Algorithms for the study of chromosomal structure variability

Natalie Sauerwald

CMU-CB-20-102

September 21, 2020

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carl Kingsford, Chair
Jian Ma
Anne-Ruxandra Carvunis
William Stafford Noble

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4554 to Carl Kingsford. Research reported in this thesis was supported by the Carnegie Mellon University Richard K. Mellon Fellowship, the National Human Genome Research Institute of the National Institutes of Health under award number R01HG007104, and the National Institute of General Medical Sciences of the National Institutes of Health under award number P41GM103712. C.K. received support as an Alfred P. Sloan Research Fellow. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: Three-dimensional structure, Algorithms, Hi-C, Topologically associating domains, Computational Biology

For the women who paved the way.

Abstract

The last two decades have introduced several experimental methods for studying three-dimensional chromosome structure, opening up a new dimension of genomics. Studies of these new data types have shown great promise in explaining some of the open questions in gene regulation, but the experiments are indirect and imperfect measurements of the underlying structure, requiring rigorous computational methods. We can now study the 3D relationships between all pairs of chromosome segments across the genome, but questions such as the variability of this structure between cell and tissue types, the predictors of structural similarity, the dynamics of this complex system, and a complete definition of the observed substructures remain unclear. This dissertation presents several approaches to improve our understanding of human genomic spatial architecture. We present a new method to quantify the variability of chromosomal substructures, called topologically-associating domains (TADs) between any pair of samples. This algorithm efficiently identifies all regions with statistically significantly similar TAD structures between the two samples. Using this method, we quantify the structural similarity within each chromosome and between chromosomes, and between cell types. We show that cancer cell lines are structurally disrupted at pan-cancer genes, but not globally. We perform extensive data analysis using this method and others to assess the consistency of TADs across a range of biological and technical conditions. This large scale study of chromosomal structural variability emphasizes the differences between chromosome structures between cell and tissue types, in contrast to the belief that genome structure is highly conserved. We quantify the influence of genetic difference and similarity, as well as technical confounders, on chromosome structural similarity in a systematic study of over 100 samples. We also apply a biophysics model to predict the dynamics of chromosomes from static data. Our predictions correlate well with several different experimental measures and known substructures. We predict the existence of long range dynamic couplings involved in gene regulation that have not been found without a dynamic model. Finally, we develop a generalized TAD-finding algorithm that can be guided towards selecting TADs for any desired property. Defining several functions around common evaluation criteria for TADs, we explore the relationships between various biological TAD properties and the computational definitions used to identify TADs. The algorithms and analysis we have developed enable rigorous study of the basic properties of this new dimension of genomics, and can continue to inform the study of TADs as more experimental data becomes available.

Acknowledgments

I have been incredibly lucky to find amazing mentors, friends, and resources to support me through this phase of graduate school. To my advisor, Dr. Carl Kingsford, I cannot thank you enough for always making time for me and respecting my ideas. Your support of my goals – both research and non-research – was invaluable on both a personal and professional level. The independence you gave me allowed me to explore my interests and abilities knowing I would be supported whenever I struggled. The members of the Kingsford group, both past and present, consistently challenged me and created a scientific environment that taught me more than I can explain. It certainly hasn't always been easy or smooth, but I wouldn't have made it here without all of your help.

I have to also acknowledge the support and help I received from CPCB administrative staff, especially Nicole. You work so hard for us, and knowing you cared and that you'd advocate for me made the department feel so welcoming through so many changes.

I'd also like to thank everyone involved in the TechNights program - leaders past and present, volunteers, staff, and most importantly the girls themselves. I learned so much from all of you and cannot wait to see the future these girls will lead us to.

To my friends both here in Pittsburgh and far away: thank you for everything. For listening to me vent, for celebrating my successes, for critiquing my work, for making me laugh, for the wine nights and wine skypes, for the animal pictures, for the inspiration, for the tether to sanity during a pandemic, for the walks in the park, for the pet therapy, for the shared meals and cookies, for the scientific discussions, for the very non-scientific discussions, for the climbing trips, for the YouTube lunch breaks, for the bike rides, for more than I could possibly list: thank you.

No one has believed in me more than my family and my partner, and I can never thank them enough for their faith in me and comfort when I needed it most.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human genome is made up of about 2 meters of DNA, tightly packed into a nucleus approximately 6 micrometers in diameter [4]. The three-dimensional structure of these DNA strands was first studied with fluorescent imaging techniques [87], and more recently with sequencing-based methods [31, 39, 69, 117]. These experimental techniques have shown that 3D genomic architecture is associated with several regulatory systems, including replicating timing [35, 74, 92] and gene regulation [8, 23, 29, 40, 58, 65], and its misalignment is associated with diverse human diseases and disorders [48, 54, 71, 75, 78]. Chromosome structure is therefore implicated in many of the cell's most important processes.

One of the most fundamental questions in biology remains how cells determine which genes to express and when to express them. Recent work has shown that three-dimensional chromosome structure can account for a large proportion of overall gene expression levels [101]. Many human diseases and disorders, including cancers [48, 54, 75, 78], neurological disorders [123], limb malformations [71], and more [119], are tied to genomic architecture, making this question clinically important as well as a question of basic biology. A full understanding of these regulatory disorders, along with new therapies, must come from a deeper understanding of their causes on a genomic level. However, the complexity of the system and many experimental biases make it very challenging to accurately describe this underlying 3D structure. Principled computational

methods are needed to identify meaningful signal from this imperfect description of an extremely complicated system.

1.1 Measurements of chromosome conformation

The first high-throughput sequencing technique for measuring chromosome structure is commonly referred to as 3C, for chromosome conformation capture [31]. This technique introduced the basis for all subsequent sequencing experiments that measure the 3D chromosome shape: cells are first fixed and subjected to cross-linking, which causes physically close segments of the genome to be stuck to each other. Restriction enzymes are then used to cut the genome up into pieces, the ends of which are ligated and then sequenced. Several variants, all based on this same basic set of steps, have since been developed for capturing different aspects or resolutions of chromosome structure.

- **3C:** The original protocol requires prior knowledge of two interacting regions. All interactions between these two loci of interest are quantified by the 3C experiment [31].
- **4C:** This protocol uses inverse PCR to capture the interactions between one genomic locus of interest and the rest of the genome [117].
- **5C:** All interactions within a given region, generally at most a megabase, are quantified [39]. This is a high resolution method to study a very specific region of interest, but has relatively low coverage and cannot be expanded to study larger genomic regions because of its reliance on primers.
- **Hi-C:** Using paired-end sequencing, Hi-C captures all interactions between all loci of the genome [69]. It does not require prior knowledge of any regions of interest, but is much lower resolution than the other methods.

All of these protocols are population-based, so they rely on aggregating results from a large number of cells. Some single-cell techniques have also been developed to study chromosome

structure, which will be discussed later.

Outside of these proximity ligation-based techniques, a few other genome-wide measurements of chromosome structure have been developed recently (review: [62]). Genome architecture mapping (GAM) combines randomized ultrathin cryosectioning and DNA sequencing to measure 3D distances between pairs of loci in a nucleus [14]. The matrices generated by GAM are highly correlated with Hi-C matrices, but GAM is able to capture multi-loci interactions in addition to pairwise contacts, making it well-suited to study hubs or clusters of chromosome segments. GAM requires fewer cells than Hi-C, but is a more complex and expensive experiment performed by fewer labs, leading to even less publicly available GAM data than Hi-C data. SPRITE is another sequencing-based assay that avoids proximity ligation by using a split-pool strategy, identifying genome-wide DNA interactions including those around nuclear speckles and the nucleolus [94]. This data largely recapitulates the structures identified with Hi-C, but additionally sheds light on organization relative to nuclear bodies.

1.1.1 Hi-C: capturing the 3D structure of the full genome

In the interest of understanding the overall structure of the entire human genome, we focus on Hi-C data. For the Hi-C protocol, after cross linking, enzyme digestion, and ligation, genomic pieces are aligned back to a reference genome to determine where in the genome the two ligated segments originated. Each time two reads from different regions of the genome are found ligated together counts as one interaction between these segments. These interaction counts are reported in Hi-C matrices, where each row and column corresponds to a segment of the genome. The length of that segment is known as the *resolution* of the data, and generally ranges from 1 kilobase (kb) to 1 megabase (Mb), depending on the sequencing depth. Each entry in the Hi-C matrix equals the number of interactions, or cross-linking events, found between the corresponding segments of the genome. Hi-C matrices can represent all interactions between segments on the same chromosome (intra-chromosomal) or interactions between two different chromosomes

(inter-chromosomal). Inter-chromosomal matrices tend to be very sparse, and while there are some notable exceptions [17], chromosomes tend to occupy their own spatial territories [29]. An example of an intra-chromosomal Hi-C matrix is given in Figure 1.1.

Hi-C data presents many challenges for analysis, from the size of the data to experimental biases and lack of true validation experiments. Depending on the resolution of the data, intra-chromosomal matrices vary in size from hundreds of thousands of rows and columns to a few thousand rows and columns. Most segments of the chromosome are only close to a limited number of other segments so Hi-C matrices are fairly sparse, which makes true interactions difficult to separate from any false contacts induced by the experiment. Beyond the data size, the dependence on a population of cells adds complication to the analysis. A Hi-C matrix does not truly describe the structure of any individual cell, but rather describes a sort of average structure across the population. Hi-C matrices therefore do not describe a specific 3D structure nor can they be easily interpreted as 3D distances, because they do not satisfy basic distance properties such as the triangle inequality. The cross-linking procedure is also likely to induce many false interactions as well as omitting many true interactions, adding errors to the difficulty of interpreting a population-averaged structure. On top of all of these challenges with the data itself, there is no independent method for validating genome-wide three-dimensional structures. There is no ground truth to compare with and identify errors or test methods against. Imaging techniques such as fluorescent in-situ hybridization (FISH) can be used to test a specific prediction about a limited number of known genomic loci, but cannot validate any genome-wide structural predictions.

While there are many complexities of Hi-C data that cannot be directly addressed, several normalization methods exist to correct for known biases in the data such as varying numbers of restriction enzyme sites within each Hi-C bin. One of the first such normalization techniques is called vanilla coverage (VC) normalization, which simply divides each entry in the Hi-C matrix by the average contact probability across the genome for locus pairs the same distance apart [69].

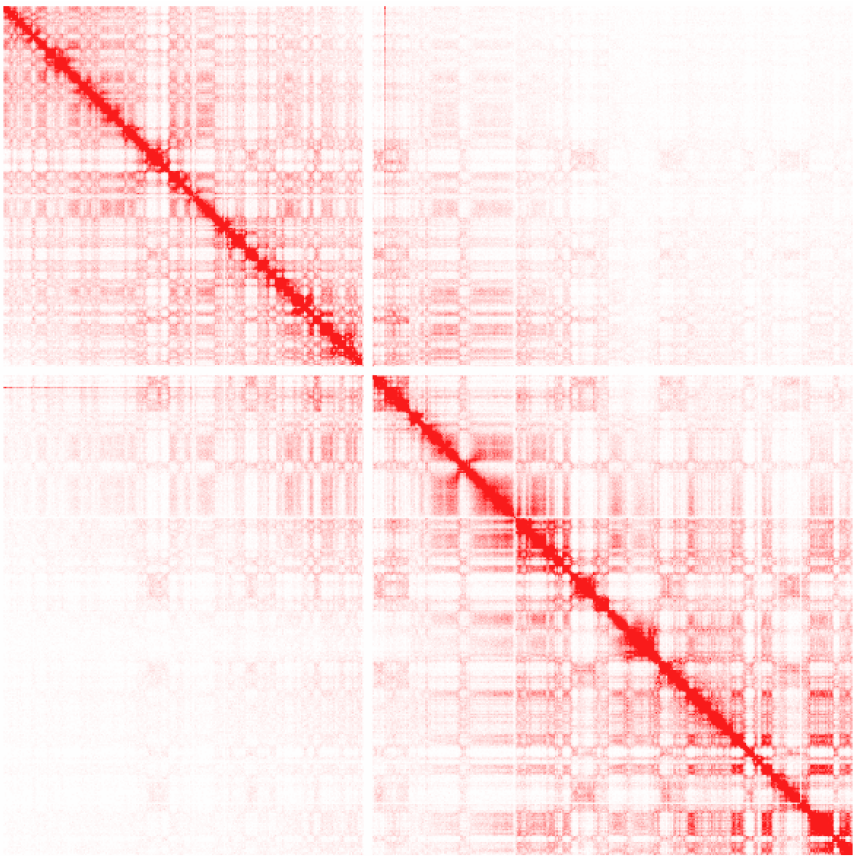


Figure 1.1: Example of a Hi-C matrix from chromosome 1 of human ES cells at 100kb, from `http://homer.ucsd.edu/homer/interactions/HiCmatrices.html`, accessed on 09/04/2020. The color intensity represents the number of interactions identified between the two genomic regions corresponding to the row and column of the matrix entry, with zero interactions showing as white.

