University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Rafał Zaborowski
Student no. 369592

Computational methods for differential analysis of chromatin contact matrices

PhD's dissertation
in COMPUTER SCIENCE

Supervisor:
dr hab. Bartosz Wilczyński
Institute of Informatics, University of Warsaw

March 2020

Supervisor’s statement

I hereby confirm that the presented thesis was prepared under my supervision and that it fulfills the requirements for the degree of PhD of Computer Science.

Date   Supervisor’s signature

Author’s statement

I hereby declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date   Author’s signature

Abstract

Computational methods for differential analysis of chromatin contact matrices

Understanding the relationships between chromatin structure and gene regulation is a fundamental problem in genetics. However, for a long time there has been little progress in the field of genome architecture except for studying low scale chromatin organization. This situation changed during last two decades due to the advancement in the development of NGS technology, which gave rise to Chromosome Conformation Capture (3C) methods. The availability of 3C techniques enabled genome organization studies on an unprecedented scale. In particular, the 3C-derived Hi-C protocol allowed researchers to interrogate millions of chromatin interactions between pairs of regions genome-wide at a very high resolution. One of the main applications of Hi-C is the differential analysis, which aim to identify the structural differences of chromatin influencing regulatory processes across various cell types, treatments or species.

In this thesis we focus on the issue of comparing Hi-C contact matrices. First, we study the problem of assessing the similarity between chromosome segmentations arising from identification of Topologically Associating Domains (TAD) - an inherent feature of mammalian Hi-C maps, which were shown to shape regulatory landscape of the genome. We present a novel distance measure called BP-score, tailored for comparison of TAD partitionings and prove that our measure satisfy metric properties. Evaluation of the BP-score on real and simulated datasets demonstrates that it performs competitive against existing approaches. Additionally, we introduce local measures of domain rearrangement and show their correlation with functional measurements.

Second, we develop a normalization-free method for discovery of Hi-C differential interactions called DiADeM. Our method introduces an intuitive definition of differential interaction, which takes into account the cross-dataset contact profile similarity. Finally, we assess DiADeMs ability to detect differential interactions using simulated contact maps and show it performs well against other available methods for Hi-C differential analysis. In summary, the tools developed by us may help researches in discovering unknown structural alterations driving regulatory mechanisms.

Streszczenie

Metody obliczeniowe w analizie różnicowej macierzy kontaktów chromatynowych

Problem zrozumienia relacji pomiędzy strukturą chromatyny, a regulacją genów ma kluczowe znaczenie w genetyce. Niestety przez wiele lat możliwe były wyłącznie badania architektury genomu w niskiej rozdzielczości lub małej skali. Sytuacja zmieniła się w ciągu ostatnich dwóch dekad głównie ze względu na postęp w rozwoju technologii NGS, która dała początek metodom 3C (ang. Chromosome Conformation Capture). Dostępność technik 3C umożliwiła badania organizacji genomu na niespotykaną dotąd skalę. W szczególności wysoko-rozdzielczy wariant metody 3C - protokół Hi-C pozwala uzyskać dane dotyczące milionów interakcji pomiędzy parami regionów chromatyny w całym genomie. Jednym z głównych zastosowań protokołu Hi-C jest analiza różnicowa, która ma na celu zidentyfikowanie różnic w strukturze chromatyny wpływających na procesy regulacji genów w różnych typach komórek, warunkach eksperymentalnych lub gatunkach.

W tej pracy koncentrujemy się na problemie porównywania macierzy kontaktów Hi-C. Po pierwsze, badamy problem oceny podobieństwa między segmentacjami chromosomów wynikającymi z identyfikacji domen topologicznych (TAD) - nieodłącznej cechy map Hi-C organizmów ssaków, które, jak wykazano, kształtują krajobraz regulacyjny genomu. Prezentujemy nową miarę odległości o nazwie BP-score, dostosowaną do porównania segmentacji TAD oraz dowodzimy, że nasza miara jest metryką. Przykładowe analizy porównawcze przeprowadzone na danych symulowanych i rzeczywistych pokazują, że odległość BP jest konkurencyjna w stosunku do innych metryk wykorzystywanych dotychczas podczas badania podobieństwa segmentacji. Dodatkowo wprowadzamy lokalne miary rearanżacji domen topologicznych i pokazujemy, że pomiary rearanżacji uzyskane przy użyciu wprowadzonych przez nas miar korelują z pomiarami ekspresji genów lub metylacji.

Po drugie, opracowujemy metodę do wykrywania różnicowych oddziaływań Hi-C o nazwie DiADeM działającą na danych nieznormalizowanych. Nasza metoda wprowadza intuicyjną definicję interakcji różnicowych, która uwzględnia podobieństwo profili kontaktów pomiędzy zestawami danych. Na koniec oceniamy zdolność naszej metody do wykrywania interakcji różnicowych przy użyciu symulowanych map kontaktów i pokazujemy, że osiąga konkurencyjne wyniki w porównaniu z innymi dostępnymi metodami służącymi do analizy różnicowej Hi-C. Podsumowując, opracowane przez nas narzędzia mogą pomóc badaczom w odkrywaniu nieznanych zmian strukturalnych wpływających na mechanizmy regulacji genów.

Keywords

Hi-C, Chromosome Segmentation, Topologically Associating Domains, Contact Matrices, Differential Analysis, Generalized Linear Models

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

Applied computing → Life and medical sciences →
→ Bioinformatics
→ Computational biology → Computational genomics
→ Genomics → Computational genomics

Tytuł pracy w języku polskim

Metody obliczeniowe w analizie różnicowej macierzy kontaktów chromatynowych

List of publications with results from the thesis:

Niskanen, Henri and Tuszynska, Irina* and **Zaborowski, Rafał*** and Heinäniemi, Merja and Ylä-Herttuala, Seppo and Wilczyński, Bartek and Kaikkonen, Minna U (2017). Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions. *Nucleic acids research*, 46(4):1724–1740. doi: 10.1093/nar/gkx1214.

Zaborowski, Rafał and Wilczyński, Bartek (2019). BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations. *Journal of Computational Biology*, 26(4):305–314. doi: 10.1089/cmb.2018.0162.

Zaborowski, Rafał and Wilczyński, Bartek (under review). DiADeM: differential analysis via dependency modelling of chromatin interactions with robust generalized linear models. *bioRxiv.* doi: https://doi.org/10.1101/654699

Acknowledgements:

First, I would like to express my gratitude to my supervisor Bartek Wilczyński for his constant support in research, optimistic attitude, patience and giving me an opportunity to explore the exciting field of Bioinformatics.

Let me thank my collaborators Henri Niskanen and Minna U. Kaikkonen from A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland for successful teamwork.

Next, I would like to express my gratitude to my family, my fiancée Justyna and all my friends for support and company during hard times, as well as many happy moments that we shared.

Finally, I would like to thank for the financial support of Polish National Center for Research and Development (ERA-NET-NEURON/10/2013 grant) and Polish National Science Centre (DEC 2015/16/W/NZ2/00314 grant).

To my grandparents, Danusia and Sławek.

Contents

CHAPTER 1
Biological Introduction

In 2003, after more than 10 years of research, the Human Genome Project was completed. The main purpose of this undertaking was to establish the sequence of human DNA and determine all of the genome fragments encoding proteins (so-called genes). The results obtained during the Human Genome Project combined with the development of Next Generation Sequencing technology gave rise to dynamic progress in genetics. To date, the sequences of approximately 100 organisms have been discovered, and sequencing has become a widely used method. Availability of reference genomes stimulated advances in phylogenetics, epigenetics, chromatin structure studies and many other branches of biology.

One of a key challenges in genetics is the study of gene regulation, which aim to demystify the circumstances that induce decoding and transfer of genetic information from certain DNA fragments. While DNA contains all the instructions required for cell functioning, it is essential to properly transfer this information in order to maintain life. In general organisms consists of various types of cells, however the genetic information contained in DNA sequence across given individual cells is almost identical. That implies the existence of mechanisms, which selectively decode DNA leading to tissue specific gene regulation and cell differentiation. Precise elucidation of gene regulation remains an open problem. Among many factors demonstrated to influence cell differentiation and functioning is the spatial organization of chromatin [ZX19].

The current chapter discusses some of the basic concepts of cell biology and methods of studying the structure of chromatin, which are relevant later in this thesis. Section 1.1 covers selected issues related to the structure and basic functions of the cell. Due to the extensiveness of the topic, it is limited to the description of eukaryotic organisms as the data analyzed in this work originates from mammalian species. Section 1.2 is a brief introduction into Next Generation Sequencing technology, which forms the basis of the chromatin structure studying techniques discussed in Section 1.3. In Section 1.4 the mechanisms of gene regulation processes and their relationship with the DNA spatial structure are described.

1.1. The Basics of Cell Biology

The cell is considered the basic building block of all living organisms and the smallest unit capable of maintaining life. From a genetic standpoint, the cell acts as a container for genome and enables the proper flow of genetic information. With respect to classification, two types of cell are distinguished: eukaryotic (with nucleus) and prokaryotic (also known as bacterial, without nucleus). This section contains basic facts about eukaryotic cells.

1.1.1 Cell Structure

Metazoan cells are complex objects composed of many elements. An outer layer of cell, which acts as barrier separating the organelles from external environment is called cell membrane. Apart from protecting the interior of the cell, another function of the membrane is to facilitate exchange of different substances, e.g. import of nutrients or export of metabolites. The core of any eukaryotic cell is the nucleus where most of genetic information is kept and transferred to daughter cells during mitosis and DNA replication or onto mRNA during transcription. Molecules of mRNA are exported from the nucleus and information is further passed to proteins inside ribosomes. Some other important organelles are mitochondria - responsible for cell respiration, or Golgi apparatus where chemical modifications of lipids and proteins takes place (Figure 1.1).

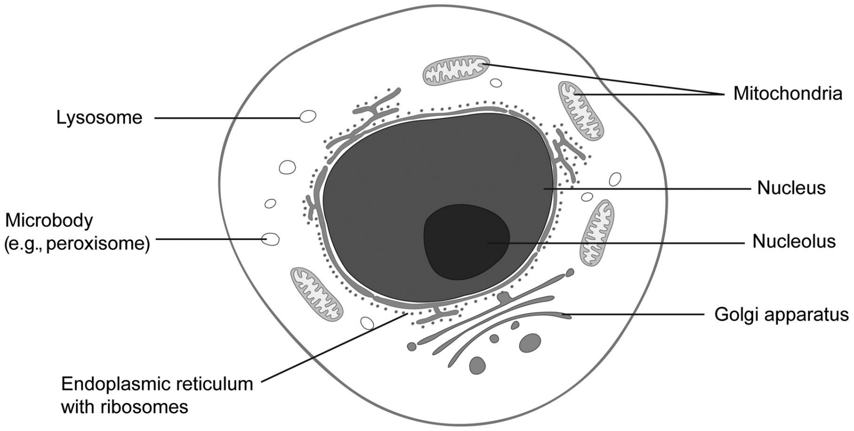


Figure 1.1. The simplified illustration of eukaryotic cell. Figure from [Gie11]. Used with permission from Oxford University Press.

1.1.2 Genome Organization

The main responsibilities of the cell nucleus is keeping the correct packing of the genetic material, i.e. the genome. The genomes of mammalian organisms consists of DNA and proteins, which form the chromatin fiber adopting complicated, hierarchical structures.

The main carrier of genetic information is the DNA (deoxyribonuceic acid) molecule. The basic compounds that make up DNA are the phosphate residue, sugar - deoxyribose, and nucleobases: Adenine, Cytosine, Guanine and Thymine, abbreviated as the first letter of their name. Each nucleobase is linked to a sugar molecule to form a nucleoside. Nucleosides combine into strand through phosphate residues. DNA usually occurs in the form of a double-stranded helix [WC+53; WSW53; FG53]. The helix structure is maintained through hydrogen bonds between complementary nucleobases of opposite strands. This means that base pairs can only form between Guanine and Cytosine or between Adenine and Thymine (Figure 1.2). The sequence of nucleobases of DNA strands determines the genetic material of an individual. Due to the function of particular fragments, DNA is divided into coding and non-coding regions. The former are matrices for protein production in a process called gene expression. The latter can perform structural functions or regulate gene expression. Other non-coding regions do not have assigned role. Under evolutionary conservation, it is estimated that about 8 to 15% of human DNA performs biochemical functions, while the total number of base pairs of coding sequences represents less than 2% of the human genome sequence [PH11; Ran+14; Con+01].

Chromosomal DNA may be a very long molecule. For example each copy of the human genome consists of approximately 3.5 billion base pairs which would measure around 2-2.4 meters in length if fully expanded while the diameter of cell nucleus is on average only a few micrometers in size. How such a long molecule fits inside such a little nucleus? It turns out that DNA binds to specific proteins to form a complex structure and achieve proper compaction level. First, the DNA double helix wraps around protein complexes called nucleosomes (Figure 1.3). Nucleosomes form so-called histone octamers, consisting of 4 pairs of histones. Short pieces of DNA between nucleosomes are called linker DNA, while the entire structure is called a chromatin fiber. Due to the degree of condensation and the associated transcriptional activity, two types of chromatin are distinguished: euchromatin and heterochromatin [Amo05]. The former is a loose, transcriptionally active structure also called a string of beads or a 11 nanometer fiber. The latter arises when the nucleosomes and linker DNA compacts forming so-called solenoid structure or 30 nanometer fiber. Chromatin can pass between both states accompanied by an increase or decrease in gene regulation respectively. The single chromatin fiber is called a chromatid. There are 24 pairs of chromatids in human cells. Each pair comprise different chromosome (or sister chromatids). In addition recent data obtained from Hi-C studies suggest a fractal structure of chromatin fiber.

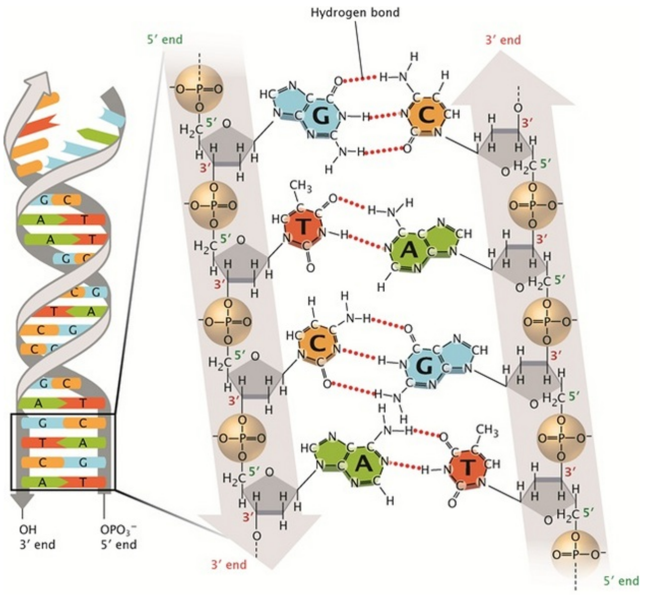


Figure 1.2. The schematic illustration of DNA double helix structure. Two sugar-phosphate strands are linked together by existence of hydrogen bonds between corresponding nucleobases. Figure from [Pra08]. Used with permission from Nature Education.

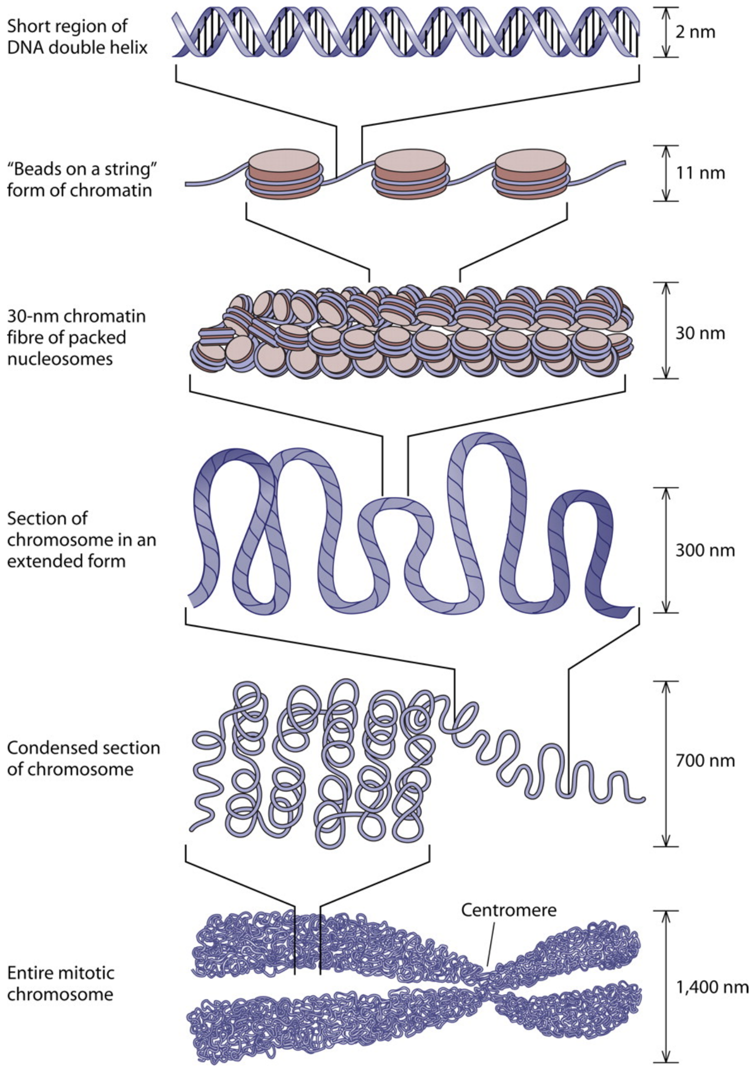

Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

Figure 1.3. The compaction of DNA into chromatin. DNA helix wraps around nucleosomes to form chromatin. The chromatin usually occupy condensed state of 30-nm fiber. This fiber is further packed into hierarchy of domains, which form chromatids and chromosome. Figure from [JV11]. Used with permission from American Society for Microbiology.

1.1.3 Central Dogma of Molecular Biology

The central dogma of molecular biology describes the flow of genetic information in a cell. Quoting Francis Crick, information can pass from nucleic acids to proteins, but once it got to protein it can not get back to DNA [Cob17]. The most important directions of genetic information flow are:

- from DNA to DNA, i.e. replication,
- from DNA to RNA, i.e. transcription,
- from RNA to protein, i.e. translation.

The flow of information in the remaining 3 directions is possible, however, it occurs only in some organisms or in specific conditions [TM+70; Ahl02; MH65].

Living organisms require their cells to divide in order to maintain life. Cell division includes copying the genetic material, i.e. DNA replication. DNA replication occurs in three stages: initiation, elongation and termination. The initiation phase consists of unwinding DNA double helix at certain chromosomal locations and constructing protein complexes called replicons. During the elongation phase, the replicon slides along DNA and copies both strands. When replication complex recognizes the termination sequence, it detaches from DNA [MS58; Pra].

One of the primary roles of DNA is to encode protein sequences, which are the basic building blocks and signaling molecules in most organisms. Proteins are synthesized from genes in two step process. First, a DNA sequence of a given gene is transferred onto a mRNA molecule during transcription. Transcription begins when the RNA polimerase and other proteins called transcription factors bind together with a certain DNA region called the gene promoter. Afterwards, the process enters the elongation phase. The double helix is unwinded and the mRNA molecule is synthesized base by base. During the termination step, transcription machinery releases DNA and strands folds back into the double helix. It must be noted that this is a brief description of transcription. The precise mechanism of this process is much more complex, involving many additional steps. The synthesized RNA undergoes various post-transcriptional processes, and then it is removed from the cell nucleus to the cytoplasm where protein synthesis takes place.

When ribosome captures mRNA, its sequence is translated into a protein. During this process, the ribosome moves along the mRNA reading subsequent nucleotide triplets (so-called codons). The amino acids used to form the peptide chain are delivered by tRNA molecules, which bind to the mRNA codon and a specific ribosome fragment. There are 2 types of codons that do not encode any amino acid - these are called start and stop codons designating the translation start and end sites respectively. During last stage of protein synthesis, the ribosome catalyzes the process of transforming the polypeptide into its native form.

Not all transcribed genes are translated to proteins. Some fragments of DNA called non-coding RNA (ncRNA) genes are responsible for synthesis of functional RNA molecules, which facilitate various cellular processes [Edd01]. Examples of ncRNAs are tRNA or ribosomal RNA mentioned above.

1.2. Next Generation Sequencing

Until the 70's, the process of discovering DNA sequences (i.e. sequencing) of even small genome organisms was severely limited. For example in 1968 Wu and Kaiser performed the first successful sequencing of lambda phage cohesive ends sequence [WK68; San01]. Researchers used tedious protocol, which allowed to establish the sequence of only 10 nucleobases length.

In 1977, Frederick Sanger and coworkers developed a method (later called Sanger sequencing), which allowed for quick and efficient sequencing of relatively large sequences [SNC77]. The method became standard in sequencing assays due to the ease of its automation, the ability to sequence long DNA fragments (up to 1000 nucleobases) and its low error rate. Sanger sequencing has been successfully applied for sequencing human mitochondrium genome (around 16.6 kbp) or phage lambda genome (around 48.5 kbp) and played a major role during Human Genome Project

[Con+01]. Despite the above mentioned advantages, the Sanger sequencing suffers from one serious flaw - it is still relatively labor intensive as it sequences one individually amplified DNA molecule at a time [SN14].

This issue led to the development of Next Generation Sequencing methods. These techniques, also called massively parallel or high-throughput, allow to sequence entire genomes at relatively low cost. For example, prior to 2008, the cost of sequencing the entire human genome using Sanger technologies was estimated at $20-25 million, which dropped to a few thousands dollars currently using NGS [Sch+19]. At present there are several commercial platforms developed for NGS methods, for example: Illumina HiSeq, Roche 454, Pacific Biosciences and many others differing in details and purpose of analysis.

However, all of them are based on similar core principles common for high-throughput methods (Figure 1.4). At the first stage, the DNA sample is randomly sheared into very short fragments (usually around 50 - 500 basepairs) and ligated with adapters (Figure 1.4A). The goal of adapters is twofold. First they serve as starting point in PCR (so called primers). Second, they hybridize with solid basis before sequencing step. Depending on the type of assay, fragments may be filtered to retain specific part of the genome like for example during Whole-Exome studies (Figure 1.4B). Remaining fragments are then amplified using PCR. Next, fragments are captured into wells, so that one well is common to a single adapter sequence. Then the sample is sequenced in a cycle-by-cycle manner (Figure 1.4C). At every cycle, the substrates are added to the sample and the newly created chain is extended by incorporating nucleotides and emitting a colored signal, which is registered. The cycle ends with a wash-out of unincorporated nucleotides. The outcome of a NGS experiment is a library of short reads (i.e. sequences), which is later mapped on reference genome to determine location of reads (Figure 1.4D).

1.3. Into Chromatin Structure

Before the discovery of 3C and derived methods, our understanding of chromatin topology remained very limited. Light microscopy studies conducted by Carl Rabl as early as the end of 19th century provided evidence of chromosome territories [Mis08; CC10]. This model assumed that chromosomes occupy separate volumes of cell nucleus rather then interweave one another's fiber. The concept of chromosome territories suggested by Carl Rabl in 1885 was eventually confirmed by Thomas and Christoph Cremer in 1980s [CC01]. Another advancement in genome architecture research was facilitated by the development of electron microscopy, which enabled the discovery of beads-on-string organization of DNA [Ann08]. In between nucleosome level and chromosome level the chromatin was observed to adopt either the condensed, repressed or relaxed, transcriptionally active states currently known as hetero- and euchromatin respectively [Lan+83; WH97; Amo05; GR05; DT12].

Current technical capabilities enable us to investigate the arrangement of chromatin at the level of long-range chromatin interactions between DNA fragments spanning as little as few hundreds of basepairs. Moreover due to rapid progress in the development of NGS technology it is possible to examine chromatin contacts genome-wide for millions of fragment pairs in single experiment.

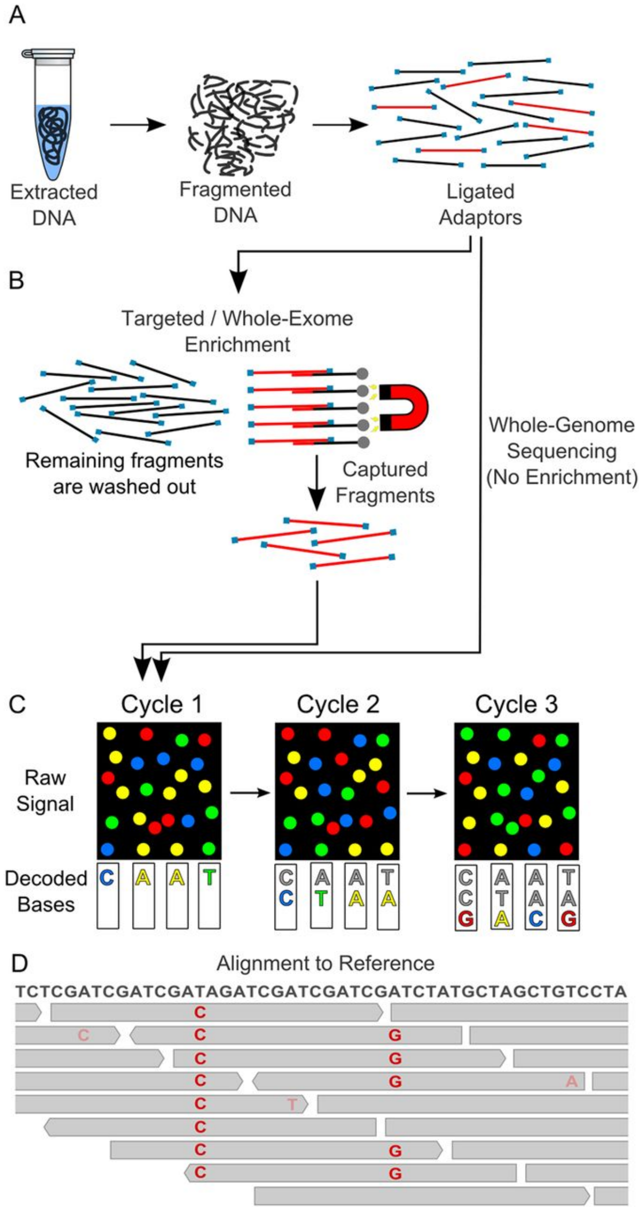

A
Extracted DNA
Fragmented DNA
Ligated Adaptors
B
Targeted / Whole-Exome Enrichment
Remaining fragments are washed out
Captured Fragments
Whole-Genome Sequencing (No Enrichment)
C
Cycle 1
Cycle 2
Cycle 3
Raw Signal
Decoded Bases
D
Alignment to Reference
TCTCGATCGATCGATAGATCGATCGATCGATCTATGCTAGCTGTCCTA

Figure 1.4. Summary of the NGS protocol. First, extracted DNA sample is fragmented into pieces and ligated with adapters. Depending on the type of assay some fragments may be filtered out. Next, the remaining DNA is amplified using PCR and captured into sequence specific wells. Finally the sequencing process is performed - bounded DNA fragments are copied by synthesizing counterparts base by base, simultaneously registering the accompanied emitted fluorescence. As a result, a library of short reads is obtained and mapped on reference genome in order to establish the most likely location of sample DNA. Figure from [SN14]. Used with permission from BMJ Publishing Group Ltd.

1.3.1 FISH Protocol

In 1969 Gall and Pardue developed foundations of techniques known as in situ hybridization [GP69]. This sort of methods allow us to localize a DNA sequence of interest in cell nuclei by hybridizing a complementary labeled sequence called the probe, which can be later identified using appropriate imaging device. Initially, radioactive probes were replaced with fluorescent ones and the resulting FISH protocol quickly became the standard in cytogenetics [O'C08].

The overall scheme of the protocol is depicted in Figure 1.5. In the beginning, probe sequences of interest are designed and labeled (Figure 1.5a, b). This step may include insertion of fluorophores (middle column) - for instant imaging or hapten (left column), that can be used to yield fluorescence at any time during later steps of the protocol. Before the hybridization step can be performed, a double helix structure must be broken into a single strand form (Figure 1.5c). After hybridization (Figure 1.5d), probes may be localized using fluorescent microscope.

FISH turned out to be extremely useful in many important experiments. For example, this method was applied to confirm the existence of chromosome territories when fluorescent probes were used to visualize individual chromosomes [SBD77; Zor+79]. FISH also played a major role in Human Genome Project during the annotation of genes on human chromosomes. Currently, FISH is often exploited in regulatory genomics to examine co-localization of regulatory elements.

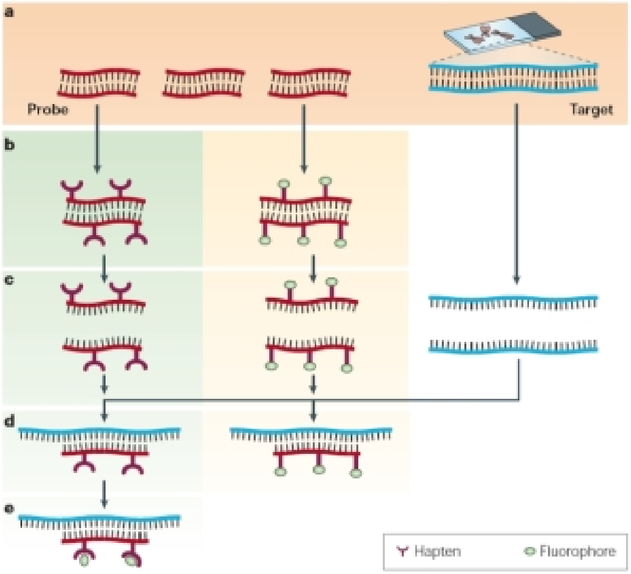


Figure 1.5. The scheme of FISH protocol. Initially, probe sequences of interest are prepared and labeled with hapten or fluorophore. Next, both the probes and target DNA sequence are melted in order to produce the single stranded form. During the subsequent hybridization, probes bind with target DNA. Finally, the probes are visualized. Figure from [SC05]. Used with permission from Springer Nature.

1.3.2 3C Protocol

Chromosome Conformation Capture (3C) is the first method used to assess genome architecture without imaging [Dek+02]. The main principles behind 3C are also common to all its derivative techniques [HZW18]. The protocol is illustrated in Figure 1.6. The first step, called chromatin cross-linking, involves covalently binding pairs of DNA regions bridged by proteins. This process is performed via treatment with formaldehyde and should only affect DNA fragments that are in close spatial proximity. Afterwards, the sample is digested with a restriction enzyme to produce short pairs of fragments. The following ligation results in chimeric molecules, which are subsequently reverse cross-linked to yield 3C templates. Under high dilution, intramolecular ligation should be favored over undesired intermolecular - leading to noise amplification. Finally, templates are interrogated with PCR or sequencing methods to quantify the frequency of interactions. In 3C, PCR primers are designed to match selected, particular ligation junction allowing to examine interactions of single pairs of DNA segments one at a time.

Therefore, 3C enables us to assess contacts in one-versus-one manner making it difficult to scale. Another bottleneck of this technique lies in its inability to detect contacts spanning more than few hundred kbp range. Fortunately, modifications of 3C can overcome such limitations.

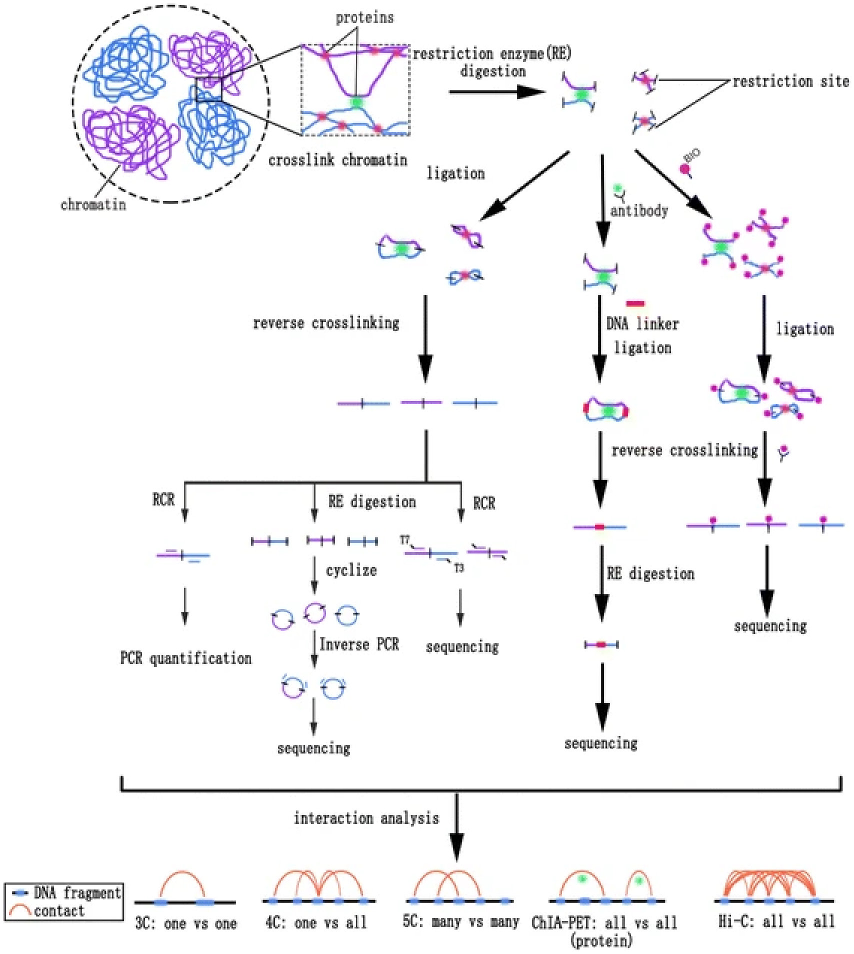


Figure 1.6. Schematic illustration of C protocols. Each presented method uses proximity based ligation to sample genomic interactions. The first step consist of cross-linking followed by restriction enzyme digestion. Next, a method dependent on ligation and subsequent reverse cross-linking is performed. Afterwards, ligation products are sheared into small hybrid fragments. Then, the obtained fragments are mapped to the reference genome using methods, that depends upon the choice of the protocol. Figure from [HZW18]. Used with permission from Springer Nature.

