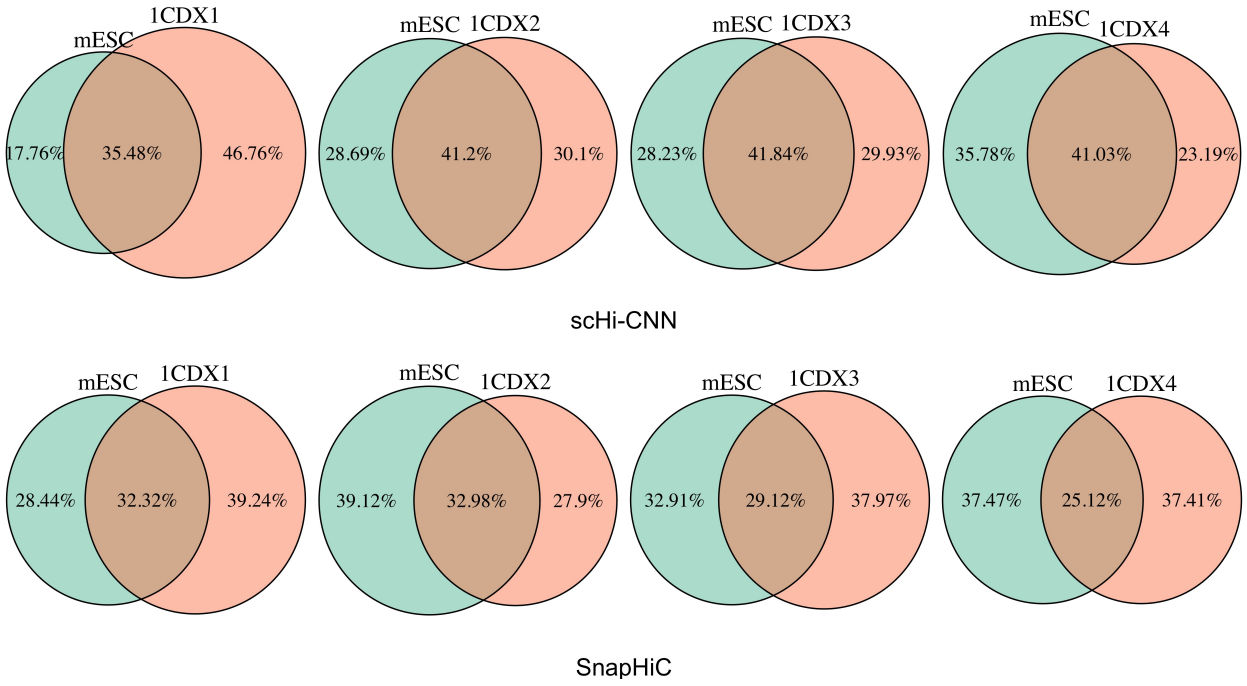

Figure 4.3. Common interactions percentages between the cell cycle and mESC datasets using scHi-CNN and SnapHiC

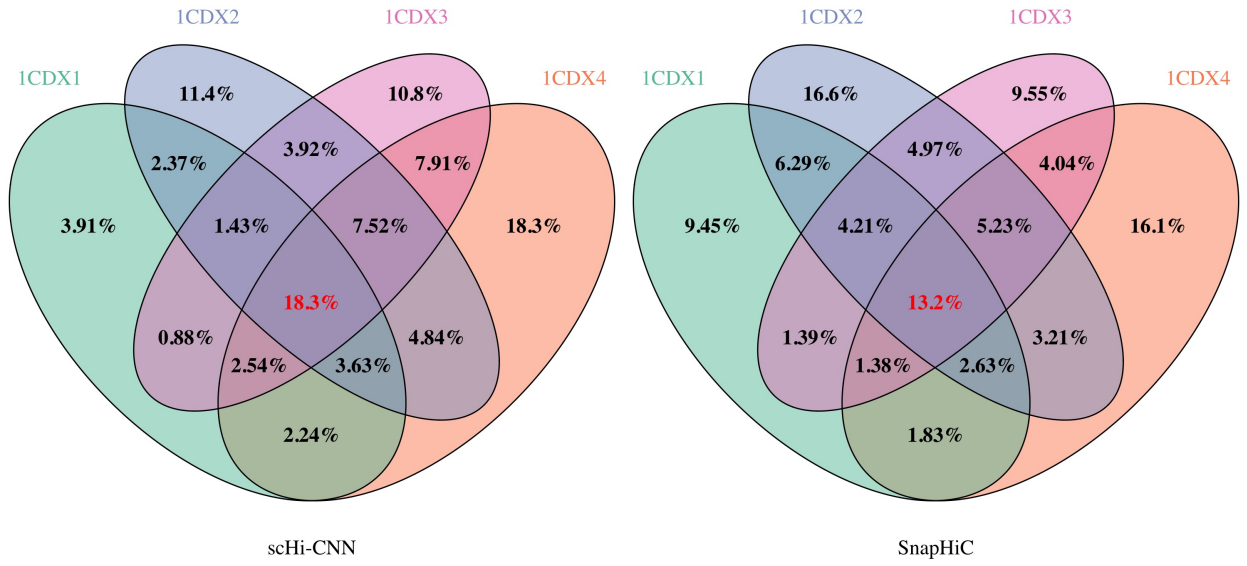

Figure 4.4. Common interactions percentages among cell cycle phases using scHi-CNN and SnapHiC

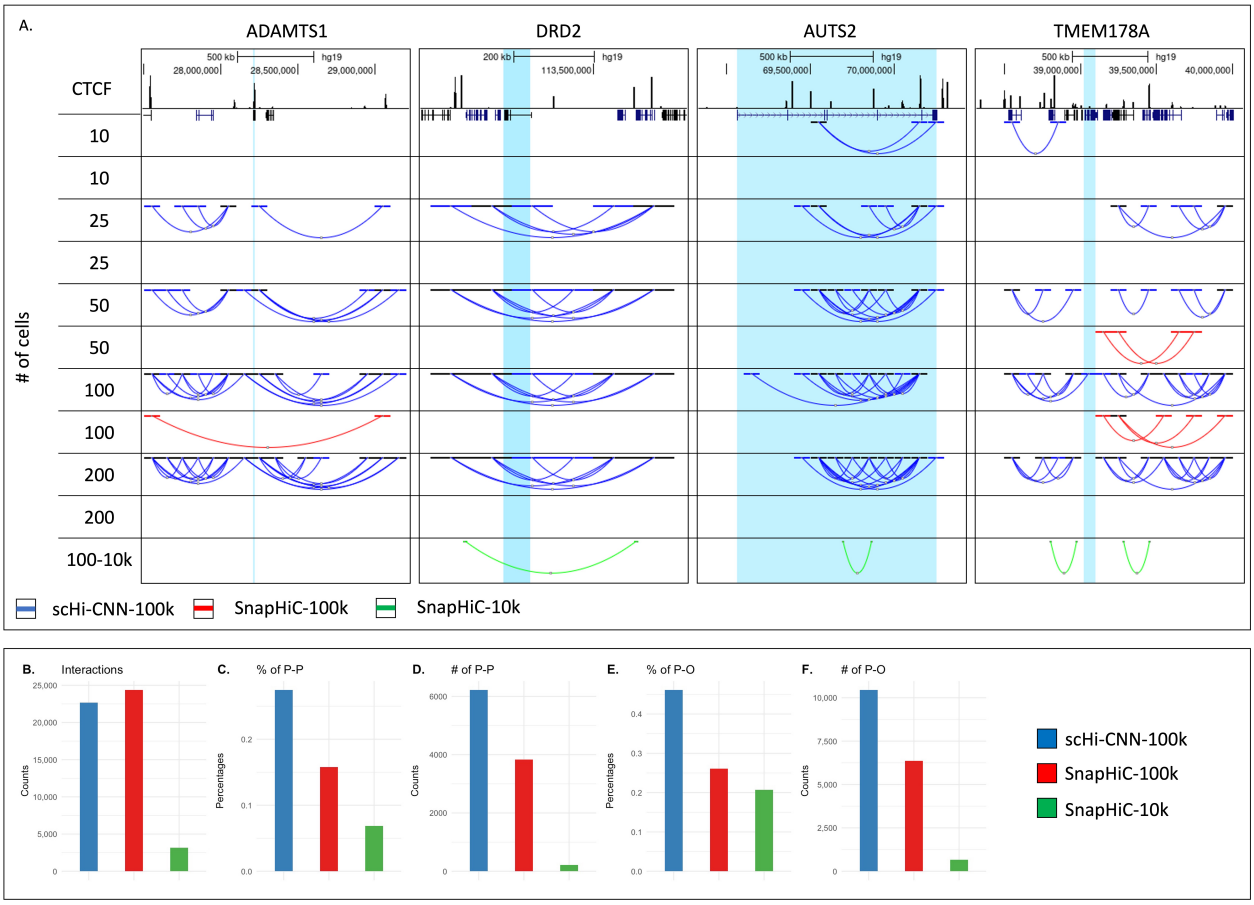

Figure 4.5. Identified significant interactions in human cortex cell lines related to known Promoter-centered interactions using scHi-CNN and SnapHiC. A. Identified significant interactions for each cell quantity using scHi-CNN and SnapHiC within the marked areas associated with the four known promoters. B. Number of significant interactions derived using scHi-CNN and SnapHiC. C and E. Percentage of overlap with known promoter-promoter interactions and promoter-other interactions. D and F. Overlapping interaction count with known promoter-promoter interactions and promoter-other interactions.

5. INTEGRATIVE ANALYSIS OF EPIGENETICS AND CHROMATIN INTERACTION DATA

5.1. Introduction

A key research field in bioinformatics involves studying how DNA is organized in three-dimensional structures inside cells and identifying crucial genomic components that play significant roles in gene expression and regulation, thus affecting cells functionality. Chromatin interactions significantly influence gene regulation by bringing corresponding regulatory elements into close proximity [30]. Chromatin interactions data are important in identifying key chromatin topological structures, such as TADs and compartments, which are essential for analyzing genomic functions. Additionally, various genetic disorders, including cancer and other pathologies, are associated with disruptions in this chromatin architecture [58] [159] [36] [113]. The study of epigenetics offers profound insights into gene activity by examining chemical modifications in DNA and histone proteins. Epigenetic markers, which impact gene regulation without altering DNA sequences, are crucial for understanding cell behavior and differences in cell types. Epigenetic changes, such as histone modifications and DNA methylation, have been linked to genomic instability, potentially leading to genetic disorders like cancer by interfering with the functions of associated genes or oncogenes [24]. Therefore, analyzing epigenetic data can reveal information on mutations or oncogenes related to genetic diseases. Furthermore, DNA methylation is particularly critical for studying cell development and disease [73]. Hence, analyzing the correlation among these factors is essential to understanding their impact on gene regulation and cellular function. Recent research suggests an interplay between epigenetic markers and chromatin structure in genomic function [56]. This correlation is observed through studies focused on analyzing and prediction of the relationship between A/B compartments, TADs and chromatin modifications [46] [118], as well as through analyzing different epigenetic domains and chromatin interactions using imaging techniques [15]. However, current studies have revealed only a limited relationship between these factors at more finer scales [56] and most of the studies focused on their individual role than the interplay. Consequently,

the influence of chromatin interactions and epigenetic markers on chromatin organization remains unclear, and exploring this through biological experiments alone presents significant challenges.

In recent years, significant advancements in the field of genomics have emerged due to the application of advanced computational models such as machine learning and deep learning. Genomics data, known for its complexity and volume, requires sophisticated computing techniques for proper analysis. Graph embedding algorithms, particularly those based on deep learning approaches, effectively transform complex real world graph structures and relationships into a lower-dimensional space which enhances the efficiency of downstream data processing such as prediction, classification, clustering and visualization [51]. The ability to transform chromatin interaction data into a graph structure, along with the characterization of epigenomic markers as features, enables the synthesis of these distinct omics data using a graph embedding strategy. This process allows a systematic evaluvation of their collective impacts on interpreting the structural organization of chromatin. In various cases, graph embedding algorithms have been applied to predict tasks related to chromatin interactions. The Sub-compartment Identifier (SCI) is an algorithm that utilizes graph embeddings to predict sub-compartments from chromatin interaction data [4]. Varrone et al. introduced a computational framework for predicting co-expression networks from chromatin conformation data. They argue that a non-linear relationship exists between chromatin conformation and gene regulation, and that gene topological embeddings contain relevant information [176]. Recently, epigenomic markers data has been used in conjunction with chromatin interaction embedding data for the annotation of chromatin domains [155]. In this approach, the LINE embedding algorithm was utilized to generates embeddings from chromatin structure data, which are subsequently aligned with epigenomic markers data for annotation. However, to the best of our knowledge, no existing studies have thoroughly assessed the role of chromatin structural information using a integrative computational methodology.