Another common technique is called Knight-Ruiz (KR) normalization, which balances symmetric matrices such that each row and column sums to one [63]. Iterative correction and eigenvector decomposition (ICE) normalization corrects for bias under the assumption that all loci have equal visibility, thereby implicitly correcting for biases such as restriction site density and DNA sequencing bias [59]. These three methods are the most commonly used for Hi-C data, but other techniques have been developed as well [56, 73, 120].

1.2 Topologically associating domains (TADs)

The first analyses of Hi-C data showed evidence of organization by substructures made up of relatively large chromosomal segments. The largest of these structures were called "compartments", which are multi-megabase-sized regions of the genome corresponding to two main classes referred to as "A" compartments and "B" compartments [69]. Genomic segments in A compartments generally contain active chromatin; these sections contain more genes, more highly expressed genes, and fewer repressive histone markers. In contrast, B compartments contain less accessible DNA, fewer genes, and more repressive histone marks. These two broad classes of compartments have since been subdivided further into five sub-compartments, two of which are active and three of which are inactive, each displaying distinct epigenomic signatures [98].

Further analysis of Hi-C data revealed smaller (generally less than 1 Mb) block structures along the diagonal of the Hi-C matrix, which were termed "topologically associating domains" (TADs) [36]. TAD structures are imprecisely defined as contiguous sections of DNA that exhibit higher contact frequencies within themselves than with outside segments. They seem to mostly cover the genome, suggesting a globular 3D chromosome structure based on these building block regions. The exact scale of TADs is unclear: they have been shown to exist at various resolutions, with a hierarchical nesting structure. Smaller contact domains have been referred to as subTADs, though a formal definition of subTADs requires additional conditions beyond a size cutoff [13]. A precise definition of TADs, perhaps including sub-classifications, is a matter of

ongoing discussion in the community [13].

While a formal definition is still debated and there is no gold standard or ground truth for TAD sets, there are several properties TADs possess that are used for evaluating the accuracy of TAD sets. It is clear that TAD boundaries are enriched for the insulator protein CCCTC-binding factor (CTCF), as well as proteins RAD21 and SMC3 which are components of the structural cohesin complex [36, 52, 98]. There is some debate about the mechanisms by which cohesin and CTCF relate to TAD structure [72, 86], but they are clearly critical to the formation or stability of TAD structures. Additionally, TAD boundaries are expected to follow patterns of enrichment or repression of specific histone markers and can even be inferred through a supervised method from this data alone [98, 112]. We also expect TADs to be fairly consistent between replicate samples. For more technical evaluations, TAD finding tools are often measured on their robustness to subsampling and different resolutions of the same Hi-C sample. Some have tried to use manual annotations, although it is also not straightforward to identify TAD boundaries by eye [30]. There have also been several methods proposed to simulate Hi-C data but identifying TADs or simulating realistic TADs in this context has also proven challenging [137].

1.2.1 Computational approaches to TAD identification

While TADs are visually apparent as block structures along the diagonal of a Hi-C matrix, it is not straightforward to define and identify them computationally. Many methods have been developed leading to conflicting results (see reviews: [30, 44, 139]). Without a ground truth to compare with, it is generally difficult to validate the accuracy of TAD sets. Evaluation therefore depends on quantifying desired TAD properties, such as the enrichment of certain factors such as CTCF, RAD21, and SMC3 at TAD boundaries or the difference between distributions of contacts within a TAD versus between TADs. Broadly speaking, these computational approaches all optimize for a definitive TAD trait such as highly insulated boundaries [28, 116], dense intra-TAD contacts [42], or strong block-like structures [67].

1.2.2 Models of genome organization

Outside of TADs, Hi-C data has been shown to follow power-law scaling in which contact frequency decreases as a function of linear distance along the chromosome [69]. For length scales up to around 10Mb, a fractal globule model based on polymer simulations with physical constraints has been suggested to explain the genomic organization [77], though multiplexed FISH imaging studies suggest it does not seem to hold at longer length scales [128]. The emerging model for TAD organization and formation is known as loop extrusion. The loop extrusion model has not yet been conclusively observed in human cells, but predictions made from this model recapitulate the observations from several structural perturbation studies. The proposed mechanism involves a loop extruding factor (suspected to be cohesin in humans) actively pulling loops of chromatin through itself until it hits a barrier element (CTCF) and drops off the chromosome, leaving a TAD structure behind [49]. Simulations of loop extrusion match patterns observed in Hi-C, and predict the effects of both cohesin and CTCF depletion on chromosome structure [86].

1.2.3 Chromosome structures in single cells

A full understanding of genome architecture and its role in cellular functioning relies on not only a population average measurement, but also on the ability to measure this structure in individual cells. Single cell Hi-C (scHi-C) was developed for this purpose [80] and has been used to quantify the structures of thousands of single cells [82, 97, 121]. scHi-C data has been used to study the dynamics of chromosome structures throughout the cell cycle [82], and to confirm the existence TAD-like structures at the single cell level [121]. scHi-C data suffers from many of the same challenges induced by cross-linking as bulk Hi-C data, including false contacts and omissions of true contacts, but the matrices are much sparser. This additional sparsity makes reliable signal detection very difficult, but scHi-C data has nonetheless shown single cell TADs that are highly variable between individual cells, with higher probability of boundaries at the same locations of

TAD boundaries identified with bulk Hi-C data [82, 121].

Recently there have been significant advances in imaging technology for measuring single cell chromosome structure. A super-resolution chromatin tracing method was used on single cells to confirm the observation of variable TAD-like structures in individual cells [19]. This study also demonstrated that cohesin depletion, which is known to remove TAD structures in bulk Hi-C data, does not cause the disappearance of the block-like structures in individual cells, but rather leads to randomized boundary locations. Another very recent study from the same group expanded their multiplexed imaging platform to provide the first chromosome- and genome-scale imaging study of chromosome structure [122]. Because imaging does not require destroying the genome as is necessary for Hi-C, this study was also able to simultaneously measure transcriptional activity and nuclear landmarks in single cells. This type of work is critical to understanding the relationship between chromosome structure and gene regulation, and imaging experiments will become the gold standard for structural measurements in the future.

1.2.4 Functional relationships of chromosome structure

Initially, TADs were suggested to be regulatory domains bringing enhancers and promoters closer in 3D space to the genes they regulate. This is in part influenced by the existence of insulator proteins at TAD boundaries, suggesting that regulatory interactions are more likely to occur within a TAD than across TAD boundaries. The loci within TADs have more highly correlated histone modifications than loci of similar genomic separation in different TADs [98]. Chromosome accessibility is also strongly related to genome architecture, as demonstrated by strong associations between Hi-C data and accessibility measurements such as DNase-seq [110]. Consistent with the idea that chromosome structure has regulatory implications, eQTLs are closer in 3D space to the genes they regulate than expected by chance [40]. Several studies additionally linked chromosome structure with replication timing [35, 74, 92], even showing that TADs display a one-to-one relationship with replication timing domains [92], suggesting an important role in multiple cel-

lular processes.

Three-dimensional genome structure has been associated with gene transcription in multiple studies [45, 58, 101]. Co-regulated genes are often found within the same TAD [83], and co-expressed genes tend to cluster together in 3D space [136]. One study demonstrated that a significant proportion of gene expression can be attributed to positional factors and predicted enhancer-promoter interactions from expression data alone [101]. Further evidence for the gene regulatory role of chromosome structure comes from its association with a variety of human diseases and disorders (review: [71]). One striking example involves the SOX9 gene: three similar genetic duplications near SOX9 can lead to either Cooks syndrome (a limb malformation), female to male sex reversal, or no phenotypic change at all. The wildly different results of similar structural variants has been attributed to their different impacts on the local TAD boundaries [33]. TAD structures have also been implicated in many neurological disorders [123] and cancers [48, 54, 75, 78]. One study demonstrated that the simple removal of a nearby insulating TAD boundary is sufficient to activate proto-oncogenes [54].

Despite this strong evidence for the role of TADs and the overall chromosome architecture in gene regulation, several other studies have suggested little to no relationship between the two [1, 33, 50, 84, 99, 100, 111]. In particular, genome-wide CTCF depletion has been shown to result in widespread loss of most TAD structures, but does not drastically alter transcription [33, 84]. In a study performed with *Drosophila*, extensive genomic rearrangements causing significant disruption to TADs and other 3D structures generally did not alter gene expression for the majority of genes [50]. Similarly, deletion of a protein required for loading of cohesin onto chromatin results in genome-wide disappearance of TADs, even in regions with no change in gene expression [111]. In cancer specifically, where many TAD disruptions have been observed across cancer types, one study showed that while structural variants can significantly impact TAD structures, only a small percentage (14%) of the boundary deletions resulted in a large difference in expression levels [1]. These results prove that the true role of chromosome structure in gene

regulation is likely much more complicated than previously suggested.

The functional role of chromosome structure is tightly linked to its level of variability; if chromosome structure does not vary across cell types or even species, it cannot be tightly linked to changes in gene expression, which does vary across these conditions. Quantification of the variability of TAD structures was previously limited to Venn diagrams [36, 98], and compared to a null model of random TAD placement [109]. There was a clear need for more precise methods to study TAD similarity across conditions, and analysis of this variability across a broad range of Hi-C data. Significant questions remain about the dynamics of chromosome structures within cell populations, which can be difficult to answer because of the nature of the Hi-C experiment, which cannot be performed in live cells, nor without disassembling the original genome structure. We therefore are limited to a static measurement of a dynamic system. Computational methods to predict these dynamics from the static Hi-C matrix could provide valuable insights without requiring additional expensive experimental methods.

A full understanding of TADs and their relationship with underlying biology requires teasing out the relationships between our computational descriptions of TADs and the biological properties we expect them to display. In order to answer the open questions related to TAD functionality, we need to unify computational and biological TAD properties under one framework. Currently, the differences in TAD finder results and their tradeoffs in matching desired biological properties point to a disconnect between computational TADs and their biological features.

1.3 Contributions

This dissertation combines method development and data analysis to study the variability, dynamics, and complexity of chromosome structure in the human genome.

- **Quantification of the structural similarity of TAD decompositions** (Chapter 2). We developed the first method to measure TAD similarity between any two sets of TADs. This method, which we call TADsim, identifies statistically significantly structurally similar

regions of TAD decompositions using the Variation of Information (VI) metric. We present this method with an initial analysis of 23 cell types, comparing the TAD structures of normal and cancer cell lines. We do not observe genome-wide structural disruption among the cancer cell lines, but we do find that there is less structural conservation among cancer cell lines at the locations of highly mutated pan-cancer genes.

- **Analysis of the variability of TADs** (Chapter 3). Using the method described in Chapter 2 along with two other measures, we performed the first large-scale comparison of TAD structures with 137 human Hi-C samples from a range of biological and experimental conditions. We quantify the influence of both technical and biological variation on TAD structures. We explore the variability of TADs between replicates, between cell types and tissues, across families, between protocol variants, and across different labs. We find support for a disordered or dynamic TAD structure, and note significant room for variation between cell types. The results are shown to be robust across resolutions and parameter choices.

- **Study of chromosome dynamics** (Chapter 4). Hi-C data is often studied as if it represents a static conformation, despite its known dynamic nature. We adapted the Gaussian Network Model (GNM) to study chromosomal dynamics and capture relationships between distant genomic regions that display dynamic coupling. The GNM is a biophysics model designed for studying protein dynamics, which uses protein contact maps similar to the information reflected in Hi-C matrices. We tailored this model to study dynamic coupling between both contiguous and distant chromosome regions, comparing the contiguous constructs to known structural elements and identifying new functional relationships between the long range couplings. The ability to study such distant interactions across the entire chromosome represents a significant step forward in understanding the link between structure and function, as we are no longer limited to analysis within a restricted distance range.

- **Comparing computational and biological TAD properties** (Chapter 5). Computational

TAD finders frequently disagree on the locations of TADs on the same data, generally presenting tradeoffs in a variety of biological properties they are assessed on. To break down this relationship between computational TAD definitions and data-driven TAD properties, we developed a flexible TAD finder with several tunable parameters, and optimized the parameters for a variety of data-driven objectives. We study the variability of resulting TAD sets within 12 different cell types, and assess the similarities between TADs optimized for different properties. While some cell types show significant consistency in TAD sets regardless of objective, others result in highly varying TAD sets. We generally find that optimizing for CTCF binding sites, CTCF ChIP-seq, and H3K36me3 ChIP-seq peaks at TAD boundaries all result in relatively similar TAD sets, while RAD21, a protein of the cohesin complex, results in inconsistent predictions. Selecting TAD finding parameters for reproducibility gives highly variable results across cell types, and directly optimizing for the difference between inter- and intra-TAD contacts gives high variable results within most cell types, often disagreeing most strongly with TADs optimized for CTCF binding sites at their boundaries. This study reveals relationships between computational TAD definitions and biological TAD properties, as well as correlations between the properties themselves.