1.3.3 4C Protocol

A modification of the 3C technique using micro-arrays is known as 4C. This protocol allows for inspection of multiple potential interactions with a DNA segment of interest (a bait) [Sim+06]. Such type of analysis is called "one versus all". During 4C, the assay templates are subjected to a second round of restriction enzyme digestion and remaining chimeric fragments are cyclicized in a subsequent ligation step (Figure 1.6). Next, an inverse PCR with bait specific primers is performed to amplify interacting regions. Baits are designed to match the DNA sequence nearby particular restriction enzyme cleavage site. This allows to capture ligation products of prespecified DNA region. The resulting library represents the fragment's of interest genomic environment, which is hybridized to micro-array and sequenced.

1.3.4 5C Protocol

An alternative variant of 3C that allows to quantify genomic contacts in many to many manner is called 5C [Dos+06]. The difference between 3C and 5C lies in the primers preparation method. The former technique requires fragment specific primers, while the latter uses universal ones. Such modification enables to examine multiple potential interactions simultaneously. Despite the major improvement with respect to 3C, 5C is also limited in detection of interactions separated by over 1 Mbp genomic distance [HM12].

1.3.5 Hi-C Protocol

As described above, the limitations of the 3C method are partially resolved by 4C and 5C. On one hand, 4C enables us to capture interactions genome-wide no matter how big the separation between DNA segments is. On the other hand, it is limited to single viewpoint. 5C however, is able to interrogate multiple interactions at once, albeit the span of detectable contacts can not exceed 1 Mbp genomic separation.

In 2009 Lieberman-Aiden and coworkers developed Hi-C protocol overcoming both mentioned issues [LA+09]. Hi-C is a powerful method, which effectively samples millions of interactions genome-wide in fully high throughput manner without any restrictions on genomic separation between interacting regions and no need to design any specific primers.

The Hi-C technique introduces one important modification in comparison with its sister methods. Before the ligation step, biotin is introduced into restriction enzyme cleavage sites of paired fragments. Later steps involve ligation and reverse cross-linking followed by DNA shearing of restriction segments into little fragments. Finally, initial ligation products are selected by specific binding of biotin-containing chimeric molecules with streptavidin beads. The resulting library is paired-end sequenced to produce a list of interacting regions later transformed to genome-wide count matrix (Figure 1.6).

Prior to constructing contact matrices a researcher must first determine a plausible binning of the genome. Bin size measures the number of base-pairs used to divide the genomic DNA on consecutive, adjoint segments. Obtained segments (bins) are then matched with collected reads. Bin size influence the power of any Hi-C based statistical analysis and therefore reasonable choice of this parameter determines the quality of Hi-C analysis. The optimum size of the bin depends on the cutting frequency of restriction enzyme and sequencing depth. If the value is too small the noise will dominate the intrinsic signal. On the other hand, large bin sizes impede the discovery of fine scale structural features like chromatin looping. Often the term bin size is used interchangeably with resolution. However, it is important to remember that high resolution indicate small bin sizes.

Although Hi-C is the most advanced of all proximity ligation methods, it requires much effort in data processing and analysis. In particular, high resolution Hi-C datasets may need significant computational resources to convert raw data into useful contact matrices. Moreover, due to high complexity of protocol Hi-C data is accompanied with noise and biases, which are difficult to remove. All this severely impedes Hi-C based research unless appropriate algorithmic and statistical approaches are employed in order to draw relevant conclusions.

1.3.6 Capture-C Protocol

The resolution of Hi-C is limited by the minimum size of restriction fragment, which depends on the cutting frequency of the restriction enzyme used. Hi-C studies achieving the resolution as high as 1kb were already successfully conducted [Rao+14]. However, increasing Hi-C resolution is very costly and results in large datasets, which are very time-intensive to process and analyze.

In some situations, instead of conducting very high resolution Hi-C, one may choose the Capture-C assay [Hug+14]. Capture-C is a derivative of the 3C technique designed to interrogate interactions in many-to-many fashion. The difference between Capture-C and the other many-to-many method 5C is that it can relatively efficiently track contacts in very high resolutions (below 1kb), thereby allowing to examine precise interactions between regulatory elements at relatively low cost. Essentially Capture-C works by first generating a 3C library with frequently cutting restriction enzyme and then shearing the library into very small fragments of approximately 300 bp. Afterwards, the obtained fragments are hybridized with capture probes (using oligonucleotide capture technology - OCT) designed to bind in locations of interest like promoters or enhancers. The application of this protocol enables to capture very short sequences and examine which DNA fragments they interact with.

1.4. Gene Regulation

Mammals are complex organisms consisting of trillions of cells of different types. The proper functioning of such intricate systems require precise orchestration of various regulatory mechanisms. How exactly does a cell know what type of process and when to initiate or which cell type it should transform into? The annotation of genes during the Human Genome Project raised many questions regarding their function. Since then, a lot of effort have been made to understand the mechanisms driving gene regulation processes.

In eukaryotes, the gene expression control is said to be combinatorial, i.e. cell specific genes are regulated by the formation of various complexes containing combinations of multiple proteins. These processes depends on chemical modifications of chromatin, which influence its compaction and spatial structure. Moreover, gene regulation can be divided on transcriptional and post-transcriptional. The former class of processes refers to transcription mechanisms controlling the production of mRNA while the latter indicate modifications of mRNA called splicing conducted prior to protein synthesis. In particular depending on cell type or its environment splicing can proceed differently leading to alternative forms of mRNA and therefore different proteins [Phi08].

1.4.1 Regulatory Elements

The prominent majority of human DNA sequence consists of non-coding regions. Certain class of such regions called regulatory elements play a crucial role in maintenance of cell processes [MEG06]. In general, regulatory elements can be divided into two classes: cis-acting and distal (Figure 1.7a). Cis-acting elements contain recognition motifs for DNA polymerase, transcription factors and distal elements. Usually, tissue specific transcription is triggered after a DNA fragment called an enhancer binds with a gene promoter region (Figure 1.7b). At the same time, silencers linked to a promoter act as repressors of given gene. Insulators on the other hand, constrain the range of regulatory elements (Figure 1.7). They isolate different regulatory regions (Figure 1.7a) from undesired mutual influence of their elements. In contrast to cell specific regulation, a large fraction of genes must be expressed at constant level across all cells to maintain basic functions needed for living [EL13]. Those genes are called housekeeping and are considered to be unregulated [Phi08].

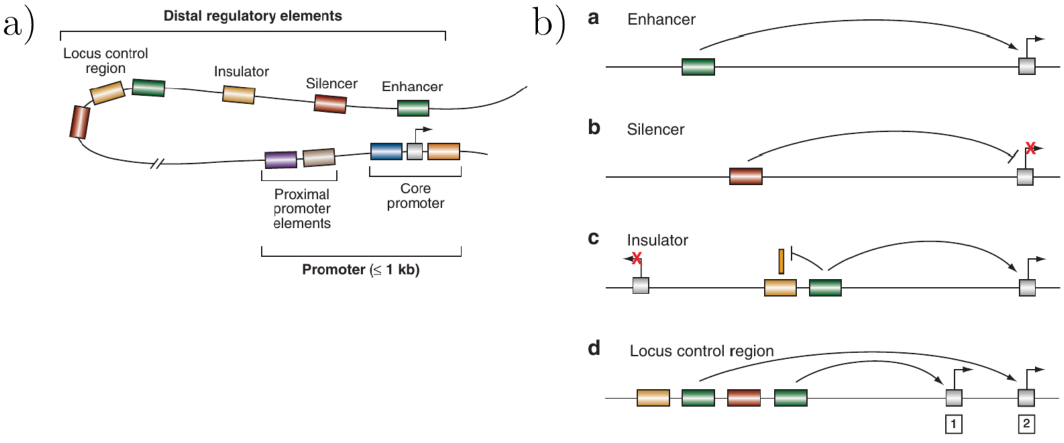


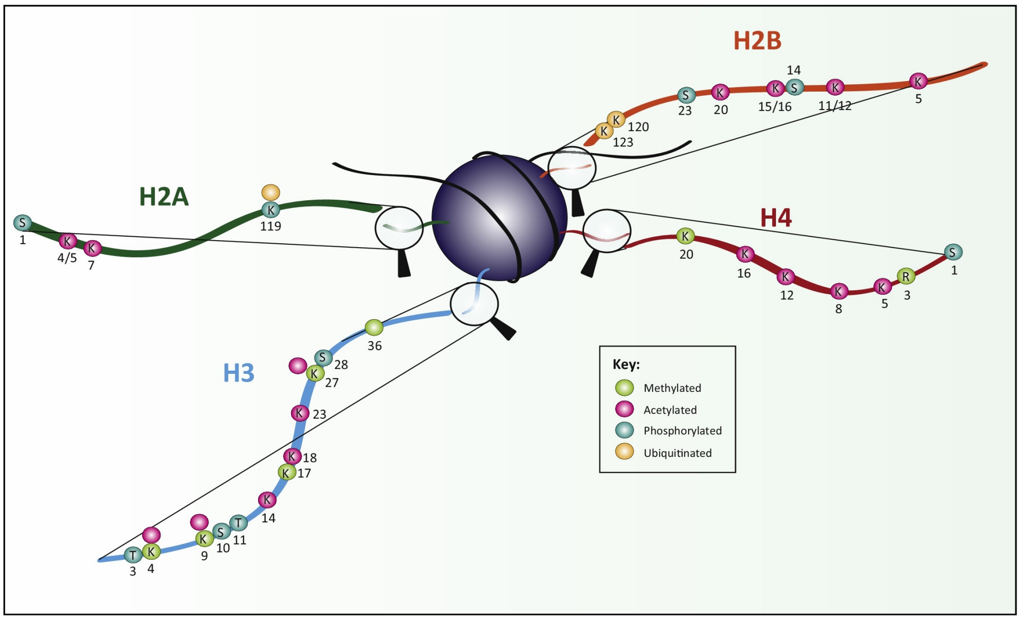
Figure 1.7. Schematic illustration of regulatory elements. a) Gene regulatory region contains promoter elements and distal elements like insulators, silencers and enhancers. Some genes regulation is initiated by the formation of loop between enhancer and promoter elements. b) The function of distal regulatory elements. Enhancers initiate transcription by looping interactions. In similar manner, silencers prevent genes from being transcribed. Insulators constrain the range of enhancers and silencers. The distal regulatory elements are usually organized into locus control regions. Figures from [MEG06]. Used with permission from Annual Reviews Inc.

1.4.2 Histone Modifcations

The same genome is responsible for the development of various cell types in multicellular organism of any individual, hence gene regulation can not be explained solely on the basis of sequence. However, the genomic function of a DNA segment can be influenced through chemical alterations of nucleobases or histone proteins.

Such mechanisms are referred to as epigenetic modifications. According to contemporary definitions: "An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" [Ber+09]. It should be stressed that this definition may refer to both cell-to-cell and organism-to-organism heritability. However in this thesis, only the former case will be considered.

The epigenetic studies are particularly important as histone modifications increase our understanding of relationships between genome structure and its functions. As have been mentioned in Section 1.1.2, chromatin can switch between loose, active and compact, repressed states. These changes are associated with adjustments in nucleosome compaction and subsequent accessibility of DNA helix. In general, compaction of nucleosomes prevent DNA transcription by physically limiting access of transcription factors to DNA helix [LDS16]. Numerous studies showed that the reversed processes are driven by chemical modifications of histones [Ral08; PS08; BK11]. More specifically, there are 4 histone proteins: H2A, H2B, H3 and H4 comprising a nucleosome core. Each of core histones contain a tail with amino-acid residues, which can be chemically modified (Figure 1.8). The fifth histone referred to as H1 is called a linker. To this day, there are many different modifications discovered with the most deeply studied being acethylations and methylations leading to gene repression and activation respectively.



Trends in Genetics

Figure 1.8. Illustration of a nucleosome. Each zoom-in shows a tail of respective core histone: H2A, H2B, H3, H4. Every tail consists of different sequence of amino-acids. Frequently modified amino-acids are highlighted with letters: K - lysine, R - arginine, S - serine, T - threonine. The color represents a type of modification. Black string coiling nucleosome represents DNA double helix. Figure from [LDS16]. Used with permission from Elsevier.

CHAPTER 2

Computational Methods for NGS Data Analysis

Sydney Brenner said in his 2002 Nobel lecture that "we are drowning in a sea of data and starving for knowledge" [Bre03; RC12]. At that time, the first reference sequence of human genome was just reported and NGS technology was under development. During the next two decades, biology was revolutionized as a consequence of huge reduction of NGS costs and rapid evolution of various experimental and computational methods allowing for easier, faster and cheaper acquisition of vast amounts of data. According to [Coo+15], the current pace of nucleotide and proteomics data generation exceeds the improvement in storage capacity thereby challenging the retention of such data in public domain and handling them on large scale. This issue raises the importance of developing more efficient compression techniques. Another challenge is to keep up with the analysis of ever-growing datasets. Many experiments require computational solutions. Currently, practically every study in genomics incorporates some specialized computational techniques and thorough statistical analysis to provide meaningful interpretation of the gathered data.

This chapter is a brief introduction into common computational approaches used in analysis of data obtained from high throughput methods. First, the problem of mapping short reads onto reference genome is discussed. That issue is of fundamental importance in most NGS related experiments. Another essential problem is the meaningful comparison of gene expression. Such experiments are always accompanied by multiple biases causing straight-forward comparisons being non informative. Therefore, an appropriate normalization of raw sequencing data is required to provide meaningful conclusions.

2.1. Basics of Short Sequence Mapping

The problem described within this section can be informally introduced as follows. Given a long sequence over an alphabet of 4 letters (a word or string) consisting of approximately a billion characters, we are given a collection of even hundred millions of short sequences extracted from a long word by selecting some letter and cutting this and rightmost, consecutive characters. Typically, the length of short words is between 50 and 500. After cutting the short sequences, some of them may have had certain letters changed due to the characteristics of the data-generating process. The task is to find the most likely cut position for every short sequence in the long one. In bioinformatics this challenge is known as short read mapping.

The problem of short read mapping gained significant attention as a result of widespread adoption of NGS technology. As pointed out in [FB09], an output of sequencing experiment, i.e. a collection of reads brings no valuable information without proper processing. That processing depends on the type of analysis including either assigning individual reads to precise positions on the reference genome or putting together all pieces into a single sequence. The former is called mapping or alignment while the latter is referred to as sequence assembling. During the remaining part of this chapter, only the mapping problem will be discussed.

2.1.1 String Matching Problem

To formally define the string matching problem (SMP), a proper notation must first be introduced. Let $\Sigma = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ represents the (nucleobase) alphabet and Σ^* the set of all finite strings over Σ. The reference sequence (usually the genome) is a word $T \in \Sigma^*$ of length n. A set $\mathcal{R} \subset \Sigma^*$ of m strings, each of length p represents a collection of reads produced by sequencing device during the experiment. The usual assumption is that $n \gg p$.

The first formulation of SMP is called exact: given T and $P \in \mathcal{R}$ find the locations of all occurrences of P in T. In the presented context, the occurrence means exact matches of P in T. Alignments based solely on the described definition are impractical in most bioinformatic applications due to frequent mismatches between reference genome and obtained reads. The reason for mismatching bases can be either caused by sequencing errors or relevant biological variation between individual genotypes including mutations, insertions, deletions. Either way, reads containing mismatches should usually not be discarded. That leads to the second formulation of SMP called approximate: given T and $P \in \mathcal{R}$ find all such substrings t of T, such that edit distance $d(P, t)$ is minimized. The second formulation may also be extended to allow for reporting reads that are within certain number of mismatches k, $d(P, t) \leq k$.

2.1.2 Selected Solutions of SMP

The naive solution to SMP would be trying to match each of the m reads with subsequences of T starting at every character ranging from 1 to $n - p$. Such procedure has a run time $\mathcal{O}(m(n - p))$, which is prohibitive given the size of genome and the large number of reads. One way to reduce string query complexity is by constructing an index of either the reference sequence or the reads. Initial alignment tools were using suffix tree or suffix array structure to index the genome. Solutions based on the former structure allowed to achieve $\mathcal{O}(p + o)$ complexity to find the pattern P within string T while those exploiting the latter had $\mathcal{O}(p + \log n + o)$ search times, where o is the number of occurrences of P in T [CS15]. However, both methods have space requirements of $\Theta(n \log n)$ bits, which at the time limited their use. The major progress followed the development of full-text minute index (FM-index) [FM00]. The finding of Ferragina and Manzini allowed to construct a structure with size linearly proportional to initial sequence and perform a string search in $\mathcal{O}(o \log^{1+\varepsilon} n)$ time while maintaining enough information to recreate the text. The constant $\varepsilon > 0$ expresses the space-time tradeoff - for example, the Bowtie tool (discussed later) fixes $\varepsilon < 0.01$ [CS15]. In general, the FM-index exhibits relevant improvement of space requirement with respect to suffix array at the acceptable expense in run time.

Many popular mappers are based on the FM-index with modifications of the exact match search procedure. The modifications are introduced in order to not exclude the reads with low number of mismatches. Unfortunately, the problem complexity increases very quickly as a function of the number of mismatches. Accordingly, many algorithms rely on heuristic approaches instead of an exhaustive search, which leads to compromise between speed and accuracy. As has been shown in [Hat+13], there are various metrics to evaluate the quality of mapping tools. The results reported therein indicate that no tool outperforms all the other in every metrics. However, Bowtie as well as BWA perform remarkably well in most benchmarks. According to the authors, BWA shows better performance than Bowtie when applied to the alignment of longer reads. The discussion below is only limited to Bowtie mapper providing a representative example of the alignment tool commonly used in numerous Hi-C studies.

The Bowtie aligner operates by first creating an FM-index of a genome and then performing read queries using a modified EXACTMATCH algorithm. The EXACTMATCH algorithm developed by Feragina and Manzini allow for fast pattern search. Construction of FM-index requires the application of the Burrows-Wheeler Transform (BWT). The BW transform creates a matrix, which rows are cyclic permutations of initial character sequence prepended with a special character indicating the start of the text. Matrix rows are then lexicographically sorted (fig. 2.1a). The algorithm uses 3 properties of BW transfrom:

1. any character in the first column is preceded by the character in the last column (in the initial string),
2. characters in the first column are lexicographically ordered,
3. LF (last-to-first) mapping, i.e. i-th occurence of a character in the last column corresponds to the i-th occurence of the same character in first column.

The above properties allow to perform fast pattern search by gradually extending the query string suffix one letter at a time and examining whether it matches any row range (fig. 2.1c). Resulting pattern occurrences correspond to row range left after inspecting every letter. That procedure enables only exact match discovery. To accommodate mismatching reads, Bowtie uses a modified algorithm with a backtracking mechanism [Lan+09]. If the extension of the suffix by one letter leads to no valid alignments, the suffix is modified by replacing one of its letter with another one. Bowtie selects a character (nucleobase) with the lowest quality score. Afterwards, the search is resumed from the replacement position.

In order to reduce excessive backtracking, Bowtie introduces double indexing, which creates two indices of the genome: the BWT of forward sequence and the BWT of reversed sequence called the mirror index. If the read is allowed to contain one mismatch, an algorithm will explore 2 cases - a mismatch in the left half of the read or the right half. The recipe for first scenario is to scan the forward index requiring exact match of the right half of the read. The second case would be using a reversed index and the reversed characters read simultaneously banning the aligner to substitute in the reversed read right half subsequence. When the alignments are allowed to contain two or more mismatches, it is not possible to fully avoid excessive backtracking, so Bowtie introduce a limit on the maximum number of permitted backtracks.

In its default options Bowtie permits two mismatches and therefore it does not guarantee to find the most likely read alignment. However, competitive runtimes and high alignment rates made this tool the primary choice in numerous sequencing projects. It is worth noting that Bowtie has been replaced with its successor

Bowtie2, which improves the overall mapping performance [LS12]. The alignment times on modern computers with multiple cores are order of hours for high depth NGS experiment. Given the time required for conducting the sequencing part, there is no incentive for further improvement of current performance.

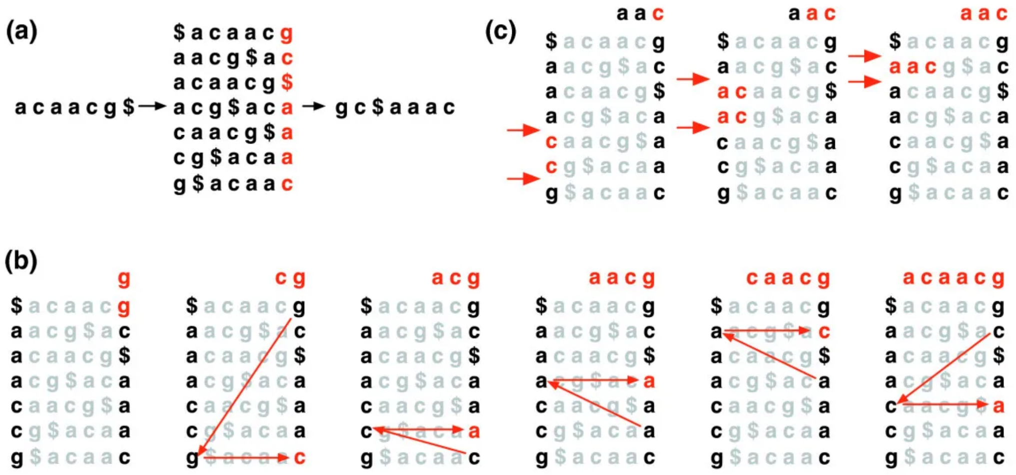


Figure 2.1. Burrows-Wheeler Transform and LF-mapping. a) Construction of BWT. The input string is appended with special, lexicographically smallest character. All cyclic permutations of the obtained string are produced, lexicographically sorted and arranged in row matrix. The BWT is the last column. b) LF-mapping. Last (L) and first (F) columns of BWT matrix suffice to recreate original text. This is possible, because any character in the first column is preceded by the character in the last column (in initial string) and i-th occurrence of a character in the last column corresponds to the i-th occurrence of the same character in first column. c) Pattern search. The LF-mapping can be used to examine the existence of a given pattern in the original sequence by gradually extending the query suffix. Figure from [Lan+09]. Used with permission from BioMed Central Ltd.

2.2. Statistical Models of Gene Expression

The study of differential gene expression (DGE) is an important part of many genomic projects and diagnostic procedures. Over- or under-expression of some genes may be an indicator of certain abnormalities in an organism [Bai+13; Loh+13; Lee+17]. Therefore, reliable methods to estimate which genes changed their expression between two measurements, taking into account the natural variability between replicate experiments, are essential. Initially, gene expression assays were conducted using the micro-array technology, which quantified the intensity of colored light emitted after the hybridization of DNA sample sequence to fragments of reference called probes. While this technology is not completely obsolete, more efficient, NGS based solutions, are now preferred. Such experiments produce reads, which are aligned to the reference genome and counted, so eventually, gene expression is measured by the absolute abundance of reads mapped to the given gene across multiple samples.

A typical gene expression experiment produces a table of counts Y_{gi} with thousands of genes $g \in \{1, ..., n\}$ and multiple samples $i \in \{1, ..., m\}$ also called libraries. Usually, each sample belongs to some group $j \in \{1, ..., k\}$ with the most standard setup consisting of 2 groups: treatment and control. Typically, we expect at least 2 replications per group. The simplest design compares 2 groups, but more complicated experiments including several conditions are often conducted. Another important characteristic of GE experiments is that the number of reads per gene is on average significantly smaller then the library size (the total number of reads).

In simpler terms, the problem of determining differentially expressed genes may be described as finding those genes for which read abundance changes upon some treatment. However, the direct comparison of read counts is not very useful due to its inherent variability even between replicated experiments. The usual solution is to treat the abundance as random variable and model the observed variation using appropriate probability distributions. Under this framework, one can conveniently re-state the problem by formulating the null hypothesis asserting that the reads in both groups were sampled from the same distribution. If the associated test rejects the null hypothesis, the gene is considered to be differentially expressed.

The main challenge of the presented approach is the choice of appropriate models for read counts. This section describes the most popular distributions for studying differential gene expression. It starts by introducing a very simple model used for count data, which is later gradually complicated in order to better suit the GE experiments characteristics.

2.2.1 Count Data Distributions

One of the most natural models used for simulating count data is the Binomial distribution. The binomial distribution models sampling process consisting of n independent repetitions of a binary outcome experiment occurring with success and failure probabilities p and $1-p$ respectively. Using the binomial density function, one can calculate the probability of y successes in n trials, i.e. occurrences of a selected outcome:

$$f(y; p, n) = \binom{n}{y} p^y (1-p)^{n-y}$$

The DEGseq package described in [Wan+09] uses binomial distribution to model the number of reads produced by gene g at group j: $y_{gj} \sim binomial(n_j, p_{gj})$. Wang and coworkers suggest estimating parameters directly from read counts table: $n_j = \sum_{g=1} Y_{gj}$ and $p_{gj} = Y_{gj}/n_j$. Their test for differential expression is based on MA transformation: $M = \log_2 y_{g1} - \log_2 y_{g2}$ and $A = (\log_2 y_{g1} + \log_2 y_{g2})/2$. Assuming the independence of y_{g1} and y_{g2}, the authors show that both M and A follows asymptotic normal distribution. This finding is used to construct a test asserting the null hypothesis $p_{g1} = p_{g2}$ versus the two-sided alternative for any gene g.

Another popular model used for simulating a discrete outcome is the Poisson distribution. The Poisson distribution can be derived by approximating a Binomial model when n grows large and p is small [Hil11]. More precisely for $n \to \infty$, $p \to 0$ such that $\mu = np$ remain constant one can show that:

$$\lim_{n\to\infty} \binom{n}{y} p^y (1-p)^{n-y} = \frac{e^{-\mu}\mu^y}{y!}$$

So the density function of Poisson distribution is expressed as:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

The Poisson model is sometimes used for modeling read abundances, because it is simple and several assumptions of this distribution are frequently satisfied when conducting GE experiments. For example, the number of reads per gene is usually small with respect to the library size, which leads to large number of trials and low success probability in the binomial setting. Conversely, the assumption of independence is unlikely to be satisfied - some genes expression may in fact influence other genes measurements. Another issue is the equality of mean and variance (equidispersion), which basically follows from the definition of Poisson distribution. As it turns out (see next section), the assumption of equidispersion is often violated in practice of NGS experiments, forcing to seek alternatives to the Poisson model.

An extension of the Poisson distribution allowing to model over-dispersion is the Negative Binomial. The Negative Binomial distribution can be derived from the Poisson distribution by introducing heterogeneity term τ [Hil11]:

$$f(y;\mu,\tau) = \frac{e^{-\mu\tau}(\mu\tau)^y}{y!}$$

Given the density of τ, one can integrate it out to obtain an unconditional distribution of y:

$$f(y;\mu) = \int_0^\infty f(y;\mu,\tau)g(\tau)d\tau$$

Assuming $\tau \sim Gamma(\theta,\theta)$, a solution to the above integral is the Negative Binomial density function:

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^\theta\left(\frac{\mu}{\theta+\mu}\right)^y$$

After substituting $\phi = 1/\theta$ ($\phi > 0$), one obtains another parametrization:

$$\begin{aligned} f(y;\mu,\phi) &= \frac{\Gamma(y+\phi^{-1})}{y!\Gamma(\phi^{-1})}\left(\frac{\phi^{-1}}{\phi^{-1}+\mu}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^y \\ &= \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}(\phi\mu+1)^{-\phi^{-1}}\left(\frac{\phi\mu}{\phi\mu+1}\right)^y \end{aligned}$$

which turns out to be very useful in modeling read abundances, because it allows to express the variance as quadratic function of the mean:

$$\mathrm{Var}(y) = \mu + \phi\mu^2$$

In the above expression, one may notice the ϕ coefficient, which captures the additional dispersion with respect to Poisson model. It is worth emphasizing that the Poisson distribution can be considered a limiting case of Negative Binomial where $\phi = 0$.

2.2.2 Generalized Linear Models

The models presented so far allow to conduct paired comparisons between treatments in count collecting experiments. However, many modern GE studies include complex designs, which are easy to formulate using linear combinations of multiple experimental conditions. In other words, a researcher wants to examine if the expression of a certain gene depends on the design of the experiment of interest. One of the most common approaches for learning dependencies between measured variables are linear models. Unfortunately, this class of methods is bounded by strict assumptions including constant variance and linear relationship between variables, which are often violated in practice of NGS data. An extension of linear models offering a lot more flexibility is known as Generalized Liner Models (GLM). A GLM can be defined using following general formula [ACH14]:

$$g(\mathbb{E}[y_i|\boldsymbol{x}_i]) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

and consists of 3 components:

- the distribution of y_i conditional on $\boldsymbol{x}_i$,
- the link function $g(\cdot)$,
- coefficient vector of linear predictors $\boldsymbol{\beta}$.

Depending on the particular instantiation of GLM, the link function can be identity, inverse, square root, logarithmic or other. Similarly, the conditional distribution of y can be selected based on the experimental setup. A popular choices, especially for count data are Binomial, Poisson or Negative Binomial distributions. Parameter estimates are usually obtained through Maximum Likelihood Estimation (MLE). For example, the MLE of $(\boldsymbol{\beta}, \phi)$ for Negative Binomial GLM are derived by differentiating log-likelihood function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\beta}, \phi) = -\sum_{i=1}^{n} \log f(y_i; \mu_i, \phi)$ with respect to model parameters leading to the following estimating equations:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{y}; \boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi)$$
$$\frac{\partial}{\partial\phi}\mathcal{L}(\boldsymbol{y}; \boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \Psi_{\phi}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi)$$

where:

$$\Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = (y_i - \mu_i)V^{-1}(\mu_i)\frac{\partial\mu_i}{\partial\eta_i}\boldsymbol{x}_i$$

and:

$$\Psi_{\phi}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \left(-\frac{1}{\phi^2}\right)\Bigg(\digamma(y_i + 1/\phi) - \digamma(1/\phi) - \log(\phi\mu_i + 1) - \frac{\phi(y_i - \mu_i)}{\phi\mu_i + 1}\Bigg)$$

Here, the parametrization for Negative Binomial distribution density function $f(y_i; \mu_i, \phi)$ is the same as the one specified in section 2.2.1 and $V(\mu_i) = \mu_i + \phi\mu_i^2$, $\digamma(u) = \frac{\partial \log\Gamma(u)}{\partial u}$ are variance component and digamma function respectively. Final estimates are obtained by solving for $(\boldsymbol{\beta}, \phi)$:

$$\begin{pmatrix} \sum_{i=1}^{n} \Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) \\ \sum_{i=1}^{n} \Psi_{\phi}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) \end{pmatrix} = \boldsymbol{0}$$

When y is distributed as Negative Binomial, the solutions are obtained with numerical methods. By far, the most popular techniques for finding $(\boldsymbol{\beta}, \phi)$ estimates are iterative methods like Fisher scoring and Newton-Raphson algorithms [Hil11].

2.2.3 Selected Methods of DGE Analysis

Various research suggests that methods based on Negative Binomial distribution are among the best performing models for read count data [YHV13; Fro+19]. Under this framework, the number of reads mapped to gene g in sample i follows the Negative Binomial distribution with mean μ_{gi} and dispersion ϕ_g:

$$Y_{gi} \sim \mathrm{NB}(N_i p_{gj}, \phi_g)$$

The mean parameter can be reexpressed as $\mu_{gi} = \mathbb{E}(Y_{gi}) = N_i p_{gj}$, where $N_i = \sum_{g=1}^{n} Y_{gi}$ is the library size and p_{gj} is the relative abundance of gene g in group j. The observed variability of read abundances in GE studies exhibits substantial over-dispersion hence the choice of Negative Binomial distribution. The above parametrization leads to following variance formula (Section 2.2.2):

$$\mathrm{Var}(Y_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2$$

After dividing both sides by μ_{gi}^2, one obtains the expression for squared coefficient of variation (CV):

$$\mathrm{CV}^2(Y_{gi}) = \frac{1}{\mu_{gi}} + \phi_g = \mathrm{TCV}^2 + \mathrm{BCV}^2$$

The coefficient of variation quantifies the relative variability of gene g in sample i. It consists of 2 terms: technical and biological coefficients of variation. The former is the technical variability, which is a result of a measurement error attributed to sequencing process. It is dependent on sequencing depth and decreases as the library size increases. The latter represents the inherent variability of transcript abundances between replicate samples. Importantly, the BCV express the variation that would persist between biological replicates even if sequencing depth could be increased indefinitely [MCS12].

The major problem concerning the estimation of ϕ_g is insufficient replication. For example, many studies consists of 2 or 3 replicates per group. As a result, the obtained samples are too few to reliably estimate parameters separately for every gene. In such situations, gene-wise dispersion estimates would be severely biased. The usual solution for this problem is to assume common dispersion for all genes, i.e. $\phi = \phi_g$ and then estimate it from the mean-variance relationship. This strategy called information sharing is based on assumption that genes originating from identical sample carry similar aspects of biological variability. This approach is adopted in 2 popular methods used for DGE analysis: DESeq and edgeR [AH10; RMS10]. Once the common dispersion is calculated, it is applied to compute the gene-wise estimates using Empirical Bayes methods. Afterwards, gene-wise mean and dispersion estimates are used to examine the differential gene expression using Fisher's exact test adapted for over-dispersed data.

As mentioned in Section 2.2.2, GE studies often include complex experimental designs consisting of multiple explanatory factors. Previously presented methods were developed for paired comparisons and are therefore unable to describe multifactor experiments. The usual approach in this situation is to employ the GLM framework and model the gene expression using linear combination, which fits the appropriate design X_{ij} linking sample i with treatment j:

$$\log(\mu_{gi}) = \sum_{j=1}^{k} X_{ij} \beta_{gj} + o_i$$

The remaining parameters are: β_{gj} - the gene-specific value of coefficient j and o_i - a sample-specific offset. In the DESeq method, the library size bias is accounted for through normalization:

$$\mu_{gi} = s_i q_{gi}$$
$$\log(q_{gi}) = \sum_{j=1}^{k} X_{ij}\beta_{gj}$$

where the normalization factor s_i is calculated with the median-of-ratios method:

$$s_i = \underset{g:Y_g^R \neq 0}{\text{median}} \frac{Y_{gi}}{Y_g^R} \text{ with } Y_g^R = \left(\prod_{i=1}^{m} Y_{gi}\right)^{1/m}$$

In the GLM framework, the null hypothesis asserts no relationship between dependent variable and the response, which in terms of model parameters may be translated to zero-valued coefficient or contrast, i.e. the linear combination of coefficients. The hypothesis is examined using the likelihood ratio test [MCS12; LHA14].