Graph embedding algorithms have shown promising outcomes in various fields, including social networks[178], computational biology [4], natural language processing [177], and recommendation systems [35], by transforming the structural integrity of graphs into latent spaces. Graph convolutional networks (GCNs) represent a major step forward compared to traditional graph embedding methods such as DeepWalk [131], LINE [167], and Node2Vec [53] in analyzing large graphs with node features. GCNs encapsulate graph information by aggregating feature data from a node's

local neighborhood, which allows for effective integration of information from its immediate surroundings [197]. Moreover, GCNs integrate node features, enabling the model to consider both feature and structural information of the neighborhood. Drawing inspiration from GCNs, GraphSage [54] was introduced as a framework for inductive representation learning on large graphs through the sampling and aggregation of features from a node's local neighborhood. GraphSage accommodates large graph data and can adjust to various graph structures. Numerous extensions and applications have been built upon GraphSAGE to leverage its capabilities, including PinSAGE [190] for handling large and complex graphs and HinSAGE [28] for heterogeneous graphs. Given GraphSage's ability to manage large graphs and relevant node feature information, it demonstrates the capability to integrate chromatin structural information to learn latent embeddings.

In this study, we investigate the impact of chromatin interactions and epigenomic data on chromatin structure and organization by integrating this information into a graph embedding model to generate embeddings. We evaluated the accuracy of the predictions of these embeddings under three distinct scenarios that disrupt the graph's structure but maintain global characteristics such as node degree and edge count. In addition, we applied a clustering approach on the generated embeddings to predict TADs like domains. The findings indicate that while epigenetic markers assist in the model's training and predictions, chromatin interactions are crucial in preserving the structural integrity of the chromatin. Although the approach is based on statistical analysis, it suggests that chromatin interactions are vital in determining the effects of chromatin architecture on genomic functions through gene regulation, with epigenetic markers serving to modulate these interactions. Moreover, our findings highlight the significance of incorporating multi-dimensional genomic data (structural, epigenetic, genetic) for a thorough understanding of genome structure and function.

5.2. Data

This study utilized three distinct datasets. The initial dataset includes data on chromatin interaction and epigenetics, derived from three cell lines associated with breast cancer: parental endocrine-sensitive ER+ MCF7 cells, tamoxifen-resistant (TAMR) cells, and fulvestrant-resistant (FASR) cells [1] [25]. This dataset incorporates chromatin interaction information generated through the Hi-C method and data on epigenetic markers collected via ChIP-Seq techniques. Specifically, we analyzed ChIP-Seq data for H3K4me3, H3K4me1, H3K27ac, H3K27me3, H2AZac markers, and

Table 5.1. Breast cancer related cell lines

Cell Line	Data type	Data format	Accession	Reference
FASR	Hi-C	allValidPairs	GSE118712	[1]
FASR	H3K27ac	Bigbed	GSE118711	[1]
FASR	H3K4me3	Bigbed	GSE118711	[1]
FASR	H3K4me1	Bigbed	GSE118711	[1]
FASR	H2AZac	Bigbed	GSE118711	[1]
FASR	CTCF	Bigbed	GSE118711	[1]
MCF7	Hi-C	allValidPairs	GSE118712	[1]
MCF7	H3K27ac	narrowPeak	ENCSR752UOD	[25]
MCF7	H3K4me3	narrowPeak	ENCSR985MIB	[25]
MCF7	H3K4me1	narrowPeak	ENCSR493NBY	[25]
MCF7	H2AZac	Bigbed	GSE118711	[1]
MCF7	CTCF	narrowPeak	ENCSR000DWH	[1]
TAMR	Hi-C	allValidPairs	GSE118712	[1]
TAMR	H3K27ac	Bigbed	GSE118711	[1]
TAMR	H3K4me3	Bigbed	GSE118711	[1]
TAMR	H3K4me1	Bigbed	GSE118711	[1]
TAMR	H2AZac	Bigbed	GSE118711	[1]
TAMR	CTCF	Bigbed	GSE118711	[1]

CTCF binding sites across all three breast cancer cell line types for a comprehensive analysis in conjunction with Hi-C data. Details related to the data from the breast cancer cell lines, including accession IDs and associated publications, are presented in Table 5.1.

We utilized chromatin interactions and epigenomic indicators from three prostate cancer cell lines for the second dataset. This dataset comprises genomic information from prostate cancer cell lines (PC3 and LNCaP) and normal human prostate epithelial cells (PrEC) [165] [14] [166]. Additionally, this dataset includes data on chromatin interactions obtained through the Hi-C method, and we obtained the H3K4me1, H3K4me3, H3K27ac epigenetic markers, along with CTCF binding sites data, gathered via ChIP-Seq for further analysis. Detailed information on the prostate cancer dataset is presented in Table 5.2.

This study incorporates a single-cell dataset to analyze the effects of the proposed approach on single versus bulk cell data with paired sequencing. The dataset includes 4,238 single human brain prefrontal cortex cells, obtained through single-nucleus methyl-3C sequencing (sn-m3C-seq) [94]. This method simultaneously captures chromatin interactions and DNA methylation information. The dataset has been made available under the accession number GSE130711. It should be noted

Table 5.2. Prostate cancer related cell lines

Cell Line	Data type	Data format	Accession	Reference
LNCAP	Hi-C	fastq	GSE73785	[165]
LNCAP	H3K27ac	fastq	GSE73785	[165]
LNCAP	H3K4me1	fastq	GSE73785	[165]
LNCAP	H3K4me3	fastq	GSE38685	[14]
LNCAP	CTCF	fastq	GSE38685	[14]
PrEC	Hi-C	fastq	GSE73785	[165]
PrEC	H3K27ac	fastq	GSE57498	[166]
PrEC	H3K4me1	fastq	GSE57498	[166]
PrEC	H3K4me3	fastq	GSE57498	[166]
PrEC	CTCF	fastq	GSE38685	[14]
PC3	Hi-C	fastq	GSE73785	[165]
PC3	H3K27ac	fastq	GSE57498	[166]
PC3	H3K4me1	fastq	GSE57498	[166]
PC3	H3K4me3	fastq	GSE57498	[166]
PC3	CTCF	fastq	GSE57498	[166]

that the datasets related to breast and prostate cancer are not part of paired experiments. In contrast, the single-cell experiment is designed as a paired experiment, capturing both chromatin interaction and DNA methylation data simultaneously.

5.3. Method

The overall processing pipeline is illustrated in Figure 5.1

5.3.1. Processing Chromatin Interaction Data

Raw and processed chromatin interaction data from bulk Hi-C cell lines were obtained from relevant repositories as mentioned in Tables 5.1, 5.2 and single cell data was obtained from the GEO accession GSE130711. Given the diversity in experimental approaches and techniques applied for these datasets, we adopted distinct processing pipelines as outlined below.

Prostate cancer cell lines consist of multiple replicates for each type. We merged the associated FASTQ files for replicates and processed them using HiCPro[150] to generate contact matrices for each cell line. Specifically, there were three replicates for normal human prostate epithelial cells (PrEC), eight for LNCaP prostate cancer cells, and two for PC3 prostate cancer cells. Following merging, the datasets were aligned to the hg19 reference genome build using the BglII restriction enzyme. Subsequently, significant interactions were identified from the ICE-normalized data us-

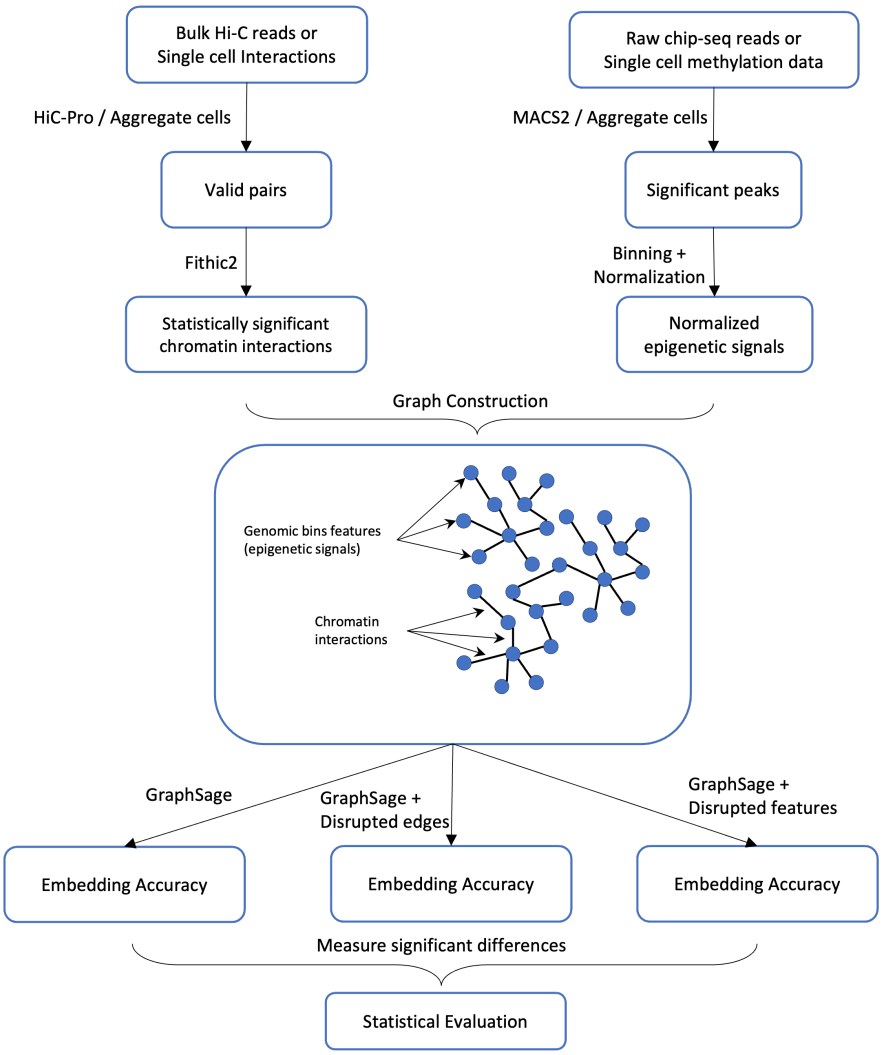

Figure 5.1. Method workflow

ing FitHiC2[85], with a specific p-value threshold. We then applied the HiC-breakfinder[37] tool to exclude interactions associated with regions potentially containing structural variants in cancer genomes.

For breast cancer cell lines, we obtained already processed chromatin interaction data from the GEO repository, as detailed in Table 5.1. Similar to the prostate cell lines, each breast cancer cell type included three replicates, and we merged the processed data for MCF7, FASR, and TAMR cell types. These datasets were processed using the HiCPro tool with the hg38 reference genome build and the NcoII restriction enzyme to generate all valid pairs files. We continued the rest of the HiC-pro pipeline to generate ICE normalized matrices and utilized FitHiC2 to identify significant interactions.

For single-cell Hi-C data generated from human prefrontal cortex cells, each cell's data was deposited separately, totaling 4,238 cells. The deposited chromatin interaction data for each cell was processed using Bismark with Bowtie1 against the hg19 reference genome. We merged all deposited contact data for each individual cell to identify significant interactions using FitHiC2 as in Figure 5.1.

5.3.2. Processing ChIP-Seq and DNA Methylation Data

We processed the ChIP-seq data related to prostate cancer cell lines, as detailed in Table 5.2, starting with raw fastq files. Initially, we assessed the quality of the reads using FastQC, followed by alignment to the reference genome using Bowtie2. We sorted the resulting output SAM files and converted them into BAM files using SAMtools. For peak detection, MACS2 was utilized to compare the ChIP-seq sample data against a control sample to identify regions showing significant enrichment of sequenced tags, consequently identifying statistically significant ChIP-seq peaks. After determining a significance threshold, we excluded less significant peaks and normalized the remaining output to integrate with Hi-C data.

The processed ChIP-seq data for breast cancer cell lines, stored in peak file format, appears in Table 5.1. We downloaded this processed data from the corresponding GEO and ENCODE repositories, then applied binning and normalization for integration with relevant Hi-C data. In a similar manner, we merged and processed the available DNA methylation data for individual cell lines and match with the resolution of Hi-C contact maps.