Together, these contributions significantly advance our understanding of the variability and dynamics of chromosome structure, and provide new algorithms to facilitate continuing study in this exciting field.

Chapter 2

Quantifying TAD variability

A central question to understanding chromosome structure is the degree to which TADs are conserved or vary between conditions. This question cannot be answered without a principled way to quantify TAD variability, so we present in this chapter the first method to quantify resemblance and identify structurally similar regions between any two sets of TADs. We present an initial analysis of 23 human Hi-C samples representing various tissue types in normal and cancer cell lines. We quantify global and chromosome-level structural similarity, and compare the relative similarity between cancer and non-cancer cells. We find that cancer cells show higher structural variability around commonly mutated pan-cancer genes than normal cells at these same locations.

A version of this chapter was published in *Bioinformatics* and presented as a proceedings talk at the *Intelligent Systems for Molecular Biology* (ISMB) conference in 2018 and is joint work with Carl Kingsford [106]. In a subsequent publication [108] we described minor modifications to this method that are reflected in this chapter, therefore it differs slightly from the original publication. TADsim source code and analysis scripts are available at `https://github.com/Kingsford-Group/localtadsim`.

2.1 Background

Three-dimensional chromosome structure has been shown to be an influential factor in diverse aspects of cellular functioning. Since the introduction of chromosome conformation capture [31] and its many variants including a high-throughput experiment permitting genome-wide structural measurements termed Hi-C [69], there have been many studies associating chromosome structure with numerous cellular processes. Among these include several studies linking chromosome structure to gene expression and regulation [23, 29, 40, 65, 107], and more specifically changes in structure have been associated with various human diseases and disabilities, including several cancers [48, 54, 75, 78], as well as deformation or malformation of limbs during development [71]. On the mechanistic side, structural components have been implicated in replication timing [7, 79, 92, 104] and associated with DNA accessibility and nuclear organization [96].

Although studies of chromosome structure have provided meaningful biological insights such as those mentioned above, many questions remain about the precise role and variability of the chromosomal architecture. In particular, one key question is the extent to which chromosome structure is conserved between cell types, or how much it differs between normal and diseased tissue, e.g., cancer tissue. A deeper understanding of the level of structural similarity across cell types would reveal mechanistic insights into the role of three-dimensional folding of the chromosomes and demonstrate the relative cell-type specificity of the arrangement, yet very limited work has been devoted to this question. We address this question through quantifying structural similarity in pairwise comparisons, and apply this method to compare chromosome structure across many cell types, as well as between cancer and normal cells.

Chromosome structure is described in terms of several different scales of components, from multi-megabase compartments to sub-megabase topologically-associating domains (TADs) and subTADs [20]. Compartments divide chromosomes into two broad categories: loosely packed, gene-rich areas termed A compartments and densely packed inactive areas termed B compartments. They can be identified in a straightforward way from the correlation matrix of the Hi-C

map [69]. TADs, visually identifiable as squares along the diagonal of the Hi-C contact map with enriched contact density, represent smaller regions that interact significantly more with other loci within the same TAD than with those outside of it [36]. Although TADs are somewhat visible in Hi-C maps, it has proven challenging to definitively classify them computationally.

TADs have been shown to correlate with several epigenetic features, including histone markers and CCCTC-binding factor (CTCF) [88]. Histone modifications have proved very tightly linked to Hi-C data, leading to several methods for identifying TADs or predicting Hi-C maps based on ChIP-seq data from a range of histone marks [15, 34, 57, 112]. Beyond epigenetics, TADs seem to be involved in several other cellular functions. TAD boundaries correlate well with replication timing domains and thus are involved in cell reproduction [35]. Lamina-associated domains (LADs), regions near the nuclear lamina associated with gene repression, also frequently coincide with TAD domains [127]. Interruption of TADs has also been shown to alter enhancer/promoter interactions [70], further implicating TAD structure in gene regulatory mechanisms.

Many methods have been developed to identify TADs, first through an HMM-based method [36], and later through optimization of various scoring functions such as InsulationScore [28] and Armatus [42]. It is not yet clear how to evaluate TAD finder accuracy with no settled ground truth, but two recent benchmarking studies evaluated the performance of 7 different TAD callers, 6 of which overlapped between the two studies, and found no clear consensus on optimal performance [30, 44].

Though there is some preliminary evidence that TAD structure is conserved across cell types [98] and possibly even species [36], this previous work has not attempted to identify the locations of structural similarity, nor which genomic features or disease states may correlate with conserved structures. Hi-C data itself is highly variable and likely full of false contacts and missing true contacts, and it is impacted significantly by the choice of data processing and normalization techniques, making it difficult to compare Hi-C maps directly [132]. Spurious differences like

coverage variance can have a strong effect on the apparent similarity of two Hi-C maps, even if the underlying structures are similar. The variability within and between chromosomes is also large, which could mask intrinsic similarity in a global metric. For these reasons, we choose to compare TAD structure rather than Hi-C measurements directly, and we seek regions of locally similar structures rather than one global measure of similarity.

We present a method to identify statistically significantly structurally similar regions of TAD structures, in two main steps. First, we use the information theoretic variation of information (VI) metric [76] to measure the similarity of all subsets of the two TAD structures, using a dynamic programming algorithm that we designed to efficiently compute this metric. We then select the statistically significant chromosomal regions among those with a locally optimal VI measure through a rigorous null model, and eliminate redundancies from this set. We apply this method to evaluate the similarity of chromosome structure across all pairwise combinations of 23 human samples, across both cancer and non-cancer conditions. The following large-scale comparison of structural concordance and variability across cell types, both globally and on the chromosomal level, identifies biologically meaningful cell type pairs with high structural similarity, and a trend of low structural similarity among cancer cells can be seen at the locations of commonly mutated pan-cancer genes.

This work is the first large-scale study of human chromosomal structural similarity, providing a framework method for future work in this domain. Our comparison of cancer and normal cells reveals insight into the three-dimensional disruptions that occur in cancer genomics, corresponding to the known changes in genome sequence from mutations and structural variants.

2.2 Methods

We introduce a method which, given two lists of TADs from different samples on one chromosome, identifies the sub-intervals in which the two TAD lists are significantly similar. This is done by optimizing a distance metric, selecting the statistically significant optima, and removing

redundant intervals with a heuristic.

2.2.1 Data

For the analysis in this chapter, Hi-C data were taken from four different studies [36, 69, 98, 125] that were published over seven years, representing 21 unique human cell types across healthy and diseased states, with 23 Hi-C samples in total, as summarized in Table 2.1. The samples were chosen to be publicly available and represent a wide array of cell types and conditions.

All data were downloaded as raw read (.fastq) files, and processed through the same Hi-C Pro (version 2.8.0) [113] pipeline into Hi-C maps, using iterative correction and eigenvector decomposition normalization [59]. All Hi-C maps were generated at 100kb resolution, the highest shared by all four studies, meaning that each point in the Hi-C matrix corresponds to the number of contacts between two chromosomal intervals of 100kb each. We call each of these 100kb segments a genomic bin. This resolution is relatively low because only the more recent studies were sequenced deeply enough for significantly higher resolution. This may affect our results in that we can only capture relatively large-scale regions of structural similarity, but these larger regions are likely to be the most robust. The TAD sets were calculated using version 2.1 of the Armatus software [42], a principled method that is extremely efficient and has performed favorably in recent benchmarking studies [30, 44]. Armatus requires one parameter, γ, which varies the resolution of TADs that are predicted, biasing the algorithm towards choosing larger or smaller domains. There is no direct relationship between the γ value and the domain sizes, so in order to ensure that all TAD sets have the same approximate median TAD size, the γ value was chosen individually for each Hi-C map and chromosome. The γ value which returned TADs at the expected median size of 880kb reported in Bonev and Cavalli [20] was used in each case.

Cell type	Description	Study	Resolution
GM06990	blood lymphocyte	[69]	100kb
K562	chronic myeloid leukemia	[69]	100kb
IMR90	lung fibroblast	[36]	40kb
hESC	embryonic stem cell	[36]	40kb
IMR90	lung fibroblast	[98]	5kb
GM12878	blood lymphocyte	[98]	1kb
HMEC	mammary epithelial	[98]	5kb
HUVEC	umbilical vein endothelial	[98]	5kb
K562	chronic myeloid leukemia	[98]	5kb
KBM7	chronic myeloid leukemia	[98]	5kb
NHEK	epidermal keratinocyte	[98]	5kb
A549	adenocarcinomic alveolar basal epithelial	ENCODE (2016)	20kb
Caki2	clear cell renal carcinoma (epithelial)	ENCODE (2016)	20kb
G401	rhabdoid tumor kidney epithelial	ENCODE (2016)	20kb
LNCaP-FGC	prostate carcinoma epithelial-like	ENCODE (2016)	20kb
NCI-H460	large cell lung cancer	ENCODE (2016)	20kb
Panc1	pancreas ductal adenocarcinoma	ENCODE (2016)	20kb
RPMI-7951	malignant melanoma	ENCODE (2016)	20kb
SJCRH30	rhabdomyosarcoma fibroblast	ENCODE (2016)	20kb
SKMEL5	malignant melanoma	ENCODE (2016)	20kb
SKNDZ	neuroblastoma	ENCODE (2016)	20kb
SKNMC	neuroepithelioma	ENCODE (2016)	20kb
T47D	ductal carcinoma	ENCODE (2016)	20kb

Table 2.1: Hi-C samples used for pairwise comparisons. Cell types listed in italics are non-cancer cell lines.

2.2.2 Overview of the approach

To compare two samples we quantify the similarity between their TAD boundary locations. A TAD is a genomic interval, consisting of a range of bins. A TAD set is then a collection of these intervals identified by a TAD caller. A TAD set can be thought of as a one-dimensional clustering for all of the genomic bins along a chromosome, where the bins within each TAD form a cluster. A natural way to compare clusterings is using a distance metric on these clusterings, and two highly similar clusterings, (i.e., TAD sets, in our case) will be identifiable by a low distance. To identify structurally similar regions, we compute the distances for all possible regions (e.g., all sub-intervals of the chromosome) and then select the regions with statistically significantly low distance. More specifically, we compute VI between the TADs in $[i, j]$ in one sample to the TADs in $[i, j]$ in another sample, for all relevant $[i, j]$.

The distances between all sub-intervals on the chromosome can be represented in an $n \times n$ matrix, where n is the length of the chromosome in bins, and every entry (i, j) represents the distance of the TAD structures in the region between genomic bins i and j. In this full matrix, the elements that are candidates for representing the most similar regions will appear as local minima in the sense that they are smaller than all eight surrounding values. These are intervals that are more similar than any neighboring interval. To determine which are significant we compute p-values for each of these local minima with a strict null model (Section 2.4). Once the statistically significant intervals have been identified, we further select only those which are dominating in the sense that every sub-interval within them has a higher distance measure. These intervals are then called significant structurally similar regions. An overview of the method is seen in Figure 2.1.