The techniques presented in this chapter are the core of many modern bioinformatic experiments relying on the NGS technology. Although the models presented so far were described in the context of gene expression, they may as well be used during other types of differential analysis incorporating count data including ChIP-seq (analysis of histone modifications) and Hi-C [NR14; LS15].

CHAPTER 3

Hi-C Contact Matrices

The development of 3C related methods and Hi-C in particular has allowed us to investigate the chromatin structure genome-wide with potentially very high resolution. Many regulatory mechanisms are driven by a complex network of interactions needed to be deciphered in order to understand genome functioning. Therefore, the emergence of Hi-C is an important step towards elucidating relationships between chromosome structure and gene regulation that are fundamental in molecular biology.

However, Hi-C is a very complex protocol with multiple steps, contributing to different biases, which amplify the overall noise to signal ratio. Moreover, precise inference of genome organization is complicated by the fact that interactions from millions of cells are sampled jointly rather than individually. In effect, a researcher obtains averaged information on an ensemble of interactions. All this makes the analysis of Hi-C data a challenging task requiring caution and appropriate statistical techniques in order to draw meaningful conclusions.

Biological aspects of the Hi-C protocol have been described in section 1.3.5. This chapter focus on the computational side of Hi-C analysis. First, the notation for contact maps is introduced along with relevant definitions. As already mentioned, Hi-C assays are accompanied by multiple sources of bias. Much attention has been dedicated to that matter resulting in variety of tools to perform normalization. Two important approaches to normalization are discussed in Section 3.2. Section 3.3 is devoted to the description of genome architecture established on the basis of Hi-C studies. Some elements of structural features were already known before the development of C methods, however Hi-C studies shed more light on elucidating the complexity of chromatin organization. The last section contains a brief introduction into the problem of comparing Hi-C contact maps. That issue is addressed in more detail during next 2 chapters as the main subject of research described within this thesis.

3.1. Notation and Definitions

A Hi-C experiment produces a library of paired-end reads, which are mapped onto the reference genome to localize their origin. Each useful pair of reads is therefore an indicator of interaction between 2 chromatin segments. Usually, the results are summarized using contact maps, i.e. matrices of interaction abundances.

Each axis of a contact map corresponds to genomic bins (see Section 1.3.5) along a certain chromosome. In general, 2 types of contact maps can be distinguished: inter-chromosomal and intra-chromosomal. The former consists of contacts between pair of different chromosomes, while the latter contains interactions from a single chromosome (Figure 3.1). From now on, we will only refer to intra-chromosomal matrices unless otherwise stated. A intra-chromosomal, unnormalized contact map is denoted by:

$$A \in \mathbb{Z}_{\geq 0}^{n,n},\ A^T = A,\ a_{ij} \coloneqq (A)_{ij}$$

where n is the number of bins in a chromosome. A k-diagonal elements of contact map A are defined as follows:

$$A_k = \{a_{ij} :\ k = |i - j|,\ a_{ij} \neq 0\}$$

The resolution parameter, or the bin size expresses the DNA segment length measured in basepairs (see Section 1.3.5) and corresponds to a unit size of matrix A is denoted with r. Although there is no golden standard procedure for selecting the best r, many studies aim to obtain resolution as high as possible, while retaining a high correlation between contact matrices replicates or their respective features [Dix+12]. Another method is to use a bin size resulting in at least 80% of all possible bins having more than 1000 contacts [AN15]. An important parameter in Hi-C analysis is the coverage of a region i:

$$s(i) = \sum_{j=1}^{N} a_{ij}$$

Finally the function:

$$d(k) = \overline{A}_k$$

is called the decay. The term is attributed to the observed relationship between the mean number of interactions and the separation distance, which exhibit rapidly decreasing behavior (see Section 3.3.1).

3.2. Contact Matrices Normalization

The complex experimental protocol of the Hi-C method charges the resulting data with various biases and artifacts. Each of the protocol's multiple steps may amplify overall noise resulting in the decline of library quality. Some remediation to this problem can be achieved by increasing the sequencing depth leading to better sampling. However, although sequencing costs are decreasing over time, the depth increase is not always feasible. Another workaround is to carefully study sources of Hi-C biases and capture their influence using appropriate statistical models, which can be taken into account later during analytical part of experiment. Such process is called contact map normalization.

From the invention of Hi-C in 2009, this subject has been extensively studied. In general, Hi-C normalization techniques can be divided on 2 categories: explicit and implicit. The former methods try to discover the exact sources of bias, study their behavior and suggest a model following it as precisely as possible. The latter approach makes certain distributional assumptions about the interaction sampling process and tries to estimate its parameters. This chapter aims to familiarize the reader with both methods.

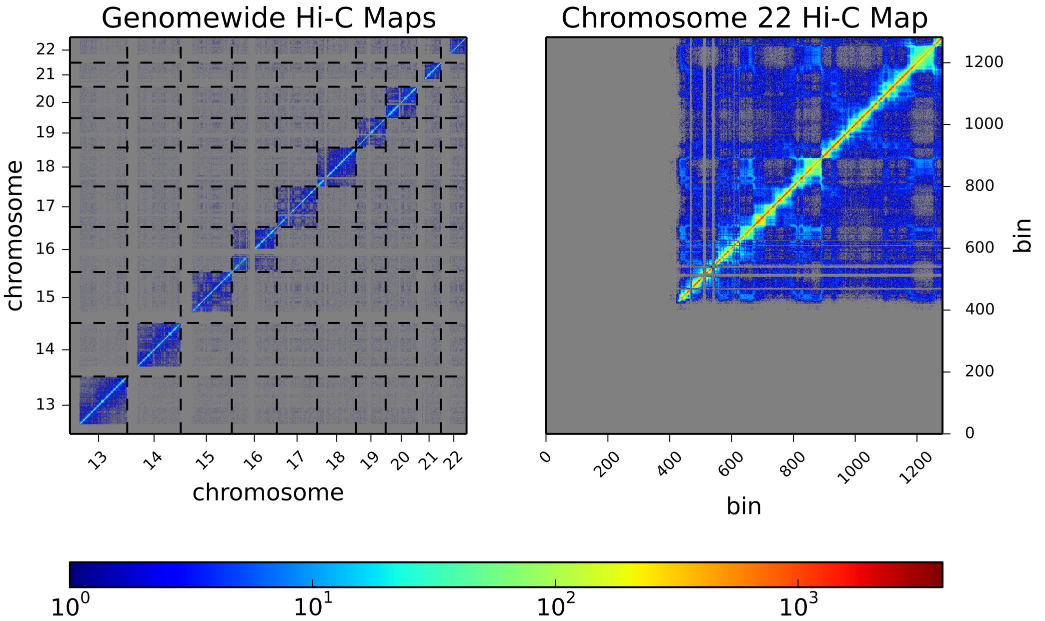


Figure 3.1. Contact maps of human IMR90 cells from the study conducted by [Rao+14]. The left matrix illustrates intra- and inter-chromosomal contact maps of chromosomes 13 to 22. Chromosome boundaries are marked with dashed vertical and horizontal lines. The right matrix is an intra-chromosomal contact map of chromosome 22. The colorbar indicates contact intensity (the number of interactions). Grey colored areas represent the absence of interactions. Large, grey strips present in most matrices correspond to centromeres, telomeres or other unmappable regions.

3.2.1 Explicit Factor Normalization

First thorough analysis of various sources of bias affecting Hi-C analysis is attributed to Yaffe and Tanay [YT11]. Their study describes several biological and technical factors, that can influence the read distribution in a misleading way. The main artifact of the Hi-C protocol are pairs of reads resulting from non-specific chromatin cleavage by a restriction enzyme leading to spurious ligation products. As shown by Yaffe and Tanay, such non-existent interactions can comprise as much as 20% of the library. Researchers suggested to identify spurious ligations by assessing the distribution of sum of distances of pairs of reads to their nearest restriction sites. For the restriction enzymes analyzed, the majority of read pairs are expected to map within 500 bp from the closest restriction sites in contrast to random ligation events characterized with uniform distribution of sum of distances (Figure 3.2a). Other factors preventing an unbiased Hi-C analysis are related with features of restriction fragments like fragment length, GC content of neighborhood surrounding restriction site or its mappability (Figure 3.2b-d). Specified characteristics can be quantified by binning fragment ends according to feature and calculating the ratio of observed $O_{\text{feat.}}[i,j]$ to total $T_{\text{feat.}}[i,j]$ possible number of contacts:

$$S_{\text{feat.}}[i,j] = (1/P_{\text{prior}}) \cdot \frac{O_{\text{feat.}}[i,j]}{T_{\text{feat.}}[i,j]}$$

Here P_{prior} is equal the to the total number of observed pairs divided by the total number of possible pairs. The number of bins is predefined to 20 for fragment length and GC content or to 5 for mappability. Importantly, *cis*-interactions are studied separately from *trans* contacts. Inspection of frequency of interactions classified by the mentioned characteristics of respective restriction fragments reveals a nonuniform distribution of read pairs (Figure 3.2b-d). This may impact Hi-C analysis as potential differences could be attributed to a varying nucleotide composition instead of relevant biological effects.

To prevent the above biases from impacting the analysis, Yaffe and Tanay suggested a multiplicative model predicting a probability of interaction between 2 restriction fragment ends a, b given their characteristics (fragment length bins a_{len} and b_{len}, GC content bins a_{gc} and b_{gc}, mappabilities $M(a)$ and $M(b)$):

$$P(X_{a,b}) = P_{\text{prior}} \cdot F_{\text{len}}(a_{\text{len}}, a_{\text{len}}) \cdot F_{\text{gc}}(a_{\text{gc}}, b_{\text{gc}}) \cdot M(a) \cdot M(b)$$

where F_{len} and F_{gc} are 2 real valued functions. Model parameters are estimated by the Maximum Likelihood method with following likelihood function:

$$\mathcal{L}(F_{\text{len}}, F_{\text{gc}}) = \prod_{\{a,b\} \in I} P(X_{a,b}) \cdot \prod_{\{a,b\} \notin I} (1 - P(X_{a,b}))$$

After initialization of F_{len} and F_{gc} with $F^0_{\text{len}} = S_{\text{len}}$ and $F^0_{\text{gc}} = S_{\text{gc}}$ the likelihood function is maximized by alternating between 2 objectives:

$$F^{n+1}_{\text{len}} = \arg\max_{F_{\text{len}}} \mathcal{L}(F_{\text{len}}, F^n_{\text{gc}}); F^{n+1}_{\text{gc}} = F^n_{\text{gc}}$$
$$F^{n+1}_{\text{gc}} = \arg\max_{F_{\text{gc}}} \mathcal{L}(F^n_{\text{len}}, F_{\text{gc}}); F^{n+1}_{\text{len}} = F^n_{\text{len}}$$

The two steps above are repeated using the BFGS algorithm until a prespecified improvement threshold of log-likelihood has been reached [NW06]. Explicit normalization of Yaffe and Tanay has been shown to greatly increase the correlation between replicate Hi-C maps and reveal valuable biological insights despite the low resolution of raw data.

Despite the high importance of the above model, its usage remain limited due to high computational costs of parameter estimation. Meanwhile, Hu and coworkers developed an alternative normalization procedure called HiCNorm, which is based on explicit modeling of biases discovered by Yaffe and Tanay [Hu+12]. Essentially HiCNorm computes fragment length x^i_j, GC content y^i_j and mappability z^i_j features for a binned region j at chromosome i similarily to the Yaffe and Tanay method. Importantly, bins refer here to the Hi-C contact map loci instead of the feature matrix. Given a contact map $A^i = \{a^i_{jk}\}$ the number of interactions between regions j and k is assumed to follow the Poisson distribution with a rate θ^i_{jk}:

$$a^i_{jk} \sim \text{Poisson}(\theta^i_{jk})$$

depending on features via log-linear relationship:

$$\log(\theta^i_{jk}) = \beta^i_0 + \beta^i_{\text{len}} \log(x^i_j x^i_k) + \beta^i_{\text{gcn}} \log(y^i_j y^i_k) + \log(z^i_j z^i_k)$$

Model parameters $\hat{\beta}^i_0$, $\hat{\beta}^i_{\text{len}}$, $\hat{\beta}^i_{\text{gc}}$ are estimated by the Poisson regression. The comparison of HiCNorm with YT normalization reveals the former method to exhibit a

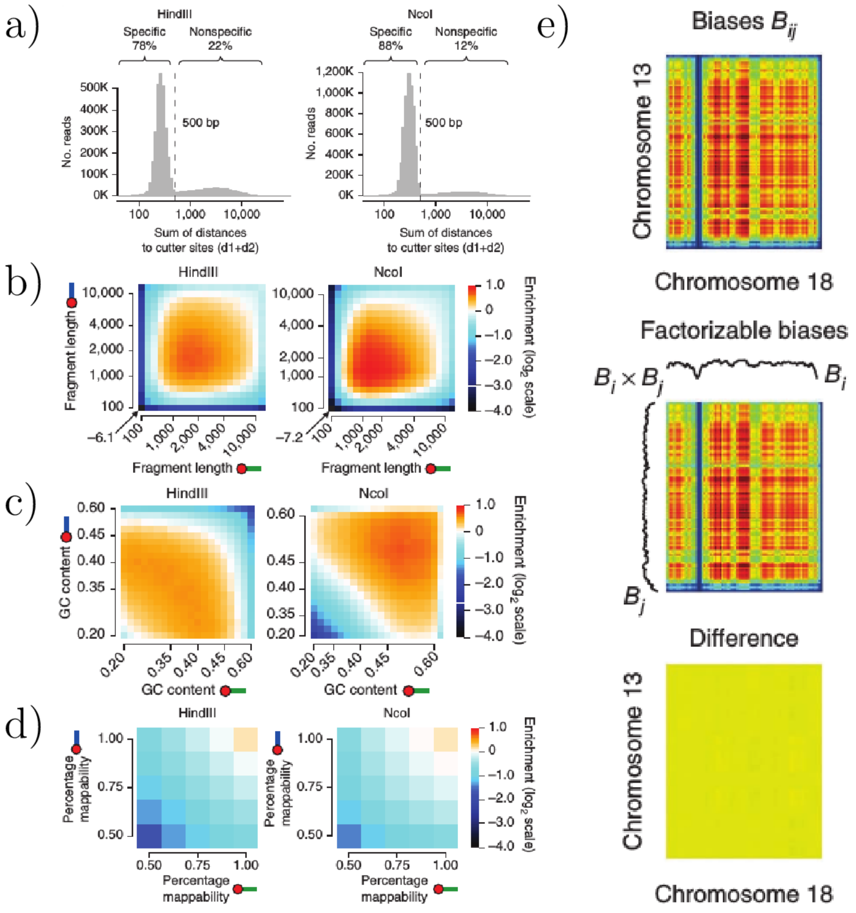


Figure 3.2. Sources of bias in Hi-C protocol. a) The distribution of sum of distances to nearest restriction sites exhibit bimodal shape. Pairs of reads mapping far from restriction sites are likely products of spurious ligations. b-d) The number of interactions is highly dependent on the length of restriction fragments, their GC content and mappability. Therefore the difference in abundances at various loci may for example represent a different nucleotide composition rather than relevant biological variation. e) The comparison of biases obtained by explicit method (top) and implicit one (middle). Figures a-d from [YT11], figure e from [Ima+12]. Used with permission from Springer Nature.

significant speed-up over the latter approach while retaining similar reproducibility improvement between replicate data.

3.2.2 Implicit Factor Normalization

A different approach to normalization of Hi-C contact maps (ICE) was suggested by Imakaev and coworkers [Ima+12]. Instead of studying and modelling individual biases separately, they proposed that the observed number of interactions O_{ij} can be factorized into a product of true contact probability T_{ij} between regions i, j with their associated biases B_i, B_j:

$$O_{ij} = B_i B_j T_{ij} \text{ s.t. } \sum_{\substack{i \\ i \notin \{j-1,j,j+1\}}}^{n-3} T_{ij} = 1$$

Although Imakev and coworkers provide a method for estimation of the vector B and the matrix T, the described problem is well-known in literature as matrix balancing and has been extensively studied. A common algorithm used to find doubly stochastic matrix T and vector B (diagonal matrix in most formulations) is due to Sinkhorn and Knopp [SK67]. Their method is based on the fixed point iteration scheme and is proven to converge, given the matrix A has total support [SK67; KR13].

The ICE normalization of replicate Hi-C data yielded similar improvement in reproducibility as that achieved by explicit methods. Remarkably, the outer product of estimated bias vector B exhibited a striking similarity to the bias matrix computed by the normalization procedure of Yaffe and Tanay (figure 3.2e).

3.3. Structural Units of Chromatin

The most direct approach towards chromatin structure research would be to obtain spatial coordinates of every nucleobase along the DNA in a given cell using some experimental technique. Collected data could be then used for thorough analysis of genome architecture similarly to protein conformation studies. Unfortunately, this straight-forward approach is far beyond current technical capabilities. Moreover, in contrast to many proteins, chromatin does not exhibit a native structure, which means we would still need to obtain structural information across thousands of cells. Instead of hypothetical direct approach, we are limited to indirect C-methods, which produce less interpretable data and are therefore prone to misguided conclusions. Nevertheless, the application of C and Hi-C studies in particular revealed many important insights on chromatin architecture, which were later confirmed by more direct experimental techniques. The next sections present major findings obtained using the Hi-C protocol and explains in more detail how these results are derived through the analysis of contact maps.

3.3.1 Contact Decay Bias

First genome-wide analysis of chromatin interactions is attributed to Lieberman-Aiden and coworkers [LA+09]. Despite relatively low resolution, with bin size equal to 1Mbp, the study provided significant remarks regarding genome architecture. One of the most distinctive features of Hi-C contact maps reported by Lieberman-Aiden and coworkers is the presence of the so-called decay bias, which results from polymer-like behavior of chromatin fiber. The existence of contact decay was demonstrated by studying the relationship between the mean contact abundance and the linear genomic separation of respective chromosomal regions (Figure 3.3). Notably, the existence of decay effect was confirmed before, using 3C and FISH studies, which provided evidence for utilizing Hi-C in genome research. The analysis of contact decays emphasizes the difference between inter- and intra-chromosomal interaction density highlighting the existence of chromosome territories. Even at large distances, the average number of intra-chromosomal contacts is higher then the mean number of contacts between different chromosomes (Figure 3.3).

3.3.2 A/B Compartments and Hierarchical Structure

Another prominent feature of contact maps are A/B compartments. They correspond to the partitioning of chromosomes into consecutive, non-overlapping intervals

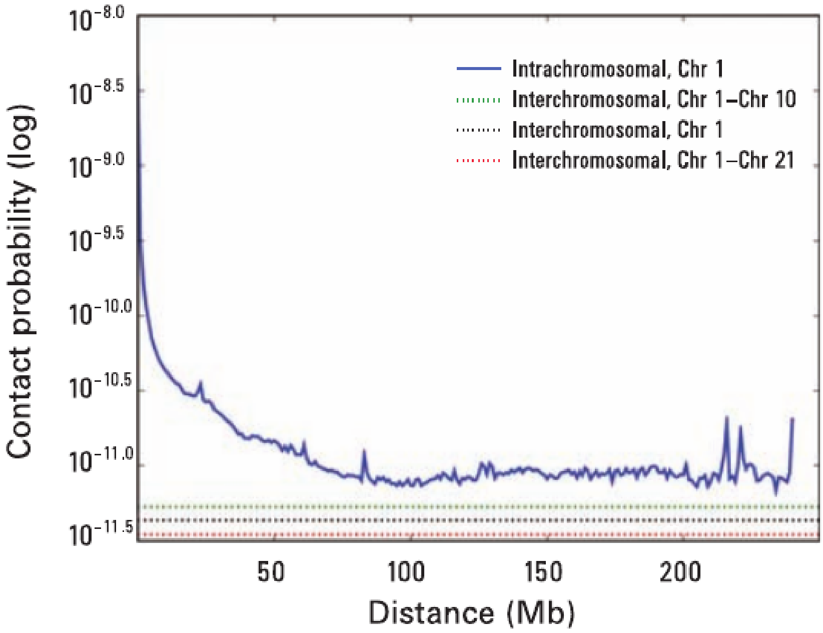


Figure 3.3. The contact probability as a function of genomic distance. Within a chromosome, the mean number of interactions diminish rapidly with the increase in genomic separation. Even at large genomic distances, the average number of interactions exceeds the one between different chromosomes. Figure from [LA+09]. Used with permission from The American Association for the Advancement of Science.

labeled as either A or B. Their distinctive property is that the number of interactions between a pair of regions within one compartment is enriched with respect to interactions across compartments. FISH experiments demonstrated that DNA fragments located in the same compartment are closer in space than when placed in opposite compartments despite smaller genomic distance. This once again validated the existence of correlation between Hi-C contacts depletion and spatial distance of interacting regions.

In order to determine the A/B labels, Lieberman-Aiden and coworkers developed a 3-step procedure. First, raw contact maps are normalized by decay rate (Figure 3.4 middle). Next, the resulting normalized matrix is converted into Pearson correlation map, which cells measure Pearson r between interaction profiles of corresponding regions (Figure 3.4 right). Finally, after applying PCA to the correlation matrix, it is partitioned based on the positive and negative values of the first principal component.

Both FISH analysis and decay rate comparison between A and B compartments indicate their high correspondence with either condensed or relaxed chromatin. Moreover, subsequent Spearman correlation analysis exhibited significant association of compartment A with the presence of genes, higher expression, accessible chromatin as well as enrichment of activating and repressing marks. These results suggest the connection of compartment A with open, actively transcribed chromatin.

Another remarkable finding reported by Liberman-Aiden and coworkers is the demonstration of hierarchical, domain-like packing of chromatin fiber supported by polymer simulations. As noticed by the researchers, the observed contact decay exhibits power law scaling in approximately 500 kb and 7 Mb genomic distance range

(Figure 3.5a). The power law type of relationship was suggested to describe polymer behaviour in general and chromatin folding in particular [DGG79]. For example, a popular model used by various researchers to simulate chromatin fiber folding called equilibrium globule exhibits power law scaling of contact decay [ML98; ML+09]. The fractal globule model used in this study extends the simple equilibrium model by introducing a hierarchy accounting for folding the polymer into a fractal-like self-similar conformation (Figure 3.5c,d). It turns out that contact decays calculated from simulations obtained using fractal globule are more similar to observed decay then those predicted by equilibrium globule (Figure 3.5a,b).

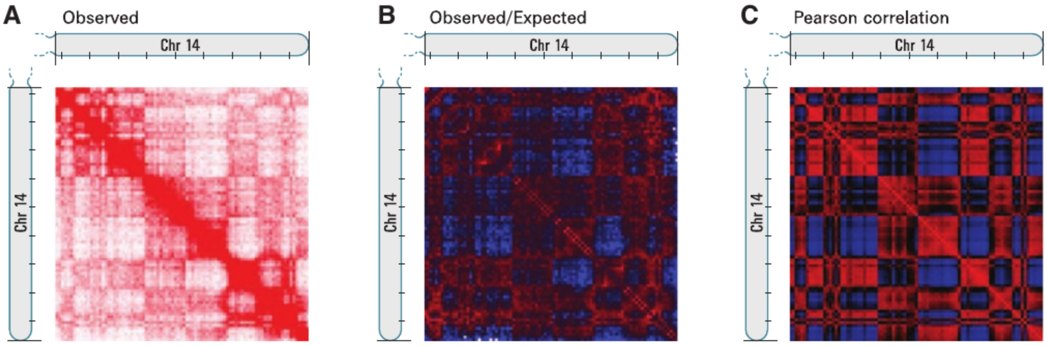


Figure 3.4. A/B compartments may be associated with plaid pattern at contact maps (left). The pattern is enhanced by first normalizing Hi-C map by contact decay (middle) and transforming the normalized matrix into correlation map of normalized interaction vectors (right). Figure from [LA+09]. Used with permission from The American Association for the Advancement of Science.

3.3.3 Topologically Associating Domains

Semi-theoretical results regarding domain-like chromatin organization obtained by Lieberman-Aiden and coworkers were confirmed 3 years later in a higher resolution Hi-C study conducted on Drosophila embryonic nuclei [Sex+12]. During the experiment carried out to explore the architecture of Drosophila genome, Sexton and coworkers developed a statistical approach to model the interaction probability, given the technical biases and genomic distances between pairs of restriction fragments. Their model included distance-scaling, fragment-dependent factor expressing how likely the regions to the left of restriction fragment are to establish contacts with the regions located on the right side of the same fragment. Therefore, a high value of scaling factor would indicate an insulator function of DNA segment preventing the spread of chromatin interactions. Further examination of fragments with the highest scaling factor facilitated systematic identification of chromosomal domains - highly contact-enriched sub-matrices located along main diagonal of Hi-C contact maps.

More evidence for domain existence in mammalian genomes was reported by Dixon and coworkers in their article from 2012 [Dix+12]. Upon examination of normalized Hi-C contact maps obtained from mouse and human cell lines researchers noticed the emergence of strong, contact-enriched, square-shaped blocks located along the main diagonals of intra-chromosomal contact maps (Figure 3.6a,b). Dixon and coworkers developed an algorithm for systematic detection of these entities and referred to them as Topologically Associating Domains (TADs). In general, TADs

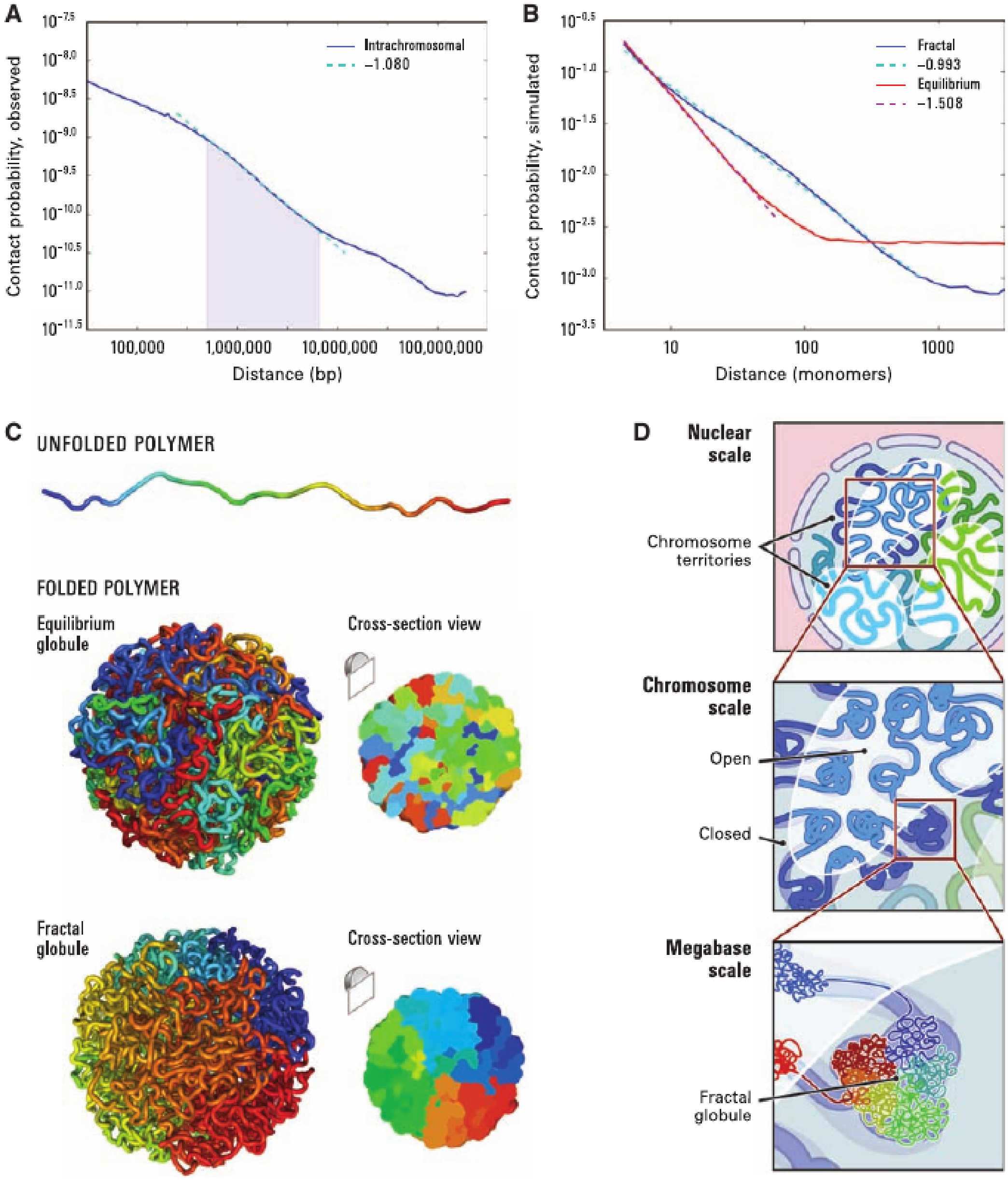


Figure 3.5. Hierarchical chromatin organization. a) The observed contact decay (solid line) and power law fit (dashed line). The fit is based on range of genomic distances marked with grey color. b) Simulated contact decays and power law fit for equilibrium (red) and fractal globule (blue). c) Illustration of equilibrium and fractal globule polymer models. d) Postulated domain-like hierarchical organization of the genome. Figure from [LA+09]. Used with permission from The American Association for the Advancement of Science.

arise from partitioning of chromosome (an interval) into subintervals according to a specific algorithm. More precisely a set of TADs T_γ consists of domains - intervals $t_i = [s_i, e_i]$ such that $1 \leq s_i < e_i \leq n$ and no two TADs t_i and t_j overlap for $i \neq j$.

For instance, the algorithm suggested by Dixon and coworkers is based on an observation that a pair of loci within a single domain is enriched with interactions in contrast to regions belonging to different TADs (Figure 3.6c). To quantify a strength of this effect for specific loci x, a statistic called Directionality Index was derived:

$$DI(x) = \left(\frac{B(x) - A(x)}{|B(x) - A(x)|} \right) \left[\frac{(A(x) - E(x))^2}{E(x)} + \frac{(B(x) - E(x))^2}{E(x)} \right]$$

Given a contact map C and a bin range k, $A(x)$ - upstream interaction bias, $B(x)$ - downstream interactions bias and $E(x)$ are defined as:

$$A(x) = \sum_{j=x-k}^{k} c_{xj},\ B(x) = \sum_{i=x}^{k} c_{ix},\ E(x) = \frac{A(x) + B(x)}{2}$$

Therefore, DI measures the ratio of upstream to downstream interaction bias and should exhibit large departures from 1 near domain boundaries as well as a change of sign. The metaparmeters k as well as the resolution were selected to maximize reproducibility of DI between 2 replicates of the experiment in each tissue. Finally, TAD boundaries are determined using Hidden Markov Model with 3 states: upstream bias, downstream bias and no bias. The MLE parameters of HMM were calculated using the Baum-Welch algorithm. Domain boundaries are called at bins transitioning from downstream bias to upstream bias state.

Functional analysis of domains discovered in ESC and IMR90 cell lines by Dixon and coworkers linked TAD boundaries with strong enrichment of CTCF protein binding and frequent occurrences of active promoters as well as housekeeping genes. Conversely, no enrichment in epigenetic marks associated with enhancers was observed. In agreement with the characteristics of insulator elements, TAD boundaries were shown to stop the spread of heterochromatin as measured by the H3K9me3 modification. Importantly, while TAD partitioning remained mostly unchanged across predifferentiated and differentiated cells, methylation pattern of specific TADs can change. This finding was also confirmed in another study investigating the structure of X-inactivation locus [Nor+12]. The same study also examined the influence of boundary deletion on transcription and reported that this leads to pathogenic mis-regulations. That result is consistent with conclusions of experiment conducted by Andrey and coworkers investigating effects of deleting a boundary region between 2 TADs known to contain different regulatory landscapes [And+13]. As a consequence, regulatory elements of a specific TAD were observed to interact with promoters of the neighboring domain leading to severe mis-regulation and disease-like phenotype. Taken together, these observations show evidence supporting the hypothesis where TADs function as separate regulatory units. According to early Hi-C studies, TAD partitioning of the genome seems to be highly conservative both between examined cell types and species (across syntenic regions). The evidence supporting this hypothesis have been also reported in another study comparing chromosome architecture of 4 species using Hi-C [Rud+15].

The discovery of TADs by Dixon and others raised the importance of systematic detection of Hi-C domains. As pointed out by [Fil+14] there may exist multiple possible TAD segmentations of the chromosome. An approach suggested by Filipova and coworkers is based on maximizing the sum of average intra-domain interaction frequency. The problem may be defined in terms of the following objective function:

$$\max \sum_{[s_i, e_i] \in T_\gamma} q(s_i, e_i, \gamma)$$

The meta-parameter γ, called resolution or scaling factor, influences the average domain size. Lower γ yields sets of larger domains and higher γ results in sets of smaller TADs. Different choice of γ will lead to various collections of TADs. The ultimate solution includes set of domains most persistent across different values of $\Gamma = \{\gamma_1, \gamma_2, ...\}$, which is defined in terms of the second objective:

$$\max \sum_{[s_i, e_i] \in T_c} p(s_i, e_i, \Gamma)$$

Here, T_c is the set of non-overlapping persistent TADs across range of γ values and $p(s_i, e_i, \Gamma)$ is the persistence of domain t_i, which measures how frequently it occurs throughout resolutions.