5.3.3. Graph Embedding Generation

Considering the genome's length and the Hi-C network's resolution, the resulting genomic graphs are often significantly larger compared to traditional graphs. Therefore, a robust embedding algorithm is essential for processing genomic graphs in unsupervised manner to generate embeddings. Based on the Hi-C datasets applied in this study, the resulting graph consists of 50,000 to 110,000 nodes, as detailed in Table 5.3 which illustrates a large and complex structure. Moreover, it is necessary to integrate epigenomic data to characterize the genomic regions as node features. After evaluating various graph embedding techniques, including traditional methods, Graph Convolutional Networks (GCNs), graph autoencoders (GAEs), and various attributed network embedding tools, we selected GraphSAGE as the graph embedding model. Compared with other GCN-based and traditional methods, GraphSage offers more scalability for large graphs due to its sampling approach.

This choice was based on its capabilities in unsupervised learning, managing larger graphs, and associating node features.

We first constructed the graph from the processed Hi-C data, where nodes represent genomic regions, and edges indicate the interactions between these regions. Thus, an edge appears in the graph only when marked as a significant interaction through tools like fithic. We utilized chip-seq and DNA methylation data, aggregated for the respective genomic regions, as node features. For model training, we split the edge space into a training set and a testing set in a 70:30 ratio. The models were trained over 1 million cycles with a learning rate of 0.001 to understand latent representations across 128 dimensions. In addition, we adjusted the relevant neural network hyperparameters to optimize model performance.

$$AGGREGATE_k^{pool} = mean(\{\sigma(\mathbf{W}_{pool}\mathbf{h}_{u_i}^k + \mathbf{b}), \forall u_i \in \mathcal{N}(v)\}) \tag{5.1}$$

GraphSAGE provides several model variants, including mean-based aggregators, LSTM-based aggregators, GCN-based aggregators, and pooling aggregators. In our experiments, we utilized GraphSAGE with the mean-pooling aggregator, as outlined in equation 5.1 [54], following an analysis of the performance of alternative aggregators. The mean-pooling aggregator applies an element-wise mean-pooling operation to collect information from a set of neighbors, each processed individually through a fully-connected neural network.

5.3.4. Identify TAD like Domains

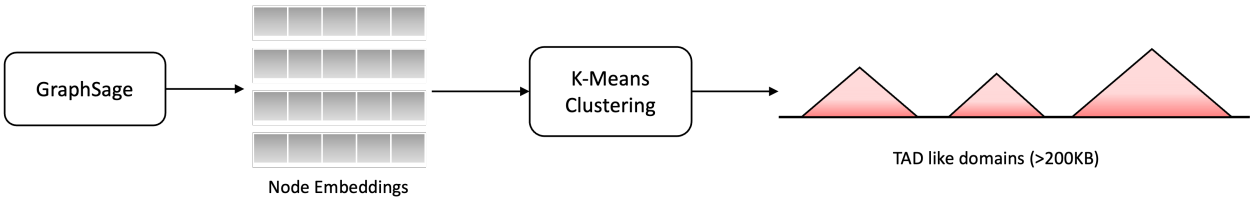

Figure 5.2. TAD like domains identification methodology.

Topologically Associating Domains (TADs) are large regions of the genome that preferentially interact within themselves, creating distinct three-dimensional structures in the nucleus. They play a crucial role in gene regulation, influencing gene expression by facilitating or restricting inter-

actions between regulatory DNA elements and their target genes. To further assess the performance of the generated embeddings, a clustering-based method was utilized for identifying TAD-like domains, drawing inspiration from TAD identification algorithms such as ClusterTAD [125]. This process began with clustering the generated embeddings through K-means clustering (k=5). Subsequently, adjacent bins within the same chromosome that were assigned to the same cluster were grouped together as identified domains similar to the approach in ClusterTAD. Previous research indicates that the size of a TAD can range from several hundred kilobases to a few megabases [50]. Based on that, domains exceeding 200KB were classified as TAD-like domains, while smaller segments were regarded as boundaries or gaps between TADs. The high level overview of the approach is represented in Figure 5.2.

This methodology was applied across three disruption scenarios to assess both the quality and quantity of the TADs identified in each. The evaluation of TAD quality involved measuring the statistical significance of differences between intra-TAD and inter-TAD interactions, following the method used in ClusterTAD. Intra-TAD interactions denote the interactions within a TAD-like domain, whereas inter-TAD interactions refer to the interactions between consecutive TADs. The average number of raw interaction counts was calculated to gather these statistics. Moreover, the average lengths of the TAD-like domains identified in each scenario were examined. Lastly, the count of TAD-like domains identified in each scenario was analyzed, alongside random clustering, to serve as a benchmark.

5.3.5. Evaluate using Statistical Measurements

The models were evaluvated through statistical metrics. We assessed the capability of the proposed model to generate precise embeddings by analyzing the model's performance across various disrupted graphs, derived from the initial graph. Our experiments included three distinct scenarios: the initial graph, the edge-disrupted graph, and the feature-disrupted graphs. In the case of edge-disrupted graphs, we rearranged the structure by shuffling the edges while maintaining the same node degrees and using an identical set of nodes. In the feature-disrupted variant, we shuffled node features across genomic regions while preserving the initial structure of the graph. We statistically assessed these disruptions to determine their collective and seperate impact on the chromatin structural integrity.

To generate an adequate sample size, we conducted model training sessions using a sufficient number of permuted graphs for each scenario. In each case, we collected the training and validation accuracy of the learned embeddings as sample populations for further analysis. We first assessed the normality of the generated samples and applied the paired t-test to compare the significant difference between the groups. For this statistical test, we generated three different sample populations based on accuracies from the initial graph, edge disruption, and feature disruption. We then compared the accuracy of the disrupted graphs against the performances of the initial graph to understand how graph structure and node features influence embedding prediction. To evaluate accuracy, we utilized the Mean Reciprocal Rank (MRR), which assesses a ranked list based on the similarity or disparity in the embedding space, as shown in Equation 5.2.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{5.2}$$

5.4. Results

5.4.1. Graph Embedding Predictions

Table 5.3 shows resultant graph structural information after processing Hi-C data from raw fastq files to identify significant chromatin interactions based on threshold p-value 0.05. The edge count reflects the number of significant interactions, while the node count indicates genomic regions containing at least one interaction. Analysis reveals a significant portion of genomic regions was discarded in post-filtering, as demonstrated by the node count. For example, in graphs associated with breast cancer cell lines, approximately only 50,000 genomic regions out of a total 150,000 (mapped at a resolution of 20,000 bins) register at least one significant interaction. Consequently, the resulting Hi-C contact matrices demonstrates a high degree of sparsity which complicates biological interpretation.

After constructing graphs for each cell line, we initially analyzed the training and validation curves of the model to verify the impact of adding additional node features on the embedding accuracy. The Figure 5.3 illustrates the training and validation mean reciprocal rank (MRR) spread across one million epochs for breast cancer and prostate cancer cell lines, highlighting the effect of epigenetic markers on chromatin organization. The relevant box plots indicating the significance in differences is represented in Figure 5.4. The data indicate that utilizing only a single epigenetic

Table 5.3. Graph information

Cell line	Number of nodes	Number of edges
FASR	50,612	133,693
MCF7	51,543	142,562
TAMR	55,880	170,786
Single cell-cortex	111,238	270,889
LNCAP	95,722	426,754
PC3	100,703	478,726
PReC	101,623	408,625

marker, the MRR peaks at approximately 0.55, whereas it increases up to to 0.8 with the inclusion of all five ChIP-seq markers for the FASR breast cancer cell line. A similar pattern can be observed in both the training and validation accuracy for LNCaP prostate cancer cell lines, as shown in Figure 5.4. We derived these results by averaging data from samples collected from models with different training and validation sets to minimize biases.

To analyze the impact of chromatin interaction and epigenetic data along with their interplay on embedding predictions, we assessed the significance of the differences between initial and disrupted model variants. The accuracy distribution for the gathered samples across three scenarios is represented for breast cancer cell lines, prostate cancer cell lines and single cells in Figure 5.5. We conducted the Shapiro-Wilk test to determine the normality of the data sets, as shown in Table 5.4. The related statistical information is provided in Table 5.5 and visualized in Figure 5.6 for breast cancer cell lines, prostate cancer cell lines and single cells. The paired t-test results across all these cell lines indicate a significant difference between the initial and disrupted graphs which highlights the interplay between chromatin and epigenetics, as supported by existing literature [56]. However, the results also reveal that graphs with edge disruptions tend to show lower prediction accuracies than those with feature disruptions which suggests that chromatin interaction data may play a pivotal role in this correlation and epigenetics data could assist in modulating these interactions. Furthermore, in the edge-disrupted graphs, we altered the local structure while preserving the global structure which highlights the significant role of local chromatin configurations, such as chromatin loops between regulatory elements and higher-order structures like TADs. While this provides a biological intrepretation through statistical analysis and graph embedding techniques,

Table 5.4. Normality test using Shapiro-Wilk Test

Cell line	Initial	Edge Disrupt	Feature Disrupt
FASR	0.528	0.340	0.970
MCF7	0.582	0.619	0.328
TAMR	0.078	0.958	0.537
Single Cell	0.298	0.039	0.488
LNCAP	0.380	0.082	0.386
PC3	0.105	0.100	0.441
PrEC	0.074	0.000	0.987

Table 5.5. Mean and Standard Deviation of the sample populations

Cell line	Initial	Edge Disrupt	Feature Disrupt
FASR	M=0.772, SD=0.015	M=0.500, SD=0.020	M=0.725, SD=0.016
MCF7	M=0.788, SD=0.015	M=0.480, SD=0.018	M=0.727, SD=0.015
TAMR	M=0.778, SD=0.015	M=0.450, SD=0.019	M=0.736, SD=0.015
Single Cell	M=0.545, SD=0.033	M=0.313, SD=0.025	M=0.477, SD=0.036
LNCAP	M=0.652, SD=0.017	M=0.401, SD=0.016	M=0.592, SD=0.017
PC3	M=0.667, SD=0.017	M=0.284, SD=0.045	M=0.571, SD=0.020
PrEC	M=0.679, SD=0.015	M=0.222, SD=0.033	M=0.568, SD=0.021

further exploration is necessary to clarify the biological significance of the data and the practical uses of the embeddings, especially in solving problems related to experimental data and limitations of existing tools.

5.4.2. Identified TAD like Domains

We first assessed the quality of the identified TAD-like domains through analysis of inter-TAD and intra-TAD interactions. The figure 5.7 demonstrates a significant difference between the counts of intra-TAD interactions and inter-TAD interactions. For qualification as a TAD-like domain, the count of intra-TAD interactions should greatly exceed that of inter-TAD interactions. The results indicate that the identified TAD-like domains exhibit a significant difference between these two interaction types across all scenarios and remain consistent in each cell line. This suggests that the identified domains meet the criteria for TAD-like domains in every scenario. Subsequently, we examined the average size of TAD-like domains, as illustrated in the figure 5.8. The violin

plots reveal that the lengths of the domains range from 200KB to 3MB, aligning with findings from previous studies.

Finally, we evaluated the number of identified TAD-like domains in each scenario, as depicted in the figure 5.9. The figure reveals that random baseline clustering discovered the fewest TAD-like domains, indicating the effectiveness of identifying TAD-like domains through corresponding graph embeddings. Original/initial graph embeddings detected a higher number of those domains compared to disrupted graph embeddings, and feature-disrupted graphs identified more TAD-like domains than edge-disrupted graphs. These findings align with those from the embedding accuracy analysis, suggesting that chromatin interactions play a crucial role in the identification of a higher number of TAD-like domains compared to epigenetic features. Moreover, the combined use of these elements leads to the identification of an even larger number of TADs, underscoring the significance of their interplay.

5.5. Discussion

The structure of chromatin is non-random, biologically significant, and represents the spatial arrangement within the nucleus, influencing gene expression and regulation. Analyzing the interplay between chromatin interactions and epigenetics is essential for understanding their impact on genomic functions. However, the specific correlation between these two factors and significance of their roles are not well defined and still remain as a question [56]. This study expects to carry out an integrative analysis of chromatin structural data through a graph embedding model to decipher the underlying patterns and relationships between chromatin interactions and epigenetic data, and to identify their importance in genomic function. Graph embedding algorithms such as GraphSAGE, a neighborhood aggregation algorithm, generates node embeddings by iteratively aggregating and transforming feature vectors of a node's neighbors. These embeddings can be utilized for downstream tasks such as link prediction, node classification, etc. In this proposed approach, we assessed the differences in performance of learning graph embeddings following disruptions to the graph structures while preserving their global integrity, to identify the key elements within the interplay between chromatin interactions and epigenetics against genomic functions.

Since the initial network demonstrates the highest validation accuracy and higher number of TAD like domain identification, it indicates that the chromatin interactions (edges) together with epigenetic information (node features) hold essential information for predicting latent low-

dimensional feature embeddings. It suggests that the physical proximity and interaction among various genomic regions (as captured by chromatin interactions) along with their epigenetic markers play a critical role in biological processes and chromatin organization. This finding aligns with current biological understanding [170] [128], which recognizes the interplay between chromatin interactions and epigenetics as influential in gene regulation and consequently, cellular functions.