As a distance measure, we use the well-established VI metric, which evaluates the level of agreement between two clusterings based on information theoretic quantities [76]. The VI of two clusterings C and C' can be computed as the normalized sum of the two conditional entropies,

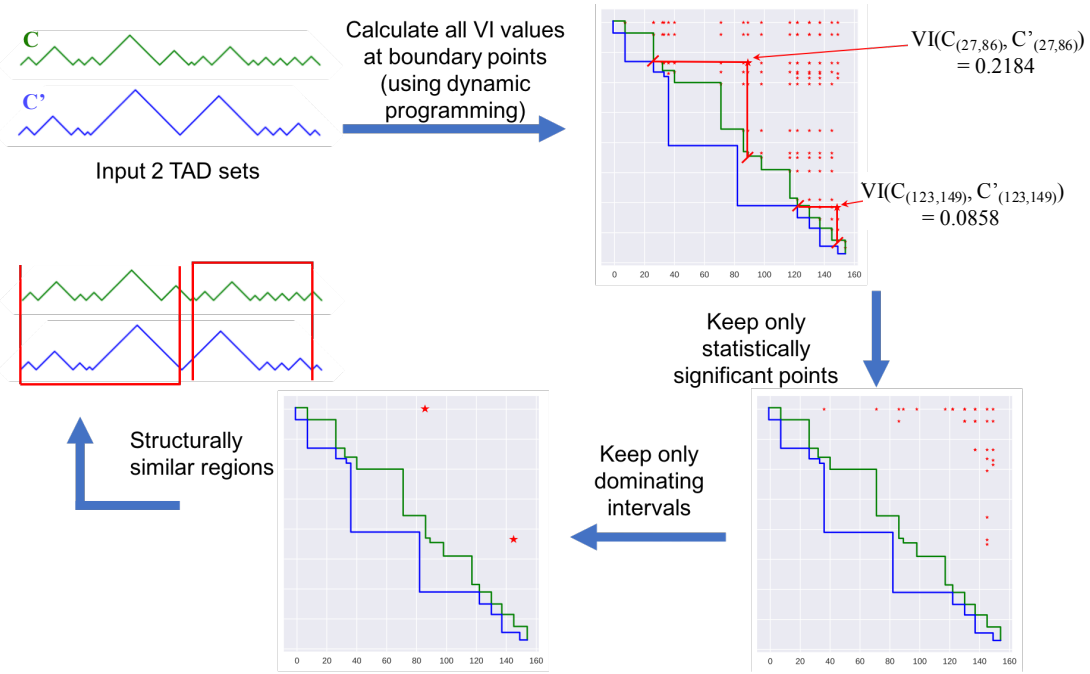


Figure 2.1: Overview of major steps to identify structurally similar intervals.

where n is the number of elements (genomic bins, in our case) in C and C', as shown below.

$$VI(C, C') = \frac{H(C \mid C') + H(C' \mid C)}{\log(n)} \tag{2.1}$$

where the conditional entropy is defined as

$$H(C|C') = \sum_{i=1}^{k} \sum_{j=1}^{k'} P(i,j) \log \frac{P(j)}{P(i,j)} \tag{2.2}$$

and C and C' contain k and k' clusters, respectively, and $P(i) = \frac{|C_i|}{n}$, $P(i,j) = \frac{|C_i \cap C'_j|}{n}$. This metric was also used by Fillipova et al. [42] to compare their TAD calls with previous methods.

In practice, rather than calculating the entire matrix of VI values for every possible chromosomal sub-interval, we only compute sub-intervals that begin and end at TAD boundaries. Although it seems intuitive that the minimum VI distance would occur exclusively at cluster boundaries, this is not strictly true, as the VI formulation holds no such theoretical guarantees. However, in 10 randomized empirical tests, we observed 100% of local minima occur at boundary points. Biologically, outside of TAD boundaries we have little understanding of fine-scale

chromosome structure, and therefore it is difficult to interpret the meaning of structural similarity away from these demarcations. We therefore calculate VI values only at TAD boundaries, and analyze this much smaller set of sub-intervals.

While some TAD callers return a partition of the chromosome with no gaps between TADs, Armatus does not explicitly require each bin to be within a TAD. This results in occasional gaps, or non-TAD domains, though they are rare; on average across all cell types and parameter values, TADs cover 92.02% of the genome. Our method does not distinguish between these non-TADs and TADs; we consider all domains in the same way. The result of this is that we are practically measuring the partition of the chromosome induced by the TAD set, rather than the exact TADs themselves, but this remains a measurement of structural similarity.

2.2.3 Dynamic programming to compute multiple VI distances

In order to further improve efficiency, we use a dynamic programming algorithm to compute VI for every pair of boundaries. The algorithm is initialized by calculating the VI for every single-TAD interval in both TAD sets. We then proceed by adding the subsequent TAD to each of these intervals, computing the VI of the new interval as a function of the VI values of the smaller two intervals composing it. After computing VI values for every interval of two TADs in each TAD set, we continue increasing the intervals by one TAD until all sub-intervals have been covered.

After initialization, at each step we have the VI of both sub-intervals to be combined into a larger interval. Let the sub-interval (i, j) be covered by TAD sets or clusterings C and C', and the sub-interval $(j+1, k)$ be covered by D and D'. We then define the sets CD and $C'D'$ as the concatenation of C and D, and C' and D', respectively, which cover (i, k). In order to compute $VI(CD, C'D')$, there are two cases to consider, illustrated in Figure 2.2. In the simpler case, there is no TAD in either TAD set that crosses the boundary at j, and the new VI is simply a rescaled sum of the previously calculated VIs.

$$VI(CD, C'D') = \frac{j-i+1}{k-i+1} VI(C, C') + \frac{k-j}{k-i+1} VI(D, D') \tag{2.3}$$

In the case of a TAD that overlaps the boundary between the two sub-intervals, one conditional entropy term can simply be rescaled as before, but we must adjust the entropy term conditioned on the TAD set including the overlapping boundary. If there is a TAD in $C'D'$ which begins at $s \leq j$ and ends at $e > j$ which we refer to as $C'D'_{se}$ (made up of C'_k, the last TAD in C' and D'_1, the first TAD in D'), the new conditional entropies are given below.

$$H(C'D'|CD) = \frac{j-i+1}{k-i+1}H(C'|C) + \frac{k-j}{k-i+1}H(D'|D) \tag{2.4}$$

$$H(CD|C'D') = \frac{j-i+1}{k-i+1}H(C|C') + \frac{k-j}{k-i+1}H(D|D') \tag{2.5}$$

$$- \frac{1}{k-i+1}\sum_a |C_a \cap C'_k| \log \frac{|C'_k|}{|C_j \cap C'_k|} \tag{2.6}$$

$$- \frac{1}{k-i+1}\sum_a |D_a \cap D'_1| \log \frac{|D'_1|}{|D_j \cap D'_1|} \tag{2.7}$$

$$+ \frac{e-s+1}{k-i+1}H(CD \mid C'D'_{se}) \tag{2.8}$$

We only compute VI at locations with a boundary in one of the two TAD sets, so we do not encounter the case in which there is an overlapping TAD in both TAD sets. The algorithm ensures that for each VI calculation, at least one TAD set will have a boundary at the point joining the two sub-intervals. In a timing test on ten randomly chosen cell type pairs and chromosomes, the dynamic programming algorithm reduced the time to compute VI at all boundary points by 58.24% (from 3.384s to 1.413s). When computing similar intervals and using a permutation test for significance (Section 2.4) between all cell types using all chromosomes, this savings is significant.

2.2.4 Identifying statistically significant sub-intervals

Once the VI values for all candidate sub-intervals have been calculated, we select the statistically significant regions through an adapted permutation test. For each sub-interval, we fix each TAD set and randomly shuffle the TADs from the other set n times, calculating the VI at each reshuffling. The p-value is then the average of the two fractions of shuffles in which a lower VI was

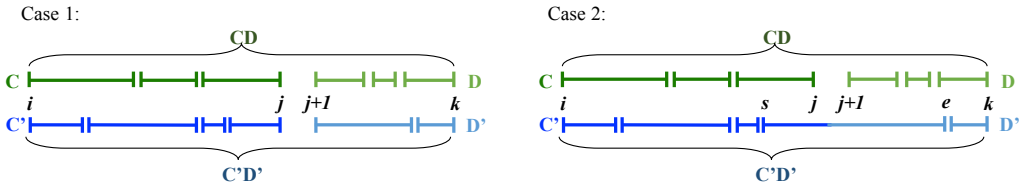


Figure 2.2: The two possible cases for dynamic programming algorithm. Case 1 shows the combination of TAD sets where both have a boundary at j, while case 2 illustrates a TAD in one set which overlaps the boundary at j.

found than the original. The number of permutations n is chosen adaptively, where we continue randomly shuffling until the p-value converges. We define convergence as the point at which the p-values for 5 consecutive iterations are equal up to at least 5 decimal places. The strictness of this null model comes from looking at each interval separately rather than shuffling the TADs across the entire chromosome at once, as well as keeping the TAD lengths fixed in the shuffling. For each interval, we are therefore calculating the likelihood of achieving a more closely matched TAD set while keeping the exact same number of TADs and their lengths. After computing this probability, we control the false discovery rate at a level of 0.05 through the Benjamini-Hochberg procedure [16], keeping only the intervals for which we cannot reject the null hypothesis at this level.

2.2.5 Dominating intervals

The set of statistically significant intervals still includes many nested intervals, so to remove redundant results we introduce the notion of dominating intervals. An interval is defined as dominating through three tests. First, it must have a p-value that passes the statistical significance tests described above. Next, we keep only the intervals that do not contain any sub-intervals with a lower VI value. Finally, if there are still intervals among this set that begin or end at the same point, we keep only the longest. Our method therefore outputs statistically significant intervals that are optimal in the sense that there is no significant sub-interval that represents a

higher similarity score. These significant, dominating intervals are the final result of the method, representing chromosomal intervals with significantly similar TAD structures.

2.2.6 Hanging TADs

One of the common artifacts of this method is a portion of an additional TAD included in the reported interval, where this full TAD at the end of the interval does not match well with its counterpart in the other TAD set. This happens because the intervals considered may fall within a TAD, creating a false TAD boundary at the start of end of the interval, and the algorithm is unable to distinguish between true TAD boundaries and those imposed by the genome segmentation. We therefore define a "hanging TAD" as a TAD at the edge of an interval that has been truncated to less than 50% of its original length. Hanging TADs can appear to be a perfect match to a true TAD and will therefore be included in the interval identified by the method (an example of this can be seen in Figure 2.3, where the hanging TADs have been circled).

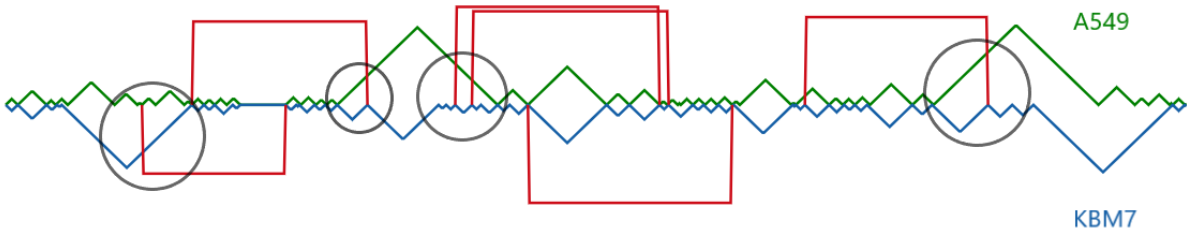


Figure 2.3: TADsim example. TAD sets from chromosome 18 are represented by triangles spanning each TAD, with TADs from A549 in green and from KBM7 in blue. The red brackets outline the significant, dominating intervals (regions of the genome covered by similar TAD structures) identified by the original TADsim method. Hanging TADs are shown by gray circles.

In order to avoid this, a preprocessing step was added to the algorithm. Sub-intervals that include hanging TADs (as defined above) are removed from consideration before testing for statistical significance, guaranteeing that they will not be included in the output. Figure 2.4

shows the output for the same input set as Figure 2.3, after removing hanging TADs. Though there are fewer total intervals identified after this modification, the area covered by both outputs is essentially the same.

2.2.7 Parallelism, concurrency, and memory optimization

The original algorithm, especially the statistical testing of intervals, is fairly computationally expensive and could take up to several hours depending on the number of TADs in the input sets. To mitigate this difficulty, we modified the implementation to make it both concurrent and memory optimized. Each execution can spawn one or more processes with each process having 1 or more threads. The independence of each permutation in the p-value calculation sub-routine allows for high concurrency that is maximized when each process has only one thread. The speedup is then directly proportional to the number of processing cores. However, once the process count becomes more than the number of cores there is only a minor improvement in the speedup which may be attributed to processor saturation. Tests were run on a 24-core/48-thread 2.6 GHz Intel Xeon E5-2690 machine. These improvements, along with the hanging TAD fix, were made with Akshat Singhal, who was a summer intern at the time.

2.3 Results

2.3.1 Comparison of TAD similarity across 253 pairs of cell types

The method described above was run on all pairwise combinations of the 23 Hi-C maps (253 pairs total), on all 22 autosomal chromosomes, resulting in an average of 5.908 significant intervals per pairwise comparison per chromosome. The average length of a region of structural similarity across all 253 pairwise comparisons is 15.25Mb, with the longest spanning almost the entirety of chromosome 2 at 219.7Mb, between NHEK and GM12878, and the shortest of length 1.4Mb, on chromosome 9 between A549 and NCI-H460. An example of the output intervals can be seen

in Figure 2.4.