The solution to the first objective can be formulated as a dynamic programming problem:

$$\mathrm{OPT}_1(l) = \max_{k<l} \{\mathrm{OPT}_1(k-1) + \max\{q(k, l, \gamma), 0\}\}$$

The quality function is proportional to scaled and centered intradomain interaction frequency:

$$q(k, l, \gamma) = u(k, l, \gamma) - \mu_u(l-k)$$

where:

$$u(k, l, \gamma) = \frac{\sum\limits_{g=k}^{l} \sum\limits_{h=g+1}^{l} a_{gh}}{(l-k)^{\gamma}}$$

and $\mu_u(l-k)$ is the mean value over all sub-matrices of length $l-k$ along the diagonal of contact matrix A. The objective function defined above will allow to split domains satisfying $q(k, l, \gamma) \leq 0$ arbitrarily without affecting the optimal score. To rule out the preceding behavior and guarantee that the algorithm can only produce a sets of TADs lacking adjacent negatively scoring domains, the following modification to the objective function was introduced:

$$\mathrm{OPT}'_1 = \max \begin{cases} \max_{k<l}\{\mathrm{OPT}_D(k-1)\} \\ \mathrm{OPT}_D(l) \end{cases}$$

where:

$$\mathrm{OPT}_D(l) = \max_{k<l} \{\mathrm{OPT}'_1(k-1) + q'(k, l, \gamma)\}$$

and:

$$q'(k, l, \gamma) = \begin{cases} q(k, l, \gamma) & \text{if } q(k, l, \gamma) > 0 \\ -\infty & \text{otherwise.} \end{cases}$$

With initial conditions $l \in \{0, 1\}$: $\mathrm{OPT}_D(l) = \mathrm{OPT}'_1(l) = 0$. The dynamic program described above can be associated with directed acyclic graph $\mathcal{G}$, which nodes correspond to $\mathrm{OPT}'_1(l)$ and $\mathrm{OPT}_D(l)$ functions and an edge connects node with all other nodes it depends on: $\{\mathrm{OPT}'_1(k)\}_{k<l}$ and $\{\mathrm{OPT}_D(k)\}_{k<l}$. An edge $e = (k, l)$ has weight $q'(k, l, \gamma)$. Thus, finding an optimal solution for $\mathrm{OPT}'_1(n)$ can be reduced to

finding the heaviest path from the corresponding node. To find the top K highest weight paths in $\mathcal{G}$, a standard procedure described in [HC05] were used.

The solution to the second objective is produced according to the following recipe. First, a set $\mathcal{T} = \bigcup_{\gamma \in \Gamma} T_\gamma$ is constructed according to objective 1. Next, to get the final collection of non-overlapping, highly persistent TADs, below algorithm is used:

$$\mathrm{OPT}_2(j) = \max\{\mathrm{OPT}_2(j-1), \mathrm{OPT}_2(c(j)) + p(s_j, e_j, \Gamma)\}$$

Here, $\mathrm{OPT}_2(j)$ is the highest scoring, non-overlapping set of TADs for the jth domain and $c(j)$ is the closest domain before j that does not overlap with domain j. Domain i persistence is defined as:

$$p(s_i, e_i, \Gamma) = \sum_{\gamma \in \Gamma} \sigma_i, \text{ s.t. } \sigma_i = \begin{cases} 1 & \text{if } [s_i, e_i] \in D_\gamma \\ 0 & \text{otherwise.} \end{cases}$$

The overall runtime of the described algorithm is $\mathcal{O}(m \log m + (n^2 + m)|\Gamma|)$, where $m = |\mathcal{T}|$. In a comparison conducted using Hi-C data from [Dix+12], the dynamic programming approach produces domains with higher mean intra-domain interaction frequency than HMM method developed by Dixon and coworkers.

It should be emphasized that so far over 20 methods for TADs identification have been developed. A comprehensive comparison of domain sets produced by different TAD calling tools reveals poor concordance among respective collections of TADs [Zuf+18]. Resulting domains were shown to differ in number and average size at various examined resolutions. As noted by the authors, a likely source of high variability between methods lies in hierarchical organization of the genome and inability of current methods to properly account for it. Although multiple tools for determination of domain hierarchy exist, describing all of them in detail would be outside of the scope of this thesis.

Finally it should be noted that TADs are not a phenomenon pervasive in every species. For example studies of Arabidopsis thaliana genome architecture revealed no evidence of TADs existence [Fen+14]. It is also possible that domain segmentation only concerns certain chromosome. Such phenomenon was observed in Caenorhabditis elegans, where the X chromosome was shown to exhibit domain organization absent across autosomes [Cra+15]. Another important issue is the phase of cell cycle, which is demonstrated to heavily influence the structure of chromosomes (Section 3.4.1).

3.3.4 Chromatin Loops

Although the existence of significantly enriched long-range Hi-C interactions have been postulated in [YT11; Dix+12; Jin+13], a more precise definition of this effect was presented in [Rao+14]. This study was the first to perform a Hi-C experiment with an unprecedented sequencing depth, allowing to achieve an astonishing 1 Kbp resolution. The high quality of the resulting data enabled the authors to investigate the chromatin looping phenomenon. Intuitively, a chromatin loop is a pair of regions exhibiting higher contact frequency between each other than to the loci on the chromosome. However, systematic discovery of loops requires precise criteria and an appropriate background model.

Rao and coworkers developed a method called HiCCUPS, which is based on assessing local neighborhood of interacting regions. The background model is calculated using 2 matrices: T_{ij} - the normalized contact map with uniform coverage

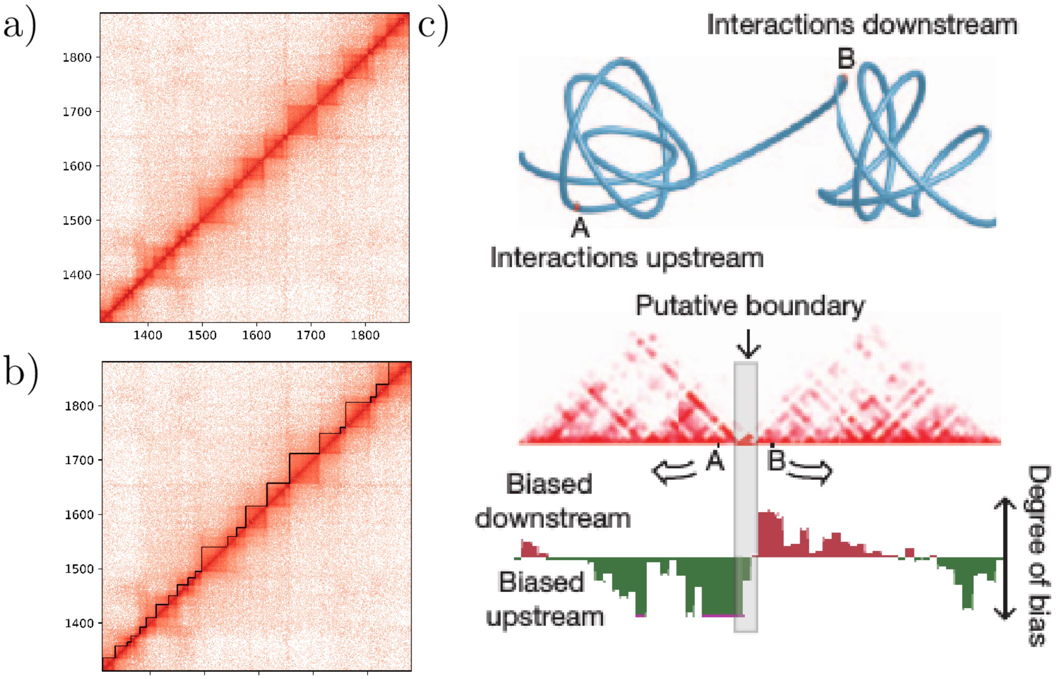


Figure 3.6. Topologically Associating Domains. a) Fragment of normalized Hi-C contact map of human embryonic stem cells (ESC) chromosome 18 from [Dix+12]. TADs emerge as strong intensity square-shape blocks of interactions along the main diagonal. b) The same contact map as in a, but additionally TADs identified using Directionality Index apporach were overlayed. c) A schematic illustration of Directionality Index approach. Figure c from [Dix+12]. Used with permission from Springer Nature.

(see Section 3.2.2) and $D_{ij} = d(|i - j|)$ - the decay matrix accounting for contact decay bias (see Sections 3.1 and 3.3.1). The local expectation for pair of loci i, j is obtained from the following formula:

$$E_{ij} = D_{ij} \frac{\sum\limits_{(i,j)\in\mathcal{N}} T_{ij}}{\sum\limits_{(i,j)\in\mathcal{N}} D_{ij}}$$

where: $\mathcal{N}$ is the pixel neighborhood surrounding the cell i, j. The neighborhood $\mathcal{N}$ can adopt various shapes (see Figure 3.7 bottom) that are introduced in order to limit false positive discoveries. For example, the lower-left filter (Figure 3.7 bottom, yellow color) prevents pixels located inside the TAD, while the horizontal (Figure 3.7 bottom, blue color) and vertical (Figure 3.7 bottom, green color) filters prevent the identification of pixels occupying TAD edges. The HiCCUPS method examines all neighborhoods illustrated in Figure 3.7 and considers the interaction at i, j as looping only if it shows significant enrichment relative to each of 4 areas. As the resulting E_{ij} value is obtained from normalized matrix it does not obey the Poisson statistic. To derive the expected raw contact count, which can be assumed to follow Poisson distribution E_{ij} is multiplied by respective bias values: $\lambda = E_{ij}B_iB_j$ obtained from implicit normalization (3.2.2). Finally, the hypothesis that the number of raw contacts at i, j, i.e. O_{ij} is enriched with respect to background model is calculated according to:

$$p_{ij} = P(Y > O_{ij}), \text{ where: } Y \sim \text{Poisson}(\lambda)$$

for all types of neighborhood $\mathcal{N}$. The obtained p-values are adjusted using Benjamini-Hochberg procedure to FDR and then thresholded.

The loops discovered by Rao and others using HiCCUPS model were then examined and shown to coincide with previously reported enhancer-promoter interactions. Additionally, it was shown that genes whose promoters occupy loci engaged in loop formation are 6 times more expressed than genes, which are not associated with any peaks. Moreover, the presence and absence of loops between different cell types was frequently accompanied by changes in gene expression. Consistently with previous reports indicating the role of CTCF and cohesin in establishing long range chromatin interactions majority of discovered loops (86%) were found to be bound with these 2 proteins [Spl+06; Hou+08; PC09]. Importantly, the observations of Rao and coworkers support a model where gene activation is mediated by enhancer-promoter looping interaction.

3.4. Introduction to Hi-C Comparative Analysis

The existence of diverse cell types repeatedly undergoing cycles of events and communicating with external environment through complex molecular machinery require specific mechanisms to respond accordingly given the current conditions. Some processes are characteristic to cell type or external environment while others are not influenced by these factors. For example, housekeeping genes are expressed at similar rates in all cell types while developmental genes are only transcribed under very specific situations. For a long time, it has been known that chromatin architecture is dynamic and correlates with gene regulation. However, the emergence of Hi-C data enabled researchers to investigate the relationships between genome conformation and cell specificity on a new level. For instance, today's standard Hi-C experiments easily achieve the resolution of 40kb. Numerous results including the ones discussed in Section 3.3.2 and Section 3.3.3 indicate that the arrangement of A/B compartments and TADs correlate with chromatin accessibility, regulatory activity, protein binding and histone modifications. One wonders whether the differences in contact abundance also correlates with changes in gene expression. After discovery of A/B compartments, TADs and looping interactions, many studies set out to explore the existence of alterations between identified structural units especially in context of regulatory processes.

3.4.1 Domain Comparative Analysis

Local chromatin architecture should, in theory, exhibit some extent of variability across different cell types due to specific goals they serve in maintaining organism functioning. Initial Hi-C studies were expected to demonstrate a high degree of variability in domain partitioning of chromosomes across different cell types. Many researchers tried to tackle this problem by assessing the extent of overlap between domain boundary locations. This approach led to the conclusion that genome-wide domain segmentation is remarkably similar across cell types and species [Dix+12;

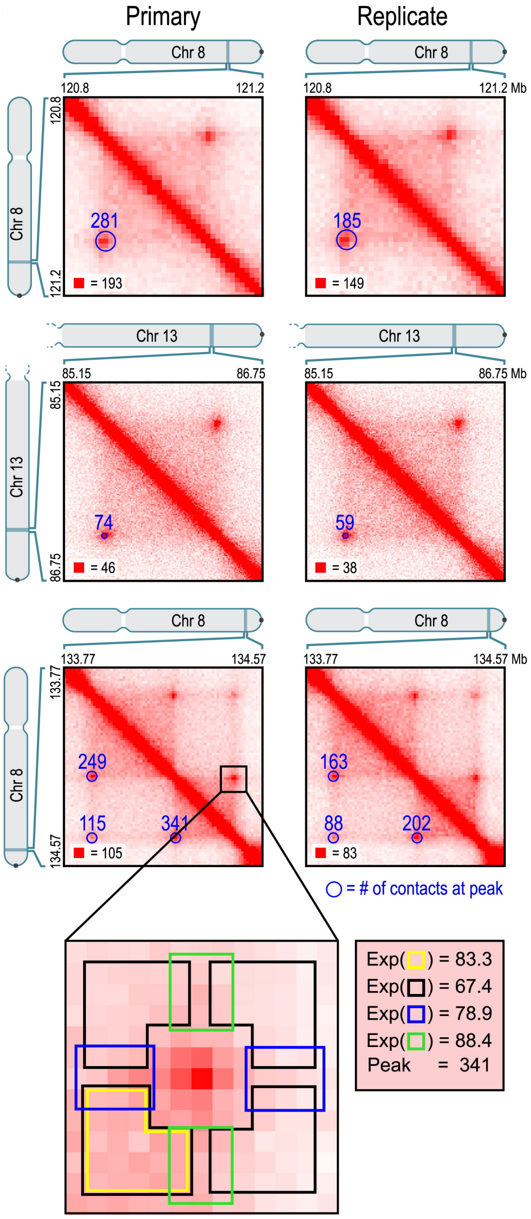


Figure 3.7. Schematic illustration HiCCUPS model. Chromatin loops are characterized by strong contact enrichment relative to the surrounding neighborhood. However the existence of TADs may bias the expected number of interactions for pixels located inside or on the edges of the TAD. In order to reduce the number of false positives, the pixel is tested against 4 background models build using different neighborhoods (bottom, marked with different colors). Figure from [Rao+14]. Used with permission from Elsevier.

LD+14; Dix+15; Cha+15]. Nevertheless, it has been later shown by us and others that TAD segmentation between cell types may significantly vary when inspected more thoroughly and using more precise metrics [ZW19a; SK18]. Another reason for this discrepancy is the choice of TAD calling algorithm. It has been shown that existing methods may produce highly variable partitionings from the same input data [Zuf+18]. To this date, the choice of proper method for TAD determination remains an open problem. Moreover, although many reports suggests the invariance of domain boundaries across cell types, they also acknowledge relevant variability of sub-TADs [DGR16]. Sub-TADs are usually obtained through the application of higher resolution data and/or alternative algorithms [Rao+14]. Recently, several groups also developed techniques for determination of whole hierarchy of domains rather than single TAD partitioning [Fra+15; WR16; WCP17; An+19]. Hierarchical methods seems to yield more reasonable results as they determine several equally possible genome partitionings. Despite this, there is still no consensus regarding the correct definition of TAD hence more in-depth studies of domains and their detection are required. Given these objections it should not be ruled out that current conclusions on TAD persistence across cell types may need to be revisited in the future. The problem of comparing TAD segmentations is further discussed in Chapter 4. Here we focus on other published approaches for comparison of Hi-C datasets, which assume the persistence of domain boundaries.

Apart from comparing the TAD boundary arrangement, several studies analyzed the concordance of A/B compartments and the differences between intra-TAD interactions. Establishing which chromosomal regions undergo compartement switch seems to be an easier task than domain calling. For example, Dixon and others suggested the following method to annotate A/B changes. Start with finding the bins that switch compartment labels across examined cell types. Then, retain bins exhibiting statistically significant variability of the first principal component as measured with ANOVA. An alternative approach is implemented in the HOMER software [Hei+10]. It first selects the regions having uncorrelated interaction profiles between Hi-C experiments. Afterwards, it scans along selected loci and outputs the longest consecutive sequence of bins changing compartment label across experiments. The application of both methods turn out to provide similar conclusions. For example, in a study conducted by Dixon and others, the authors examined the influence of A/B switch on gene expression during stem cell differentiation by comparing Hi-C data of embryonic stem cells (ESC) to 5 other cell types derived from ESC. They concluded that the transition from A to B (inactivating) was accompanied by reduced gene expression, whereas genes located within activated compartments (B to A) exhibited enriched expression when compared with stable chromatin segments (retaining their label). The results obtained by us during a study comparing endothelial cells with embryonic and mesendoderm cells are in agreement with conclusions of Dixon and others. Addtionally our analysis of histone modification patterns indicated upregulation of H3K27me3 (repressive) mark within closed compartment and enrichment of H3K27ac (active) mark inside open compartments (Figure 3.8c, top). Although the overall changes are subtle, they are statistically significant.

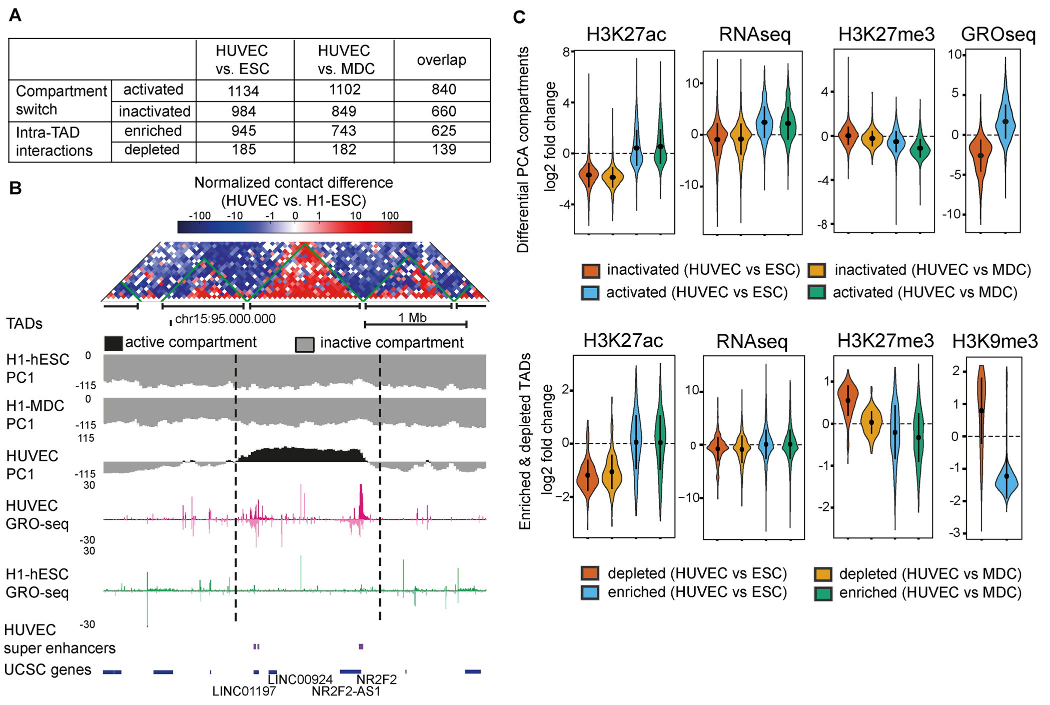
The problem of testing which domains exhibit significant variability in the number of intra-TAD interactions across Hi-C experiments is usually addressed with the help of the bootstrapping procedure. The usual approach is to first construct the difference map from a pair of Hi-C normalized matrices by subtracting respective cell values. The obtained matrix is then used to produce a randomized one by permuting every diagonal of the difference map. This procedure is repeated 1000 times to produce a collection of matrices and consequently a null distribution of median difference for each domain. Finally, every TAD is examined for significant departure (enrichment or depletion) from the null distribution leading to a p-value estimate. At the end, p-values need to be adjusted for multiple hypothesis testing.

Conclusions resulting from the comparison between the density of interactions within TADs indicate a certain relationship between regulation and changes in chromatin contact abundance. For example, a study conducted by Dixon and coworkers reports that a large proportion of TADs gain or lose contacts during cellular differentiation. The portion of TADs linked with significant change of interaction density was determined to range from 30% to even 70% depending on cell type. Further analysis of epigenetic traits demonstrated that binding of active epigenetic marks such as DHS, H3K27ac and CTCF correlated positively with changes in domain interaction frequency while repressive chromatin modifications such as H3K27me3 and H3K9me3 exhibited negative correlation with changes in TAD interactions. Consistent with the changes in epigenetic marks, gene expression measurements shown up-regulation of genes located within contact-gaining TADs and down-regulation inside interaction depleted domains. Similarly, in a study inspecting an influence of hormone treatment on T47D cell line system, Le Dily and others demonstrate the significant correlation between gene expression and changes in internal TAD contact frequency [Tru+95; Cha+15]. The conclusions obtained during our study examining the chromatin architecture of endothelial cells (EC) are in agreement with those results. We discovered that approximately 35% of all TADs undergo significant enrichment or depletion of interactions with respect to embryonic or mesendoderm cells (Figure 3.8a). Similarly to Dixon and others, we observed a positive correlation between changes in intra-domain contact frequency and active epigenetic marks as well as gene expression (Figure 3.8b and c bottom).

Lastly, it should be emphasized that the majority of Hi-C studies are concerned with the structure of interphase chromosomes. During interphase, the chromatin fiber is relatively decondensed, and adopts cell-type specific 3D structure. Multiple studies confirmed the existence of A/B compartments, TADs and looping interactions as apparent features emerging from Hi-C maps. However, the interaction landscape changes significantly upon entering the metaphase (mitotic chromsome) [Nau+13]. The study of Naumova and coworkers shows that mitotic chromosomes are virtually devoid of A/B compartments as indicated by flattening of first principal component vector during transition from interphase to metaphase. Similarly to A/B compartment loss the switch between cell cycle phases is accompanied by the reduction of TAD segmentation signal. Additional studies, including inspection of contact decays revealed different pattern of chromatin folding for interphase and metaphase chromosomes. The former adopts a fractal-globule conformation while the latter acquire a linearly-organized longitudinally compressed array of consecutive chromatin loops. Importantly, the structure of mitotic chromosomes retain a high level of similarity across different cell types. These results were also confirmed in the more recent study in different mammalian species [Gib+18].

3.4.2 Long-Range Interactions Comparison

The progress in development of Hi-C methods allowed for a rapid increase of sequencing depth resulting in higher resolution of contact maps. Increasing the Hi-C resolution shifted researchers attention towards lower scale chromatin phenomena like looping interactions. A good example of this trend are the results from the study of Rao and coworkers discussed in Section 3.3.4. During analysis of long-range chromatin interactions conducted on GM12878 cell line, researchers discovered approximately 10000 loops, which in 30% cases (versus 7% expected by chance) appeared to link promoters with enhancers. The authors also examined differences in chromatin looping across cell types by searching for loops absent exclusively in



		HUVEC vs. ESC	HUVEC vs. MDC	overlap
Compartment switch	activated	1134	1102	840
	inactivated	984	849	660
Intra-TAD interactions	enriched	945	743	625
	depleted	185	182	139

Figure 3.8. a) The number of regions changing compartment from A to B (inactivated) and from B to A (activated) during cell differentiation from ESC through MDC and terminating on HUVEC. Similarly, the number of TADs significantly enriched/depleted for contacts in HUVEC is indicated. b) The fragment of normalized interaction difference map of chromosome 15 along with UCSC Genome browser image indicating compartmentalization (principal component 1) and gene expression signal (GRO-seq). The middle TAD is associated with enrichment of contacts, activating compartment switch and up-regulation. c) Violin plots demonstrating the relationship of gene expression and histone modifications with the type of compartment switch and TAD. Figure from [Nis+17]. Used with permission from Oxford University Press.

either of the compared datasets. Using the HiCCUPS model (described in Section 3.3.4) Rao and others discovered 557 loops in GM12878 that were absent in IMR90. Likewise 510 loops were annotated in IMR90, which were not present in GM12878. Detailed analysis of gene expression demonstrated the relationship between existence of cell-specific loops and up-regulation of genes associated with them. For example, the inspection of loops specific to GM12878 indicated the promoters of 43 highly up-regulated (>50-fold) genes, but of only one gene that was markedly up-regulated in IMR90. Similarly, the promoters of 94 genes related with loops characteristic to IMR90 were found to be markedly up-regulated versus 3 promoters up-regulated in GM12878.

Another approach for discovery of long-range interactions, which seems to be especially useful when dealing with lower resolution Hi-C data, is searching for inter-TAD contacts. The inter-TAD contacts can be discovered using a method developed during our study on endothelial cells (EC, HUVEC line). To begin with, a contact map must be converted into TAD-wise (square) matrix T of size l, where l indicate the

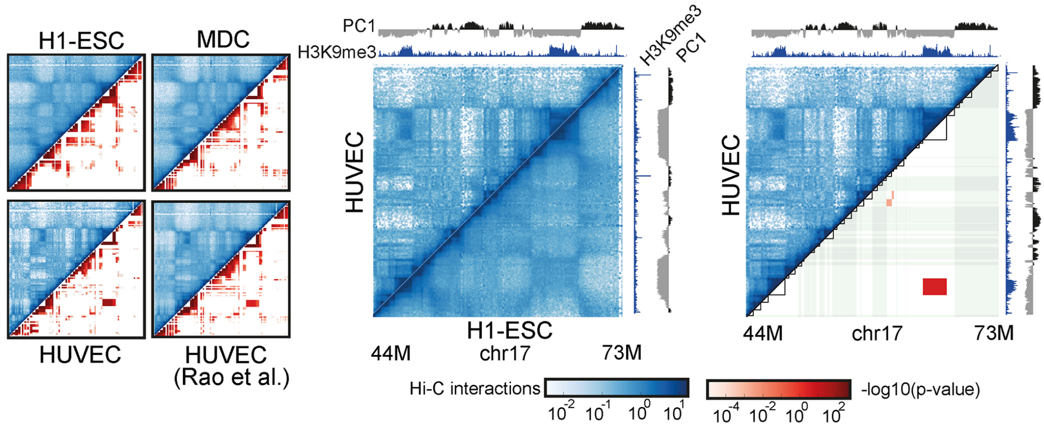


Figure 3.9. Long-range TAD interactions. Left: interaction frequency (above diagonal) and significance (below diagonal) for analysed datasets. Middle: interaction frequency comparison between ESC and HUVEC. Additionally, the compartmentalization and H3K9me3 profiles are indicated on top (HUVEC) and to the right (ESC). Right: HUVEC interaction frequency map and significance map after filtering out non-significant LRIs. Figure from [Nis+17]. Used with permission from Oxford University Press.

number of domains and T_{ij} represents the total sum of interactions between domain i and j. The p-value associated with TADs i, j is calculated using the hypergeometric test:

$$p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}}$$

$$\text{pval} = 1 - \sum_{i=0}^{k-1} p(i, M, n, N)$$

where: $k = T_{ij}, M = \sum_{i,j} T_{ij}, n = \sum_i T_{ij}, N = \sum_j T_{ij}$. The application of hypergeometric test on ESC, MDC and HUVEC-obtained contact maps resulted in annotation of 60000-80000 long-range domain interactions (LRIs) with approximately 60% of these interactions shared between all 3 examined datasets (Figure 3.9). Further analysis of EC-specific LRIs demonstrated that emergence of loops is often assisted by enrichment of repressive chromatin marks. This is also supported by the location of many LRI-participating TADs, which were observed to reside within inactive compartments more frequently than expected by chance. Interestingly, the engagement of domains in LRIs was observed to increase gene expression for TADs overlapping activated compartments. Conversely, when TADs reside inside inactivated compartment, they are more likely to exhibit up-regulation when not participating in LRI formation.

The methods discussed so far are based on discovering chromatin loops separately for each analyzed contact map and examining the simultaneous existence of determined interactions across range of studied datasets. However, this approach can diminish the accuracy of the study as such procedure does not take into account the sample biases and relationships between interaction profiles. Additionally, the methods depending on comparing the existence of annotated loops between contact maps do not express the effect strength and thereby impede the assessment of interaction relevance. A more precise approach could pool all available information and model the joint variability including datasets similarity. This topic is expanded on in Chapter 5.

CHAPTER 4
Chromosome Segmentation Comparison

One of the most intriguing questions in regulatory genomics is the problem of identification of the differences between chromatin conformation, which influence the distinct functioning of cells across various tissues or conditions. The discovery of TADs followed by evidence of their correlation with gene expression and epigenetic traits raised the importance of these entities in regulatory processes. Preliminary results encouraged numerous studies exploring TAD segmentation across different species, cell lines and conditions. The current chapter discusses the problem of differential analysis of global and local TAD segmentation, the solutions and their applicability in genomic research. Finally, a new method tailored for comparison of chromosome partitioning is introduced.

4.1. Introduction

The comparative analysis of Hi-C data may be conducted at various granularities including A/B compartments, TADs, loops (Section 3.4) or even individual interactions (Section 5). In order to study the differences between chromosome partitionings, a researcher must first determine TAD segmentation using an arbitrarily selected algorithm. Some domain calling techniques allow for gaps between TADs, while other produce consecutive sets of ungapped TADs. Ultimately, the goal is to measure the similarity between pair of partitions of some chromosomal interval or to evaluate the difference between certain contact frequencies.

4.2. TADs Similarity Measures

A TAD set determines specific chromosome segmentations and can be considered as either collections of intervals or boundaries. Depending on which interpretation is used, one may assess chromosome partition similarity by either computing the boundary overlap or the intersection of segments induced by them. Additionally, the choice of the appropriate similarity measure relies on whether a global or a local segmentation comparison is to be performed. The former tests are mostly conducted to check structural similarity of different cell lines, compare quality of replication or benchmark TAD calling algorithms. The latter may be used to establish regions responsible for differential activity between cell lines or aid tracking structural rearrangements of chromosomes.

4.2.1 Boundary Oriented Comparison

The simplest way to compare the similarity of two TAD boundary sets is by counting boundary overlaps. A higher number of overlaps indicates a larger similarity. As the number of boundaries in both sets need not to be equal, some sort of normalization is required. Therefore, given a chromosome, i.e. a set of N bins and two boundary sets $A, B \subseteq \{i|\ i \in \mathbb{Z}, 0 < i < N\}$ the proper distance measure between them may be expressed using Jaccard Index:

$$\mathrm{JI}(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The overlap set is defined as $A \cap B = \{i|\ i \in A, i \in B\}$. Usually, Hi-C experiments are noisy and even TAD datasets from 2 technical replicates may exhibit significant variability of boundary locations. Therefore, it is common to relax the definition of boundary overlap to: $A \cap B = \{i|\ \exists_j |i - j| \leq \Delta, i \in A, j \in B\}$, where $\Delta \ll N$ is some tolerance taking into account boundary shift in a replicate experiment. This approach has been used in numerous studies suggesting a high level of similarity between chromosome segmentation of different cell lines, species and indicating a low structural variation during the stimuli treatment [Dix+12; LD+14; Bar+15; Fra+15].

An alternative approach for detection of cell type specific boundaries is based on the Directionality Index described in Section 3.3.3. The method developed by Dixon and coworkers [Dix+12] calculates the similarity between two 20-bin vectors of Directionality Index between A and B centered on the detected boundary. The resulting similarity score (Spearman correlation) is compared with random control to determine if a boundary is cell type specific.

The main disadvantage of the 2 methods described above lies in the inability to distinguish between small and large boundary shifts - both of them have equal impact on the overall similarity score, provided that they exceed Δx. Additionally, the DI-based approach is not applicable for the comparison of chromosome segmentations derived from different TAD calling techniques.

4.2.2 Domain Oriented Comparison

The flaws accompanying the above described methods can be avoided by adopting a different approach, which is based on comparing domain overlap instead of boundaries. TAD set can be considered as a result of some bin clustering procedure. A very popular measure used to compare 2 clusterings is called Variation of Information (VI) [Mei03]. VI is based on a concept from information theory and requires the definition of probability distribution over clusterings. In the domain oriented approach, a clustering of bin set X with cardinality $|X| = N$ is a set $\{X[1], X[2], ..., X[n]\}$ of n non-empty subsets of X called domains such that their union equals X. The probability that a randomly selected bin from clustering X belongs to domain i is expressed as $p_X(i) = \frac{|X[i]|}{N}$. Given 2 clusterings A and B and two clusters $A[i]$, $B[j]$ their overlap is defined as set $A[i] \cap B[j] = \{k | k \in A[i] \wedge k \in B[j]\}$. The probability that a randomly selected bin k will fall into the overlap of i-th and j-th domain equals: $p_{A,B}(i, j) = \frac{|A[i] \cap B[j]|}{N}$. Then, the conditional entropy of segmentation A given B equals:

$$H(A|B) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{A,B}(i, j) \log \frac{P_B(j)}{P_{A,B}(i, j)}$$

and Variation of Information between segmentations A and B can be calculated according to the formula:

$$VI(A,B) = H(A|B) + H(B|A)$$

Some important properties of VI include:

- VI satisfies metric properties,
- the value of VI depends only on the relative sizes of the clusters, but not directly on the number of points in the data set,
- the following upperbound is true for all N: $VI(A,B) \leq \log N$,
- $VI(A,B)$ can be computed in $\mathcal{O}(N + n_A n_B)$ time.

The comprehensive discussion on VI and its properties is contained in [Mei03].

4.2.3 TADsim

Another approach to segmentation comparison is to search for local segmentations preserving high similarity. This could be useful in analysis of regulatory landscapes exhibiting differential activity. An algorithm for the determination of locally similar TAD sets was suggested by [SK18]. Their approach is based on the VI metric and Dynamic Programming to evaluate similarities of local TAD partitioning. The algorithm of Sauerwald and Kingsford consists of 3 steps.

During the first step, the VI distance matrix between all subintervals in 2 sets of boundaries is calculated. The entry i, j of this matrix represents the distance between subintervals starting at bin i and ending at bin j. To speed up calculations, a Dynamic Programming algorithm is employed instead of individually computing each subinterval distance.

Next, VI scores of every subinterval are compared with randomized distribution in order to select the significant ones. The random distribution of distances is obtained by fixing subinterval of interest in one set, permuting domains in another set 1000 times and calculating the resulting VI scores. This randomized distribution is used to compute the probability of obtaining a matching at least as good as the one observed. As either of the 2 sets of TADs can be shuffled, the reported p-value is the average taken over both random distributions. The resulting p-values are controlled for false discovery rate using the Benjamini-Hochberg procedure.