The fact that graphs with disrupted features (where node features represent epigenetic markers) maintain some level of predictive accuracy, less than the initial but more than networks with disrupted edges (chromatin interactions), suggests that while epigenetic markers are important, their specific association to specific genomic regions (nodes) might not be as critical as the local structure of chromatin interactions in generating embeddings and identifying TAD like domains. This scenario indicates that epigenetic markers itself do not play as specific or vital role as the precise organization of chromatin interactions.

The lower performance of edge-shuffled graphs (where chromatin interactions are disrupted while maintaining global graph properties) suggests that the local structure and specific connections between regions (local chromatin architecture) are more important for the biological processes than the global structure alone. This might indicate that specific interaction patterns, such as enhancer-promoter interactions or insulator functions, and higher order chromatin structures such as TADs or compartments are crucial for understanding the regulatory mechanisms at play.

Combining these observations, we can conclude that in the context of the chromatin architecture, specific chromatin interactions (and the local genomic architecture they represent) are crucial and likely govern key biological processes by facilitating or restricting access to regulatory elements. The epigenetic context, while important, may act more as a modulator rather than the primary driver, enhancing or diminishing the effects based on the chromatin context.

However, integrating analysis methods that merge multi-omics data with Hi-C data introduces several significant challenges. One primary challenge is the difference in resolution between Hi-C data, which typically ranges from 1 kb to 1 MB, and one-dimensional (1D) chromatin data, such as ChIP-seq, which provides a much finer resolution of 100 bp to 1 kb. This variance complicates the effective integration of information across different scales and may lead to the loss of essential high-resolution details. Although we have aggregated the ChIP-seq signals into genomic

bins to align with the resolution of Hi-C experiment data, this approach can result in the underrepresentation of certain aspects of ChIP-seq data.

Additionally, assessing the relationship between proposed models and their biological relevance is often complicated due to the incomplete nature of processed experimental data. Although bulk Hi-C experiments can generate a vast amount of reads, from millions to billions, the count of biologically meaningful interactions is likely to be considerably lower once noise is removed and biases are corrected using tools like FitHiC and HiCCUPS. For instance, in the breast cancer dataset, the resulting graphs contain about 50,000 nodes (genomic bins) at a 20k bin resolution, even though the total number of genomic regions across the entire genome could at about 150,000 regions at the same resolution. This sparse nature of significant interactions poses a major challenge for comprehensive genome-wide analysis. The filtering process is essential for identifying meaningful interactions considering the limitations of experimental methods, but it leads to an incomplete genomic representation of the chromatin interaction network.

Despite its theoretical nature, our research establishes a solid groundwork for further empirical studies investigating chromatin structural data and its correlation to genomic functions. The specific local structures of chromatin interactions, as opposed to global properties, have proven critical which highlights the importance of enhancing the resolution and accuracy of chromatin interaction maps produced by Hi-C and other 3C-based technologies in future genomic research. More precise interaction maps can offer in-depth views into the physical proximity of different genomic regions and their possible regulatory connections. Moreover, in genetic studies, especially those related to cancer or pathologies, integrating data on chromatin organization and interactions could lead to better identification of genomic regions that are crucial for diseases. Utilizing these approaches allows for a more profound comprehension of chromatin dynamics and their associations with gene regulation, disease mechanisms, and cellular functions. By introducing this conceptual approach, our work supports further exploration into the intricate relationship between chromatin architecture and epigenetics. We expect these insights to enhance future studies on the interplay between chromatin interactions and epigenetic mechanisms.

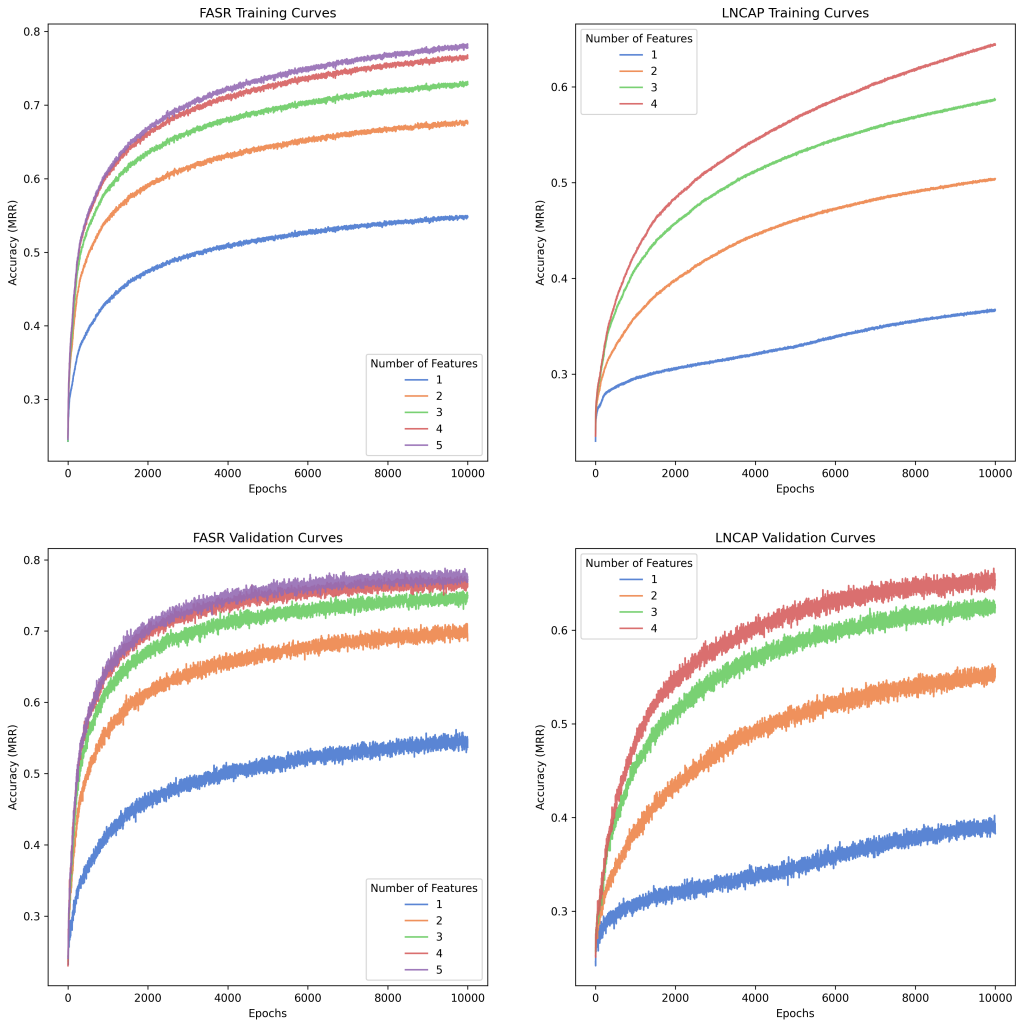

Figure 5.3. Accuracy variation with different number of features (chip-seq) markers for breast cancer and prostate cancer cell lines.

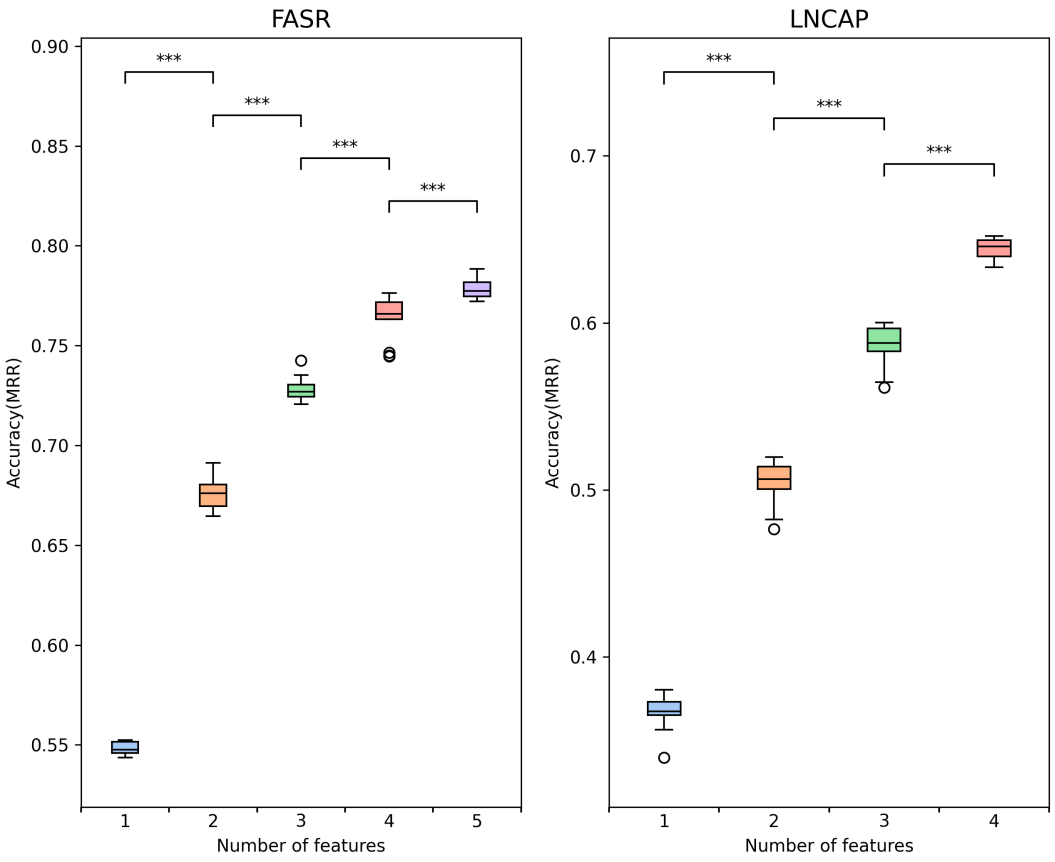

Figure 5.4. Accuracy variation with the increasing number of Chip-seq markers and significance difference related to FASR and LNCAP cell types. Asterisk '***' represent pvalue<0.001

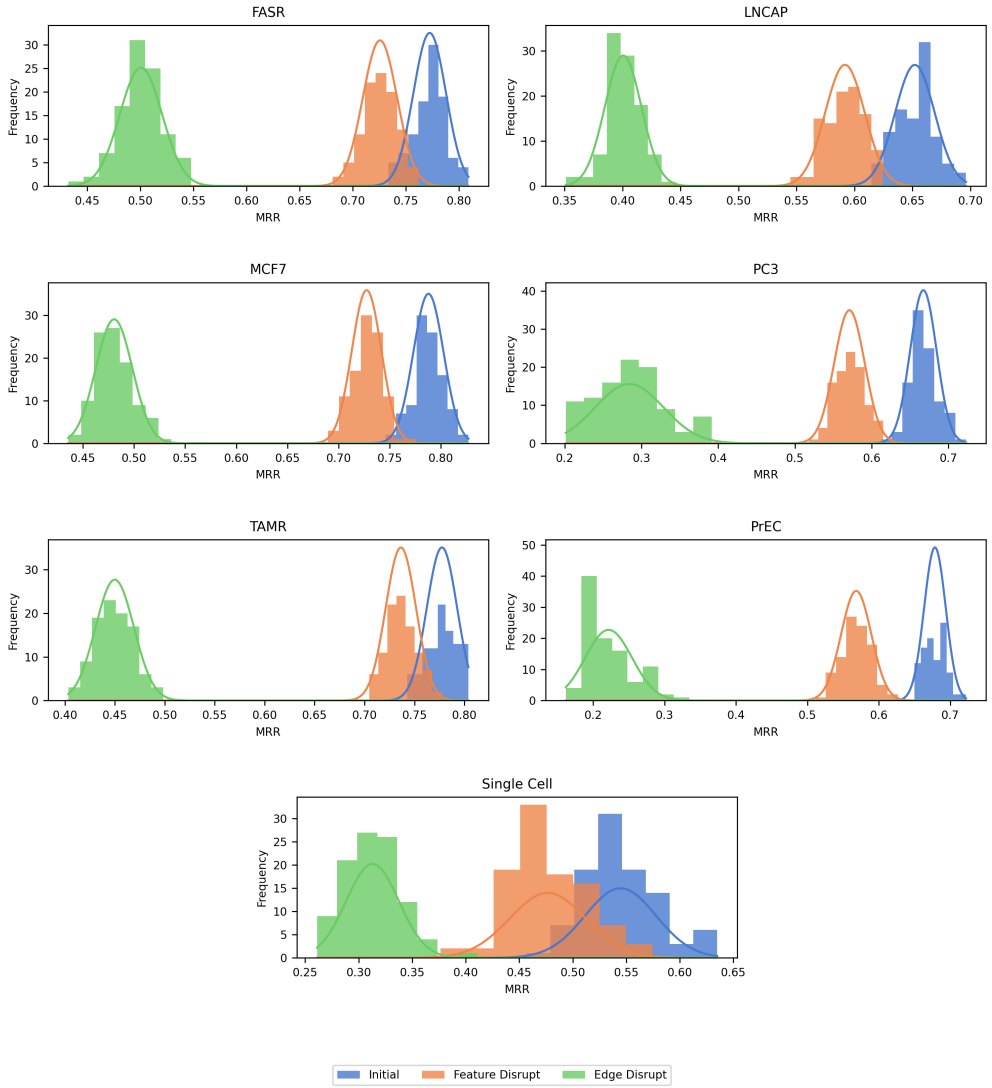

Figure 5.5. MRR variation with the shuffling of edges and features in Breast cancer cell lines; FASR, MCF7, TAMR, prostate cancer cell lines; LNCAP, PC3, PrEC and single cell lines.