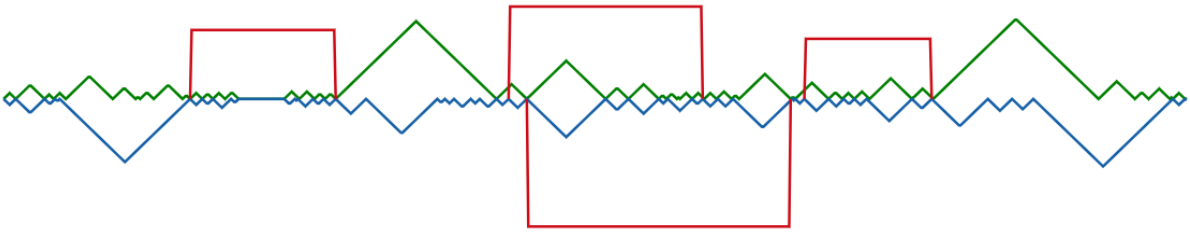


Figure 2.4: A sample output of our method, from chromosome 18 of A549 (green) and KBM7 (blue), with the significant, dominating intervals marked by red brackets. The blank space with no TADs in either set corresponds to the centromere, where no reads can be mapped in the Hi-C data.

We can compare the relative conservation and variability of chromosomal regions by looking at the results at the chromosome level. We say that a genomic bin is structurally conserved in one pairwise comparison if it is contained within one of the significant, dominating intervals. On average, each 100kb genomic bin is structurally conserved in 115.02 out of 253 possible pairwise comparisons, though this varies significantly by location. Figure 2.5 shows, at each genomic bin, the number of cell type pairs in which the bin was contained in a significant structurally similar interval, across two representative chromosomes. We expect the centromere to be conserved in all cell types, and it does appear as a highly conserved element though not in every pairwise comparison. The reason for this is our significance test, which ensures that no single-TAD interval will be considered significant. There must therefore be enough structural similarity in the regions flanking the centromere to deem any interval spanning the centromere significant. Outside of the centromere, overall variability of this bin-level similarity measure is fairly high. There appear to be chromosomal regions that are extremely similar across most cell types, while others share almost no similarity between any of the pairs we studied.

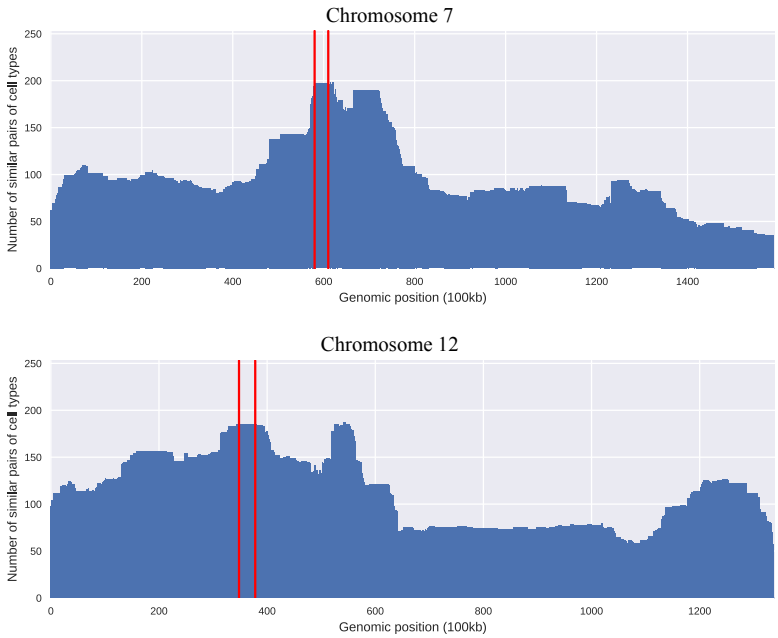


Figure 2.5: Structural conservation by genomic location for several chromosomes. The height at each genomic bin represents the number of cell type pairs in which the bin was contained in a significant structurally similar interval. The red lines show the approximate location of the centromere, where reads cannot be mapped and therefore almost all Hi-C maps should be empty in this region, resulting in the appearance of a highly conserved structural element. The significance threshold enforces a minimum number of TADs that must be included in a significant interval, so there are some cell type pairs which differ enough in structure around the centromere that it does not appear as a conserved element in these comparisons.

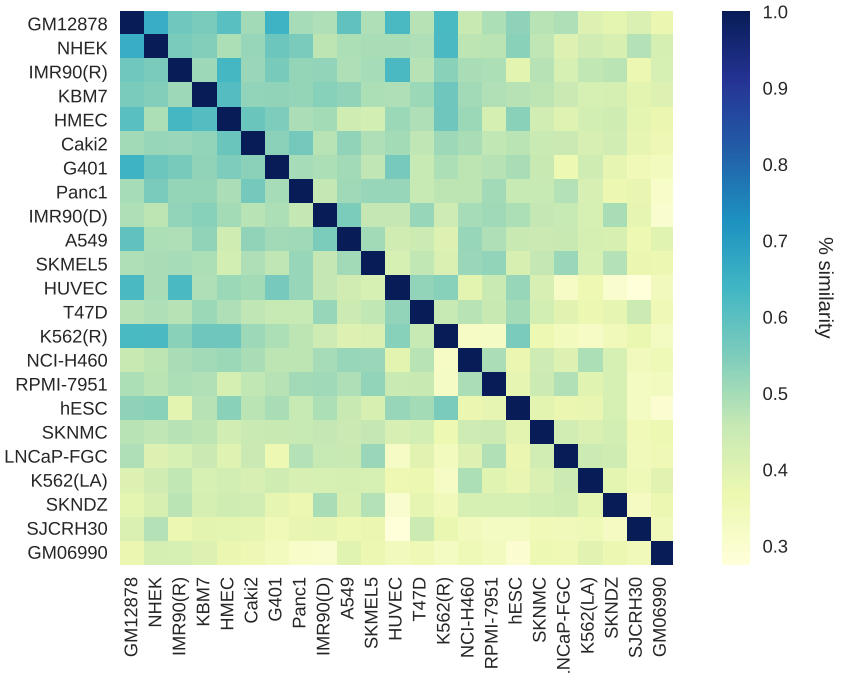


Figure 2.6: Heat map of the genome-wide percent similarity between all pairs of cell types studied. Rows are ordered by highest to lowest average pairwise % similarity, calculated by summing the values across each row, dividing by the number of rows, and sorting by this average.

2.3.2 Quantifying genome-wide and chromosome-level similarity

The identified structurally similar regions can be further used to measure the genome-wide and chromosome-level similarity. The percent similarity between two genomes (or two chromosomes) was defined as the percentage of the genome (or chromosome) covered by a significant, dominating interval between each pair of cell types. The full set of pairwise percent similarity values is presented as a heat map in Figure 2.6. The top ten pairs in terms of percent similarity are shown in Table 2.2.

The two IMR90 samples rank somewhat highly (52.41%, ranked 35 out of 253) in terms of percent similarity, but the two K562 samples are very dissimilar (32.14%, ranked 246 out of 253). This could be explained by the markedly low average similarity of both [69] cell types (K562 and GM06990) with all other cell types; both rank in the bottom four of average similarity. This is

Cell type pair	Cell type 1 description	Cell type 2 description	% similar
GM12878, NHEK	blood lymphocyte	epidermal keratinocyte	66.02
G401, GM12878	rhabdoid tumor kidney epithelial	blood lymphocyte	64.64
IMR90 (R), HMEC	lung fibroblast	mammary epithelial	63.24
GM12878, K562(R)	blood lymphocyte	chronic myeloid leukemia	62.56
K562(R), NHEK	chronic myeloid leukemia	epidermal keratinocyte	62.33
GM12878, HUVEC	blood lymphocyte	umbilical vein endothelial	62.30
IMR90 (R), HUVEC	lung fibroblast	umbilical vein endothelial	62.13
HMEC, KBM7	mammary epithelial	chronic myeloid leukemia	60.84
GM12878, HMEC	blood lymphocyte	mammary epithelial	60.15
A549, GM12878	adenocarcinomic alveolar basal epithelial	blood lymphocyte	59.22

Table 2.2: Top 10 cell type pairs in percent similarity. For the cell types which could come from two different samples, the initial of the first author of the data source is in parentheses.

the oldest data set we use, so the data may contain more errors or stronger batch effects than the more recently generated samples. If we instead compare the K562 data from the Rao et al. [98] study to KBM7, which comes from the same cancer type (chronic myeloid leukemia), we see a similarity of 57.26%, which ranks them 12 out of 253 pairs. There is some biological similarity and functional connection between the cell type pairs near the top of the structural similarity measure. The fourth most similar pair (GM12878 and K562) consists of a blood lymphocyte cell line and a chronic myeloid leukemia cell line of lymphoblast morphology, so these come from the same tissue and cell lineage. However, many of the most similar pairs have no apparent biological justification.

Though there are no previous methods quantifying structural similarity to which we can compare, two previous studies counted the number of TAD boundaries (computed using different methods) that they considered overlapping between certain pairs of cell types. Dixon et al.

[36], using their method referred to as DomainCaller, reported data indicating a Jaccard index (similarity score between 0 and 1) of 0.52165 between their IMR90 and hESC cell types. Our method reports a comparable similarity score of 0.4902 (fraction similarity across the genome, also between 0 and 1), despite using different methods for data normalization and TAD calling. Rao et al. [98] similarly reported the number of shared TAD boundaries between pairs of cell types including GM12878, which was sequenced much more deeply than the others. Using their own TAD calling method, they identified significantly more TADs in GM12878 than any other cell type because of the higher resolution of the data, so overall their data gave Jaccard indices ranging from 0.2129 to 0.3033 for comparisons of GM12878 to each of IMR90, HMEC, HUVEC, K562, KBM7, and NHEK. However, because there are more GM12878 TADs than any other cell type, this comparison is somewhat skewed. Simply looking at the fraction of each cell type's shared TAD boundaries with GM12878 to its own overall number of TAD boundaries gives similar TAD boundary fractions in the 0.499 to 0.6688 range. In our analysis, these same cell type pairs ranged in percent genomic similarity levels from 0.5552 to 0.6603. Again, this is using yet another TAD caller and data normalization method, but the level of similarity measured seems to be fairly robust to all of these differences.

At the chromosomal level, these percent similarities and even the ranking of pairwise similarity can vary significantly. Similarity levels averaged over all pairwise comparisons per chromosome vary from 33.70% on chromosome 1 to 69.05% on chromosome 22. For an individual pair, similarity can cover an entire chromosome as in the case of the Caki2 and HMEC which are 100% similar on chromosome 1. In contrast, some pairs have almost no similarity on a chromosome, such as SKMEL5 and the IMR90 sample from [36], which have 0.963% similarity on chromosome 1. Box plots of the distribution of overall similarity among all normal-normal and cancer-cancer cell type pairs are shown in Figure 2.7, where several chromosomes, such as 3 and 20, stand out as being particularly more structurally similar among normal cell type pairs than cancer cell type pairs.

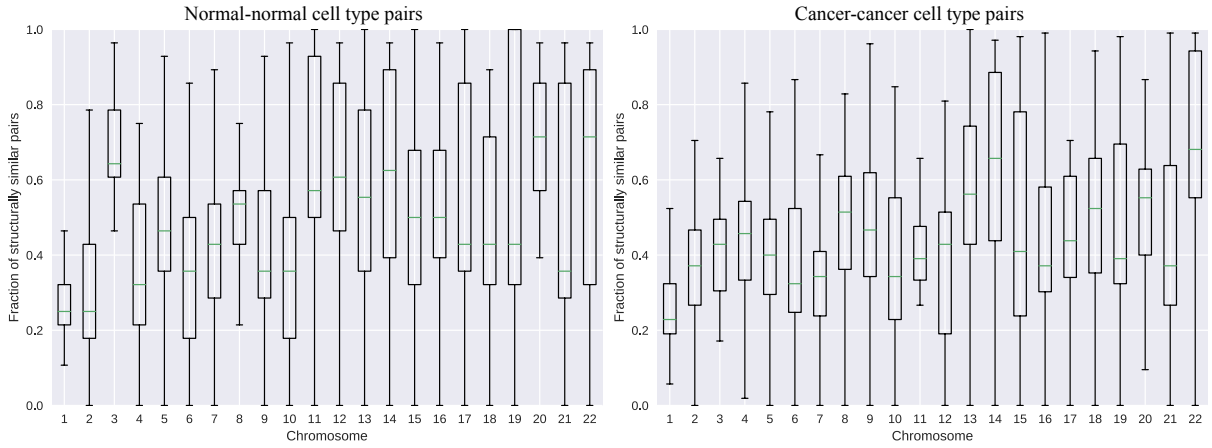


Figure 2.7: Box plots showing the distributions of similarity measures across chromosomes, in pairs of cancer cell types and normal cell types. The distribution is over all genomic bins of the given chromosome, and the value at each genomic bin is the fraction of (cancer-cancer or normal-normal) pairs for which the bin is contained in a significant dominating interval.