The last step is required to remove nested intervals and select the final set of non-overlapping significant intervals. It consists of 3 parts. First, only statistically significant subintervals are chosen. Next, intervals not containing any subintervals with a lower VI score are selected. If there are still some intervals left that begin or end at the same position at this point, the longest one is selected as the resulting interval.

4.3. BP Score

In this section we introduce a new distance measure for assessing chromosome segmentation similarity. Our measure is called BP score and it satisfies triangle inequality. The name is attributed to bipartite graph of domains and their overlaps induced by 2 partitionings of the same chromosome (Figure 4.1a,b). First, we define the measure and present the proof of its metric properties. Then, the behavior of the BP score is examined by comparing it with Jaccard Index and Variation of Information. Lastly, we introduce local similarity measures derived from the BP score and VI and discuss their applicability.

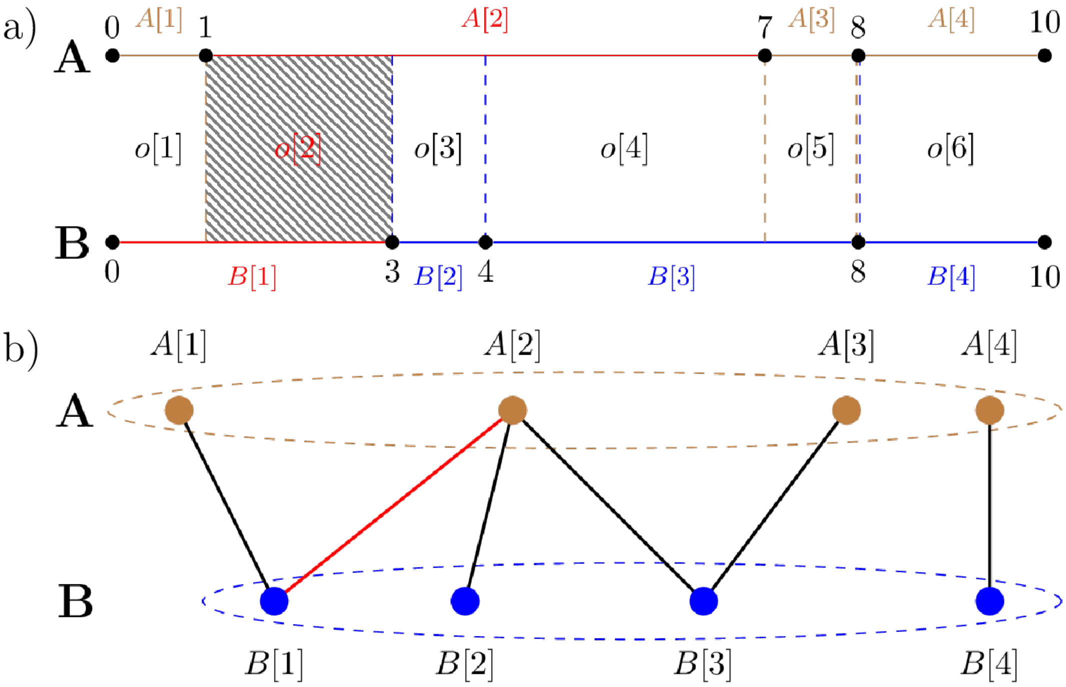


Figure 4.1. Chromosome segmentations. a) Two segmentations and the overlaps between their respective domains. The sample overlap between 2 domains marked with red color is indicated using hatch pattern. b) The segmentations can be illustrated using a bipartite graph which nodes corresponds to domains and edges indicate the existence of the overlap between a pair of domains. The parts of the graph correspond to compared partitionings and were highlighted using brown (domains set A) and blue (domains set B) ellipses respectively. The red edge represents the overlap between 2 domains highlighted in a) using the red color.

4.3.1 Notation and Definitions

A basic concept when comparing TAD partitionings is a segment.

Definition 1. *A segment is a semi-closed non-empty discrete interval:* $(a, b] = \{x \in \mathbb{N}^+ \mid a < x \leq b\}$ *with the following standard relations:*

(i) equality: $(a, b] = (c, d] \iff a = c \wedge b = d$

(ii) subset: $(a, b] \subseteq (c, d] \iff a \geq c \wedge b \leq d$

(iii) intersection:

$$(a, b] \cap (c, d] \iff \begin{cases} \varnothing, & \text{if } a \geq d \vee c \geq b \\ (\max(a, c), \min(b, d)], & \text{otherwise} \end{cases}$$

Segments may refer to chromosomes, TADs and their overlaps. Every chromosome can be partitioned into collection of non-overlapping, consecutive segments - TADs or non-TAD regions. For simplicity, from now on, the notation will be restricted to a single chromosome assuming that any partition refers to the same chromosome. This notation can be naturally extended to multiple chromosomes, for example, by assuming segmentations of concatenated chromosomes with fixed boundaries between them. The capital letters are used to distinguish partitions, and indexes refer to sorted segments (i.e. domains): $X\left[i\right], Y\left[j\right]$ (Figure 4.1a).

Definition 2. *The following functions are defined on segments:*

(i) segment start: $s\left(\left(a,b\right]\right) = a$,

(ii) segment end: $e\left(\left(a,b\right]\right) = b$,

(iii) segment length: $\left|\left(a,b\right]\right| = b - a$,

As stated above, all partitions $(X, Y, ...)$ refer to a single chromosome, so $|X| = |Y| = ... = N$ denote its length.

Definition 3. *Intersection of two partitions X and Y induces a partition called the segmentation $o_{X,Y}$ with the following properties:*

(i) $\forall_i \forall_j X\left[i\right] \cap Y\left[j\right] \neq \varnothing \implies X\left[i\right] \cap Y\left[j\right] \in o_{X,Y}$

(ii) $\forall_i \forall_{j>i} s\left(o_{X,Y}\left[i\right]\right) < s\left(o_{X,Y}\left[j\right]\right)$

When considering a triplet of partitions X,Y,Z, two types of segments may be distinguished: atomic and non-atomic (or divisible). A segment $(a, b]$ is called atomic and denoted $o[i]$ if there is no other segment $o_{X,Y}[k]$, $o_{X,Z}[l]$ or $o_{Y,Z}[m]$ that is shorter than $(a, b]$ and included in $(a, b]$. Otherwise, the segment is called non-atomic and denoted $o_{X,Y}[i]$. For example, in Figure 4.2 segment $o_{A,C}[3]$ is not atomic - it can be further partitioned into $o[3]$ and $o[4]$, both of which are atomic.

Definition 4. *The function $f_{X,Y}\left(i\right)$ gives the original segment from partition X, that encompasses the segment $o[i]$. More formally:*

$$f_{X,Y}\left(i\right) = \left(a,b\right] \;\; s.t. \;\; \left(a,b\right] \in X \wedge o_{X,Y}\left[i\right] \subseteq \left(a,b\right]$$

The same function denoted by $f_{Y,X}$ is used to find segments in the second partition.

Definition 5 (BP score)**.** *Given 2 partitions X and Y s.t. $|X| = |Y|$ their BP score is defined as:*

$$d\left(X,Y\right) = 1 - \frac{1}{N}\sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_{X,Y}(i)|, |f_{Y,X}(i)|\right)} \tag{4.1}$$

Theorem 1. *The function $d(X, Y)$ is a metric.*

Proof. Let us introduce 3 segmentations (Figure 4.2). From now on, the notation of A,B,C will be used to distinguish between 3 partitions used to construct this proof and X,Y whenever referring to any pair of partitions from set A,B,C s.t. $X \neq Y$. To show that $d(X, Y)$ is a distance function 3 properties must be proved:

- $d(X, X) = 0$,

- $d(X, Y) = d(Y, X)$,
- $d(A, B) \leq d(A, C) + d(B, C)$ for any A,B,C

Claim 1. *Function d satisfies: $d(A, A) = 0$ for any A.*

Proof.

$$\begin{aligned} d\left(A, A\right) &= 1 - \frac{1}{N} \sum_{i}^{n} \frac{\left|o_{A,A}[i]\right|^2}{\max\left(\left|f_{A,A}(i)\right|, \left|f_{A,A}(i)\right|\right)} \\ &= 1 - \frac{1}{N} \sum_{i}^{n} \frac{\left|A[i]\right|^2}{\left|f_{A,A}(i)\right|} \\ &= 1 - \frac{1}{N} \sum_{i}^{n} \frac{\left|A[i]\right|^2}{\left|A[i]\right|} \\ &= 1 - \frac{1}{N} \sum_{i}^{n} \left|A[i]\right| \\ &= 1 - \frac{1}{N} \cdot N = 0 \end{aligned} \tag{4.2}$$

□

Claim 2. *Function $d(X, Y)$ is symmetric for any A,B.*

Proof.

$$\begin{aligned} d\left(A, B\right) &= 1 - \frac{1}{N} \sum_{i}^{n} \frac{\left|o_{A,B}[i]\right|^2}{\max\left(\left|f_{A,B}(i)\right|, \left|f_{B,A}(i)\right|\right)} \\ &= 1 - \frac{1}{N} \sum_{i}^{n} \frac{\left|o_{B,A}[i]\right|^2}{\max\left(\left|f_{B,A}(i)\right|, \left|f_{A,B}(i)\right|\right)} = d\left(B, A\right) \end{aligned} \tag{4.3}$$

□

Claim 3. *Consider 3 domain sets A,B,C. Function d satisfies:*

$$d\left(A, B\right) \leq d\left(A, C\right) + d\left(B, C\right) \tag{4.4}$$

Remark 1. *Refer to Figure 4.2 as an example.*

Start with expanding 4.4 with 4.1:

$$\begin{aligned} 1 - \frac{1}{N} \sum_{i}^{n_{A,B}} \frac{\left|o_{A,B}[i]\right|^2}{\max\left(\left|f_{A,B}(i)\right|, \left|f_{B,A}(i)\right|\right)} \leq 1 - \frac{1}{N} \sum_{i}^{n_{A,C}} \frac{\left|o_{A,C}[i]\right|^2}{\max\left(\left|f_{A,C}(i)\right|, \left|f_{C,A}(i)\right|\right)} \\ + 1 - \frac{1}{N} \sum_{i}^{n_{B,C}} \frac{\left|o_{B,C}[i]\right|^2}{\max\left(\left|f_{B,C}(i)\right|, \left|f_{C,B}(i)\right|\right)} . \end{aligned} \tag{4.5}$$

Using the multinomial theorem [Hll77], the divisible segments can be substituted with atomic ones in the following way:

$$\left|o_{X,Y}[k]\right|^2 = \left(\sum_{i}^{n(k)} \left|o[i]\right|\right)^2 = \sum_{i}^{n(k)} \left|o[i]\right|^2 + 2 \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \left|o[i]\right| \cdot \left|o[j]\right|, \tag{4.6}$$

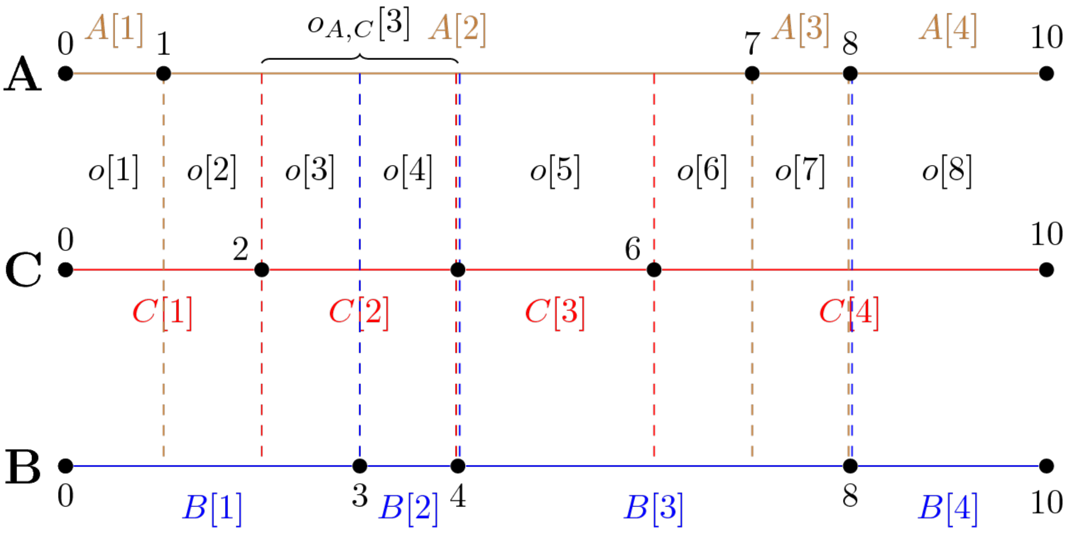


Figure 4.2. Comparison of 3 partitionings of the same chromosome. In this setting the segment $o_{A,C}[3]$ generated by segmentations A and C is non-atomic and consists of atomic segments $o[3]$ and $o[4]$.

where $n(k)$ is the number of atomic segments in a divisible segment $o_{X,Y}[k]$. Obviously, $n(k)$ also depends on X and Y, but for simplicity, it is left out of the notation here assuming it follows from the formula. Using equation 4.6, one can rewrite the inequality 4.5:

$$\begin{aligned}
&N - \sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_{A,B}(i)|, |f_{B,A}(i)|\right)} \\
&\qquad - 2 \sum_{k}^{n_{A,B}} \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A,B}(k)|, |f_{B,A}(k)|\right)} \leq \\
&N - \sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_{A,C}(i)|, |f_{C,A}(i)|\right)} \\
&\qquad - 2 \sum_{k}^{n_{A,C}} \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A,C}(k)|, |f_{C,A}(k)|\right)} + \\
&N - \sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_{B,C}(i)|, |f_{C,B}(i)|\right)} \\
&\qquad - 2 \sum_{k}^{n_{B,C}} \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{B,C}(k)|, |f_{C,B}(k)|\right)} .
\end{aligned} \tag{4.7}$$

Here, n is the number of atomic segments and $n_{X,Y}$ is the number of divisible segments induced by X,Y partitioning. As atomic segments are common for A, B and C the subscript in n can be omitted.

Definition 6 (islands of segments)**.** *Define a family of segments* $I_k = \{o[i] \mid o[i] \subseteq$

$C[k]\}$ *referred to as islands of segments, as satisfying the following condition:*

$$\underset{o[i]\in I_k}{\forall}\ \underset{\substack{o[j]\in I_k\\ i\neq j}}{\exists}\left[f_{A,C}(i)=f_{A,C}(j)\vee f_{B,C}(i)=f_{B,C}(j)\right]$$

Intuitively, each island of segments results in a sum of product terms on the right side of the inequality 4.7.

Claim 4. *Any atomic segment* $o[i]$ *satisfies:*

$$\begin{aligned}\forall_{1\leq i\leq n}\Bigg[&\frac{|o[i]|^2}{\max\left(|f_{A,B}(i)|,|f_{B,A}(i)|\right)}\geq\frac{|o[i]|^2}{\max\left(|f_{A,C}(i)|,|f_{C,A}(i)|\right)}\\ &\vee\frac{|o[i]|^2}{\max\left(|f_{A,B}(i)|,|f_{B,A}(i)|\right)}\geq\frac{|o[i]|^2}{\max\left(|f_{B,C}(i)|,|f_{C,B}(i)|\right)}\Bigg]\end{aligned}\tag{4.8}$$

Proof.

1. using definition of $f_{X,Y}(i)$ substitute:
   - $f_{A,B}(i)=A[u]=f_{A,C}(i)$
   - $f_{B,A}(i)=B[v]=f_{B,C}(i)$
   - $f_{C,A}(i)=C[z]=f_{C,B}(i)$
2. if $|A[u]|>|B[v]|$, then: $|A[u]|\leq\max\left(|A[u]|,|C[z]|\right)$
3. otherwise $|A[u]|\leq|B[v]|$ and: $|B[v]|\leq\max\left(|B[v]|,|C[z]|\right)$

□

This allows us to split the squared terms from the right hand side of the inequality 4.7 and merge them into 2 groups (S - smaller, R - remaining), both of cardinality n:

- S is the sum of terms from either A,C or B,C, such that each term satisfies condition 2 (if it is a term from A,C) or condition 3 (if it is a term from B,C),
- R are remaining terms, i.e. they may not satisfy the above conditions.

We can rewrite the inequality 4.7:

$$\begin{aligned}N-&\sum_{i}^{n}\frac{|o[i]|^2}{\max\left(|f_{A,B}(i)|,|f_{B,A}(i)|\right)}\\ &-2\sum_{k}^{n_{A,B}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{|o[i]|\cdot|o[j]|}{\max\left(|f_{A,B}(k)|,|f_{B,A}(k)|\right)}\leq\\ N-S&-2\sum_{k}^{n_{A,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{|o[i]|\cdot|o[j]|}{\max\left(|f_{A,C}(k)|,|f_{C,A}(k)|\right)}+\\ N-R&-2\sum_{k}^{n_{B,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{|o[i]|\cdot|o[j]|}{\max\left(|f_{B,C}(k)|,|f_{C,B}(k)|\right)}.\end{aligned}\tag{4.9}$$

Now:

$$\sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_{A,B}(i)|, |f_{B,A}(i)|\right)} \geq S,$$

so we can write:

$$\begin{aligned} N - \sum_{i}^{n} \frac{|o[i]|^2}{\max\left(|f_A(i)|, |f_B(i)|\right)} \\ - 2 \sum_{k}^{n_{A,B}} \sum_{i}^{n(k)} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A,B}(k)|, |f_{B,A}(k)|\right)} \leq N - S. \end{aligned} \tag{4.10}$$

After using 4.10 to simplify 4.9, what is left is to show that:

$$\begin{aligned} R + 2 \sum_{k}^{n_{A,C}} \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A,C}(k)|, |f_{C,A}(k)|\right)} \\ + 2 \sum_{k}^{n_{B,C}} \sum_{i}^{n(k)-1} \sum_{j}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{B,C}(k)|, |f_{C,B}(k)|\right)} \leq N. \end{aligned} \tag{4.11}$$

Note that no two segments $o_{A,C}[k]$ and $o_{B,C}[l]$ can generate two different atomic segments $o[i]$, $o[j]$ that would be properly included in them.

Claim 5. *There are no 2 segments $o_{A,C}[k]$ and $o_{B,C}[l]$, such that for two different indices i,j $(i < j)$: $o[i] \subset o_{A,C}[k] \wedge o[j] \subset o_{A,C}[k]$ and $o[i] \subset o_{B,C}[l] \wedge o[j] \subset o_{B,C}[l]$.*

Proof. assume that there exist nonatomic segment $o_{A,C}[u]$ and atomic segments $o[i]$, $o[j]$ s.t. $o[i] \subseteq o_{A,C}[u] \wedge o[j] \subseteq o_{A,C}[u]$. That would imply:

- $s(o_{A,C}[u]) \leq s(o[i]) < e(o[i]) \leq s(o[j]) < e(o[j]) \leq e(o_{A,C}[u])$
- also as both atomic segments $o[i]$, $o[j]$ are contained in $o_{A,C}[u]$, they can not be induced by partitioning between A and C. This means that there exist segments $o_{B,C}[v]$ and $o_{B,C}[v+1]$ (atomic or not) s.t. $e(o_{B,C}[v]) = e(o[i])$ and $s(o_{B,C}[v+1]) = s(o[j])$.

This last statement implies that $o[j] \not\subseteq o_{B,C}[v]$. □

The sums of product terms from the inequality 4.11 can be re-expressed as a total of m groups P_k corresponding to islands I_k:

$$P_k = \{(i,j) \mid i \neq j,\ o[i] \in I_k,\ o[j] \in I_k\}$$

It will also be helpful to simplify the notation:

1. as islands of segments are considered, one can replace $f_{C,A}(k)$ and $f_{C,B}(k)$ with $f_C(k)$,
2. for any $(i,j) \in P_k$ the notation $f_{A \vee B}(i)$ is introduced such that:

$$f_{A \vee B}(i) = \begin{cases} f_{A,C}(i), & \text{if } f_{A,C}(i) = f_{A,C}(j) \\ f_{B,C}(i), & \text{otherwise} \end{cases}$$

Rewriting the inequality 4.11, gives:

$$R + 2\sum_{k}^{m}\sum_{(i,j)}^{|P_k|} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} \leq N. \tag{4.12}$$

The number of elements in P_k can be upperbounded by:

$$|P_k| \leq \binom{n}{2}$$

This allows for upperbounding the left hand side of the inequality 4.12:

$$\begin{aligned} R + 2\sum_{k}^{m}\sum_{\substack{i,j\\ j>i}}^{|P_k|} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} \leq \\ R + 2\sum_{k}^{m}\sum_{i}^{n(k)-1}\sum_{\substack{j\\ j>i}}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} . \end{aligned} \tag{4.13}$$

R can be split into 2 groups:

1. R_1 segments $o[i]$ such that: $\exists_{(u,v)\in P_k} i = u \vee i = v$,
2. R_2 remaining segments,

and rewritten as:

$$\begin{aligned} R = R_1 + R_2 = \sum_{k}^{m}\sum_{i}^{n(k)} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} \\ + \sum_{i}^{n_r} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} . \end{aligned} \tag{4.14}$$

Now, put equation 4.14 into the right hand side of the inequality 4.13:

$$\begin{aligned} \sum_{i}^{n_r} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} + \sum_{k}^{m}\Bigg[\sum_{i}^{n(k)} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} \\ +2\sum_{i}^{n(k)-1}\sum_{\substack{j\\ j>i}}^{n(k)-i} \frac{|o[i]| \cdot |o[j]|}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)}\Bigg] . \end{aligned} \tag{4.15}$$

In order to further upperbound the left hand side of inequality 4.15 we need to select the minimum possible denominator. It can be easily shown that it is minimum when:

$$\max(|f_{A\vee B}(i)|, |f_C(k)|) = |f_C(k)| . \tag{4.16}$$

This follows from the definition of islands of segments as each atomic segment $o[i] \in I_k$ also satisfies: $o[i] \subset C[k]$ meaning $|f_{A\vee B}(i)| \leq f_C(k)$. The latter upperbound let us write again:

$$\begin{aligned}
&\sum_{i}^{n_r} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)} + \sum_{k}^{m}\left[\sum_{i}^{n(k)} \frac{|o[i]|^2}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)}\right.\\
&\qquad\qquad\left. +2\sum_{i}^{n(k)-1}\sum_{\substack{j\\ j>i}}^{n(k)-i} \frac{|o[i]|\cdot|o[j]|}{\max\left(|f_{A\vee B}(i)|, |f_C(k)|\right)}\right]\\
&\leq \sum_{i}^{n_r}\frac{|o[i]|^2}{|f_C(k)|} + \sum_{k}^{m}\left[\frac{1}{|f_C(k)|}\left(\sum_{i}^{n(k)}|o[i]|^2 + 2\sum_{i}^{n(k)-1}\sum_{\substack{j\\ j>i}}^{n(k)-i}|o[i]|\cdot|o[j]|\right)\right]\\
&= \sum_{i}^{n_r}\frac{|o[i]|^2}{|f_C(k)|} + \sum_{k}^{m}\left[\frac{1}{|f_C(k)|}\left(\sum_{i}^{n(k)}|o[i]|\right)^2\right].
\end{aligned} \tag{4.17}$$

The last step is to substitute $|f_C(k)|$ with atomic segments:

1. $|f_C(k)| = |o[i]|$ for atomic segments,
2. $|f_C(k)| = \sum_i^{n(k)} |o[i]|$ for divisible segments.

Finally, rewriting 4.17 yields:

$$\begin{aligned}
&\sum_{i}^{n_r}\frac{|o[i]|^2}{|f_C(k)|} + \sum_{k}^{m}\left[\frac{1}{|f_C(k)|}\left(\sum_{i}^{n(k)}|o[i]|\right)^2\right]\\
&\leq \sum_{i}^{n_r}\frac{|o[i]|^2}{|o[i]|} + \sum_{k}^{m}\left[\frac{1}{\sum\limits_{i}^{n(k)}|o[i]|}\left(\sum_{i}^{n(k)}|o[i]|\right)^2\right]\\
&= \sum_{i}^{n_r}|o[i]| + \sum_{k}^{m}\left[\sum_{i}^{n(k)}|o[i]|\right]\\
&= \sum_{i}^{n_r}|o[i]| + \sum_{i}^{n-n_r}|o[i]| = N,
\end{aligned} \tag{4.18}$$

which ends the proof, since obviously $N \leq N$.

□

4.3.2 Comparison with Existing Approaches

The performance of JI, BP and VI metrics for comparing chromosome segmentation was evaluated on simulated and real datasets. The artificial dataset A was obtained by drawing TADs based on real Hi-C domain dataset. In order to sample TADs, we first examined the distribution of real domain lengths and modeled it using the Negative Binomial distribution. Then using the obtained fit, we simulated a TAD set A. The other TAD set (B) was generated by copying A and shifting existing boundaries or introducing new ones yielding 3 chromosome segmentation categories:

1. ε-matching - every boundary in B is shifted at most ε bins left or right with respect to its initial location (Figure 4.3a),

2. new boundary at ε - introduction of new boundary in B no more than ε bins left from the existing boundary (Figure 4.3b),

3. binary boundary additions - $2^n - 1$ boundaries are inserted inside every TAD in B according to the binary interval partitioning scheme (Figure 4.3c).

The described segmentations capture various (dis)similarity cases between pairs of partitionings that we expect to see in real data. First category is similar to a pair of TAD segmentations originating from 2 technical replicates data. The overall domain coincidence is large but some boundaries may not overlap as the data acquisition is accompanied by noise. The second case may represent 2 types of situations. When ε is small, the matching resembles partitioning produced while using the TAD detection algorithms yielding gapped segmentation like the Directionality Index or the Dynamic Programing approach. When comparing the output of algorithms with a segmentation of a very similar sample determined by an ungapped algorithm, their distance is expected to be low. Larger values of ε will lead to more dissimilar matching cases. The binary boundary additions scenario will represent high inconsistency between two segmentations except for the situation when $n = 0$.

The performance of all 3 metrics for every scenario and 3 parameter values are presented in Figure 4.4. Higher distances of ε shift than binary boundary additions matching for the JI distance indicate the limitation of this metric in quantifying the true chromosome segmentation similarity in contrast to the BP score and the VI distances. Additionally, the BP score reports a larger difference between scenario 1 or 2 versus 3 as opposed to VI. This seems to be a desired behavior as the third case represents the highest degree of reorganization in contrast to case 1 and 2 given small values of ε possibly reflecting technical variability. Moreover, the value of the BP score have an intuitive interpretation for the binary boundary additions scheme. For example, when $n = 1$, every domain in A overlaps 2 equal domains in B, which corresponds to the score of: $1 - \frac{1}{2} = 0.5$. If $n = 2$ we have 4 domains in B per 1 domain in A leading to the score of: $1 - \frac{1}{4} = 0.75$. This scheme can be continued for $n > 2$. Another interesting issue is the behavior of discussed metrics when comparing scenario 1 with 2. Although each metric exhibits a higher distance for respective ε, in the first case then second one, the differences are the smallest for the BP score. For instance, when using the BP score, the distance in scenario 1 ($\varepsilon = 1$) is smaller than in scenario 2 ($\varepsilon = 3$) as opposed to VI. We argue that if ε is small, this effect may be desired, as subtle perturbations of domain locations should provide more certainty regarding the high degree of similarity between chromosome partitionings as compared with an insertion of new TADs (like in scenario 2).

One of the potential applications of chromosome segmentation metric is to assess the structural similarity between replicate data or cell lines. The replicate comparison may serve to quantify the quality of the replication experiment or as a baseline for measuring the similarity of different tissues partitioning. The comparative analysis of the TAD segmentation across cell lines can be valuable in functional assays of chromatin. To investigate the performance of studied metrics on real Hi-C data, a set of 6 different cell lines was selected from publicly available resources having 2 (5 cell lines) and 4 (1 cell line) technical replicates. All collected datasets comprised of 22 chromosomes (excluding X and Y). Each sample TADs were determined using the Dynamic Programing approach [Fil+14], resulting domain sets were paired and their respective distances were calculated. Pairs were initially assigned to one of 2 categories: within cell type or between cell type. The examination of dis-

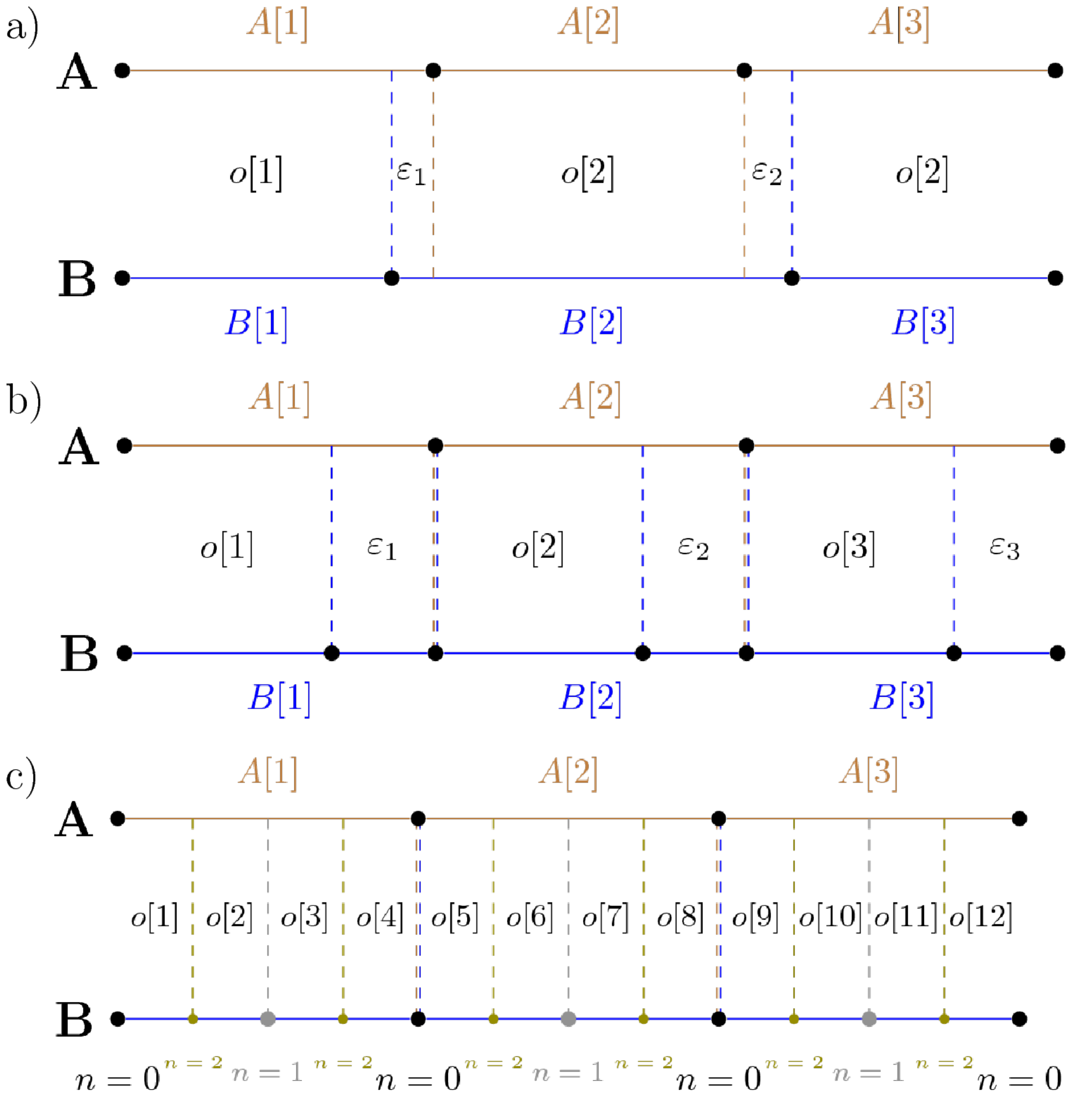


Figure 4.3. Various matching scenarios. a) In ε-matching a boundary in B is randomly shifted at most ε bins left or right with respect to their location in A. b) New boundary at ε inserts a boundary in B at randomly choosen location not further than ε bins left from their existing boundary. c) Binary boundary additions introduce boundaries according to the binary interval partitioning scheme and represents an example of highly reorganized matching.

tance distributions reveals that all metrics discriminate between two groups under consideration, that is, the distances within cell type segmentation comparisons are consistently lower than between cell type pairs.