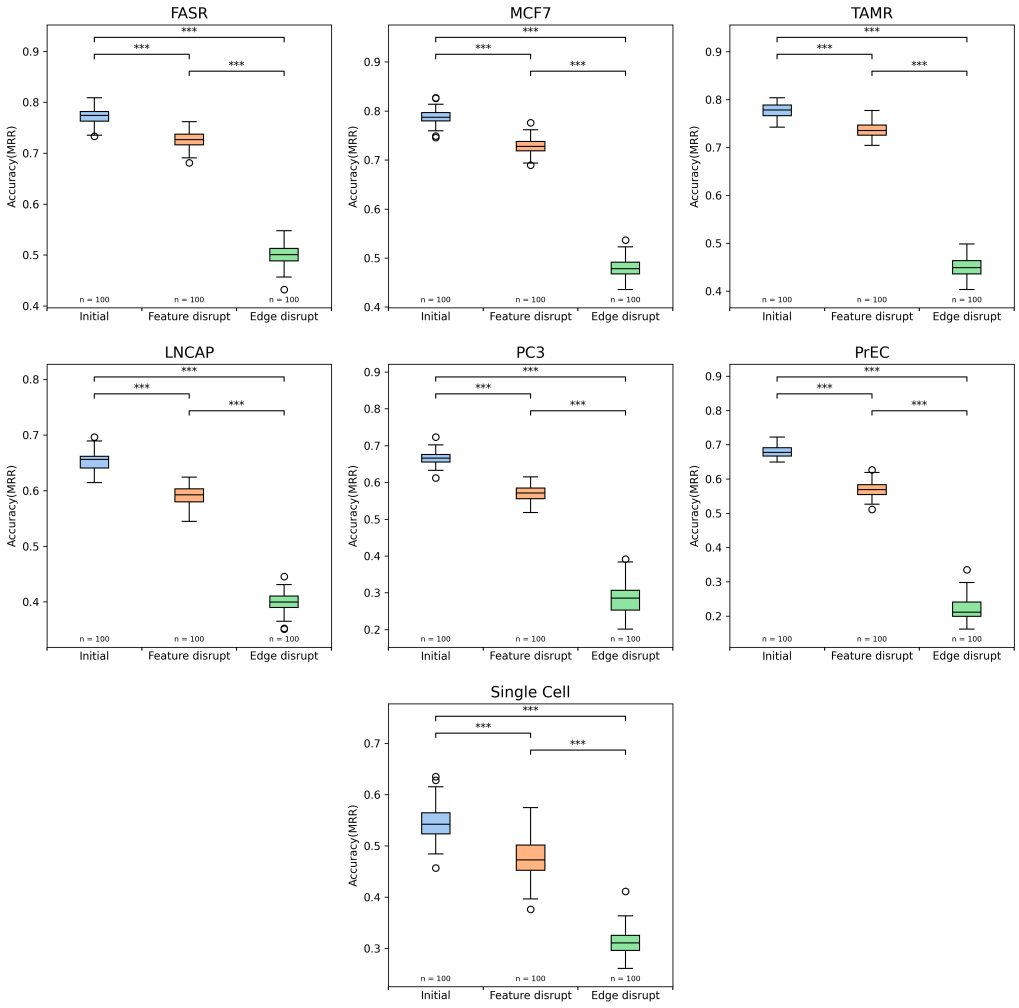

Figure 5.6. Accuracy distribution and significance in differences in Breast cancer cell lines; FASR, MCF7, TAMR, prostate cancer cell lines; LNCAP, PC3, PrEC and single cell lines. Asterisk '***' represent pvalue<0.001

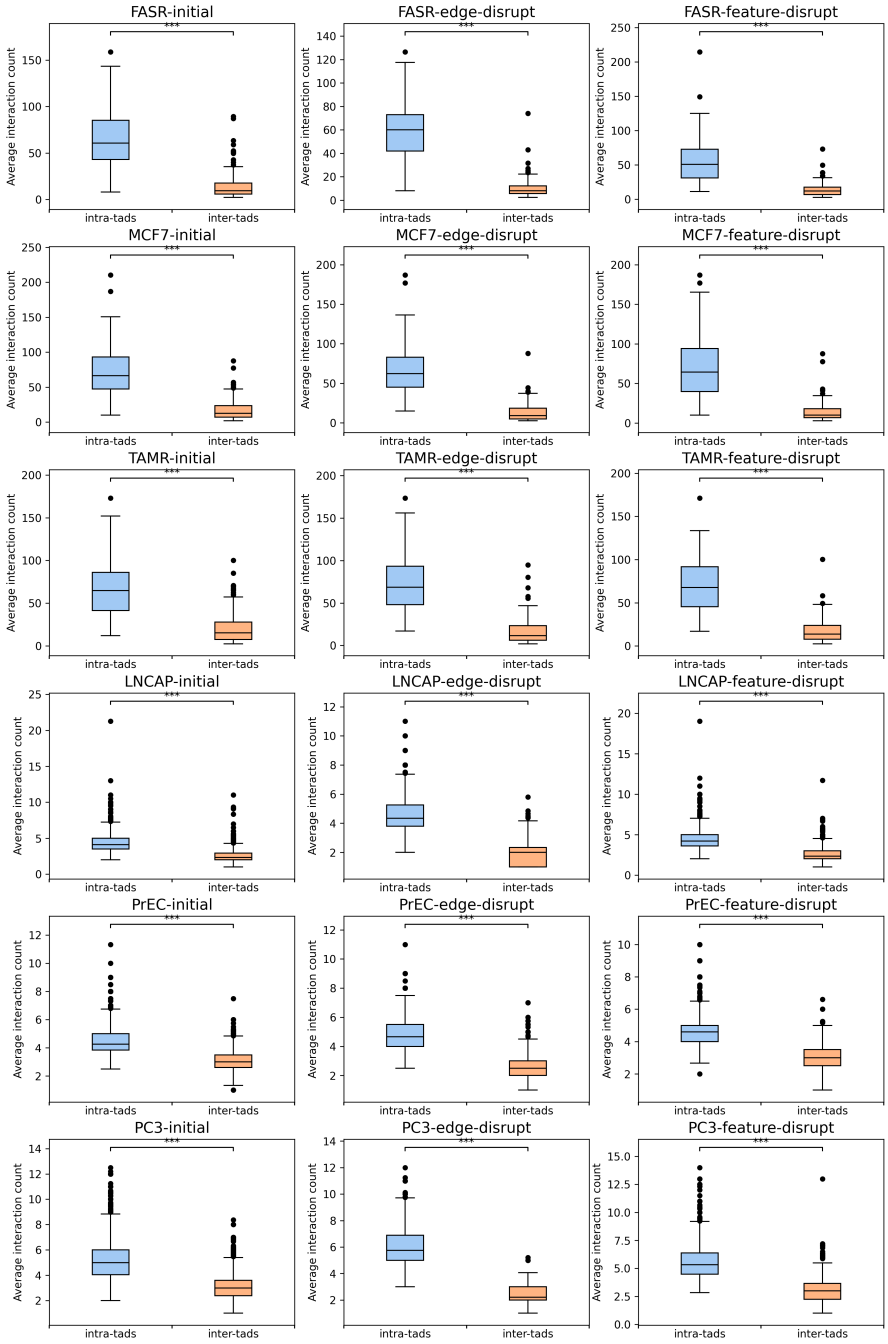

Figure 5.7. Intra-TAD interactions vs Inter-TAD interactions in Breast cancer and Prostate cancer cell lines. Asterisk '***' represent pvalue<0.001

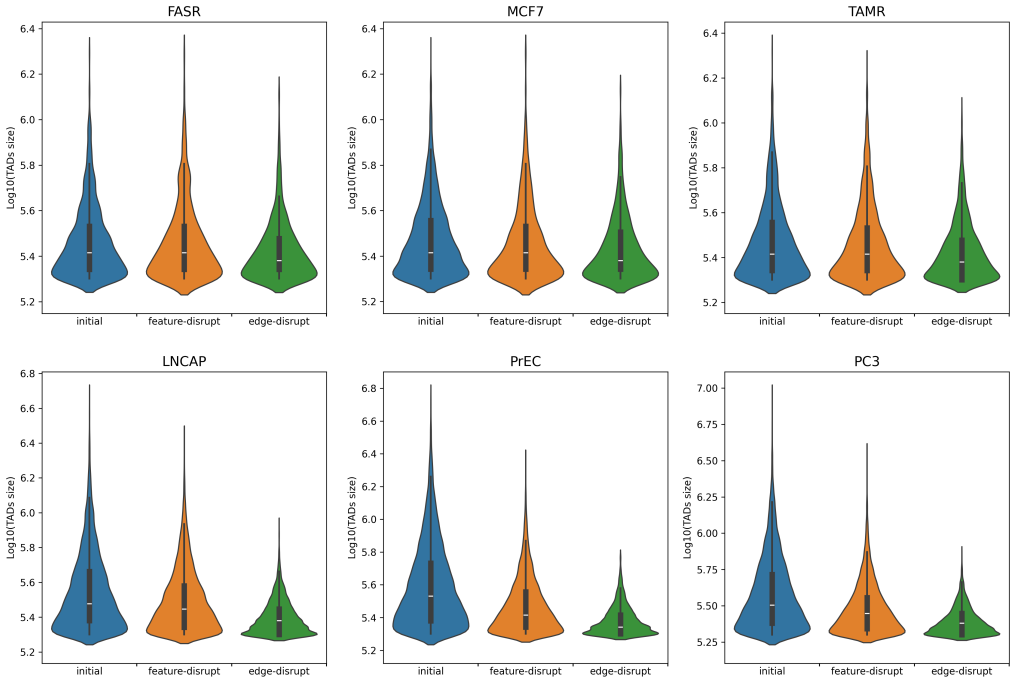

Figure 5.8. Sizes of TAD like domains.