2.3.3 Comparing structural conservation between cancer and non-cancer cell type pairs

Several studies have shown that chromosome structure can be disrupted in a broad range of cancer types [48, 54, 75, 78], and the comparison method above can give a genome-wide view of structural similarity among cell type pairs of all combinations of normal and cancer cell types. Among the 21 unique cell types in our data set, 14 come from cancer cell lines and the other 7 are non-cancerous (see Table 3.1). Including the two duplicate cell types, this gives 28 pairs of two normal cell types, and 105 pairs of two cancer cell types. Globally, the normal-normal pairs show slightly higher average structural conservation, but the difference is not significant: 44.17% average similarity among cancer-cancer pairs, and 49.02% similarity among normal-normal pairs.

However, we find that there is more structural conservation at the regions around established pan-cancer genes in normal-normal cell type pairs than in cancer-cancer pairs, which may point

to the structural disruption that occurs in conjunction with cancer mutations. Looking at the top 10 most commonly mutated pan-cancer genes from a large-scale study of data from The Cancer Genome Atlas [61], we can see that the structure around most of these genes is more conserved among normal cell type pairs than cancer pairs (Figures 2.8 and 2.9). Figure 2.8 shows the distributions for each chromosome of the percent similarity among cancer-cancer pairs subtracted from normal-normal pairs. A value above zero indicates higher structural similarity among normal-normal cell type pairs. Despite 10 out of 22 chromosomes having lower than zero average difference, 9/10 cancer genes are located on chromosomes with a positive average value. In addition, we note that three of these genes are located on chromosome 3, which has the highest average difference between structural similarity in normal-normal pairs compared to cancer-cancer pairs. The prevalence of mutations in cancer cells on genes located on chromosome 3 and the disruption caused by the mutations may result in variable structural changes in cancer cells.

Looking more closely at these ten gene locations, we note that normal-normal pairs are more structurally similar at nine of these ten gene locations (Figure 2.9). Over all human gene loci, 57.77% show a higher fraction of structurally similar normal-normal pairs than cancer-cancer pairs, which gives a probability of 0.03441 (using the hypergeometric test) of pulling at least 9/10 random genes with higher normal-normal structural conservation, suggesting that the pattern of Figure 2.9 is statistically significant. If we further restrict the null model to the probability of finding at least 9/10 genes from the same chromosomes as our pan-cancer genes, the p-value increases to 0.1425, which is expected based on the distributions shown in Figure 2.8. Though this value is above the traditional 0.05 p-value cutoff, the combination of results suggests a role for 3D structure disruption around mutated genes in cancer cell types. In order to validate and confirm this conclusion, we would need more Hi-C samples.

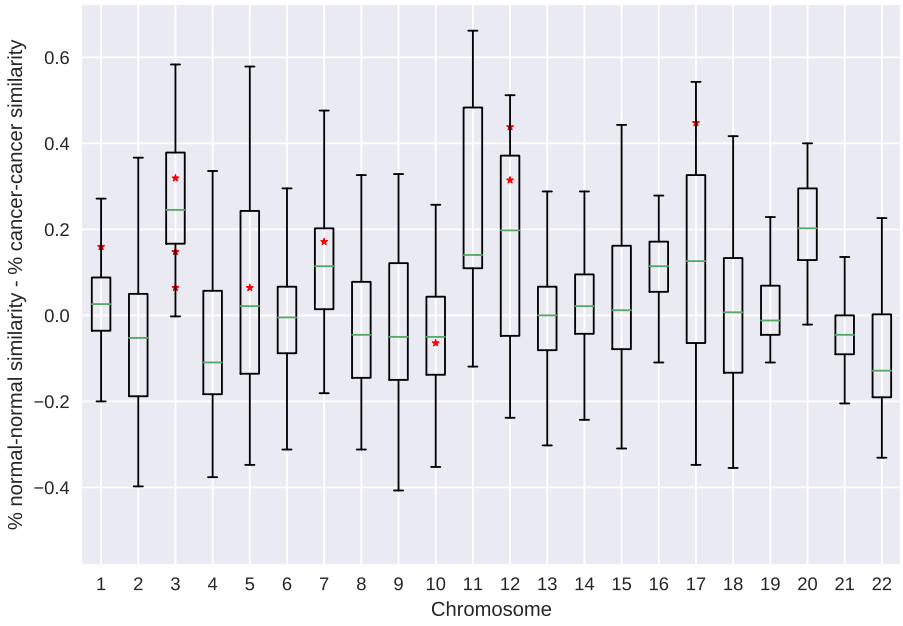


Figure 2.8: Box plot showing the chromosome-level distributions of differences between level of structural similarity at all genes in normal cell type pairs and cancer cell type pairs. The red stars represent the differences observed at the ten most commonly mutated pan-cancer genes from [61].

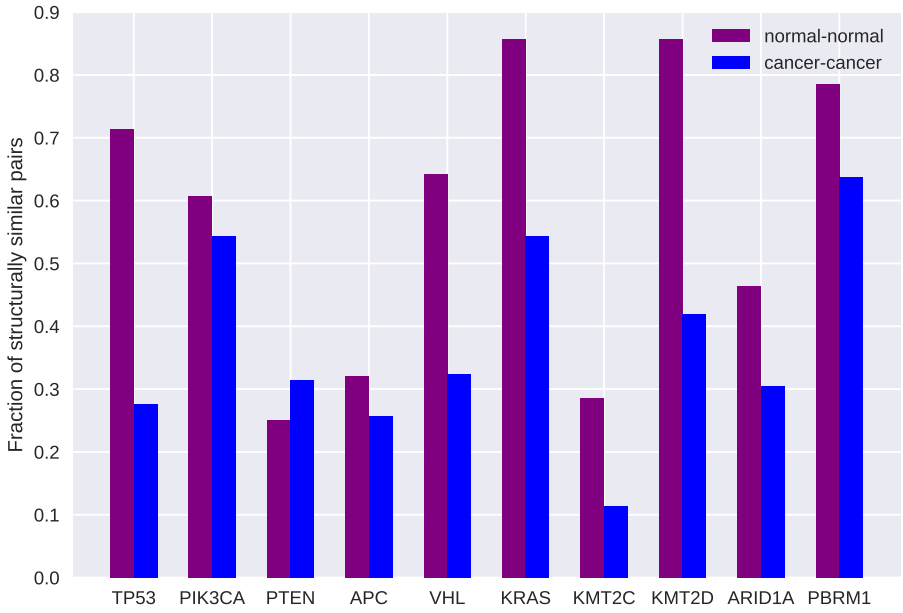


Figure 2.9: Relative conservation of cancer-cancer and normal-normal cell type pairs at ten prominent pan-cancer gene locations. For the cases in which the gene spans multiple bins, the bin for the gene location was chosen as the bin containing the gene's midpoint.

2.4 Discussion and Conclusions

We have presented the first method to quantify local chromosomal structural similarity and have used it to perform a large-scale comparison of TAD structure across 23 human samples from both cancer and noncancerous conditions. We note the variability among structural components both globally and by chromosome, as well as between cancer and normal cell types. This led to the new observation that pan-cancer gene locations show more structural variability among cancer cells than among normal cells.

Though the analysis was performed using only TADs from the Armatus software at one individually-optimized parameter setting, our results are in line with the levels of structural similarity reported by other studies using other TAD finders and pre-processing pipelines. Further study in this area will involve testing the robustness of these results to the choice of TAD caller, as well as the Hi-C data resolution and normalization. Another tunable aspect of this method is the choice of distance metric, for which we used VI. Though VI is a well-established and general metric for calculating clustering similarity, there are many other metrics which fit the same criteria.

Beyond the methodological choices, our results are somewhat dictated by the available Hi-C samples. Hi-C is a fairly expensive and time-consuming protocol, so the amount of data available is much smaller than other genomic data types such as RNA-seq. We selected samples from prominent studies in the field, but without more data it is difficult to determine whether chromosome structure can be tissue-specific or cancer type-specific, or any number of other possibilities. As more data becomes available, the robustness of the results of such a structural comparison will significantly increase.

Given the set of samples we used, it is difficult to determine the level of batch effects or other protocol-specific differences influencing our results. The extremely low similarity values for both samples from the Lieberman-Aiden et al. [69] study seem to suggest some batch effects or protocol-specific variations, but otherwise the similarity clustering did not simply group cell

types from the same studies. This concern could be further studied or mitigated with more Hi-C samples.

Another concern with the data is specific to the cancer samples, which are likely to be highly mutated and contain genomic structural variants. Despite this, we still map them to the reference (non-cancer) genome. Some of the areas where we see structural differences across cancer cells may simply be due to an inability to map reads with high mutation levels, rather than a variation in three-dimensional structure. Through further advances in long-read technology and genome mapping and assembly, it may become easier to avoid these concerns and study three-dimensional structure more directly. Some work has begun in this area, combining structural variant detection with Hi-C data [24].

Our method and analysis represents a first step towards understanding the conservation and changes in chromosome structure across human cell types and disease states. We provide the first genome-wide structural comparison of cancer and non-cancer genes, as well as a systematic pairwise analysis of similarity across 23 human cell types. As Hi-C data becomes more widely available and reliable, the ability to compare and identify structurally similar or variable regions may provide even more insight into the mechanisms and influence of chromosome architecture on gene regulation and cellular functioning.

Chapter 3

Analysis of TAD variability

In this chapter, we analyze 137 Hi-C samples from 9 studies under 3 measures (including TADsim, introduced in Chapter 2) to quantify the effects of various sources of biological and experimental variation in TAD sets. We observe significant variation in TAD sets between both non-replicate and replicate samples, and provide initial evidence that this variability does not come from genetic sequence differences. The effects of experimental protocol differences are also measured, demonstrating that samples can have protocol-specific structural changes, but that TADs are generally robust to lab-specific differences. This study represents a systematic quantification of key factors influencing comparisons of chromosome structure, suggesting significant variability and the potential for cell-type-specific structural features, which has previously not been systematically explored. The lack of observed influence of heredity and genetic differences on chromosome structure suggests that factors other than the genetic sequence are driving this structure, which plays an important role in human disease and cellular functioning.

A version of this chapter appeared in *NAR Genomics and Bioinformatics* and is joint work with Akshat Singhal and Carl Kingsford [108]. The scripts to reproduce the analyses presented here are available at `https://github.com/Kingsford-Group/localtadsim/tree/master/analysis`.

3.1 Introduction

While it is recognized that the three-dimensional structure of the chromosome is an integral part of many key genomic functions, we lack a full understanding of the variability of this structure across biological sources or experimental conditions. Changes in chromosome structure at specific genomic regions and under certain conditions have been implicated in a variety of human diseases and disabilities, including many cancers [48, 54, 75, 78], deformation or malformation of limbs during development [71], and severe brain anomalies [119]. In healthy cells, genome shape is heavily linked to key processes such as gene regulation and expression [23, 29, 40, 65, 101], replication timing [7, 79, 92, 104], and DNA accessibility and nuclear organization [27, 96, 131]. Despite the clear importance of these structures, there has been no systematic study of the expected variation of topologically associated domains (TADs) genome-wide.

Features of genome-wide, three-dimensional chromosome structure can be measured by Hi-C [69], a high-throughput variant of the chromosome conformation capture protocol [31]. The experiment involves cross-linking and ligating spatially close genomic segments, then aligning them back to the genome to find their genomic positions. The output of this experiment is a matrix in which the rows and columns represent segments of the genome along a chromosome, and each matrix entry records the pairwise interaction frequency of the genome fragments of the associated row and column. These values reflect 3D proximity, quantifying the frequency of contact between every pair of genomic segments.

A hierarchical architecture has emerged from these Hi-C matrices, in which chromosome structure is composed of several different scales of components, from multi-megabase compartments to sub-megabase TADs and sub-TADs [20, 42]. TADs represent chromosomal regions with significantly higher interaction frequency among segments within the TAD than with those outside of it [36]. TADs are considered to be a primary structural building block of chromosome architecture [38], and several methods have been developed to computationally identify

them [28, 36, 42, 85, 98, 129].

One challenge in the interpretation of TADs is that we have little understanding of the variability of TAD structures under different conditions. While some work has compared aspects of Hi-C data quality, TADs in particular were not considered [133]. No other Hi-C study has used more than 23 samples of different conditions, and even large data repositories such as ENCODE and the 4D Nucleome contain no more than 30 human Hi-C samples on their own. As more Hi-C data has become available recently, it is now possible to perform a substantial analysis of the relative consistency or variability of TADs across a variety of human cell conditions, by combining Hi-C samples from many studies and resources. Previous work has suggested that TADs are largely conserved across human cell types and possibly even species, however the degree of this conservation is unclear and has been tested in only small sets of samples [36, 98].