As it is visible in Figure 4.5a, the distributions of distances exhibit substantial overlap. Careful examination of individual segmentations revealed that the majority of unexpected values can be attributed to ESC lines from different laboratories (higher variability) and to ESC versus MDC comparisons (lower variability than expected). After moving these pairs to individual groups (Figure 4.5)b, it seems more credible to place them into different categories than initially, since ESC versus MDC comparisons seem to be drawn from the within-cell-type distribution, whereas ESC

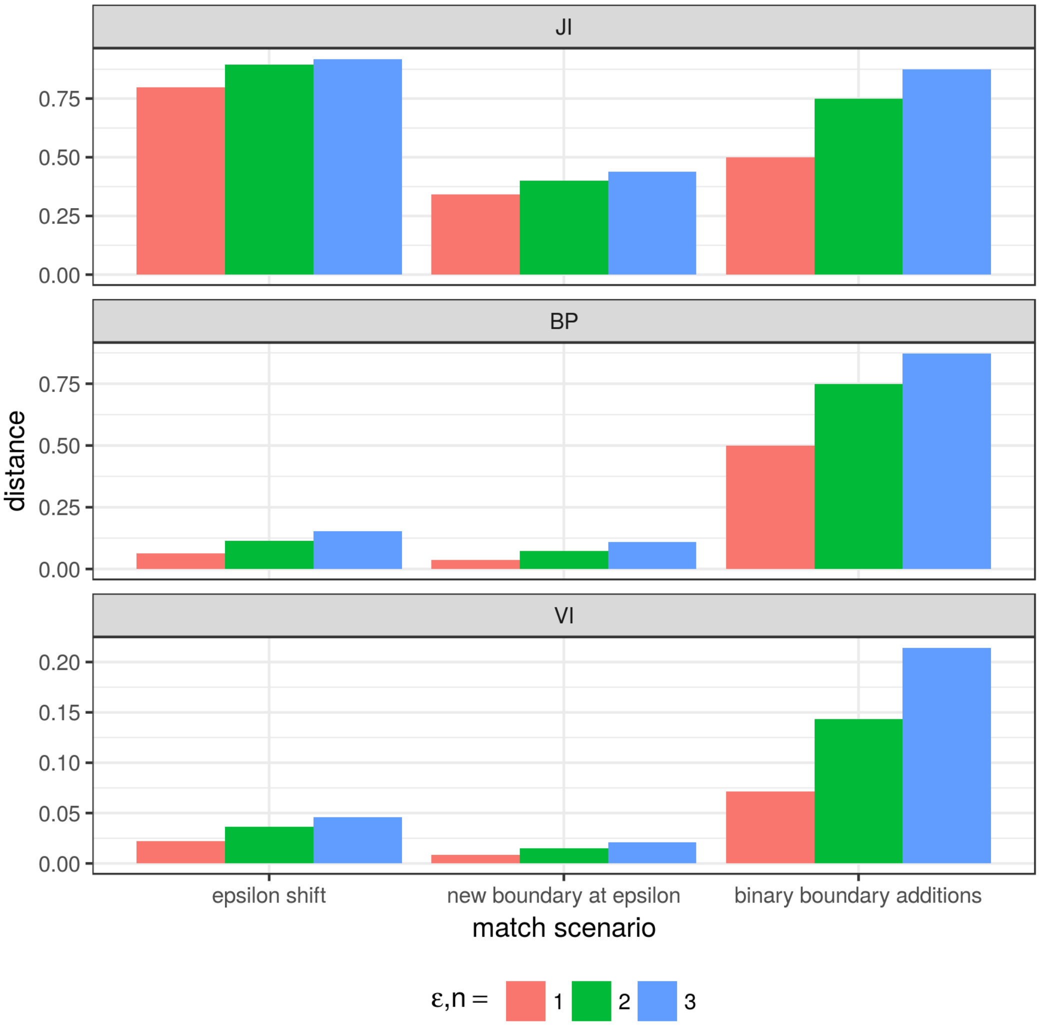


Figure 4.4. The performance of 3 metrics on simulated data comparison. The JI distance fails to recover true similarity as it reports the highest distances for ε-matching. The BP and VI distances capture the high dissimilarity of the binary boundary additions scenario and high similarity in remaining cases. The VI distance is scaled by $\log_2 N$, so it ranges from 0 to 1.

cells from different laboratories exhibit a much higher variance than expected from the replicates. This observation complies with the results on the differential chromatin organization between cell lines reported in [Dix+15]. Although all metrics indicate significant discrimination, it is clear that both BP and VI exhibit a higher significance of the difference and separation between the two groups than JI. Interestingly, the results presented in Figure 4.5 indicate some pairs of cell populations that exhibit substantial mismatch of respective TAD segmentations. This contrasts with the general view on persistence of TADs reported in numerous studies and mentioned in Chapter 3.

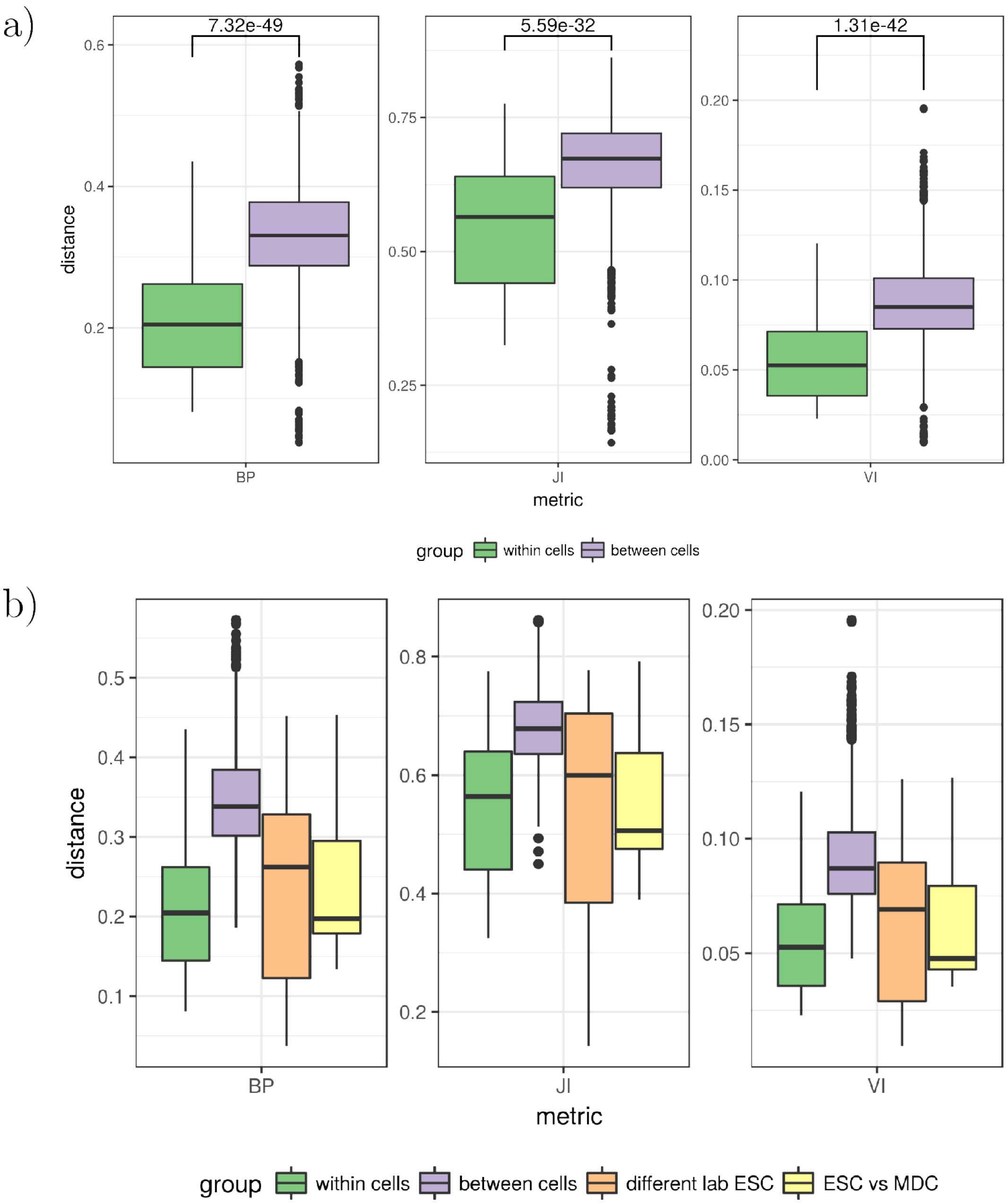


Figure 4.5. Real Hi-C datasets comparison. a) Boxplots represent distributions of pairwise segmentation comparisons obtained using 3 metrics separately for every chromosome. Each pair is assigned to either within cells (usually technical replicates data) or between cells group. b) Same as a, but pairs obtained associated with different laboratory ESC data or ESC versus MDC comparison were pulled out to separate groups due to their unexpected variability.

4.3.3 Local Measures of Similarity

The global similarity of chromosome segmentations is not always of primary interest. Frequently, researchers are concerned with finding locally reorganized regions of chromosomes given two sets of TADs. A dissimilarity between local segmentations can be for example examined for co-incidence with differentially expressed genes or methylation patterns. To compare local segmentations, the local measures associated with BP distance and VI distance may be proposed. Local BP score of i-th segment is defined as:

$$d_{A,B}^{\mathbf{BP}}(i) = 1 - \frac{|o[i]|}{\max\left(|f_{A,B}(i)|, |f_{B,A}(i)|\right)}$$

The value of the local BP score lies between 0 and 1. The smaller value of $d_{A,B}^{\mathbf{BP}}(i)$, the larger overlap between $f_{A,B}(i)$ and $f_{B,A}(i)$. The local measure related to VI would be local Mutual Information (MI) of segment i expressed with following formula:

$$d_{A,B}^{\mathbf{MI}}(i) = -p_{A,B}(i) \cdot \log_2\left(\frac{p_{A,B}(i)}{p_A(i) \cdot p_B(i)}\right)$$

The two local measures defined above are presented on Figure 4.6a,b illustrating the segmentation of human chr1: 4000-12840 kb fragment, ESC and MSC cells with color intensity reflecting the similarity. Figure 4.6b shows that in the local MI score, segments tend to be first ordered by their length (descending) and then by the overlap between domains (ascending from match to mismatch). The behavior of the local BP score seems to be easier to interpret, as for example the perfectly overlapping segments 22-28 in Figure 4.6a are consistently scored by the local BP score, while the local MI assigns some of them (for example segment 25 in Figure 4.6b) scores similar to rearranged ones (like segments in the 14-20 range).

Increasing evidence suggests the role of TADs in defining regulatory landscapes and limiting the activity of some regulatory elements [And+13] [Nor+12]. A straightforward way to investigate this phenomenon would be by examining a correlation between the local score and the differential gene expression or methylation pattern genome wide. To do this all-by-all Hi-C datasets pairing of available 6 cell lines were created. Then, every pair was assigned local BP and MI scores as well as a fold change of gene expression and methylation for every gene. The differential gene expression was derived based on publicly available datasets. As the number of genes were high, they were aggregated into 30 quantiles according to the local score value in order to reduce noise. Finally, to examine the correlation between gene expression or methylation fold change and the domain rearrangement score, the median quantile local score versus median fold change was inspected.

The relationship between the local rearrangement score and the gene expression fold change turned out to be significant for 5 out of 15 pairings in case of the local BP score and 11 out of 15 pairings for the local MI score as measured by the Spearman correlation at significance level 0.05. Unexpectedly, in the case of the local BP score, 8 out of 15 pairings exhibit negative correlation between the rearrangement score and the fold change. For the local MI score, this correlation is positive for each pairing as expected. However as both small and large values of this score may represent high and low domain overlap, it is difficult to unambigously interpret the outcome. In the case of the local BP score, the results are closer to what seems to be expected for the methylation data where 13 out of 15 pairings exhibit positive and significant Spearman correlation with mean coefficient value of 0.85. Taken together, this result demonstrates that simple analysis of relationship between differential gene expression and TAD segmentation doesn't provide conclusive findings and a more in-depth research need to be conducted in order to elucidate the alleged influence of variability in local TAD partitioning on regulatory mechanisms. A possible reason for the lack of relationships between the local score and gene expression may be partly related to the low quality of the data used. Another issue is the coarse level of analysis. The combination of local score with other predictors like histone modifications data or gene type (housekeeping or not) could have the potential for more insightful conclusions. Apart from that, the problem of searching for locally TAD-rearranged regions needs more research. Although [SK18] developed a method for discovery of locally (dis)similar TAD segmentations, little is known about alterations in TAD partitionings and their relationship with functional features of the genome.

4.4. Conclusions

This chapter introduces the new metric of chromosome segmentation similarity called BP-score. The comparison of BP-score performance with other known measures show that it can be successfully applied to examine the similarity between different Hi-C datasets and draw useful conclusions. In particular, potential applications of BP-score include assessment of replication quality or tracking the structural similarity across various cell lines.

Additionally, two local similarity measures are presented here narrowing down the assessment of chromatin segmentation differences to sub-chromosomal regions. It turns out that local structural similarity correlates with some functional measurements, especially in the case of DNA methylation.

In summary, we have described 3 new measures of similarity for comparison of chromosome segmentation and proven that our global measure - the BP score, satisfies metric properties. This results improve commonly used approaches for measuring the chromosome segmentation similarity by counting boundary overlaps used in multiple studies [Dix+12; LD+14; Bar+15; Fra+15]. The results described in this chapter were published in [ZW19a].

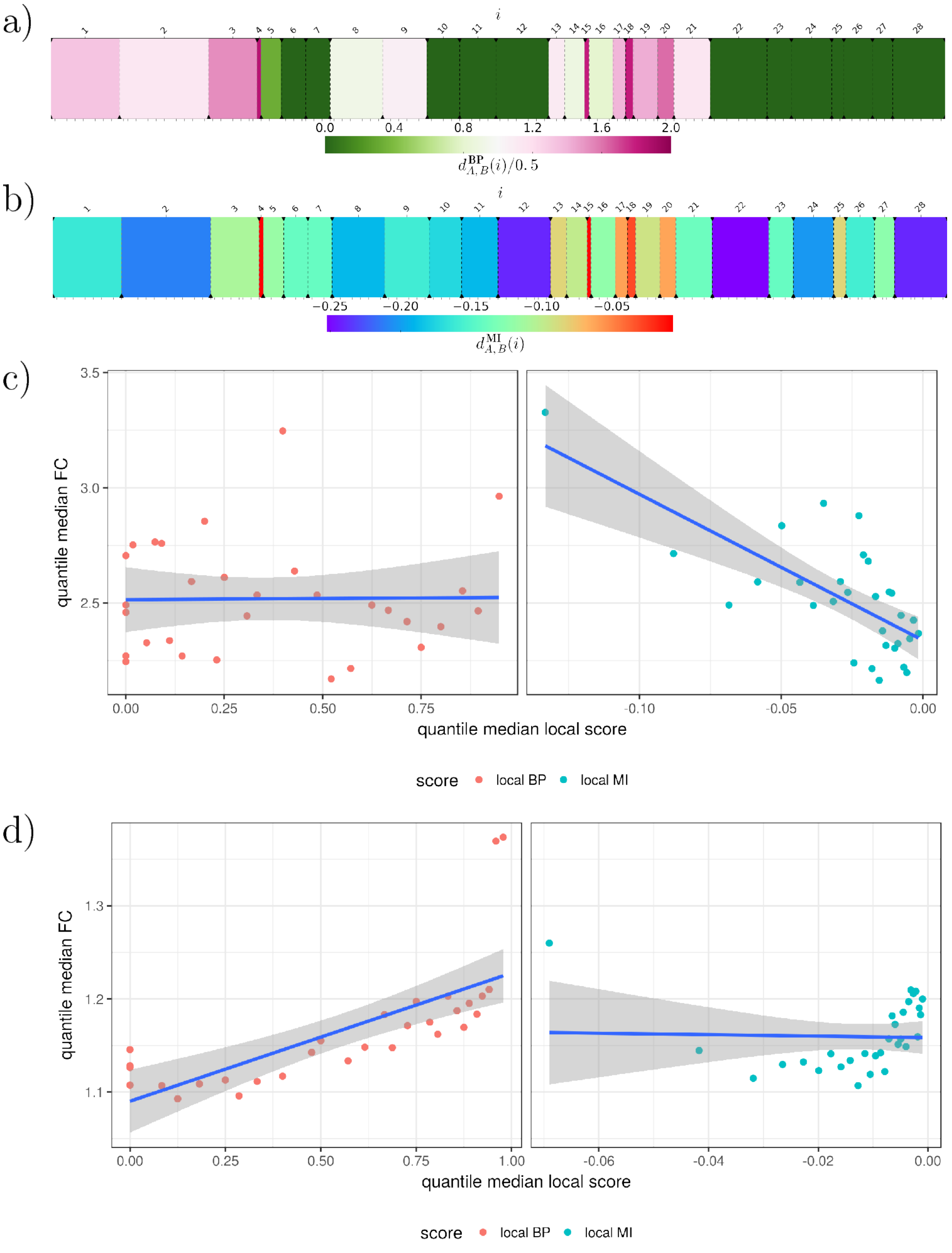


Figure 4.6. Local measures of similarity. a) Graphical representation of TAD segmentation between ESC and MSC cell lines of human chromosome 1: 4,000–12,840 kbp region (40 kb resolution) using a local BP score. The green color illustrates a match between two TADs, whereas the red color indicates a mismatch. b) Same as in a, but using local MI score. This time the color scale quantifies a match between two domains and the segment length. For this reason a perfect match between 2 domains may have various values of local score (for example domains 22-28) and hence to indicate that another color palette was used. c) The relationship between median local score of domain overlaps and median fold change of expression of genes residing inside them divided on 30 quantiles (domain overlaps were grouped into 30 quantiles based on their local score). d) The same as in c, but for methylation fold change data.

CHAPTER 5

Hi-C Differential Analysis

The differential analysis of Hi-C data aims to quantify and compare the contact difference between two or more experiments of interest. Due to occurrence of various biases, direct comparisons of interaction abundances are difficult to interpret. The usual solution is to model the contact variability at pairs of loci and seek for deviations using the hypothesis testing framework. As a result, a significance map is obtained where each entry in the matrix represents the magnitude of interaction difference for the respective pair of regions between the compared experiments.

This chapter presents a new method for Hi-C differential interactions discovery. The first section discusses common issues encountered when comparing contact abundances between Hi-C matrices. The influence of biases and normalization techniques on this process is addressed. Next, selected methods designed for Hi-C differential analysis are described. Finally a new software package called DiADeM (differential analysis via dependency modeling) is presented and the results of its application are compared with existing approaches.

5.1. Introduction

Unbiased comparison of interaction abundances between Hi-C datasets requires a proper treatment of coverage and contact decay. The signal associated with both mentioned factors can vary significantly even for technical replicate repetitions of a single experiment. Therefore, proper addressing of this issue is essential to unravel biologically relevant differences. A natural way to correct the coverage bias is by the application of Hi-C normalization procedures. However, as demonstrated in the next section, this approach is not satisfactory and can lead to other problems. Interestingly raw Hi-C data seems to exhibit significant correlations between vectors of equally distant pairs of regions even between different cell lines. The relationship between contact abundances leads to simple and intuitive definition of Hi-C differential interaction, which can be leveraged to construct a background model for discovery of differential interactions.

5.1.1 Influence of Biases on Differential Analysis

The first source of complications in Hi-C differential analysis is the varying coverage. Usually, high coverage is desired at all chromosomal locations as this increases signal to noise ratio. In practice, sequencing depth is costly. Therefore various Hi-C experiments will end up having different coverages depending on experimental budget.

As a result, researchers are often faced with an issue of comparing contact maps with diverse coverages (Figure 5.1a). This situation poses a problem, because direct comparison of corresponding entries between two matrices will lead to biased results. To avoid this kind of pitfall, contact maps are usually normalized using one of the widely adopted methods. A very popular choice is, the Iterative Correction described in section 3.2.2, which performs well in removing coverage bias (Figure 5.1b). Unfortunately, such normalization may amplify issues related with other sources of bias as it does not account for contact decay.

Divergent contact decays indicate a distinct global chromatin compaction. Such differences can span orders of magnitude as the average number of contacts at a given distance in the first map can be easily 10 times larger then the second one, thereby masking any biologically relevant local variability (Figure 5.2). Normalization methods used to successfully remove coverage bias does not solve the problem in this case, because they fail to remove the contact decays difference. The situation turns out to be even more difficult as various normalization procedures may lead to opposite patterns of relative contact decays (Figure 5.2 middle and bottom row). For example, the relationship between raw data decays of human IMR90 and MSC cell lines display much stronger signal for the former dataset within the prevalent range of the separation distance. When contact maps are subjected to normalization, this pattern changes considerably. If HiCnorm is used, the contact intensity of the MSC dataset begins to exceed that of IMR90 around the 80th diagonal. However, when ICE is applied to correct Hi-C matrices, then 2 decays cross each other at the 6th diagonal. Both methods show consistent behaviour for higher ranges of contact decays (Figure 5.2, third column, rows 2 and 3), but they diverge significantly within close separation distances (Figure 5.2, second column, rows 2 and 3), where the most reliable interaction data is collected.

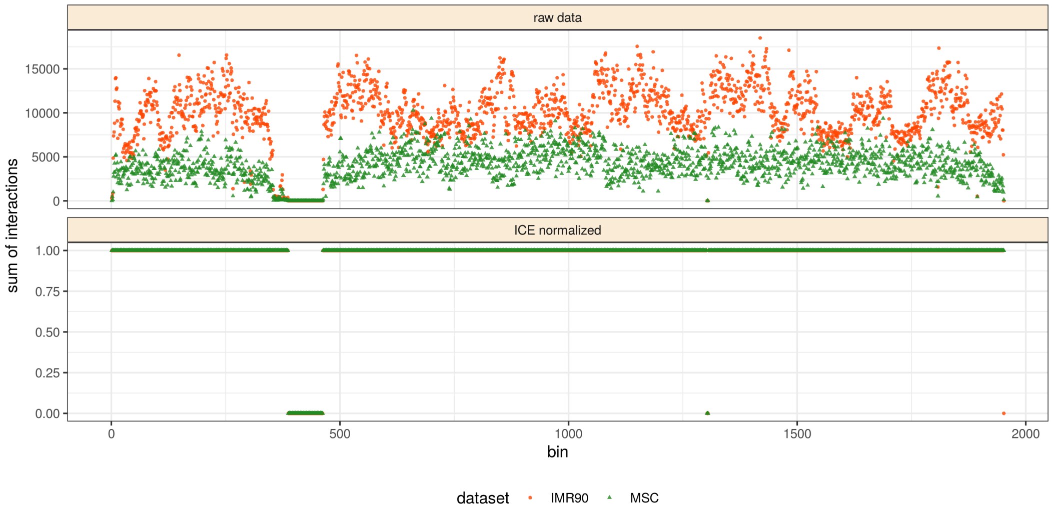


Figure 5.1. The comparison of coverages in 2 Hi-C experiments conducted in different laboratories and using distinct cell lines. Top: Raw data. Bottom: After ICE normalization.

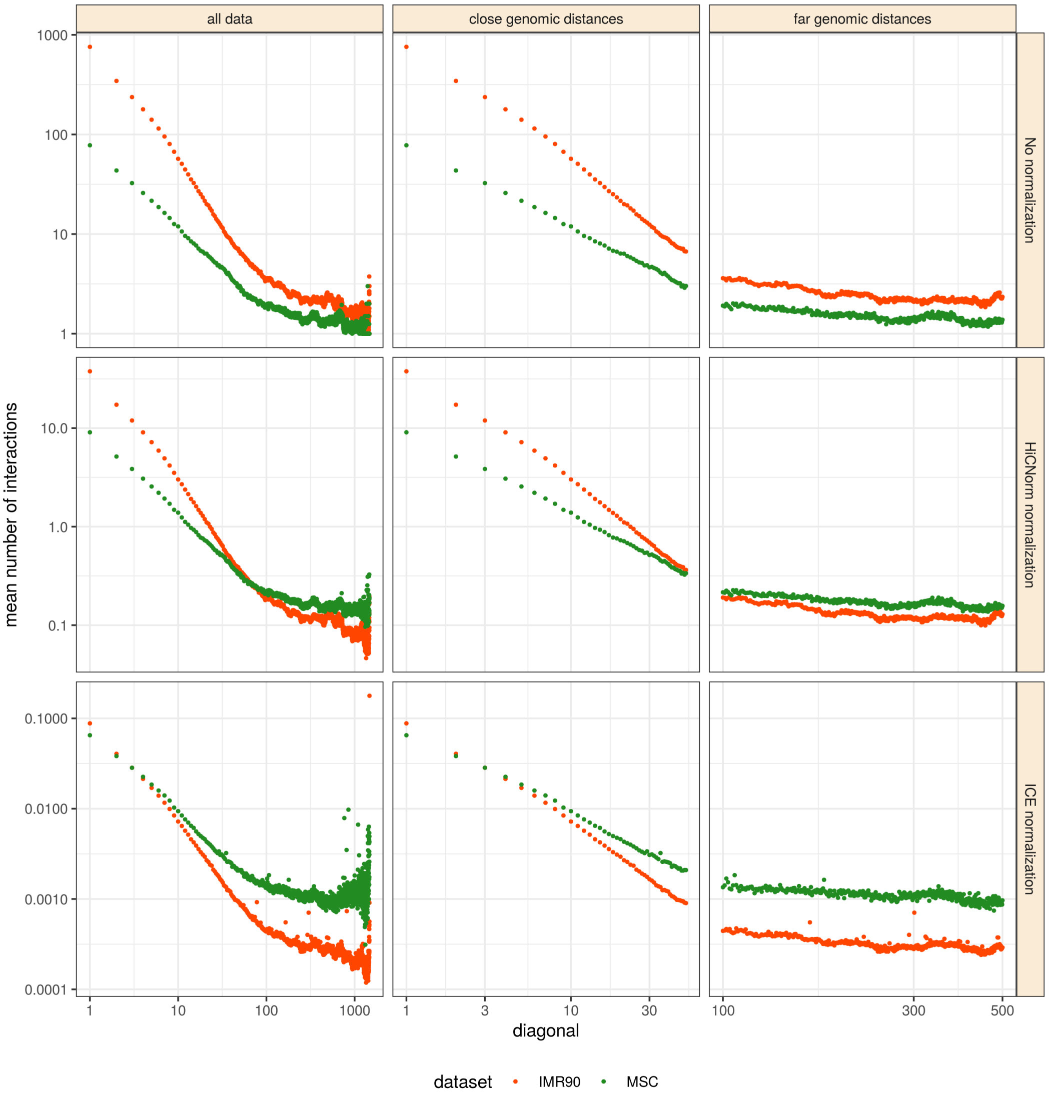


Figure 5.2. The influence of normalization on decay bias. The first row illustrates the existence of the decay phenomenon in raw Hi-C data. Colors represent cell type. Rows 2 and 3 depict how contact decay changes upon normalization using HiCnorm and ICE. Columns 2 and 3 are zoom in of the range of genomic distances from column 1. Note the logarithmic scale.

5.1.2 Correlation Between Interaction Patterns

Although the differences between contact decays of various raw Hi-C datasets may reach orders of magnitude, the correlations among pairs of interaction patterns at fixed genomic separation indicate a high degree of similarity. This observation appears to be pervasive across different replicates, experiments and most surprisingly cell lines. The similarity between contact intensity profiles measured by Pearson, Spearman or Kendall correlation turns out to be significant in most cases for pairs of regions separated by up to a few megabases. The maximum separation preserving statistical significance ranges depending on the dataset quality and the chromosome size. Typically, around 10% of diagonals follow described behavior. Despite only a low fraction of decay exhibiting significant between-diagonal correlation, it should be emphasized that the majority of contacts are located at this particular genomic distance spectrum. This observation is also in agreement with a previously reported feature of Hi-C data - the power law distribution of contact decay. Even though the importance of the Power-law distribution itself might not be essential, it is indeed one of the distributions matching experimental data [LA+09].

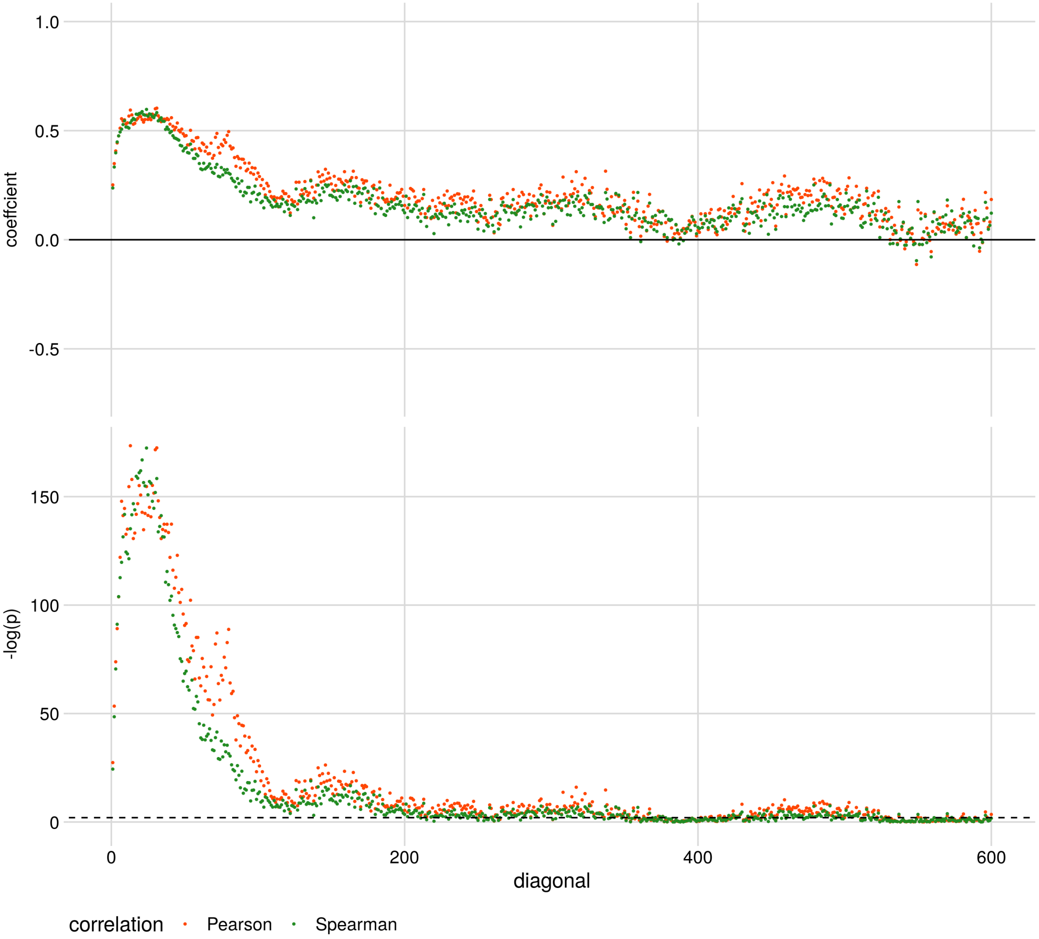


Figure 5.3. The correlations between respective pairs of regions of interaction profiles at given genomic distance. At low genomic distances where most of Hi-C data is collected, the interaction profiles between cell types are significantly similar as indicated by: various correlation coefficients (top) and b) respective p-values (bottom).

5.2. Available Methods Overview

The problem of detecting differential chromatin contacts gained significant attention in recent years. The large popularity of this issue is mostly owed to the importance of understanding the interplay between various regulatory elements. For example, the analysis of differential interactions could be useful in discovering new non-coding functional regions of chromatin. Until now, numerous approaches have been proposed. The current section discusses selected published results on Hi-C differential analysis.

5.2.1 diffHiC

Hi-C experiments are in fact a particular type of NGS assays, hence some approaches to compare chromatin contact abundances can be borrowed from the field of differential gene expression analysis. One of the most popular methods for the discovery of Hi-C differential interactions is diffHic, an R package based on edgeR framework mentioned in section 2.2.3 [LS15]. The main difference between edgeR and diffHic is that the latter uses the bin pair concept instead of a gene. DiffHic requires replication and therefore only experiments with at least 2 repetitions per group can be analyzed using this method.

The framework allows to perform a read alignment and many interaction filtering strategies. Prior to modelling the experimental design, the data is normalized to get rid off library specific biases. The removal of biases is performed by first fitting a LOESS regression against the MA relationship derived from technical replicates of specific experimental group and then adjusting offsets of every bin pair using a fitted trend (Figure 5.4). Similarily to edgeR, diffHic models read abundances for specific bin pair using a GLM:

$$\mathbb{E}[Y_{bi}] = \mu_{bi} = \sum_{j=1}^{p} X_{ij}\beta_{bj} + o_{bi}$$

Here, Y_{bi} denotes the interaction abundance for bin pair b in sample i, X_{ij} refer to the element of design matrix for i-th sample originating from treatment j and o_{bi} is offset term encompassing sequencing depth and normalization factors. The distribution of counts for bin pair b in sample i is assumed to follow the quasi Negative Binomial distribution, which variance can be specified as:

$$\mathrm{Var}[Y_{bi}] = \sigma_b^2 \left(\mu_{bi} + \phi_b \mu_{bi}^2\right)$$

The coefficient ϕ_b represents the NB dispersion for bin pair b and its value is estimated by fitting an abundance dependent trend to the NB dispersions across all bin pairs. Once ϕ_b is obtained, σ_b^2 can be estimated by performing a robust empirical Bayes procedure. Finally, individual bin pairs may be tested against differential interactions using the Quasi-Likelihood F-test and corrected for multiple hypothesis testing.

It is worth noting that a similar approach based on modeling interaction counts with Negative Binomial GLM was also adopted in multiHiCcompare method [SCD19]. Essentially, the concepts used therein resemble those applied in diffHic with the main difference being the introduction of genomic distance as an additional predictor in modeling contact abundances.