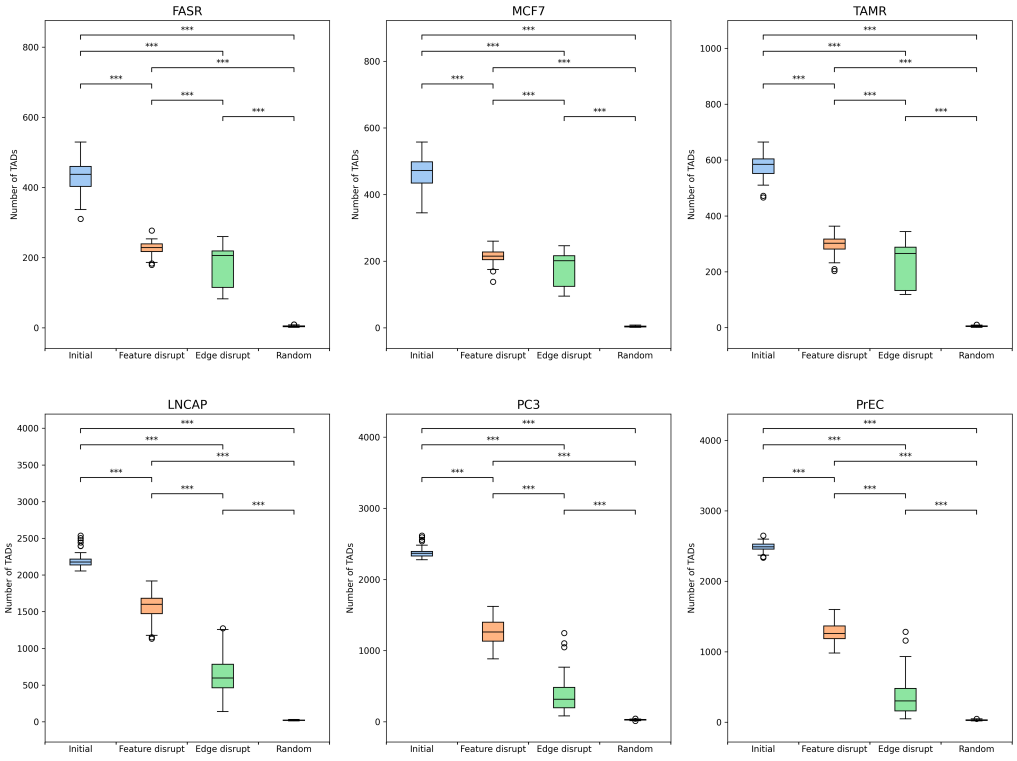

Figure 5.9. Number of identified TAD like domains. Asterisk '***' represent pvalue<0.001

6. CONCLUSION AND FUTURE WORK

Chromatin structural data, when integrated into advanced computational strategies, holds the potential to uncover biological insights beyond the reach of experimental technologies and traditional methodologies. Experimental data related to chromatin interactions, generated by technologies such as Hi-C, can reveal regulatory mechanisms within the genome and their impact on cellular functionality. These findings can uncover the role of chromatin organization in the development of genetic diseases such as cancer. However, this data often consists of various limitations, including low signal-to-noise ratio, data complexity, and sparsity. To address these challenges, further research is required to explore and experiment with different computational algorithms and data structures utilizing the existing datasets. In addition, in comparison to other omics datasets, chromatin interaction information, produced through specialized experimental techniques like 3C-based methods, provides distinct data representations, formats, and visualizations. This enables the exploration of advanced computational strategies, algorithms and data structures suitable to handle those types of data. In this dissertation, we introduce novel computational methodologies and tools designed to enhance the processing pipeline for chromatin interaction data by applying data mining and machine learning principles, taking into account the characteristics of the data and its representations.

Analyzing chromatin interactions in single cells is crucial for capturing cellular heterogeneity and understanding the structural variations across different cell types, along with their impact on cell growth and development. However, chromatin interactions data in single cells are extremely sparse which makes it challenging for downstream processing. To address this limitation, we conducted a study to curate significant inter-chromosomal interactions using single-cell interaction data generate through Hi-C technology. We were the first to implement a computational strategy for this purpose along with a publicly accessible tool. In our proposed methodology, we represented inter-chromosomal interactions as a network and identified significant interactions through statistical measurement. Additionally, we demonstrated the biological significance of the interactions identified. Due to the limited availability of similar tools, we perform benchmark using tools designed for Bulk datasets, highlighting the need for developing tools that are designed for specific

data types to uncover distinctive hidden patterns. However, the availability of single-cell Hi-C data is limited, and such datasets are typically sparse. Despite our tool's ability to detect interactions in high resolution, the sparse nature of existing datasets limited further experiments on that. Thus, we recommend the development of imputation-based computational methods to analyze chromatin interactions at the single-cell level. However, our tool proves valuable for extracting significant inter-chromosomal interactions and serves as foundational for future studies.

Another issue we identified relates to the initial stages of the Hi-C processing pipeline, specifically the alignment of raw reads with the reference genome. Most tools focus only on uniquely mapped reads, leading to the discard of a large portion of multi-mapped reads, which could represent a significant fraction of the raw data. Hi-C data possess distinct characteristics that allow for the development of tools aimed at curating multi-mapped reads. Accordingly, we introduced a computational approach to recover multi-mapped reads by utilizing a heuristic method that takes the proximity to the restriction enzyme in Hi-C experiments into account. This approach has been integrated into existing pipelines, demonstrating that the recovery of multi-mapped reads improves Hi-C processing workflows. Moreover, it could be beneficial to analyze the potential for recovering unaligned reads by leveraging features specific to chromatin interaction data.

Building upon our previous research efforts aimed at developing computational tools for analyzing single-cell resolution data, we implemented an imputation algorithm based on nearest neighbors to derive significant inter-chromosomal chromatin interactions at the single-cell level. This approach consists of three main steps: imputation, normalization, and filtering for significant interactions. Given the extreme sparsity of single-cell chromatin interactions as observed in the existing datasets, we applied a method based on k-nearest neighbors for imputation, combined with statistical tests, to extract significant interactions. At the time of this implementation, there was only a single computational tool available serving this specific need, which we used as a benchmark. However, the existing tool was designed to identify chromatin interactions with high resolution and consists of data constraints. Considering the limitations associate with the available single cell datasets, there is a clear need for a more robust computational approach to identify interactions at a more relaxed resolution suitable for most of the existing datasets. Our proposed methodology demonstrates higher performance in detecting more meaningful and significant interactions when compared to the existing tool, especially when evaluating with other omics data. This method

holds potential for further enhancement, including the ability to derive differential interactions across different cell types and to integrate with other genomic data types such as RNA-seq and ATAC-seq, offering a more comprehensive understanding of regulatory elements.

In the final section, we analyzed the interplay between chromatin interaction data and epigenetics. Chromatin interaction information and epigenetic modifications are correlated, impacting gene regulation and genomic functionality. However, the significance of these distinct data types remains unclear and presents challenges for assessment using only experimental techniques. We integrate these two data types using a graph embedding approach and assessed the impact of these elements by disrupting the graph structure. Our results suggest that chromatin interaction data could be crucial in driving genomic functionality, while epigenetics might be modulating these interactions, subsequently influencing cellular functions. Our evaluation was based solely on statistical measurements, without incorporating biological interpretation due to limitations in experimental data. Thus, additional research is necessary to integrate these findings with biological insights. Furthermore, although the generated embeddings were proven useful, their precise interpretation remains to be discovered.

We anticipate these proposed methods along with publicly accessible tools, to lay the foundation for future studies on chromatin architecture and its association with cellular functions, leading to the identification of new therapeutic approaches for genetic disorders and diseases.

REFERENCES

[1] Joanna Achinger-Kawecka, Fatima Valdes-Mora, Phuc-Loi Luu, Katherine A Giles, C Elizabeth Caldon, Wenjia Qu, Shalima Nair, Sebastian Soto, Warwick J Locke, Nicole S Yeo-Teh, et al. Epigenetic reprogramming at estrogen-receptor binding sites alters 3d chromatin landscape in endocrine-resistant breast cancer. *Nature communications*, 11(1):320, 2020.

[2] Kadir C Akdemir, Victoria T Le, Justin M Kim, Sarah Killcoyne, Devin A King, Ya-Ping Lin, Yanyan Tian, Akira Inoue, Samirkumar B Amin, Frederick S Robinson, et al. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature genetics*, 52(11):1178–1188, 2020.

[3] Chiara Anania and Darío G Lupiáñez. Order and disorder: abnormal 3d chromatin organization in human disease. *Briefings in Functional Genomics*, 19(2):128–138, 2020.

[4] Haitham Ashoor, Xiaowen Chen, Wojciech Rosikiewicz, Jiahui Wang, Albert Cheng, Ping Wang, Yijun Ruan, and Sheng Li. Graph embedding and unsupervised learning predict genomic sub-compartments from hic chromatin interaction data. *Nature communications*, 11(1):1173, 2020.

[5] Yaser Atlasi and Hendrik G Stunnenberg. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, 18(11):643–658, 2017.

[6] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014.

[7] Ferhat Ay, Evelien M Bunnik, Nelle Varoquaux, Sebastiaan M Bol, Jacques Prudhomme, Jean-Philippe Vert, William Stafford Noble, and Karine G Le Roch. Three-dimensional modeling of the p. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome research*, 24(6):974–988, 2014.

[8] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

[9] Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, 18(1):107, 2011.

[10] Jonathan A Beagan and Jennifer E Phillips-Cremins. On the existence and functionality of topologically associating domains. *Nature genetics*, 52(1):8–16, 2020.

[11] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi–c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, 2012.

[12] Shay Ben-Elazar, Zohar Yakhini, and Itai Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic acids research*, 41(4):2191–2201, 2013.

[13] Rameen Beroukhim, Xiaoyang Zhang, and Matthew Meyerson. Copy number alterations unmasked as enhancer hijackers. *Nature genetics*, 49(1):5–6, 2017.

[14] Saul A Bert, Mark D Robinson, Dario Strbenac, Aaron L Statham, Jenny Z Song, Toby Hulf, Robert L Sutherland, Marcel W Coolen, Clare Stirzaker, and Susan J Clark. Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer cell*, 23(1):9–22, 2013.

[15] Alistair N Boettiger, Bogdan Bintu, Jeffrey R Moffitt, Siyuan Wang, Brian J Beliveau, Geoffrey Fudenberg, Maxim Imakaev, Leonid A Mirny, Chao-ting Wu, and Xiaowei Zhuang. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586):418–422, 2016.

[16] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, et al. Multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3):557–572, 2017.

[17] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[18] Chanaka Bulathsinghalage and Lu Liu. Network-based method for regions with statistically frequent interchromosomal interactions at single-cell resolution. *BMC bioinformatics*, 21(14):1–15, 2020.

[19] Chanaka Bulathsinghalage and Lu Liu. A heuristic strategy for multi-mapping reads to enhance hi-c data. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1–6. IEEE, 2021.

[20] Chanaka Bulathsinghalage and Lu Liu. schi-cnn: a computational method for statistically significant single-cell hi-c chromatin interactions with nearest neighbors. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 94–99. IEEE, 2023.

[21] Mark Carty, Lee Zamparo, Merve Sahin, Alvaro González, Raphael Pelossof, Olivier Elemento, and Christina S Leslie. An integrated model for detecting significant chromatin interactions from high-resolution hi-c data. *Nature communications*, 8(1):15454, 2017.

[22] Giancarlo Castellano, François Le Dily, Antonio Hermoso Pulido, Miguel Beato, and Guglielmo Roma. Hic-inspector: a toolkit for high-throughput chromosome capture data. *bioRxiv*, page 020636, 2015.

[23] Sumantra Chatterjee and Nadav Ahituv. Gene regulatory elements, major drivers of human disease. *Annual review of genomics and human genetics*, 18:45–63, 2017.

[24] Jae Duk Choi and Jong-Soo Lee. Interplay between epigenetics and genetics in cancer. *Genomics & informatics*, 11(4):164, 2013.

[25] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[26] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp,

et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

[27] Yulin Dai, Chao Li, Guangsheng Pei, Xiao Dong, Guohui Ding, Zhongming Zhao, Yixue Li, and Peilin Jia. Multiple transcription factors contribute to inter-chromosomal interaction in yeast. *BMC systems biology*, 12(8):140, 2018.

[28] Cs Data61. Stellargraph machine learning library. *Publication Title: GitHub Repository. GitHub*, 2018.

[29] James R Davie and Deborah N Chadee. Regulation and regulatory parameters of histone modifications. *Journal of cellular biochemistry*, 72(S30–31):203–213, 1998.

[30] Ann Dean. In the loop: long range chromatin interactions and gene regulation. *Briefings in functional genomics*, 10(1):3–10, 2011.

[31] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4d nucleome project. *Nature*, 549(7671):219–226, 2017.

[32] Job Dekker and Edith Heard. Structural and functional diversity of topologically associating domains. *FEBS letters*, 589(20):2877–2884, 2015.

[33] Job Dekker and Tom Misteli. Long-range chromatin interactions. *Cold Spring Harbor perspectives in biology*, 7(10):a019356, 2015.

[34] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[35] Yue Deng. Recommender systems based on graph embedding techniques: A review. *IEEE Access*, 10:51587–51633, 2022.

[36] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

[37] Jesse R Dixon, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nature genetics*, 50(10):1388–1398, 2018.

[38] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[39] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363, 2010.

[40] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell systems*, 3(1):99–101, 2016.

[41] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas SP Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell systems*, 3(1):95–98, 2016.

[42] Emmanuelle Fabre and Christophe Zimmer. From dynamic chromatin architecture to dna damage repair and back. *Nucleus*, 9(1):161–170, 2018.

[43] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9(1):14, 2014.

[44] William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suvà, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016.

[45] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandão, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110, 2017.

[46] Jean-Philippe Fortin and Kasper D Hansen. Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16:1–23, 2015.

[47] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15(9):2038–2049, 2016.

[48] G Fundenberg, G Getz, M Meyerson, and L Mirny. High-order chromatin architecture determines the landscape of chromosomal alterations in cancer. *Nature precedings, hdl*, 10101, 2011.

[49] Luca Giorgetti, Rafael Galupa, Elphège P Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, 2014.

[50] Yixiao Gong, Charalampos Lazaris, Theodore Sakellaropoulos, Aurelie Lozano, Prabhanjan Kambadur, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos. Stratification of tad boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature communications*, 9(1):542, 2018.