An initial method to compare TADs between cell types was previously applied to compare normal versus cancer human cell types [106], but that study did not investigate other potential sources of TAD variability and only compared 23 different cell or tissue types. We instead systematically quantify several sources of variability that have not been previously studied, using over three times as many different cell conditions, and three metrics.

We quantify the influence of both technical and biological variation on TAD structures across several experimental and biological conditions in the first study of over 100 Hi-C experiments. We observe that 10–70% of combined TAD boundaries differ between replicates, regardless of sequencing depth or contact coverage, pointing to a potentially dynamic or disordered arrangement. Across 69 samples of different cell lines and tissue types, we observe $\sim$20–80% unshared TAD boundaries, suggesting that there can be fairly large differences in TAD sets across biological conditions, in contrast to previous claims of extensive TAD conservation [36, 98, 109]. We find that samples of the same cell or tissue type have elevated structural similarities, suggesting that biological function is a key driver of structural similarity. Though it is commonly believed that TADs do not vary much across cell types and possibly even species, we observe signifi-

cant TAD variation across human cell and tissue types. By analyzing the structural similarity of sets of parents and their children, as well as tissue samples taken from different individuals, we observe that the genetic sequence differences between individuals and the genetic sequence similarities between parents and their children have little impact on TAD structural similarity. Of the possible sources of technical variation considered in this work, the choice of *in situ* (in nucleus) ligation versus dilution (in solution) ligation protocols has the greatest influence on Hi-C and TAD structures. In contrast, we demonstrate that Hi-C measurements and corresponding TAD calls are robust to other technical differences such as the choice of restriction enzyme and lab-specific variations.

3.2 Materials and methods

3.2.1 Data

A total of 76 human Hi-C samples were processed from sequencing reads (.fastq files) downloaded from various publicly available sources (Sequence Read Archive (SRA) [66], ENCODE [125], Gene Expression Omnibus (GEO) [12], or 4DN portal [32]). Normalized Hi-C matrices were computed from the reads through the HiC-Pro pipeline [113], and each sample was tested for quality at 100kb resolution. Using the criteria suggested by Ay and Noble [6] and Rao *et al*. [98] (at least 80% of all bins must contain more than 1000 contacts), we found 7 experiments which could not be analyzed at 100kb resolution or less (Table 3.2), leaving 69 human Hi-C data sets (137 including all replicate samples) representing 52 unique cell types or biological sources from 9 studies. The details of these experiments, including accession numbers, are found in Table 3.1. All samples were normalized using iterative correction and eigenvector decomposition (ICE) [59], and all analyses presented here were performed at 100kb resolution, unless otherwise noted. For analyses that do not explicitly compare replicates, all aligned reads from each replicate of a sample were merged and processed into a single combined Hi-C matrix for optimal data

quality.

From the Hi-C matrices, TADs were computed using Armatus [42], a commonly used method for identifying TADs efficiently. Armatus takes one parameter γ, which controls the expected TAD size. For every sample and chromosome, Armatus was run with γ values ranging from 0 to 1 at intervals of 0.1, and the γ value was chosen to ensure a distribution of TADs with median as close as possible to the expected median TAD size of 880kb [20] on each sample and chromosome.

3.2.2 Comparison measures

In order to comprehensively compare chromosome structures, we use three different measures: HiCRep [132], Jaccard Index (JI), and TADsim [106]. HiCRep measures similarity between Hi-C matrices directly, and both JI and TADsim compare similarity of predicted TADs. All three measures were computed on all 2346 pairs of non-replicate samples, in addition to all 83 replicate pairs.

HiCRep was designed to assess the reproducibility of replicates or the similarity of two Hi-C matrices. This measure uses a stratum-adjusted correlation coefficient to reliably compute a statistical similarity score between two Hi-C matrices, explicitly accounting for both the strong distance dependence found in Hi-C and the known domain structure [132]. This method returns a value that represents the overall similarity of the full Hi-C matrix, and distinguishes between replicate and non-replicate samples significantly better than simple correlation coefficients. We ran HiCRep on all intra-chromosomal matrices of our samples and averaged over all chromosomes to get a single value per cell type pair. HiCRep requires a smoothing parameter h, which was selected for each comparison according to the heuristic optimization procedure provided by the software, which chooses the minimum h value at which the score begins to converge. We allow a range of 0 to 3, which is expanded from the 0 to 2 range shown in HiCRep's documentation example, to allow more options while maintaining computational efficiency.

The Jaccard Index (JI), a simple set similarity metric, is defined as the size of the intersection of two sets A and B divided by the size of the union of the sets: $JI(A, B) = |A \cap B|/|A \cup B|$. When comparing TADs, the two sets A and B represent the two lists of TAD boundaries, as used in Forcato *et al*. [44]. The resulting JI value is an easily interpretable number representing the fraction of shared boundaries between the two TAD sets.

While JI is an effective way to compare boundary locations, it does not take into account the total overlap between TAD interiors. We therefore also adopted a measure from Sauerwald and Kingsford [106], which presented a method to identify structurally similar regions between two TAD sets. The measure used here, which we will call "TADsim," is the fraction of the genome covered by structurally similar regions identified by the method described in Chapter 3.

3.2.3 Statistical comparisons

Distributions of similarity values under all three measures were checked for statistical significance through the Mann-Whitney *U* test, also called the Mann-Whitney-Wilcoxon (MWW) test. This nonparametric statistical test assesses the null hypothesis that a randomly selected value from one sample is equally likely to be less than or greater than a randomly selected value from the other sample. The alternative hypothesis can then be formulated as a randomly selected value from one distribution being likely to be greater than (or less than) a randomly selected value from the other distribution.

Without knowing the underlying distribution of structural similarity values, a nonparametric statistical test is required for all of our comparisons. The Kolmogorov-Smirnov two-sample test (KS test) is another commonly used nonparametric test, but it does not include any assessment of which distribution is greater than the other. The KS test is additionally sensitive not only to differences in the median or mean between two distributions, but any differences in their shapes, dispersion, or skewness as well. We therefore chose the MWW test for these analyses, given that we specifically are testing for the difference in relative magnitudes of the values in the

distributions, rather than differences their overall shapes.

3.3 Results

3.3.1 Structural similarity of replicate samples

By quantifying the similarity of all 83 replicate pairs in our data, we find that the TAD sets of replicate pairs are significantly more similar than those of non-replicate pairs (Figure 3.1**a–c**, $p < 10^{-20}$ for all comparison measures), in contrast to previous work that suggested much lower TAD reproducibility between replicates [44]. We note that this discrepancy can be explained by the fact that Forcato et al. [44] used a different pre-processing method which results in many fewer aligned reads than HiC-Pro and therefore significantly fewer Hi-C contacts, decreasing the reproducibility they observe. Using the same data analyzed by Forcato *et al* at the same resolution and the same Armatus parameters they used, but processed instead with HiC-Pro, gives JI values consistent with those we found on the larger data set analyzed for this work.

Among the samples studied in this work, replicates share an average of 62.77% of their TAD boundaries, which is consistent with other previous studies on different data using different methods (Dixon *et al*. [36]: 62.28%, 73.73% and Rao *et al*. [98]: 61.88%). This leaves almost 40% of TAD boundaries that vary across replicates. Between non-replicate pairs, almost 60% TAD boundaries are not shared on average, which contradicts the common notion that TADs are highly conserved between human cell types. These levels of variability also hold at a higher resolution of 40kb (Figure 3.1**d–f**), though the sample size is much smaller due to the limited number of samples with replicates sequenced deeply enough to be analyzed independently at 40kb. The variability we observe could not be explained by limitations of sequencing depth, as we found that reproducibility is only weakly correlated with sequencing coverage (see Figure 3.2). If these relatively low similarity values for replicates reflect the true level of variability rather than the uncertainty of TAD identification, this points to a dynamic or disordered structure,

as suggested by a recent imaging study [89], and a much higher level of TAD variation than previously thought.

3.3.2 Variability across tissues and individuals

The chromosome structure of tissue samples has not been as widely studied as that of cell lines, but these structures may provide valuable insight into tissue-specific genome spatial organization. Among our set of 69 Hi-C experiments, 13 different human tissues are represented, and there are 16 pairs of the same tissue type taken from different donor individuals. The similarity values of the chromosome structures of these pairs are statistically indistinguishable from those of replicate samples (Figure 3.3**a–c**; HiCRep: $p = 0.4792$, JI: $p = 0.1300$, TADsim: $p = 0.09559$). There is much less variation across individuals than across tissue types (Figure 3.3**a–c**; HiCRep: $p = 1.5577 \times 10^{-6}$, JI: $p = 3.876 \times 10^{-6}$, TADsim: $p = 1.017 \times 10^{-7}$), suggesting that individual genetic differences have less influence on chromosome structure than the biological function of the sample.

Our analysis suggests that around 40% of TAD boundaries are shared between different tissue types, consistent with the findings of Schmitt *et al*. [109]. While this is significantly more than expected given random TAD boundary locations, it leaves room for large differences in the TAD sets of different tissue samples. In order to determine whether TAD structure is more similar across tissues than across cell lines, we compared the similarities between tissue types to the background distribution consisting of all non-replicate pairs with at least one cell line. Two of our three measures suggest that there is elevated conservation between tissues compared with cell lines, but the two TADsim distributions are statistically similar (Figure 3.3**a–c**, HiCRep: $p = 2.120 \times 10^{-19}$, JI: $p = 0.004806$, TADsim: $p = 0.4235$). The average JI value between tissue samples of 41.6% implies that while there is a significant level of similarity among chromosome structures of different tissue types, close to 60% of TAD boundaries vary between different tissue samples. This level of variability between tissue types may indicate the existence of tissue-

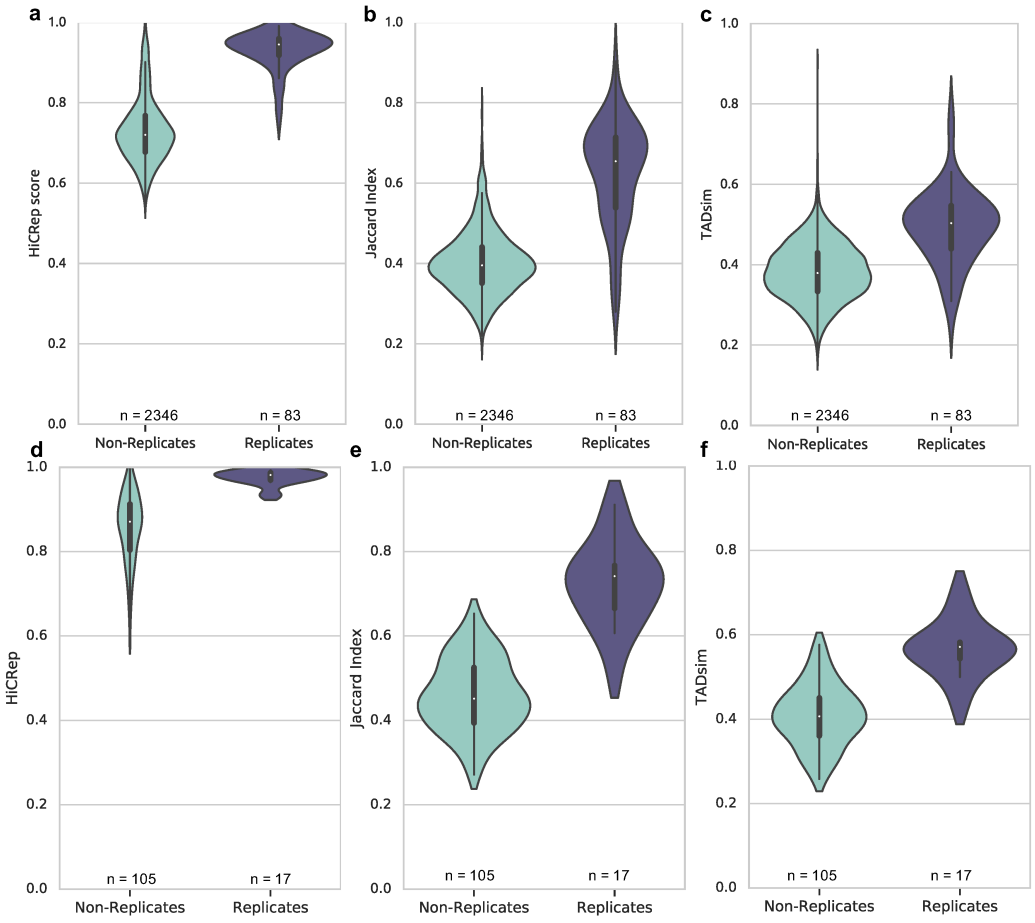


Figure 3.1: **Hi-C and TAD reproducibility.** The violin plots show distributions of HiCRep (**a,d**), Jaccard Index (**b,e**), and TADsim (**c,f**) values on pairs of either replicates or non-replicates, at 100kb (**a,b,c**) and 40kb resolution (**d,e,f**). All of these plots show a statistically significant ($p < 10^{-9}$) elevation of similarity among replicate pairs, demonstrating that both Hi-C matrices and TADs are reproducible. Only 15 samples had replicates which passed the criteria for analysis at 40kb, resulting in a much smaller sample size for these comparisons.