5.2.2 FIND

An interesting approach to the discovery of differential chromatin contacts is to compare spatial neighborhoods for pairs of regions as implemented in FIND method [DCZ18]. The study conducted by Djekidel and coworkers reports that the intensity

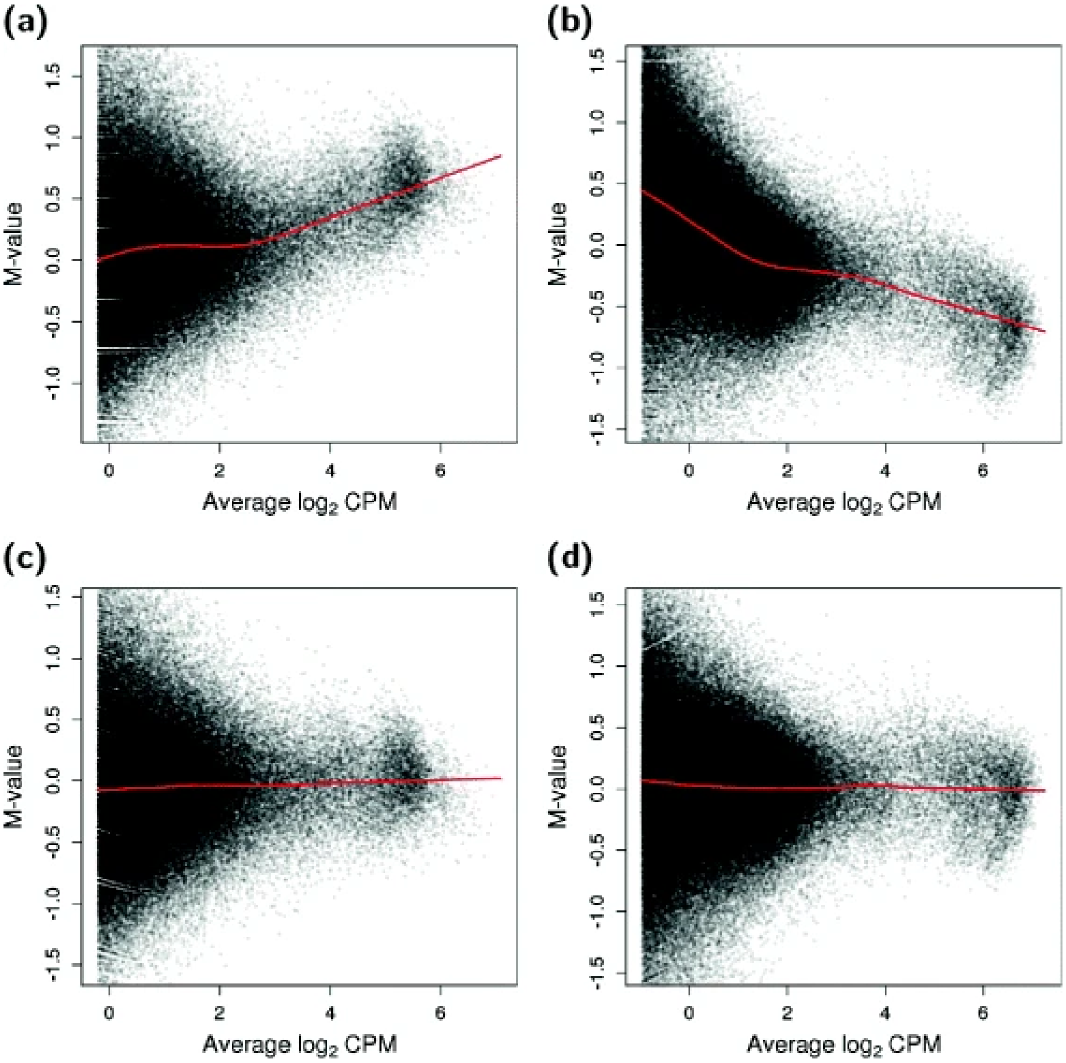


Figure 5.4. The illustration of trended biases. Each point represents the relationship between $\log_2$-count-per-million interactions averaged over 2 repetitions of given cell line Hi-C versus library size-adjusted $\log_2$-fold change (M-value) between the same replicates. The panels depict trends before (a,b) and after (c,d) normalization. The datasets studied are ERG-treated cells (a,c) and ESC cells (c,d). Figure from [LS15].

of an interaction is strongly dependent on its nearest neighbors. Therefore, comparison of contact abundances in the area surrounding the interacting chromatin regions should provide insight regarding the similarity of their local organization. In the FIND model, the neighborhood of interacting loci i, j constitutes the squared window fragment of the Hi-C contact map having a width w, which is centered on i, j. Intuitively, if the interaction i, j is not differential, then the distribution of (euclidean) distance between i, j and its k-nearest neighbor (examined across replicates) should be similar between both conditions. To examine the neighborhoods, point distribution in the surroundings of i, j is approximated with a homogeneous Poisson process. Under this assumption, the probability of observing the kth nearest neighbor at the distance x_{ik} from (i, j, μ) in the nth Hi-C replicate can be expressed as [BA14]:

$$f(x_{n,k}) = \frac{(4\lambda(\mu)\pi)^k}{3^{k-1}(k-1)!} x_{n,k}^{k-1} \exp\left(-\lambda(\mu)\frac{4\pi}{3}x_{n,k}^3\right)$$

The parameter $\lambda(\mu)$ needs to be estimated for each cell in neighborhood of i, j separately. This could be done given n_c replicates of treatment c using the maximum likelihood estimator:

$$\hat{\lambda}_k^{(c)}(\mu) = \frac{n_c k - 1}{\frac{4\pi}{3}\sum_{n=1}^{n_c} x_{n,k}^3}$$

After obtaining $\hat{\lambda}_k^{(1)}(\mu)$ and $\hat{\lambda}_k^{(2)}(\mu)$, one can test the following null hypothesis H_0 : $\hat{\lambda}_k^{(1)}(\mu) = \hat{\lambda}_k^{(2)}(\mu)$ versus a two sided alternative. The test is easy to perform because Djekidel and coworkers show that under the null hypothesis, the ratio of $\hat{\lambda}_k^{(1)}(\mu)$ and $\hat{\lambda}_k^{(2)}(\mu)$ follows Fisher distribution with $2n_1k$ and $2n_2k$ degrees of freedom.

In order to examine whether a paired region i, j is significantly enriched or depleted in interactions between 2 condtions for fixed neighborhood w, $w^2 - 1$, cells are tested against the null hypothesis specified above. To combine resulting p-values and obtain final significance for i, j interaction r-th, the ordered p-value method is used, which decides a series of tests as significant if at least r of them are significant [ST14].

5.2.3 HiCcompare

Occasionally, the replication for a specific Hi-C experiment is not available thereby preventing the usage of methods such as diffHic or FIND. One alternative approach which enables us to conduct comparative analysis for 2 samples without replicates is HiCcompare [Sta+18]. This method is based on joint normalization of paired datasets. The main concept in HiCcompare are MD plots - an analog of MA plots. In MD-plot interactions are divided by genomic distance (D - the predictor) and the log fold change between 2 contact intensities of corresponding Hi-C maps entries ($M = \log_2(IF_2/IF_1)$ - the response). According to the authors, this relationship captures relevant between-dataset biases. To further normalize the data, HiCcompare estimates the function $f(D)$ from MD relationship by the application of locally weighted polynomial regression. Given $f(D)$, normalized interaction frequencies (IF) can be calculated according to following formulas:

$$\log_2\left(\hat{IF}_{1D}\right) = \log_2(IF_{1D}) + f(D)/2$$
$$\log_2\left(\hat{IF}_{2D}\right) = \log_2(IF_{2D}) - f(D)/2$$

In order to compare individual interactions, the normalized MD plot is constructed and each value M_i associated with pair of regions i is centered and scaled by the chromosome-wide mean $\bar{M}$ and standard deviation σ_M to provide the normally distributed variable Z_i:

$$Z_i = \frac{M_i - \bar{M}}{\sigma_M}$$

Finally, to test if the pair of regions i is significantly enriched or depleted in interactions, a Z-test is performed. In the end, obtained p-values are adjusted for multiple hypothesis testing.

5.2.4 SELFISH

Another approach based on comparing interaction neighborhoods was proposed by [AAL19]. The method called SELFISH uses the local self-similarity measure, which was shown to outperform many classical image descriptors when searching for similar patterns not sharing common image properties [SI07]. The main assumption of the SELFISH model is as follows: if i, j is a differential chromatin interaction between contact maps A and B, then it should be accompanied by contact difference at i, j as well as its surrounding area. Inspecting the neighborhood of i, j rather than individual pixels should therefore reduce the risk of discovering single high value differences, which are likely resulting from noise. To quantify the information contained in neighborhood (impact region) of i, j, the convolution of contact map $\hat{A}$ and Gaussian filter with radius r_k is calculated resulting in matrix $G^{\hat{A}}_{r_k}$. Before $G^{\hat{A}}_{r_k}$ is obtained, the matrix A need to be normalized in order to remove contact decay bias:

$$\hat{A}(i,j) = \frac{A(i,j) - \mu_d}{\sigma_d}$$

To examine various impact radii, i.e. the size of the neighborhood being affected, $G^{\hat{A}}_{r_k}$ is calculated with varying k. Finally, an estimate of influence, which i, j impose on its surrounding area of size k can be quantified by a vector of values at pixel i, j taken over every matrix G_{r_k}: $\Gamma_{\hat{A}}(i,j) = (G^{\hat{A}}_{r_1}(i,j), G^{\hat{A}}_{r_2}(i,j), ..., G^{\hat{A}}_{r_n}(i,j))$. According to the authors, the difference between $\Gamma_{\hat{A}}$ and $\Gamma_{\hat{B}}$ can not be used to determine whether a pixel is associated with significant change of contact intensity because of the biases in the interaction frequencies. Instead, a first order derivative of Γ is used to compare the impact regions and mitigate the influence of biases:

$$\frac{d\Gamma}{dr}(i,j,k) \approx \Delta\Gamma(i,j,k) = G_{r_{k+1}}(i,j) - G_{r_k}(i,j)$$

In the last step, to test if difference at i, j is significant, a corresponding p-value is computed according to following formula:

$$P^k_{A,B}(i,j) = \mathbf{Pr}\left(X > (\Delta\Gamma_A(i,j,k) - \Delta\Gamma_B(i,j,k))\right)$$

Here $X \sim N(\mu, \sigma)$ and both μ and σ are estimated from the distribution of $\Delta\Gamma_A - \Delta\Gamma_B$ separately for each k. It is worth noting that in contrast to the FIND method, SELFISH is able to manage comparisons of Hi-C experiments without replication.

5.3. DiADeM Method

Evident correlations of interaction profiles described in section 5.1.2 suggests that the majority of chromatin contact patterns exhibit high degree of similarity even across different cell types. Consequently, only a small fraction of interactions can be considered as relevant differential contacts. Those observations are the basis of the DiADeM (Differential Analysis via Dependency Modeling) method, which assumes that the prevalent portion of Hi-C contacts are non-differential and therefore may be used to model biological variation.

5.3.1 Diagonal Interaction Patterns

In order to investigate the relationship across interaction patterns of various Hi-C datasets, we compared contact abundances between pairs of contact maps at respective regions. A simple method to study the dependency between two Hi-C profiles would be to inspect paired interaction sets $\mathcal{D}_k$, which relate a number of contacts in matrix A versus B at every cell i, j:

$$\mathcal{D}_k = \{(a_{ij}, b_{ij}) \mid k = |i - j|, a_{ij} \neq 0, b_{ij} \neq 0\}$$

Close examination of paired interaction sets revealed linear relationship between inspected datasets for large range of genomic distances (Figure 5.5). We noted that for multiple analyzed pairs of cell type specific datasets, the conditional variance of the observed linear trend increased along with interaction growth as indicated by the existence of funnel-like pattern. This phenomenon is referred to as heteroscedasticity and is known to impede modeling of correlated data due to violation of the constant variance assumption, which in turn prevents the application of a simple linear regression technique. A common solution to model the linear relationships in presence of heteroscedasticity is the application of the GLM framework.

Interestingly, a number of paired interactions seems to disobey the general pattern described above. For example, the right-bottom panel on Figure 5.5 includes outlying points, which indicate unusually large number of interactions in contact map B given the number of interactions in matrix A. We consider this phenomenon as a proxy of differential chromatin interaction at this stage.

5.3.2 Robust Regression

In order to effectively model the relationships described in the previous section, two issues need to be properly taken care of: the existence of outliers and heteroscedasticity. The first problem can be remedied with robust regression, which uses estimating methods resistant to the presence of unusual observations. The most common among such methods uses M-estimators. The definition of M-estimator as specified by [HR81] is following:

Definition 7. *Any estimate T_n defined by:*

$$\sum_{i=1}^{n} \Psi(x_i; T_n)$$

where $\Psi(x;\theta) = \frac{\partial}{\partial\theta}\rho(x;\theta)$ and ρ is an arbitrary function is called an M-estimate (or maximum likelihood type estimate).

The choice of $\rho(x;\theta) = -\log f(x;\theta)$ provides ordinary ML estimate. In case of the simple linear model discussed so far an M-estimate of regression are defined as:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) \tag{5.1}$$

which leads to following estimating equation:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right) = 0$$

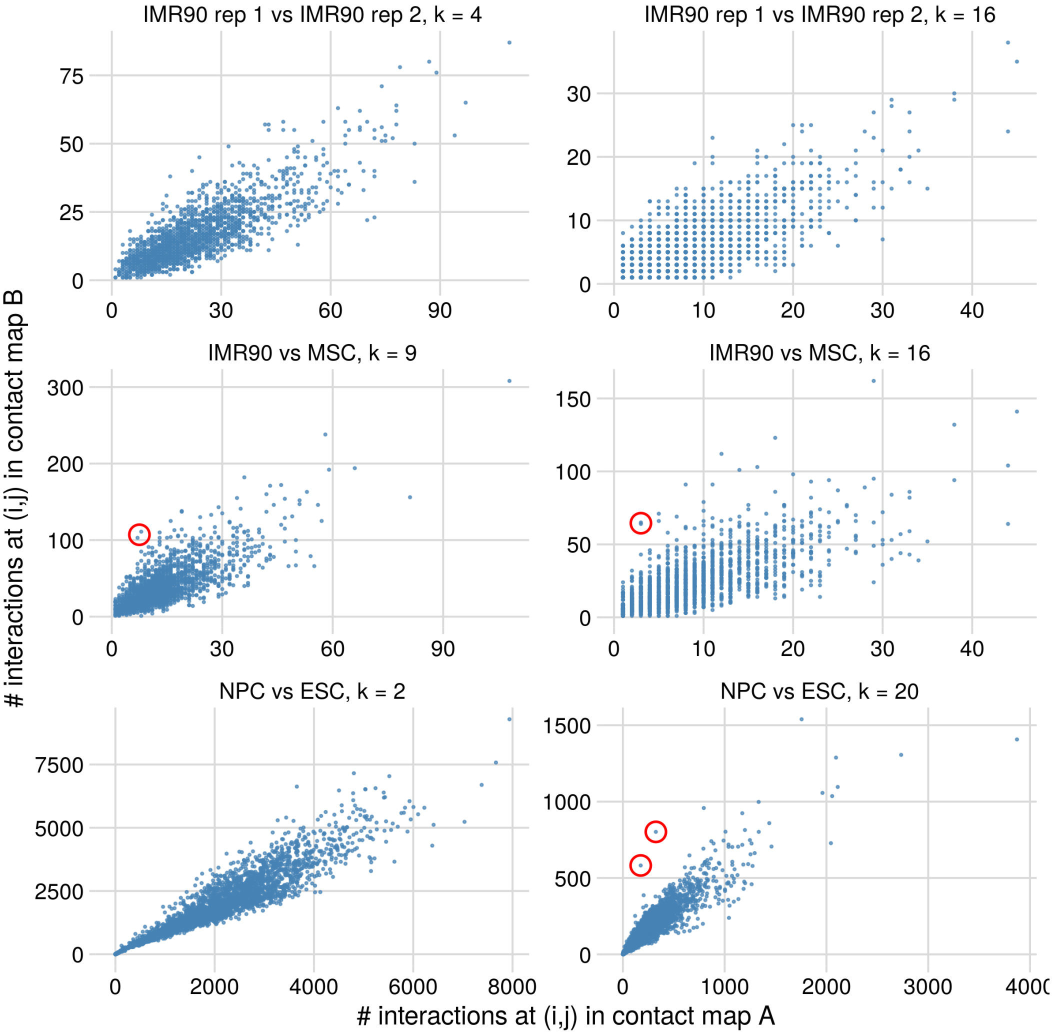


Figure 5.5. Relationship between interaction profiles of various pairs of Hi-C contact maps at selected genomic distance. Each point represents a different pair of regions at genomic distance (diagonal) k. The horizontal axis indicates a number of interactions in contact map A while the vertical axis expresses contact abundance in matrix B. Note the existence of outlying points (indicated with red circles), especially in NPC vs ESC comparison, $k = 20$.

where: $r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ is i-th residual and σ is scale parameter, which can either be an external estimate or estimated simultaneously. The usual assumptions regarding ρ function are: $\rho(r)$ is a non decreasing function of $|r|$, with $\rho(0) = 0$ and strictly increasing for $r > 0$ where $\rho(r) < \rho(\infty)$ [KS11]. The role of ψ function is to weight residuals and therefore reduce the influence of outliers on parameter estimates. In general there are 2 types of ψ functions:

- soft re-descending - residual weight decreases as r_i value increases,
- hard re-descenders - once the value of residual exceeds some threshold its weight equals zero.

Alternative expression for the above estimating equation is as follows:

$$\sum_{i=1}^{n} w_i r_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = 0$$

where: $w_i = \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right) / \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right)$ and usually it is solved by the application of Iteratively Re-weighted Least Squares (IRLS) algorithm. The IRLS starts with OLS estimates of $\boldsymbol{\beta}^{(0)}$ and proceeds by repeatedly improving the current estimate of $\boldsymbol{\beta}^{(t)}$ by assigning smaller weights to more deviating observations. The procedure is repeated until a convergence in terms of $\boldsymbol{\beta}$. Apart from estimating regression coefficients, the IRLS algorithm is also used to determine outliers - the observations, which were assigned zero weight.

The approach described above performs well when error terms are independent and identically distributed. However, in the presence of heterscedasticity, it acts poorly as the down-weighting of observations is done without prior consideration of their conditional variance [SC88]. A solution to this problem was suggested by [KS11] who derived a novel SMDM estimator, which accounts for the existence of outliers and heteroscedasticity at the same time. The SMDM-estimator determines model parameters using a procedure consisting of following steps:

1. S-estimation,
2. M-estimation,
3. D-estimation,
4. M-estimation.

Informally, the S-estimator of regression is an estimate of $\boldsymbol{\beta}$ derived from a scale statistic in an implicit way [RY84]. More precisely, S-estimate of $\boldsymbol{\beta}$ is a minimizer of a robust M-estimate of scale of the residuals, as follows [HR81]. For every value of $\boldsymbol{\beta}$, we estimate $\hat{\sigma}(\boldsymbol{\beta})$ by solving:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta$$

for σ, where $0 < \delta < 1$ is a constant depending on the distribution of X. The function ρ may not be the same as in equation 5.1, but is usually from the same family. The S-estimate of $\boldsymbol{\beta}$ is then defined as the value $\hat{\boldsymbol{\beta}}_S$, which minimizes $\hat{\sigma}(\boldsymbol{\beta})$:

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}(r(\boldsymbol{\beta}))$$
$$\hat{\sigma}_S = \hat{\sigma}(r(\hat{\boldsymbol{\beta}}_S))$$

The application of S-estimate followed by M-estimate is referred to as MM-estimate. The MM-estimates are defined as a local minimum of 5.1 obtained using an iterative procedure starting from S-estimate of regression $\hat{\boldsymbol{\beta}}_S$ and using $\hat{\sigma}_S$ as scale parameter [KS11].

Finally, the D-estimate refers to the novel scale estimator derived by [KS11] and called Design Adaptive Scale Estimate. The scale parameter σ is estimated by solving following estimating equation for σ_D:

$$\sum_{i=1}^{n} \tau_i^2 w\left(\frac{r_i}{\tau_i \sigma_D}\right)\left[\left(\frac{r_i}{\tau_i \sigma_D}\right)^2 - \kappa\right]$$

where: w is a weighting function as before, κ ensures Fisher consistency and τ_i are correction factors designed to reflect the heteroskedasticity of the distributions of the residuals r_i, which depends on the leverage of i-th observation h_i and selected ψ function. The distribution of residuals r_i is unknown and therefore it is approximated with *von Mises expansion* of β [KS11].

During the last step (M-estimate), a previously estimated scale parameter $\hat{\sigma}_D$ is plugged into 5.1 in order to determine the ultimate $\hat{\beta}$. As this process is performed with an IRLS algorithm, it is possible to obtain weights related with each point and evaluate the outliers. After filtering out unusual observations, the remaining data is used to fit the Negative Binomial regression that models the biological variability devoid of differential signal.

5.3.3 Robust Negative Binomial GLM

An alternative for the approach described in Section 5.3.2 would be direct estimation of model parameters using robust estimators. In 2014 Aeberhard and coworkers derived M-estimators for Negative Binomial regression [ACH14]. The estimates for $\boldsymbol{\beta}$ and ϕ may be obtained by solving 2 equations. First, a robust estimate of $\boldsymbol{\beta}$ is obtained as a solution to:

$$\sum_{i=1}^{n} U_{\boldsymbol{\beta},i}(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n}\left[\psi(r_i)V^{-\frac{1}{2}}(\mu_i)\frac{\partial\mu_i}{\partial\eta_i}w(\boldsymbol{x}_i)\boldsymbol{x}_i - a_i(\boldsymbol{\beta})\right] = \mathbf{0} \tag{5.2}$$

Here $r_i = (y_i - \mu_i)V^{-\frac{1}{2}}(\mu_i)$ is Pearson Residual, $w(\boldsymbol{x}_i)$ is the weight limiting influence of possible leverage points and $a_i(\boldsymbol{\beta}) = \mathbb{E}[\psi(r_i)]V^{-\frac{1}{2}(\mu_i)}\frac{\partial\mu_i}{\partial\eta_i}w(\boldsymbol{x}_i)\boldsymbol{x}_i$ is a correction term, guaranteeing the Fisher consistency of the model. The robustness is provided by the application of ψ functions as described in Section 5.3.2. A common choice of ψ includes Huber ψ-function defined as:

$$\psi_{\text{Huber}}(r;c) = \max(-c, \min(c, r))$$

and Tukey bi-weight function:

$$\psi_{\text{Tukey}}(r;c) = \begin{cases} ((r/c)^2 - 1)^2 r & |r| \leq c \\ 0 & |r| > c \end{cases}$$

The parameter c is called the tuning constant and it need to be manually adjusted. Various values of c will lead to different ratios between efficiency of estimator and its robustness against outliers. Usually, the optimal value for the tuning constant is selected based on extensive simulation studies where model parameter values are known in advance. In a study conducted by Aeberhard and coworkers, the authors deduce $c = 4$ as optimal trade-off between both mentioned properties of an estimator. The analysis of performance for several examined ψ functions suggests that the choice of Tukey bi-weight leads to the least biased results. Similar conclusions were also reported in another study applying the model developed by Aeberhard and coworkers to differential gene expression experiments [LL18].

Equation 5.2 can be solved numerically by using the Fisher scoring method similarly to ordinary ML estimates of Negative Binomial regression. According to the authors, a unique solution is not guaranteed when using re-descending functions. However, in practice, no issues were observed while using this class of ψ function with MLE as initial parameter values. Having estimated $\boldsymbol{\beta}$ an estimate for ϕ may be obtained as solution to following equation:

$$\sum_{i=1}^{n} U_{\phi,i}(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n} \left[\frac{\psi(r_i)}{r_i} \Psi_\phi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) w(\boldsymbol{x}_i) - b_i(\phi) \right] = 0$$

where $\Psi_\phi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi)$ is the estimating equation of ordinary MLE specified in section 2.2.2 and $b_i(\phi) = \mathbb{E}\left[\frac{\psi(r_i)}{r_i} \Psi_\phi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) w(\boldsymbol{x}_i)\right]$ is another Fisher consistency term. Additionally the authors derive an asymptotic distribution for parameter estimates, so their confidence intervals may also be obtained.

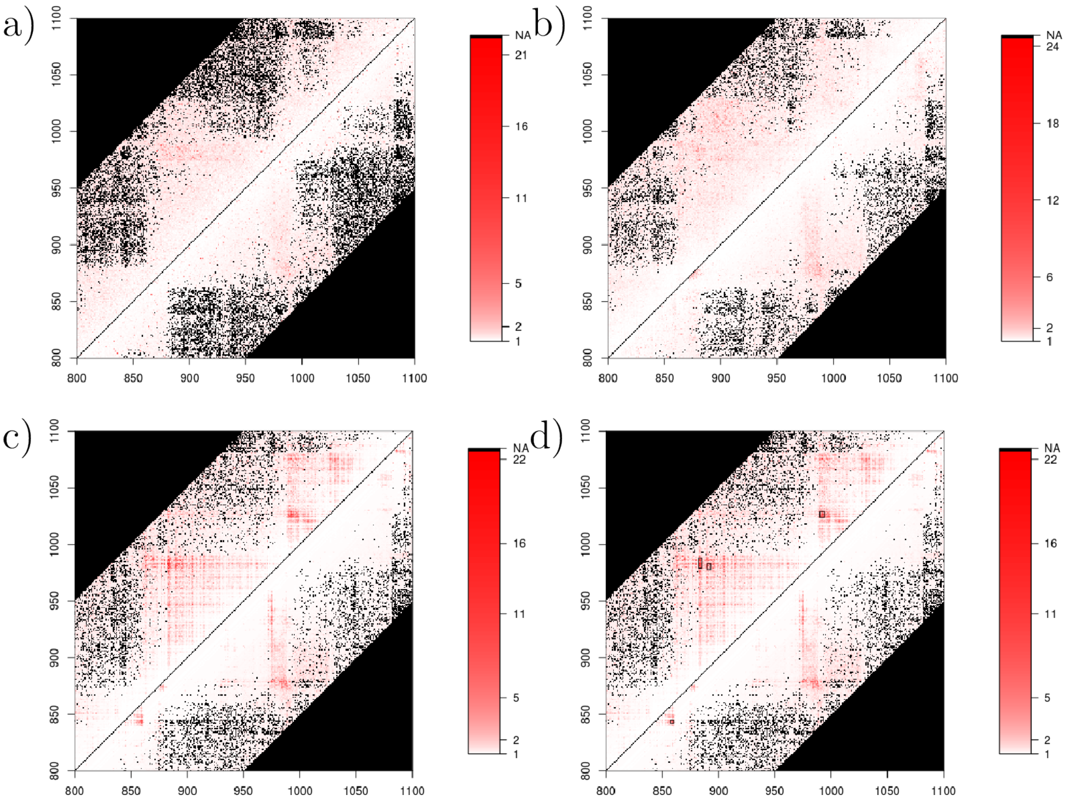


Figure 5.6. Significance maps. Each cell indicate the enrichment probability ($-\log p$). The cells above the main diagonal measure the enrichment of B with respect to A while below the main diagonal are the enrichments of A versus B. For better clarity, only a fraction of the chromosome is shown. a) IMR90 bootstrapped data. b) IMR90 biological replicates data. c) IMR90 versus MSC. d) Same as in c, but with annotated long-range differentially interacting regions.

5.3.4 DiADeM Model

The DiADeM method models decay relationships described in section 5.1.2. The main assumption of the model is that, diagonal-wise, contact abundances between two analyzed contact maps exhibit significant correlation. Accordingly, differential chromatin interactions are associated with points not fitting the observed linear relationship rather than the largest absolute difference. Essentially, our method can be summarized as a 3-step procedure:

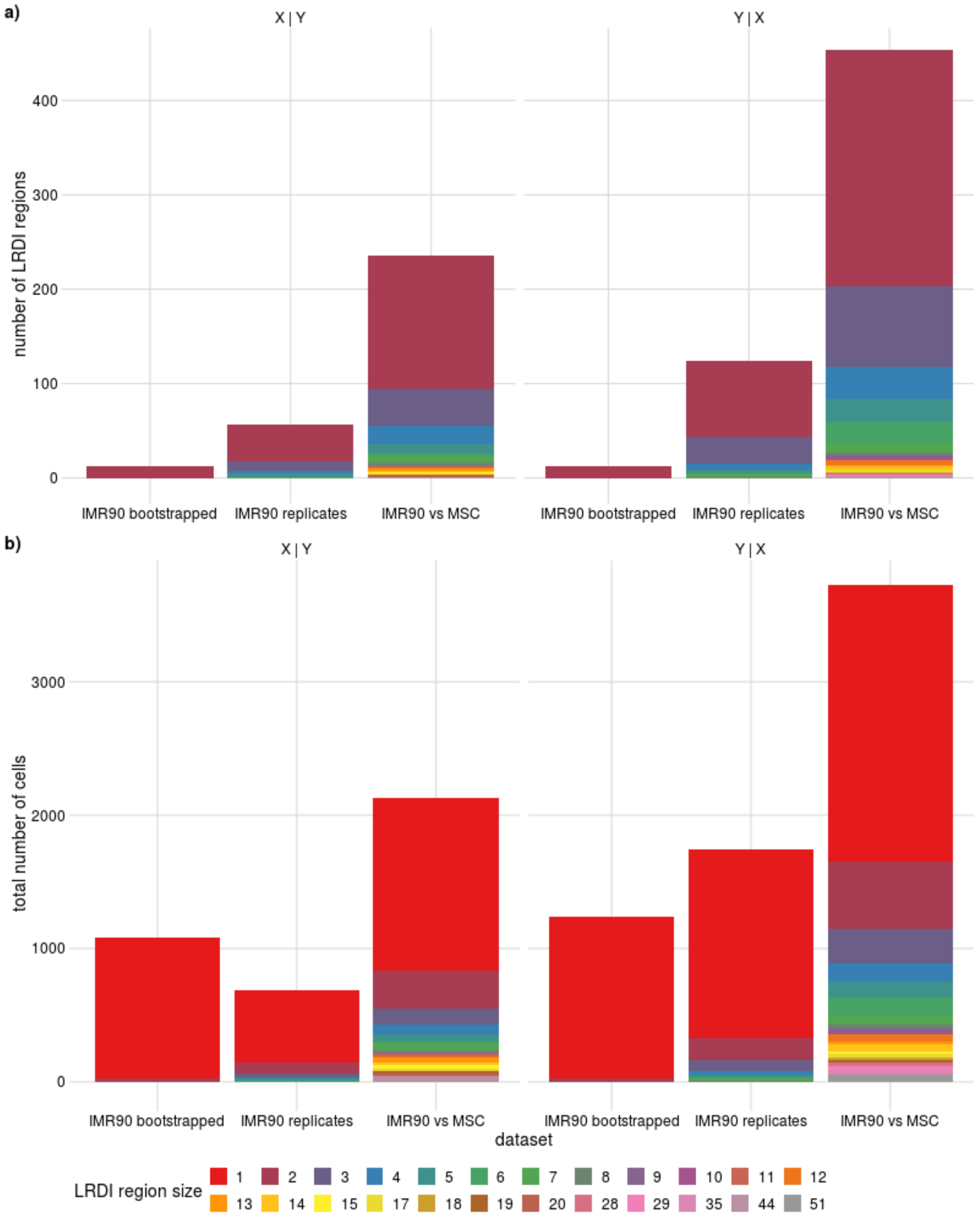


Figure 5.7. Abundances of different sizes of connected components determined on significance maps across various pairs of contact maps. a) The vertical axis indicates the number of connected components of a given size (i.e. containing certain number of differential interactions) at specific pair of cells comparison between the two contact maps (indicated on the horizontal axis). The connected component of size equal to 1 was removed for clarity. b) Similar to a, but now the vertical axis corresponds to connected component size times the number of connected components of this size detected in a specific comparison.

1. aggregate paired interaction sets based on their bivariate distribution similarity,
2. fit Negative Binomial regression for each aggregated paired interaction set,
3. given the model, determine the significance of interaction difference for every pair of regions.

Step number 1 aims to determine sets $\widetilde{\mathcal{D}}_p = \mathcal{D}_k \cup \mathcal{D}_{k+1} \cup ... \cup \mathcal{D}_l$ such that the bivariate distributions of paired interaction sets $\mathcal{D}_{i\in\{k,...,l\}}$ are sufficiently similar to each other. More precisely, the similarity of 2 n- and m-element samples X and Y obtained from multivariate distributions F_X and F_Y can be formulated in terms of the following hypothesis test:

$$\mathbf{H_0}\text{: } F_X = F_Y \text{ vs } \mathbf{H_1}\text{: } F_X \neq F_Y$$

Correspondingly, the rejection of the null hypothesis occurs for the samples that are not sufficiently similar. Testing for equality of multivariate distributions is a well studied problem in statistics. In particular, for a univariate setting, numerous tests are available. For example, the very popular Kolmogorov-Smirnov test or the Wald-Wolfowitz runs test. However, their generalizations to higher dimensions are nontrivial due to the fact that there is no obvious extension of test statistic. An interesting approach for calculating an analogue of runs statistic in multivariate samples was suggested by [FR79]. Their method relies on computing the minimum spanning tree (MST) of pooled sample points and removing edges connecting nodes assigned to different samples resulting in test statistic R - the number of disjoint subtrees. Unfortunately, the devised test is not distribution-free as the null distribution of R depends on structural features of MST. An improvement was was made by [Sch86], who proposed to count the number of occurrences, where the point and its k-nearest neighbors belong to the same sample. The resulting nearest neighbor test is distribution-free, although the exact distribution of the test statistic is not known and only an asymptotic form was given. A remarkable work of [Ros05] provided an exact, distribution-free test based on an inter-point distance. In order to compute the test statistic, the inter-point distances are calculated and used to construct the optimal non-bipartite matching, i.e. a matching, which minimizes the sum of within-pair distances. Then, the number of cross-matches i.e. different sample pairs is used as test statistic. The run time of finding the optimum non-bipartite matching is $\mathcal{O}(N^3)$ where N is the total number of points across two tested paired interaction sets. The density function of this test statistic includes multiple factorial terms making the computations of cdf infeasible for large number of observations. In practice, these two issues turn out to make this statistic unusable given the usual sizes of Hi-C matrices. For this reason and also the availability of software implementation, we decided to choose another test also based on inter-point distances [SR+04]. Its related test statistic is called energy distance (or E-statistic) and can be calculated as:

$$\mathcal{E}_{n,m}(X,Y) = \frac{nm}{n+m}\Bigg(\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\boldsymbol{x}_i - \boldsymbol{y}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \\ - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|\Bigg)$$

In the above formula, X, Y are n, m-element samples from multivariate distributions F_X, F_Y, respectively. High values of E-statistic indicate larger evidence against the null hypothesis. The E-statistic test is distribution free; however, it is not exact, so the null distribution is obtained by drawing random permutations of pooled samples [SR+04].

In order to reduce the computational burden of testing the equality of paired interaction set distributions, we adopted the following simplification. Instead of testing $l-k$ sample hypothesis $\mathbf{H_0}$: $F_k = F_{k+1} = ... = F_l$ a separate test is performed for each consecutive i starting at $k+1$ until $\mathbf{H_0}$: $F_k = F_i$ may be rejected, which gives pool p including diagonals k to $l = i-1$. This procedure is justified by the existence of a contact decay effect as for significant majority of diagonals $i < j < k$ if $\mathcal{D}_j$ will be substantially shifted with respect to $\mathcal{D}_i$ then so will be $\mathcal{D}_k$. The main reason for the grouping procedure is to improve the accuracy of the model estimation process in a subsequent step.