[51] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018.

[52] Stefan Gröschel, Mathijs A Sanders, Remco Hoogenboezem, Elzo de Wit, Britta AM Bouwman, Claudia Erpelinck, Vincent HJ van der Velden, Marije Havermans, Roberto Avellino, Kirsten van Lom, et al. A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell*, 157(2):369–381, 2014.

[53] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[54] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[55] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1):55–86, 2007.

[56] Man-Hyuk Han, Dariya Issagulova, and Minhee Park. Interplay between epigenome and 3d chromatin structure. *BMB reports*, 56(12):633, 2023.

[57] Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[58] Rajini R Haraksingh and Michael P Snyder. Impacts of variation in the human genome on gene regulation. *Journal of molecular biology*, 425(21):3970–3977, 2013.

[59] Chao He, Xiaowo Wang, and Michael Q Zhang. Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. *Nucleic acids research*, 42(11):6935–6944, 2014.

[60] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[61] Vincent J Henry, Anita E Bandrowski, Anne-Sophie Pepin, Bruno J Gonzalez, and Arnaud Desfeux. Omictools: an informative directory for multi-omic data analysis. *Database*, 2014, 2014.

[62] Erica M Hildebrand and Job Dekker. Mechanisms and functions of chromosome compartmentalization. *Trends in biochemical sciences*, 45(5):385–396, 2020.

[63] Denes Hnisz, Jurian Schuijers, Charles H Li, and Richard A Young. Regulation and dysregulation of chromosome structure in cancer. *Annual Review of Cancer Biology*, 2:21–40, 2018.

[64] Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, 2016.

[65] Hao Hong, Shuai Jiang, Hao Li, Guifang Du, Yu Sun, Huan Tao, Cheng Quan, Chenghui Zhao, Ruijiang Li, Wanying Li, et al. Deephic: A generative adversarial network for enhancing hi-c data resolution. *PLOS Computational Biology*, 16(2):e1007287, 2020.

[66] Françoise S Howe, Harry Fischl, Struan C Murray, and Jane Mellor. Is h3k4me3 instructive for transcription activation? *Bioessays*, 39(1):1–12, 2017.

[67] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, 9(1):e1002893, 2013.

[68] Tzu-Chiao Hung, David M Kingsley, and Alistair N Boettiger. Boundary stacking interactions enable cross-tad enhancer–promoter communication during limb development. *Nature Genetics*, pages 1–9, 2024.

[69] Yih-Chii Hwang, Chiao-Feng Lin, Otto Valladares, John Malamon, Pavel P Kuksa, Qi Zheng, Brian D Gregory, and Li-San Wang. Hippie: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics*, 31(8):1290–1292, 2014.

[70] Yih-Chii Hwang, Chiao-Feng Lin, Otto Valladares, John Malamon, Pavel P Kuksa, Qi Zheng, Brian D Gregory, and Li-San Wang. Hippie: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics*, 31(8):1290–1292, 2015.

[71] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999, 2012.

[72] Eva Jablonka and Marion J Lamb. The changing concept of epigenetics. *Annals of the New York Academy of Sciences*, 981(1):82–96, 2002.

[73] Sanne M Janssen and Matthew C Lorincz. Interplay between chromatin marks in development and disease. *Nature Reviews Genetics*, 23(3):137–153, 2022.

[74] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290, 2013.

[75] Peter A Jones and Stephen B Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007.

[76] Daniel Jost and Cédric Vaillant. Epigenomics in 3d: importance of long-range spreading and specific interactions in epigenomic maintenance. *Nucleic acids research*, 46(5):2252–2264, 2018.

[77] Inkyung Jung, Anthony Schmitt, Yarui Diao, Andrew J Lee, Tristin Liu, Dongchan Yang, Catherine Tan, Junghyun Eom, Marilynn Chan, Sora Chee, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature genetics*, 51(10):1442–1449, 2019.

[78] Vera B Kaiser and Colin A Semple. When tads go bad: chromatin structure and nuclear organisation in human disease. *F1000Research*, 6, 2017.

[79] Balaraman Kalyanaraman, Gang Cheng, Micael Hardy, Olivier Ouari, Marcos Lopez, Joy Joseph, Jacek Zielonka, and Michael B Dwinell. A review of the basics of mitochondrial bioenergetics, metabolism, and related signaling pathways in cancer cells: Therapeutic targeting of tumor mitochondria with lipophilic cationic compounds. *Redox biology*, 14:316–327, 2018.

[80] Omar L Kantidze, Katerina V Gurova, Vasily M Studitsky, and Sergey V Razin. The 3d genome as a target for anticancer therapy. *Trends in molecular medicine*, 26(2):141–149, 2020.

[81] Omar L Kantidze, Artem V Luzhin, Ekaterina V Nizovtseva, Alfiya Safina, Maria E Valieva, Arkadiy K Golov, Artem K Velichko, Alexander V Lyubitelev, Alexey V Feofanov, Katerina V Gurova, et al. The anti-cancer drugs curaxins target spatial genome organization. *Nature communications*, 10(1):1441, 2019.

[82] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

[83] Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matt Schwartz, Charles W Sugnet, Daryl J Thomas, et al. The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.

[84] Stefanie Kaufmann, Christiane Fuchs, Mariya Gonik, Ekaterina E Khrameeva, Andrey A Mironov, and Dmitrij Frishman. Inter-chromosomal contact networks provide insights into mammalian chromatin organization. *PloS one*, 10(5):e0126125, 2015.

[85] Arya Kaul, Sourya Bhattacharyya, and Ferhat Ay. Identifying statistically significant chromatin contacts from hi-c data with fithic2. *Nature protocols*, 15(3):991–1012, 2020.

[86] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS computational biology*, 16(9):e1008173, 2020.

[87] Nils Krietenstein, Sameer Abraham, Sergey V Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A Mirny, et al. Ultrastructural details of mammalian chromosome architecture. *Molecular cell*, 78(3):554–565, 2020.

[88] Kai Kruse, Sven Sewitz, and M Madan Babu. A complex network framework for unbiased statistical analyses of dna–dna contact maps. *Nucleic acids research*, 41(2):701–710, 2012.

[89] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[90] Bryan R Lajoie, Job Dekker, and Noam Kaplan. The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75, 2015.

[91] Kun Lan, Dan-tong Wang, Simon Fong, Lian-sheng Liu, Kelvin KL Wong, and Nilanjan Dey. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, 42:1–20, 2018.

[92] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[93] Charalampos Lazaris, Stephen Kelly, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos. Hic-bench: comprehensive and reproducible hi-c data analysis designed for parameter exploration and benchmarking. *BMC genomics*, 18(1):22, 2017.

[94] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou, Sahaana Chandran, Angeline Rivkin, Anna Bartlett, Joseph R Nery, Conor Fitzpatrick, Carolyn O'Connor, Jesse R Dixon, et al. Simultaneous profiling of 3d genome structure and dna methylation in single human cells. *Nature methods*, 16(10):999–1006, 2019.

[95] Jae K Lee, Paul D Williams, and Sooyoung Cheon. Data mining in genomics. *Clinics in laboratory medicine*, 28(1):145–166, 2008.

[96] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141, 2014.

[97] Celine Lévy-Leduc, Maud Delattre, Tristan Mary-Huard, and Stéphane Robin. Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics*, 30(17):i386–i392, 2014.

[98] Daofeng Li, Deepak Purushotham, Jessica K Harrison, Silas Hsu, Xiaoyu Zhuo, Changxu Fan, Shane Liu, Vincent Xu, Samuel Chen, Jason Xu, et al. Washu epigenome browser update 2022. *Nucleic acids research*, 50(W1):W774–W781, 2022.

[99] Guohong Li and Danny Reinberg. Chromatin higher-order structures and gene regulation. *Current opinion in genetics & development*, 21(2):175–186, 2011.

[100] Guoliang Li, Xiaoan Ruan, Raymond K Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, Yufen Goh, Joanne Lim, Jingyao Zhang, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, 2012.

[101] Pingjing Li, Hong Liu, Jialiang Sun, Jianguo Lu, and Jian Liu. Hibrowser: an interactive and dynamic browser for synchronous hi-c data visualization. *Briefings in Bioinformatics*, 24(5):bbad283, 2023.

[102] Xinjun Li, Fan Feng, Hongxi Pu, Wai Yan Leung, and Jie Liu. schictools: a computational toolbox for analyzing single-cell hi-c data. *PLoS computational biology*, 17(5):e1008978, 2021.

[103] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[104] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[105] Dejun Lin, Giancarlo Bonora, Galip Gürkan Yardımcı, and William S Noble. Computational methods for analyzing and modeling genome structure and organization. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 11(1):e1435, 2019.

[106] Jie Liu, Dejun Lin, Galip Gürkan Yardımcı, and William Stafford Noble. Unsupervised embedding of single-cell hi-c data. *Bioinformatics*, 34(13):i96–i104, 2018.

[107] Lu Liu and Jianhua Ruan. Utilizing networks for differential analysis of chromatin interactions. *Journal of bioinformatics and computational biology*, 15(06):1740008, 2017.

[108] Qiao Liu, Hairong Lv, and Rui Jiang. hicgan infers super resolution hi-c data with generative adversarial networks. *Bioinformatics*, 35(14):i99–i107, 2019.

[109] Tong Liu and Zheng Wang. Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data. *Bioinformatics*, 35(21):4222–4228, 2019.

[110] Tong Liu and Zheng Wang. Deepchia-pet: Accurately predicting chia-pet from hi-c and chip-seq with deep dilated networks. *PLOS Computational Biology*, 19(7):e1011307, 2023.

[111] Aaron TL Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):258, 2015.

[112] Xin Luo, Yuting Liu, Dachang Dang, Ting Hu, Yingping Hou, Xiaoyu Meng, Fengyun Zhang, Tingting Li, Can Wang, Min Li, et al. 3d genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell*, 184(3):723–740, 2021.

[113] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of

topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[114] Jian Ma and Zhijun Duan. Replication timing becomes intertwined with 3d genome organization. *Cell*, 176(4):681–684, 2019.

[115] Borbala Mifsud, Inigo Martincorena, Elodie Darbo, Robert Sugar, Stefan Schoenfelder, Peter Fraser, and Nicholas M Luscombe. Gothic, a probabilistic model to resolve complex biases and to identify real interactions in hi-c data. *PloS one*, 12(4):e0174744, 2017.

[116] T Mohandas, RS Sparkes, and LJ Shapiro. Reactivation of an inactive human x chromosome: evidence for x inactivation by dna methylation. *Science*, 211(4480):393–396, 1981.

[117] Tapan Kumar Mohanta, Awdhesh Kumar Mishra, and Ahmed Al-Harrasi. The 3d genome: from structure to function. *International Journal of Molecular Sciences*, 22(21):11585, 2021.

[118] Raphaël Mourad and Olivier Cuvier. Predicting the spatial organization of chromosomes using epigenetic data. *Genome biology*, 16:1–3, 2015.

[119] Ryan M Mulqueen, Dmitry Pokholok, Brendan L O'Connell, Casey A Thornton, Fan Zhang, Brian J O'Roak, Jason Link, Galip Gürkan Yardımcı, Rosalie C Sears, Frank J Steemers, et al. High-content single-cell combinatorial indexing. *Nature biotechnology*, 39(12):1574–1580, 2021.

[120] Rabih Murr. Interplay between different epigenetic modifications and mechanisms. *Advances in genetics*, 70:101–141, 2010.

[121] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59, 2013.

[122] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61, 2017.

[123] Paul A Northcott, Catherine Lee, Thomas Zichner, Adrian M Stütz, Serap Erkek, Daisuke Kawauchi, David JH Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, et al. Enhancer

hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–434, 2014.

[124] Petr Novak, Taylor Jensen, Marc M Oshiro, George S Watts, Christina J Kim, and Bernard W Futscher. Agglomerative epigenetic aberrations are a common event in human breast cancer. *Cancer research*, 68(20):8616–8625, 2008.

[125] Oluwatosin Oluwadare and Jianlin Cheng. Clustertad: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from hi-c data. *BMC bioinformatics*, 18:1–14, 2017.

[126] Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M Flyamer, Aleksandra A Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A Oksuz, and Sergey V Venev. Cooltools: enabling high-resolution hi-c analysis in python. *BioRxiv*, pages 2022–10, 2022.

[127] Vera Pancaldi. Network models of chromatin structure. *Current Opinion in Genetics & Development*, 80:102051, 2023.

[128] Vera Pancaldi, Enrique Carrillo-de Santa-Pau, Biola Maria Javierre, David Juan, Peter Fraser, Mikhail Spivakov, Alfonso Valencia, and Daniel Rico. Integrating epigenomic data and 3d genomic structure with a new measure of chromatin assortativity. *Genome biology*, 17:1–19, 2016.

[129] Jonas Paulsen, Geir Kjetil Sandve, Sveinung Gundersen, Tonje G Lien, Kai Trengereid, and Eivind Hovig. Hibrowse: multi-purpose statistical analysis of genome-wide chromatin 3d organization. *Bioinformatics*, 30(11):1620–1622, 2014.

[130] Cheng Peng, Liang-Yu Fu, Peng-Fei Dong, Zhi-Luo Deng, Jian-Xin Li, Xiao-Tao Wang, and Hong-Yu Zhang. The sequencing bias relaxed characteristics of hi-c derived data and implications for chromatin 3d modeling. *Nucleic acids research*, 41(19):e183–e183, 2013.

[131] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[132] Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, 2004.

[133] AS Quina, M Buschbeck, and L Di Croce. Chromatin structure and epigenetics. *Biochemical pharmacology*, 72(11):1563–1569, 2006.

[134] Sehrish Rafique, Jeremy S Thomas, Duncan Sproul, and Wendy A Bickmore. Estrogen-induced chromatin decondensation and nuclear re-organization linked to regional epigenetic regulation in breast cancer. *Genome biology*, 16:1–19, 2015.

[135] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263, 2017.

[136] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[137] Khalid Raza. Application of data mining in bioinformatics. *arXiv preprint arXiv:1205.1125*, 2012.

[138] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC bioinformatics*, 12(1):414, 2011.

[139] M Jordan Rowley, M Hafiz Rothi, Gudrun Böhmdorfer, Jan Kuciński, and Andrzej T Wierzbicki. Long-range control of gene expression via rna-directed dna methylation. *PLoS genetics*, 13(5):e1006749, 2017.

[140] Rebeca San Martin, Priyojit Das, Renata Dos Reis Marques, Yang Xu, Justin M Roberts, Jacob T Sanders, Rosela Golloshi, and Rachel Patton McCord. Chromosome compartmentalization alterations in prostate cancer cell lines model disease progression. *Journal of Cell Biology*, 221(2):e202104108, 2021.

[141] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012.

[142] Sergio Sarnataro, Andrea M Chiariello, Andrea Esposito, Antonella Prisco, and Mario Nicodemi. Structure of the human chromosome interaction network. *PloS one*, 12(11):e0188201, 2017.

[143] Satish Sati and Giacomo Cavalli. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126:33–44, 2017.

[144] Michael EG Sauria, Jennifer E Phillips-Cremins, Victor G Corces, and James Taylor. Hifive: a tool suite for easy and efficient hic and 5c data analysis. *Genome biology*, 16(1):237, 2015.

[145] Marc W Schmid, Stefan Grob, and Ueli Grossniklaus. Hicdat: a fast and easy-to-use hi-c data analysis tool. *BMC bioinformatics*, 16(1):277, 2015.

[146] Florian Schmidt, Fabian Kern, and Marcel H Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics & chromatin*, 13(1):1–17, 2020.

[147] Anthony D Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L Barr, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, 17(8):2042–2059, 2016.

[148] Yi Xiang See, Benny Zhengjie Wang, and Melissa J Fullwood. Chromatin interactions and regulatory elements in cancer: from bench to bedside. *Trends in Genetics*, 35(2):145–158, 2019.

[149] François Serra, Davide Baù, Mike Goodstadt, David Castillo, Guillaume J Filion, and Marc A Marti-Renom. Automatic analysis and 3d-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors. *PLoS computational biology*, 13(7):e1005665, 2017.

[150] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):259, 2015.

[151] Muthu K Shanmugam, Frank Arfuso, Surendar Arumugam, Arunachalam Chinnathambi, Bian Jinsong, Sudha Warrier, Ling Zhi Wang, Alan Prem Kumar, Kwang Seok Ahn, Gautam Sethi, et al. Role of novel histone modifications in cancer. *Oncotarget*, 9(13):11414, 2018.

[152] Shikhar Sharma, Theresa K Kelly, and Peter A Jones. Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36, 2010.

[153] Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116, 2012.

[154] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. Topdom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research*, 44(7):e70–e70, 2015.

[155] Neda Shokraneh, Mariam Arab, and Maxwell Libbrecht. Integrative chromatin domain annotation through graph embedding of hi-c data. *Bioinformatics*, 39(1):btac813, 2023.

[156] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348–1354, 2006.

[157] Marta Smyk, Przemyslaw Szafranski, Michał Startek, Anna Gambin, and Paweł Stankiewicz. Chromosome conformation capture-on-chip analysis of long-range cis-interactions of the sox9 promoter. *Chromosome Research*, 21:781–788, 2013.

[158] Michael Song, Xiaoyu Yang, Xingjie Ren, Lenka Maliskova, Bingkun Li, Ian R Jones, Chao Wang, Fadi Jacob, Kenneth Wu, Michela Traglia, et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nature genetics*, 51(8):1252–1262, 2019.

[159] Malte Spielmann and Stefan Mundlos. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays*, 35(6):533–543, 2013.

[160] Erik Splinter, Frank Grosveld, and Wouter de Laat. 3c technology: analyzing the spatial organization of genomic loci in vivo. In *Methods in enzymology*, volume 375, pages 493–507. Elsevier, 2003.

[161] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59, 2017.

[162] Nicolas Stransky, Céline Vallot, Fabien Reyal, Isabelle Bernard-Pierrot, Sixtina Gil Diez De Medina, Rick Segraves, Yann De Rycke, Paul Elvin, Andrew Cassidy, Carolyn Spraggon, et al. Regional copy number–independent deregulation of transcription in cancer. *Nature genetics*, 38(12):1386–1396, 2006.

[163] Inderpreet Sur and Jussi Taipale. The role of enhancers in cancer. *Nature Reviews Cancer*, 16(8):483–493, 2016.

[164] Miho M Suzuki and Adrian Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature reviews genetics*, 9(6):465–476, 2008.

[165] Phillippa C Taberlay, Joanna Achinger-Kawecka, Aaron TL Lun, Fabian A Buske, Kenneth Sabir, Cathryn M Gould, Elena Zotenko, Saul A Bert, Katherine A Giles, Denis C Bauer, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome research*, 26(6):719–731, 2016.

[166] Phillippa C Taberlay, Aaron L Statham, Theresa K Kelly, Susan J Clark, and Peter A Jones. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies dna methylation of enhancers and insulators in cancer. *Genome research*, 24(9):1421–1432, 2014.

[167] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.

[168] Hideki Tanizawa, Osamu Iwasaki, Atsunari Tanaka, Joseph R Capizzi, Priyankara Wickramasinghe, Mihee Lee, Zhiyan Fu, and Ken-ichi Noma. Mapping of long-range associations

throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research*, 38(22):8164–8177, 2010.

[169] Preeti Thareja and Rajender Singh Chhillar. A review of data mining optimization techniques for bioinformatics applications. *Int. J. Eng. Trends Technol*, 68(10):58–62, 2020.

[170] Thomas O Tolsma and Jeffrey C Hansen. Post-translational modifications and chromatin dynamics. *Essays in Biochemistry*, 63(1):89–96, 2019.

[171] Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. *Nucleic acids research*, 42(7):e52–e52, 2014.

[172] Oana Ursu, Nathan Boley, Maryna Taranova, YX Rachel Wang, Galip Gurkan Yardimci, William Stafford Noble, and Anshul Kundaje. Genomedisco: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, 34(16):2701–2707, 2018.

[173] Anne-Laure Valton and Job Dekker. Tad disruption as oncogenic driver. *Current opinion in genetics & development*, 36:34–40, 2016.

[174] Nynke L Van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. Hi-c: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, page e1869, 2010.

[175] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

[176] Marco Varrone, Luca Nanni, Giovanni Ciriello, and Stefano Ceri. Exploring chromatin conformation and gene co-expression through graph embedding. *Bioinformatics*, 36(Supplement_2):i700–i708, 2020.

[177] Shikhar Vashishth. Neural graph embedding methods for natural language processing. *arXiv preprint arXiv:1911.03042*, 2019.

[178] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234, 2016.

[179] Siyu Wang, Jinbo Xu, and Jianyang Zeng. Inferential modeling of 3d chromatin structure. *Nucleic acids research*, 43(8):e54–e54, 2015.

[180] Michael Weber, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schuebeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, 2005.

[181] Caleb Weinreb and Benjamin J Raphael. Identification of hierarchical chromatin domains. *Bioinformatics*, 32(11):1601–1609, 2015.

[182] Steven Wingett, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. Hicup: pipeline for mapping and processing hi-c data. *F1000Research*, 4, 2015.

[183] Joachim Wolff, Rolf Backofen, and Björn Grüning. Robust and efficient single-cell hi-c clustering with approximate k-nearest neighbor graphs. *Bioinformatics*, 37(22):4006–4013, 2021.

[184] Joachim Wolff, Leily Rabbani, Ralf Gilsbach, Gautier Richard, Thomas Manke, Rolf Backofen, and Björn A Grüning. Galaxy hicexplorer 3: a web server for reproducible hi-c, capture hi-c and single-cell hi-c data analysis, quality control and visualization. *Nucleic acids research*, 48(W1):W177–W184, 2020.

[185] Xinyu Wu, Anlan Jiang, Jixin Wang, Shiyang Song, Yaping Xu, Qian Tang, Shirong Zhang, Bing Xia, Xueqin Chen, Shenglin Ma, et al. Lungcancer3d: A comprehensive database for integrating lung cancer chromatin architecture with other multi-omics. *bioRxiv*, pages 2021–10, 2021.

[186] Qing Xie, Chenggong Han, Victor Jin, and Shili Lin. Hicimpute: A bayesian hierarchical model for identifying structural zeros and enhancing single cell hi-c data. *PLoS computational biology*, 18(6):e1010129, 2022.

[187] Kyle Xiong and Jian Ma. Revealing hi-c subcompartments by imputing inter-chromosomal chromatin interactions. *Nature communications*, 10(1):5069–5069, 2019.

[188] Koon-Kiu Yan, Galip Gürkan Yardımcı, Chengfei Yan, William S Noble, and Mark Gerstein. Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps. *Bioinformatics*, 33(14):2199–2201, 2017.

[189] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.

[190] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983, 2018.

[191] Miao Yu, Armen Abnousi, Yanxiao Zhang, Guoqiang Li, Lindsay Lee, Ziyin Chen, Rongxin Fang, Taylor M Lagler, Yuchen Yang, Jia Wen, et al. Snaphic: a computational pipeline to identify chromatin loops from single-cell hi-c data. *Nature methods*, 18(9):1056–1059, 2021.

[192] Baohong Zhang, Xiaoping Pan, George P Cobb, and Todd A Anderson. micrornas as oncogenes and tumor suppressors. *Developmental biology*, 302(1):1–12, 2007.

[193] Hongen Zhang, Paul Meltzer, and Sean Davis. Rcircos: an r package for circos 2d track plots. *BMC bioinformatics*, 14:1–5, 2013.

[194] Hui Zhang, Ruiqin Zheng, Yunlong Wang, Yu Zhang, Ping Hong, Yaping Fang, Guoliang Li, and Yuda Fang. The effects of arabidopsis genome duplication on the chromatin organization and transcriptional regulation. *Nucleic acids research*, 47(15):7857–7869, 2019.

[195] Jingyao Zhang, Huay Mei Poh, Su Qin Peh, Yee Yen Sia, Guoliang Li, Fabianus Hendriyan Mulawadi, Yufen Goh, Melissa J Fullwood, Wing-Kin Sung, Xiaoan Ruan, et al. Chia-pet analysis of transcriptional chromatin interactions. *Methods*, 58(3):289–299, 2012.

[196] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature biotechnology*, 40(2):254–261, 2022.

[197] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[198] Yan Zhang, Lin An, Jie Xu, Bo Zhang, W Jim Zheng, Ming Hu, Jijun Tang, and Feng Yue. Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature communications*, 9(1):1–9, 2018.

[199] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of computational biology*, 20(11):831–846, 2013.

[200] Ye Zheng, Ferhat Ay, and Sunduz Keles. Generative modeling of multi-mapping reads with mhi-c advances analysis of hi-c studies. *eLife*, 8:e38070, 2019.

[201] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J Sejnowski, Jesse R Dixon, and Joseph R Ecker. Robust single-cell hi-c clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, 116(28):14011–14018, 2019.

[202] Xin Zhou, Rebecca F Lowdon, Daofeng Li, Heather A Lawson, Pamela AF Madden, Joseph F Costello, and Ting Wang. Exploring long-range genome interactions using the washu epigenome browser. *Nature methods*, 10(5):375, 2013.