Having grouped paired interaction sets, Negative Binomial regression coefficients are estimated separately for each $\widetilde{\mathcal{D}}_p$. As discussed in section 5.3.2, in order to properly estimate the model, outlying observations need to be managed carefully. One solution is to use IRLS to determine unusual points and discard them. Another method would be to use robust NB GLM described in section 5.3.3. After obtaining parameter estimates $(\hat{\beta}_p, \hat{\phi}_p)$, the significance of enrichment at certain location i, j in matrix B with respect to the same location in A is calculated according to the formula:

$$\mathrm{p}_{ij} = 1 - P(Y < b_{ij}), \text{ where: } Y \sim NB(\hat{\beta}_p a_{ij}, \hat{\phi}_p)$$

The resulting p-values are then corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure and resulting q-values can be shown on significance maps (figure 5.6a-c) [BH95]. The results of DiADeM on different cell line Hi-C datasets (figure 5.6c) can be contrasted with bootstrap replicates data (figure 5.6a) or biological replicates of the same cell data (figure 5.6b). The latter 2 datasets are expected to exhibit a much weaker differential signal than different cell types, which is reflected in significance maps.

5.3.5 Long Range Differential Interactions

The model presented so far determines differential significance of individual pairs of regions i, j between two contact maps. However, inspection of significance maps of different cell types reveal that highly significant pairs of loci tend to gather in larger groups of cells. In contrast, the number and size of such clusters in bootstrapped or replicate data comparisons seems to be much smaller. In order to increase the confidence in the detection of biologically relevant differential interactions, DiADeM performs a simple procedure for aggregation of significant paired interactions, which allow to report larger clumps of significant cells.

When the p-value threshold is specified, each cell of a significance map can be labeled as either significant or not, creating a binary matrix. Such matrix can be easily converted to an incidence list of undirected graph $\mathcal{G}$, whose vertices correspond to consecutive bins of chromatin fiber and edges constitute differential interactions. Therefore, the problem of detecting clusters of significant bin pairs reduces to finding connected components of $\mathcal{G}$. When $\mathcal{G}$ is undirected, as is the case, it suffices to traverse through each non-visited vertex and recursively apply Depth First Search

(DFS) or Breadth First Search (BFS) algorithm in order to determine connected components (see Algorithm 1). An example outcome of the described algorithm is presented on Figure 5.6d, where components containing at least 5 significant pixels were shown.

After the aggregation of significant cells, the number and size of resulting clusters can be compared across datasets presented on figure 5.6. Figure 5.7a illustrate the abundance of various cluster sizes while figure 5.7b depicts the total number of significant cells inside each cluster size at every type of datasets. Notably, the differences are very pronounced and indicate that different cell lines data contain a higher number of clusters as well as their size is larger. This observation might be helpful in order to determine the minimum size of biologically relevant cluster.

Algorithm 1 Connected Components Search

```
Require: Incidence list of $\mathcal{G}$: $L$
  procedure DFS($V, i$)
      $C \leftarrow \varnothing$                          ▷ connected component set
      if $V[i]$ then
          $V[i] \leftarrow$ false
          $C \leftarrow C \cup \{i\}$
      end if
      for $j \in L[i]$ do
          if $V[j]$ then
              $C \leftarrow C \cup$ DFS($V, j$)
          end if
      end for
      return $C$
  end procedure

  $V[1, ..., n] \leftarrow$ true                     ▷ list of unvisited vertices
  $l \leftarrow [\,]$                               ▷ list of connected component sets
  for $i \in \{1, ..., n\}$ do
      if $V[i]$ then
          $C \leftarrow$ DFS($V, i$)
          append($l, C$)                  ▷ append set $C$ as the last element of $l$
      end if
  end for
  return $l$
```

5.3.6 Comparison with Existing Approaches

In order to benchmark DiADeM's performance, it was tested on simulated data and the results were compared with those produced by tools described in section 5.2. First, some of the mentioned methods used the simulation procedure developed in FIND to produce artificial Hi-C data and assess the classification quality. This type of simulated data, although employed here, turned out to not resemble real Hi-C datasets in terms of contact coverage and decay (Figure 5.8). Therefore, an additional simulation protocol based on DiADeM model was suggested. The approach proposed here, is to take a raw interaction set (Hi-C matrix) as input and a DiADeM model trained on a selected pair of contact maps and produce a corresponding replicate interaction set. Afterwards, artificial differential interactions are introduced into paired contacts set by multiplying randomly selected regions in either set by the specified fold change. Every simulated dataset was produced in 2 replicates. Some of the discussed tools don't rely on replication and in such cases, the replicates were pooled. As the simulated dataset is highly unbalanced, i.e. the number of differential interactions comprise less then 10% of all contacts, the Precision-Recall performance metric is preferred over the ROC curve [DG06]. The simulation was repeated 20 times and for each simulated dataset the area under Precision-Recall curve (PRAUC) was used as a final measure of performance.

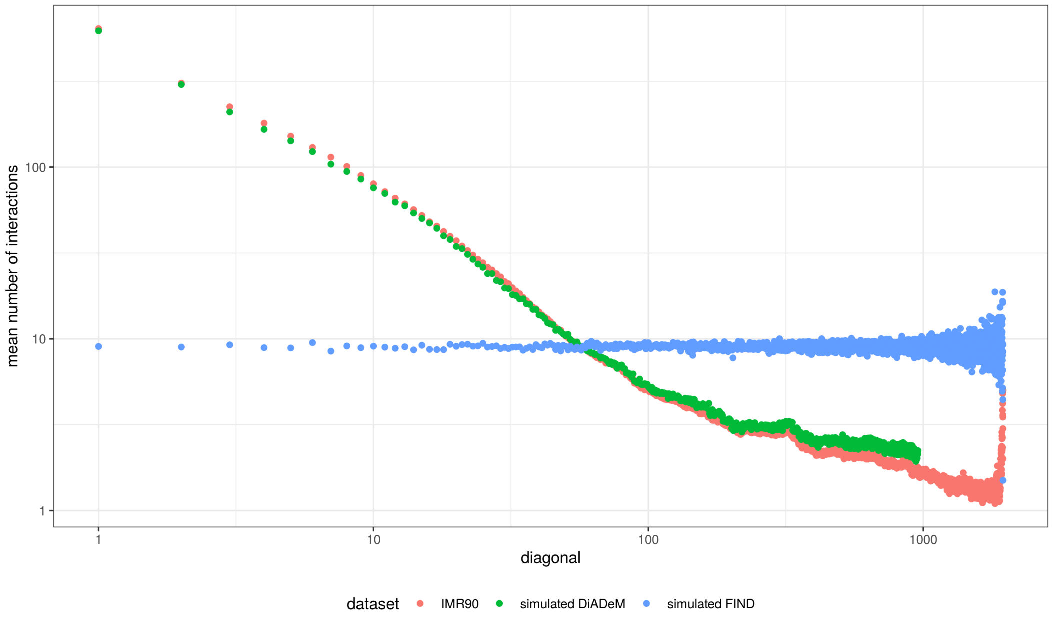


Figure 5.8. The comparison of contact decays for real and 2 types of simulated Hi-C datasets. The FIND-simulated contact decay is highly different from the IMR90 contact decay although it was produced from this Hi-C dataset. In contrast DiADeM-simulated contact decay is similar to IMR90.

The results illustrated in Figure 5.9 show that for most types of simulated data, as well as fold change values, DiADeM performs better than the other methods as indicated by higher PRAUC values. For simulated datasets obtained from the FIND method, the performance of DiADeM is similar to multiHiCcompare. In case of a low fold change value, all tools perform poorly although FIND and multiHiCcompare (FIND simulated data) or SELFISH (DiADeM simulated data) exhibit better results.

5.4. Conclusions

This chapter discusses the influence of normalization on the differential Hi-C analysis, as well as the common approaches for finding Hi-C differential contacts. Firstly, the results presented here indicate that normalization methods can impact such analysis leading to systematic biases and their direction may depend upon the choice of

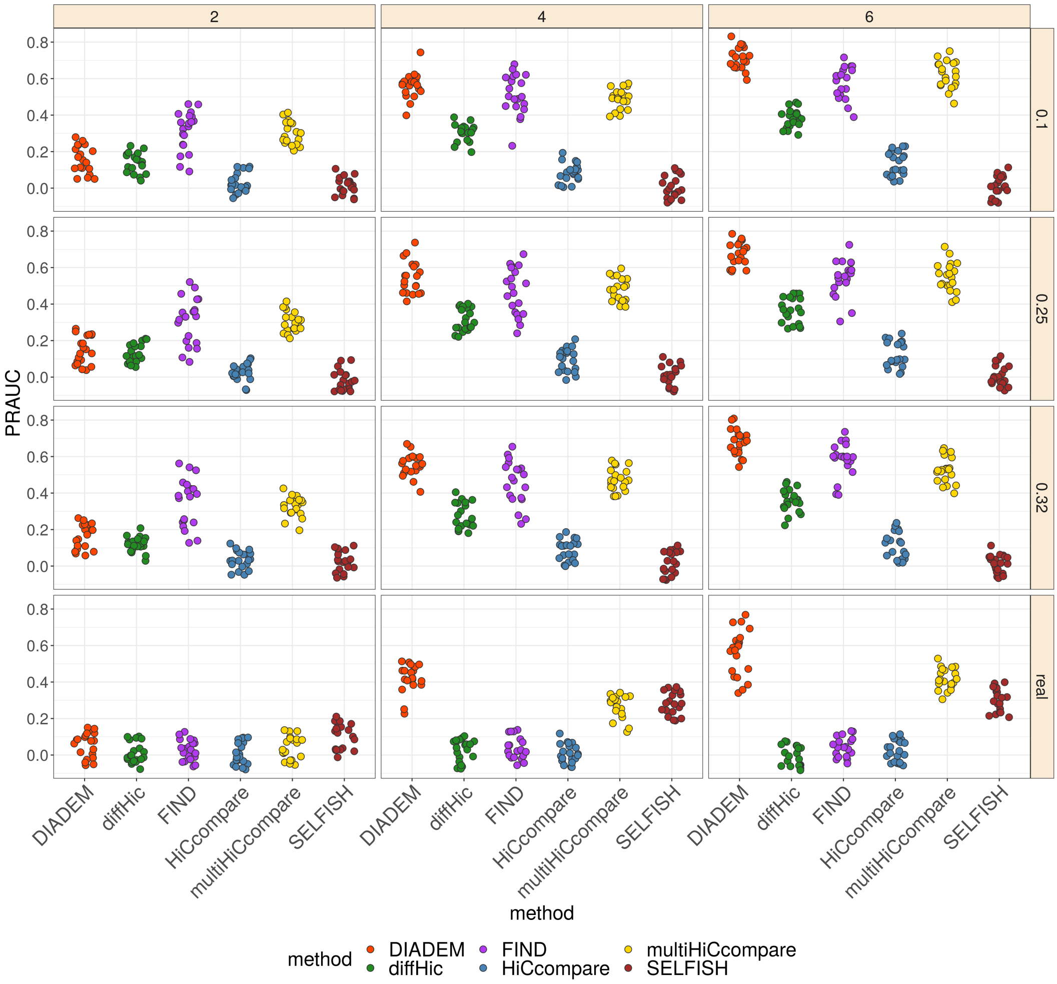


Figure 5.9. The comparison of performance of various methods for Hi-C differential interactions detection measured with area under Precision-Recall curves. Rows represents the dispersion between pairs of datasets while columns indicate fold-change between the introduced differential interactions.

normalization technique and genomic distance. Second, a new method for determination of Hi-C differential interactions is suggested. The new method called DiADeM is based on modeling the similarity of semi-diagonal interaction profiles between pairs of Hi-C datasets. Despite the fact that such relationship only exists at ranges of close genomic distances, it accounts for most of the contact map interactions. The observations we describe here, which are the basis of the DiADeM model, seem to be representative across various pairs of Hi-C datasets.

Finally, the performance of the new method was assessed using various datasets. The comparison of results obtained from bootstrapped, single cell and different cell line replicates, confirm the model's ability to discriminate between random and relevant biological variation. Comprehensive benchmarks of DiADeM on simulated datasets together with other commonly available approaches for Hi-C differential analysis indicate that DiADeM performs well across a wide range of conditions. Additionally, the new method presented here is equipped with a procedure for aggregation of significant differential interactions, which may facilitate functional analysis, for example during gene expression annotation. The results described in this chapter are described in preprint [ZW19b].

CHAPTER 6

Summary

This thesis presents several approaches for the comparative analysis of chromatin architecture using Hi-C contact matrices. We describe current methods used to analyze and compare Hi-C data, discuss their related limitations as well as derive some new results.

In Chapter 3, Section 3.4 we present results from our study conducted in collaboration with biologists from Kaikkonen group. We show the evidence of a relationship between changes in chromatin interactions and differential gene expression as well as epigenetic marks. The results we obtained were published in [Nis+17]. The simple approach we used for studying contact differences became a motivation for the development of more specific methods, which are introduced in Chapters 4 and 5.

Chapter 4 discusses the problem of comparing domain segmentations of a chromosome. The main result contained therein includes the development of a measure (the BP score) for studying the discrepancies in alignment between two domain partitionings. Moreover, we show that the BP score satisfies metric properties and performs competitively against alternative approaches. Additionally, we report two measures for the assesment of local chromosomal reorganization. Both theoretical and applied findings were published in [ZW19a].

Finally, in Chapter 5 we introduce the DiADeM model for the discovery of long-range differential chromatin interactions. The method we suggest proposes an intuitive definition of differential interaction and is shown to perform well against multiple existing approaches. Our model is described in a manuscript available on-line [ZW19b], which is currently in peer-review process.

In summary, the methods developed within this dissertation offer a potential to explore unusual structural features of chromatin during Hi-C comparative analysis, which have the potential to shed some light on unknown regulatory mechanisms mediated by alterations in genome conformation. More functional studies on high quality datasets would be required to determine how accurate our methods are in discovering the relevant connections between chromatin structure and regulation.

Bibliography

[AAL19] Abbas Roayaei Ardakany, Ferhat Ay, and Stefano Lonardi. “Selfish: discovery of differential chromatin interactions via a self-similarity measure”. In: *Bioinformatics* 35.14 (July 2019), pp. i145–i153.

[ACH14] William H Aeberhard, Eva Cantoni, and Stephane Heritier. “Robust inference in the negative binomial regression model with an application to falls data”. In: *Biometrics* 70.4 (2014), pp. 920–931.

[AH10] Simon Anders and Wolfgang Huber. “Differential expression analysis for sequence count data”. In: *Genome biology* 11.10 (2010), R106.

[Ahl02] Paul Ahlquist. “RNA-dependent RNA polymerases, viruses, and RNA silencing”. In: *Science* 296.5571 (2002), pp. 1270–1273.

[Amo05] Shannon Amoils. “The road to silence”. In: *Nature Reviews Molecular Cell Biology* 6.8 (2005), p. 593.

[An+19] Lin An et al. “OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries”. In: *Genome Biology* 20.1 (2019), pp. 1–16.

[AN15] Ferhat Ay and William S Noble. “Analysis methods for studying the 3D architecture of the genome”. In: *Genome biology* 16.1 (2015), p. 183.

[And+13] Guillaume Andrey et al. “A switch between topological domains underlies HoxD genes collinearity in mouse limbs”. In: *Science* 340.6137 (2013), p. 1234167.

[Ann08] A Annunziato. “DNA packaging: nucleosomes and chromatin”. In: *Nature Education* 1.1 (2008), p. 26.

[BA14] Jasmine Burguet and Philippe Andrey. “Statistical comparison of spatial point patterns in biological imaging”. In: *PLoS One* 9.2 (2014), e87759.

[Bai+13] Jane PF Bai et al. “Strategic applications of gene expression: from drug discovery/development to bedside”. In: *The AAPS journal* 15.2 (2013), pp. 427–437.

[Bar+15] A Rasim Barutcu et al. “Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells”. In: *Genome biology* 16.1 (2015), p. 214.

[Ber+09] Shelley L Berger et al. “An operational definition of epigenetics”. In: *Genes & development* 23.7 (2009), pp. 781–783.

[BH95] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[BK11] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications". In: *Cell research* 21.3 (2011), p. 381.

[Bre03] Sydney Brenner. "Nobel lecture: nature's gift to science". In: *Bioscience reports* 23.5 (2003), pp. 225–237.

[CC01] Thomas Cremer and Christoph Cremer. "Chromosome territories, nuclear architecture and gene regulation in mammalian cells". In: *Nature reviews genetics* 2.4 (2001), p. 292.

[CC10] Thomas Cremer and Marion Cremer. "Chromosome territories". In: *Cold Spring Harbor perspectives in biology* 2.3 (2010), a003889.

[Cha+15] Tamir Chandra et al. "Global reorganization of the nuclear landscape in senescent cells". In: *Cell reports* 10.4 (2015), pp. 471–483.

[Cob17] Matthew Cobb. "60 years ago, Francis Crick changed the logic of biology". In: *PLoS biology* 15.9 (2017), e2003243.

[Con+01] International Human Genome Sequencing Consortium et al. "Initial sequencing and analysis of the human genome". In: *nature* 409.6822 (2001), p. 860.

[Coo+15] Charles E Cook et al. "The European Bioinformatics Institute in 2016: data growth and integration". In: *Nucleic acids research* 44.D1 (2015), pp. D20–D26.

[Cra+15] Emily Crane et al. "Condensin-driven remodelling of X chromosome topology during dosage compensation". In: *Nature* 523.7559 (2015), pp. 240–244.

[CS15] Stefan Canzar and Steven L Salzberg. "Short read mapping: an algorithmic tour". In: *Proceedings of the IEEE* 105.3 (2015), pp. 436–458.

[DCZ18] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. "FIND: difFerential chromatin INteractions Detection using a spatial Poisson process". In: *Genome research* 28.3 (2018), pp. 412–422.

[Dek+02] Job Dekker et al. "Capturing chromosome conformation". In: *science* 295.5558 (2002), pp. 1306–1311.

[DG06] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[DGG79] Pierre-Gilles De Gennes and Pierre-Gilles Gennes. *Scaling concepts in polymer physics*. Cornell university press, 1979.

[DGR16] Jesse R Dixon, David U Gorkin, and Bing Ren. "Chromatin domains: the unit of chromosome organization". In: *Molecular cell* 62.5 (2016), pp. 668–680.

[Dix+12] Jesse R Dixon et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398 (2012), p. 376.

[Dix+15] Jesse R Dixon et al. "Chromatin architecture reorganization during stem cell differentiation". In: *Nature* 518.7539 (2015), p. 331.

[Dos+06] Josée Dostie et al. "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements". In: *Genome research* 16.10 (2006), pp. 1299–1309.

[DT12] Andrea Du Toit. "Chromatin: defining heterochromatin". In: *Nature Reviews Molecular Cell Biology* 13.11 (2012), p. 684.

[Edd01] Sean R Eddy. "Non–coding RNA genes and the modern RNA world". In: *Nature Reviews Genetics* 2.12 (2001), p. 919.

[EL13] Eli Eisenberg and Erez Y Levanon. "Human housekeeping genes, revisited". In: *TRENDS in Genetics* 29.10 (2013), pp. 569–574.

[FB09] Paul Flicek and Ewan Birney. "Sense from sequence reads: methods for alignment and assembly". In: *Nature methods* 6.11s (2009), S6.

[Fen+14] Suhua Feng et al. "Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis". In: *Molecular cell* 55.5 (2014), pp. 694–707.

[FG53] Rosalind E Franklin and Raymond George Gosling. "The structure of sodium thymonucleate fibres. I. The influence of water content". In: *Acta Crystallographica* 6.8-9 (1953), pp. 673–677.

[Fil+14] Darya Filippova et al. "Identification of alternative topological domains in chromatin". In: *Algorithms for Molecular Biology* 9.1 (2014), p. 14.

[FM00] Paolo Ferragina and Giovanni Manzini. "Opportunistic data structures with applications". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE. 2000, pp. 390–398.

[FR79] Jerome H Friedman and Lawrence C Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests". In: *The Annals of Statistics* (1979), pp. 697–717.

[Fra+15] James Fraser et al. "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation". In: *Molecular systems biology* 11.12 (2015).

[Fro+19] Kimon Froussios et al. "How well do RNA-Seq differential gene expression tools perform in a complex eukaryote? A case study in Arabidopsis thaliana". In: *Bioinformatics* 1 (2019), p. 6.

[Gib+18] Johan H Gibcus et al. "A pathway for mitotic chromosome formation". In: *Science* 359.6376 (2018), eaao6135.

[Gie11] Mark van der Giezen. "Mitochondria and the rise of eukaryotes". In: *Bioscience* 61.8 (2011), pp. 594–601.

[GP69] Joseph G Gall and Mary Lou Pardue. "Formation and detection of RNA-DNA hybrid molecules in cytological preparations". In: *Proceedings of the National Academy of Sciences* 63.2 (1969), pp. 378–383.

[GR05] Nick Gilbert and Bernard Ramsahoye. "The relationship between chromatin structure and transcriptional activity in mammalian genomes". In: *Briefings in Functional Genomics* 4.2 (2005), pp. 129–142.

[Hat+13] Ayat Hatem et al. "Benchmarking short sequence mapping tools". In: *BMC bioinformatics* 14.1 (2013), p. 184.

[HC05] Liang Huang and David Chiang. "Better k-best parsing". In: *Proceedings of the Ninth International Workshop on Parsing Technology.* Association for Computational Linguistics. 2005, pp. 53–64.

[Hei+10] Sven Heinz et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Molecular cell* 38.4 (2010), pp. 576–589.

[Hil11] Joseph M Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

[Hll77] David Lee Hlliker. "On the Multinomial Theorem". In: (1977).

[HM12] Ofir Hakim and Tom Misteli. "SnapShot: chromosome conformation capture". In: *Cell* 148.5 (2012), 1068–e1.

[Hou+08] Chunhui Hou et al. "CTCF-dependent enhancer-blocking by alternative chromatin loop formation". In: *Proceedings of the National Academy of Sciences* 105.51 (2008), pp. 20398–20403.

[HR81] Peter J Huber and Elvezio M Ronchetti. "Robust statistics john wiley & sons". In: *New York* 1.1 (1981).

[Hu+12] Ming Hu et al. "HiCNorm: removing biases in Hi-C data via Poisson regression". In: *Bioinformatics* 28.23 (2012), pp. 3131–3133.

[Hug+14] Jim R Hughes et al. "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment". In: *Nature genetics* 46.2 (2014), p. 205.

[HZW18] Jinlei Han, Zhiliang Zhang, and Kai Wang. "3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering". In: *Molecular Cytogenetics* 11.1 (2018), p. 21.

[Ima+12] Maxim Imakaev et al. "Iterative correction of Hi-C data reveals hallmarks of chromosome organization". In: *Nature methods* 9.10 (2012), p. 999.

[Jin+13] Fulai Jin et al. "A high-resolution map of the three-dimensional chromatin interactome in human cells". In: *Nature* 503.7475 (2013), p. 290.

[JV11] An Jansen and Kevin J Verstrepen. "Nucleosome positioning in Saccharomyces cerevisiae". In: *Microbiol. Mol. Biol. Rev.* 75.2 (2011), pp. 301–320.

[KR13] Philip A Knight and Daniel Ruiz. "A fast algorithm for matrix balancing". In: *IMA Journal of Numerical Analysis* 33.3 (2013), pp. 1029–1047.

[KS11] Manuel Koller and Werner A Stahel. "Sharpening wald-type inference in robust regression for small samples". In: *Computational Statistics & Data Analysis* 55.8 (2011), pp. 2504–2515.

[LA+09] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *science* 326.5950 (2009), pp. 289–293.

[Lan+09] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome biology* 10.3 (2009), R25.

[Lan+83] John P Langmore et al. "Low angle x-ray diffraction studies of chromatin structure in vivo and in isolated nuclei and metaphase chromosomes". In: *The Journal of cell biology* 96.4 (1983), pp. 1120–1131.

[LD+14] François Le Dily et al. "Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation". In: *Genes & development* 28.19 (2014), pp. 2151–2162.

[LDS16] Moyra Lawrence, Sylvain Daujat, and Robert Schneider. "Lateral thinking: how histone modifications regulate gene expression". In: *Trends in Genetics* 32.1 (2016), pp. 42–56.

[Lee+17] Bernard Kok Bang Lee et al. "DeSigN: connecting gene expression with therapeutics for drug repurposing and development". In: *BMC genomics* 18.1 (2017), p. 934.

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[LL18] Jun Li and Alicia T Lamere. "DiPhiSeq: robust comparison of expression levels on RNA-Seq data with large sample sizes". In: *Bioinformatics* 35.13 (2018), pp. 2235–2242.

[Loh+13] Sabine Lohmann et al. "Gene expression analysis in biomarker research and early drug development using function tested reverse transcription quantitative real-time PCR assays". In: *Methods* 59.1 (2013), pp. 10–19.

[LS12] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), p. 357.

[LS15] Aaron TL Lun and Gordon K Smyth. "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data". In: *BMC bioinformatics* 16.1 (2015), p. 258.

[MCS12] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic acids research* 40.10 (2012), pp. 4288–4297.

[MEG06] Glenn A Maston, Sara K Evans, and Michael R Green. "Transcriptional regulatory elements in the human genome". In: *Annu. Rev. Genomics Hum. Genet.* 7 (2006), pp. 29–59.

[Mei03] Marina Meilă. "Comparing clusterings by the variation of information". In: *Learning theory and kernel machines*. Springer, 2003, pp. 173–187.

[MH65] BJ McCarthy and JJ Holland. "Denatured DNA as a direct template for in vitro protein synthesis." In: *Proceedings of the National Academy of Sciences of the United States of America* 54.3 (1965), p. 880.

[Mis08] Tom Misteli. "Chromosome territories: The arrangement of chromosomes in the nucleus". In: *Nature Education* 1.1 (2008), p. 167.

[ML+09] Julio Mateos-Langerak et al. "Spatially confined folding of chromatin in the interphase nucleus". In: *Proceedings of the National Academy of Sciences* 106.10 (2009), pp. 3812–3817.

[ML98] Christian Münkel and Jörg Langowski. "Chromosome structure predicted by a polymer model". In: *Physical Review E* 57.5 (1998), p. 5888.

[MS58] Matthew Meselson and Franklin W Stahl. "The replication of DNA in Escherichia coli". In: *Proceedings of the national academy of sciences* 44.7 (1958), pp. 671–682.

[Nau+13] Natalia Naumova et al. "Organization of the mitotic chromosome". In: *Science* 342.6161 (2013), pp. 948–953.

[Nis+17] Henri Niskanen et al. "Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions". In: *Nucleic acids research* 46.4 (2017), pp. 1724–1740.

[Nor+12] Elphège P Nora et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398 (2012), p. 381.

[NR14] Olga Nikolayeva and Mark D Robinson. "edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology". In: *Stem Cell Transcriptional Networks*. Springer, 2014, pp. 45–79.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[O'C08] Clare O'Connor. "Fluorescence in situ hybridization (FISH)". In: *Nature Education* 1.1 (2008), p. 171.

[PC09] Jennifer E Phillips and Victor G Corces. "CTCF: master weaver of the genome". In: *Cell* 137.7 (2009), pp. 1194–1211.

[PH11] Chris P Ponting and Ross C Hardison. "What fraction of the human genome is functional?" In: *Genome research* 21.11 (2011), pp. 1769–1776.

[Phi08] Theresa Phillips. "Regulation of transcription and gene expression in eukaryotes". In: *Nature Education* 1.1 (2008), p. 199.

[Pra] L Pray. *Semi-conservative DNA replication: Meselson and Stahl. Nature Education 1 (1)(2008).*

[Pra08] Leslie Pray. "Discovery of DNA structure and function: Watson and Crick". In: *Nature Education* 1.1 (2008), p. 100.

[PS08] Theresa Phillips and K Shaw. "Chromatin remodeling in eukaryotes". In: *Nature Education* 1.1 (2008), p. 209.

[Ral08] Amy Ralston. "Examining Histone Modifications with Chromatin Immunoprecipitation and Quantitative PCR". In: *Nature Education* 1.1 (2008), p. 118.

[Ran+14] Chris M Rands et al. "8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage". In: *PLoS genetics* 10.7 (2014), e1004525.

[Rao+14] Suhas SP Rao et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7 (2014), pp. 1665–1680.

[RC12] Adam P Rosebrock and Amy A Caudy. *The future of deciphering personal genomes? The flies (and yeast and worms) still have it.* 2012.

[RMS10] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[Ros05] Paul R Rosenbaum. “An exact distribution-free test comparing two multivariate distributions based on adjacency”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.4 (2005), pp. 515–530.

[Rud+15] Matteo Vietri Rudan et al. “Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture”. In: *Cell reports* 10.8 (2015), pp. 1297–1309.

[RY84] Peter Rousseeuw and Victor Yohai. “Robust regression by means of S-estimators”. In: *Robust and nonlinear time series analysis*. Springer, 1984, pp. 256–272.

[San01] Fred Sanger. “The early days of DNA sequences”. In: *Nature medicine* 7.3 (2001), p. 267.

[SBD77] Stephen M Stack, David B Brown, and WC Dewey. “Visualization of interphase chromosomes”. In: *Journal of cell science* 26.1 (1977), pp. 281–299.

[SC05] Michael R Speicher and Nigel P Carter. “The new cytogenetics: blurring the boundaries with molecular biology”. In: *Nature reviews genetics* 6.10 (2005), p. 782.

[SC88] Shankar Subramanian and Richard T Carson. “Robust regression in the presence of heteroskedasticity”. In: *Advances in Econometrics* 7 (1988), pp. 85–138.

[SCD19] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. “multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments”. In: *Bioinformatics* (2019).

[Sch+19] Katharina Schwarze et al. “The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom”. In: *Genetics in Medicine* (2019), pp. 1–10.

[Sch86] Mark F Schilling. “Multivariate two-sample tests based on nearest neighbors”. In: *Journal of the American Statistical Association* 81.395 (1986), pp. 799–806.

[Sex+12] Tom Sexton et al. “Three-dimensional folding and functional organization principles of the Drosophila genome”. In: *Cell* 148.3 (2012), pp. 458–472.

[SI07] Eli Shechtman and Michal Irani. “Matching Local Self-Similarities across Images and Videos.” In: *CVPR*. Vol. 2. Minneapolis, MN. 2007, p. 3.

[Sim+06] Marieke Simonis et al. “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C)”. In: *Nature genetics* 38.11 (2006), p. 1348.

[SK18] Natalie Sauerwald and Carl Kingsford. “Quantifying the similarity of topological domains across normal and cancer human cell types”. In: *Bioinformatics* 34.13 (2018), pp. i475–i483.

[SK67] Richard Sinkhorn and Paul Knopp. “Concerning nonnegative matrices and doubly stochastic matrices”. In: *Pacific Journal of Mathematics* 21.2 (1967), pp. 343–348.

[SN14] Ricardo Parolin Schnekenberg and Andrea H Németh. “Next-generation sequencing in childhood disorders”. In: *Archives of disease in childhood* 99.3 (2014), pp. 284–290.

[SNC77] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[Spl+06] Erik Splinter et al. “CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus”. In: *Genes & development* 20.17 (2006), pp. 2349–2354.

[SR+04] Gábor J Székely, Maria L Rizzo, et al. “Testing for equal distributions in high dimension”. In: *InterStat* 5.16.10 (2004), pp. 1249–1272.

[ST14] Chi Song and George C Tseng. “Hypothesis setting and order statistic for robust genomic meta-analysis”. In: *The annals of applied statistics* 8.2 (2014), p. 777.

[Sta+18] John C Stansfield et al. “HiCcompare: an R-package for joint normalization and comparison of HI-C datasets”. In: *BMC bioinformatics* 19.1 (2018), p. 279.

[TM+70] Howard M Temin, S Mizutami, et al. “RNA-dependent DNA polymerase in virions of Rous sarcoma virus.” In: *Nature* 226 (1970), pp. 1211–1213.

[Tru+95] Mathias Truss et al. “Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter in vivo.” In: *The EMBO journal* 14.8 (1995), pp. 1737–1751.

[Wan+09] Likun Wang et al. “DEGseq: an R package for identifying differentially expressed genes from RNA-seq data”. In: *Bioinformatics* 26.1 (2009), pp. 136–138.

[WC+53] James D Watson, Francis HC Crick, et al. “Molecular structure of nucleic acids”. In: *Nature* 171.4356 (1953), pp. 737–738.

[WCP17] Xiao-Tao Wang, Wang Cui, and Cheng Peng. “HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions”. In: *Nucleic acids research* 45.19 (2017), e163–e163.

[WH97] CL Woodcock and RA Horowitz. “Electron microscopy of chromatin”. In: *Methods* 12.1 (1997), pp. 84–95.

[WK68] Ray Wu and AD Kaiser. “Structure and base sequence in the cohesive ends of bacteriophage lambda DNA”. In: *Journal of molecular biology* 35.3 (1968), pp. 523–537.

[WR16] Caleb Weinreb and Benjamin J Raphael. “Identification of hierarchical chromatin domains”. In: *Bioinformatics* 32.11 (2016), pp. 1601–1609.

[WSW53] Maurice Hugh Frederick Wilkins, Alex R Stokes, and Herbert R Wilson. “Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids”. In: *Nature* 171.4356 (1953), p. 738.

[YHV13] Danni Yu, Wolfgang Huber, and Olga Vitek. “Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size”. In: *Bioinformatics* 29.10 (2013), pp. 1275–1282.

[YT11] Eitan Yaffe and Amos Tanay. “Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture”. In: *Nature genetics* 43.11 (2011), p. 1059.

[Zor+79] Christian Zorn et al. “Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus: distribution in interphase and metaphase”. In: *Experimental cell research* 124.1 (1979), pp. 111–119.

[Zuf+18] Marie Zufferey et al. “Comparison of computational methods for the identification of topologically associating domains”. In: *Genome biology* 19.1 (2018), p. 217.

[ZW19a] Rafał Zaborowski and Bartek Wilczyński. “BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations”. In: *Journal of Computational Biology* 26.4 (2019), pp. 305–314.

[ZW19b] Rafał Zaborowski and Bartek Wilczyński. “DiADeM: differential analysis via dependency modelling of chromatin interactions with robust generalized linear models”. In: *bioRxiv* (2019).

[ZX19] Hui Zheng and Wei Xie. “The role of 3D genome organization in development and cell differentiation”. In: *Nature Reviews Molecular Cell Biology* (2019), p. 1.