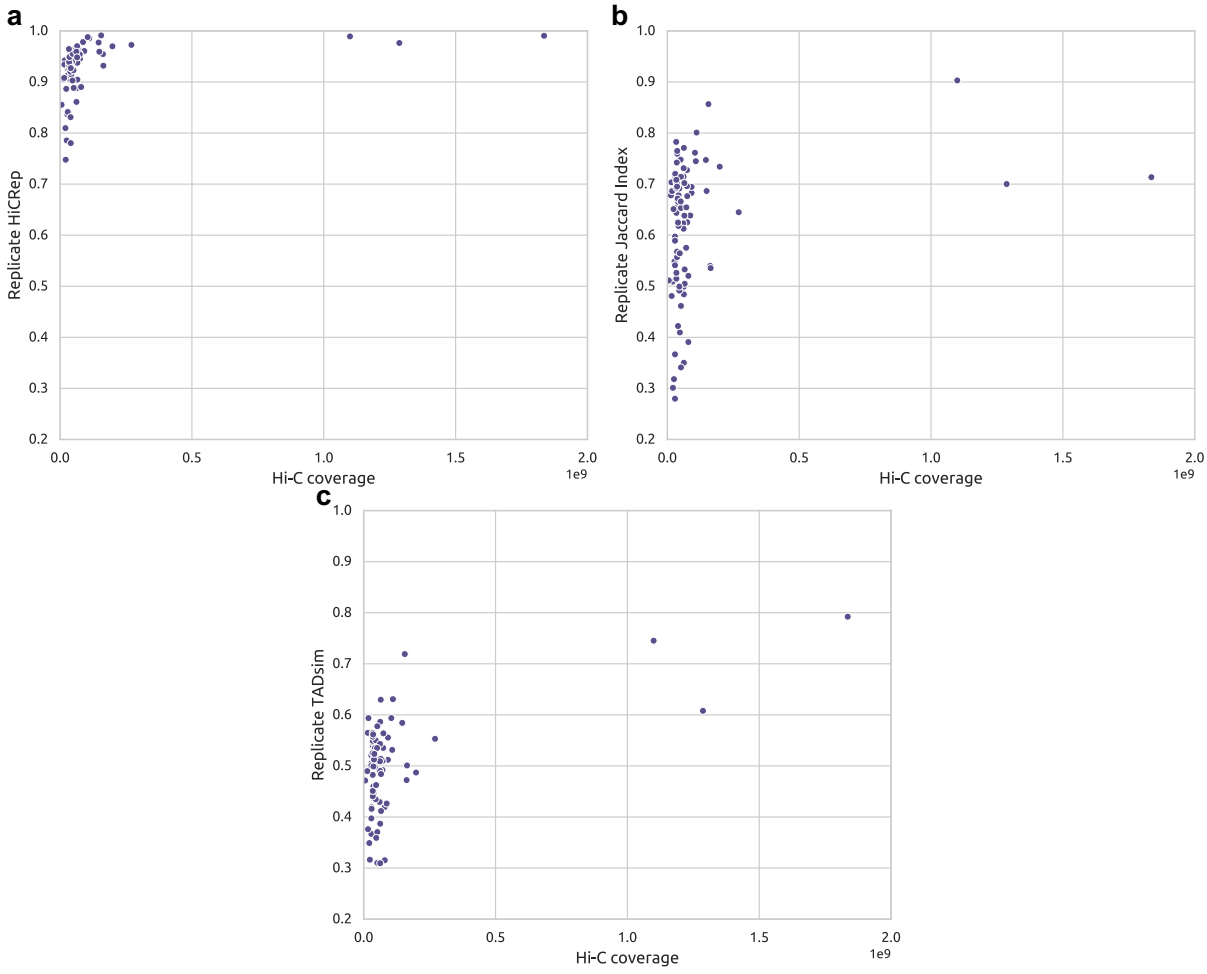


Figure 3.2: **Reproducibility versus Hi-C coverage.** Reproducibility is quantified by the similarity value of each replicate pair. We compute a contact count for each replicate sample by adding all non-normalized contacts on all intra-chromosomal matrices, and Hi-C coverage is defined as the smaller of the two contact counts for the replicate samples being compared. Across all three measures, especially the two quantifying TAD reproducibility, there is a low correlation with coverage though very high coverage experiments tend to have high reproducibility. **a** HiCRep; Spearman $\rho = 0.6378$ **b** JI; Spearman $\rho = 0.2407$ **c** TADsim; Spearman $\rho = 0.2310$.

specific structural features, rather than significant conservation of TADs between tissue types.

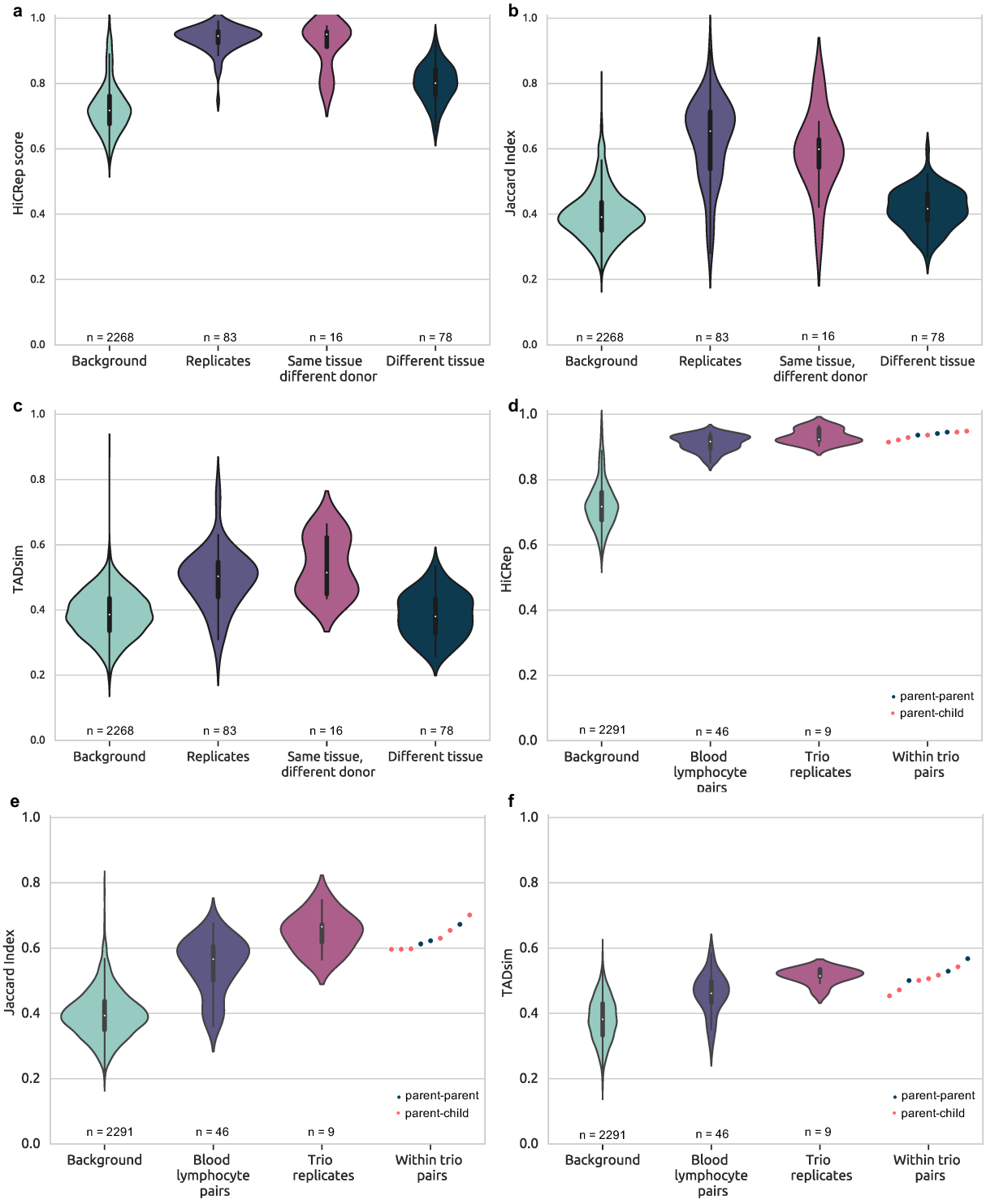
Figure 3.3: **Biological sources of TAD variation. a,b,c:** Comparisons between and within tissue samples using **a** HiCRep, **b** JI and **c** TADsim. Each figure shows four violin plots representing distributions of similarity values of the background (non-replicate pairs), replicates, pairs from the same tissue type but different donor individuals, and pairs from different tissue types. **d,e,f:** Comparisons with three trio samples of blood lymphocyte cells using **d** HiCRep, **e** JI and **f** TADsim. The background distribution consists of all non-replicate pairs, and the blood lymphocyte pair distribution shows all similarity values of two blood lymphocyte samples outside of the trios. The trio replicates refer to the similarity values of the replicate pairs from within each individual of the trio samples. The scattered points on the right side of each figure represent all within-trio comparisons, colored by family relationship.

3.3.3 Family relationships do not seem to influence TAD similarity

Hi-C measurements from blood lymphocyte cells of three parent-parent-child triplets (trios) permit a glimpse into the heritability of chromosome structure. We find that unrelated individuals (parents) share just as much structural similarity as each parent and their child (Figures 3.3**d–f**). We therefore see no evidence that chromosome structure is determined by genetic similarity, at least in blood lymphocyte cells. The similarity values within trios are generally much higher than the background of non-replicate comparisons, however they are similar to the distribution of pairs of blood lymphocytes, so this is likely a result of the shared cell type rather than genetic similarity. As with the tissue data, the biological source (cell or tissue type) seems to be a much stronger driver of structural similarity than genetic similarity.

3.3.4 Variations across Hi-C protocols

In order to investigate technical sources of variation, we compare several common variations in the Hi-C protocol, and test whether they affect the similarity of the TADs that are identified.


a
HiCRep score
1.0
0.8
0.6
0.4
0.2
0.0
n = 2268
n = 83
n = 16
n = 78
Background
Replicates
Same tissue different donor
Different tissue
b
Jaccard Index
Same tissue, different donor
c
TADsim
d
HiCRep
parent-parent
parent-child
n = 2291
n = 46
n = 9
Blood lymphocyte pairs
Trio replicates
Within trio pairs
e
Jaccard Index
f
TADsim

There are two main protocol variants that differ in the cross-linkage step. *In situ* Hi-C [98] (also termed "in nucleus Hi-C" [81]) involves cross-linking the DNA within the nucleus, while dilution Hi-C (or "in solution Hi-C") performs cross-linking in a dilute solution. Each protocol also requires the choice of a restriction enzyme, which could be any of four common options: HindIII, MboI, NcoI, and DpnII. While there has been some study of the differences in Hi-C data resulting from *in situ* and dilution protocols [81, 98], the influence on TAD sets and the effect of the restriction enzyme had not been systematically studied previously.

In situ and dilution Hi-C reproducibility

Across all replicate pairs (12 in situ, 71 dilution), the intra-chromosomal Hi-C matrices of *in situ* replicates are statistically significantly more similar than dilution replicates (Figure 3.4**a**, $p = 1.180 \times 10^{-5}$). However, the TAD sets of *in situ* replicates only show statistically significantly higher similarity than those of dilution replicates under the JI measure (Figures 3.4**b,c**; JI: $p = 0.02703$, TADsim: $p = 0.1547$). TADs capture only relatively short-range interactions, and it therefore appears that the difference between *in situ* and dilution Hi-C is not as significant a factor in TAD reproducibility as in overall Hi-C matrix reproducibility. It has been previously shown that *in situ* Hi-C matrices are more reproducible than dilution Hi-C matrices [81, 98], specifically with respect to long-range and inter-chromosomal contacts.

Comparing *in situ* and dilution samples

In order to study whether both *in situ* and dilution protocols result in the same structures, we compared samples across protocols. Among pairs of the same cell type, mixed protocol pairs, where one sample came from *in situ* and one from dilution, were consistently statistically significantly less similar than the single protocol pairs, in which both samples came from the same protocol (Figures 3.4**d,e,f**, HiCRep: $p = 0.0003423$, JI: $p = 0.03967$, TADsim: $p = 0.02002$). The chromosomal structures identified from these two protocol variants are therefore not en-

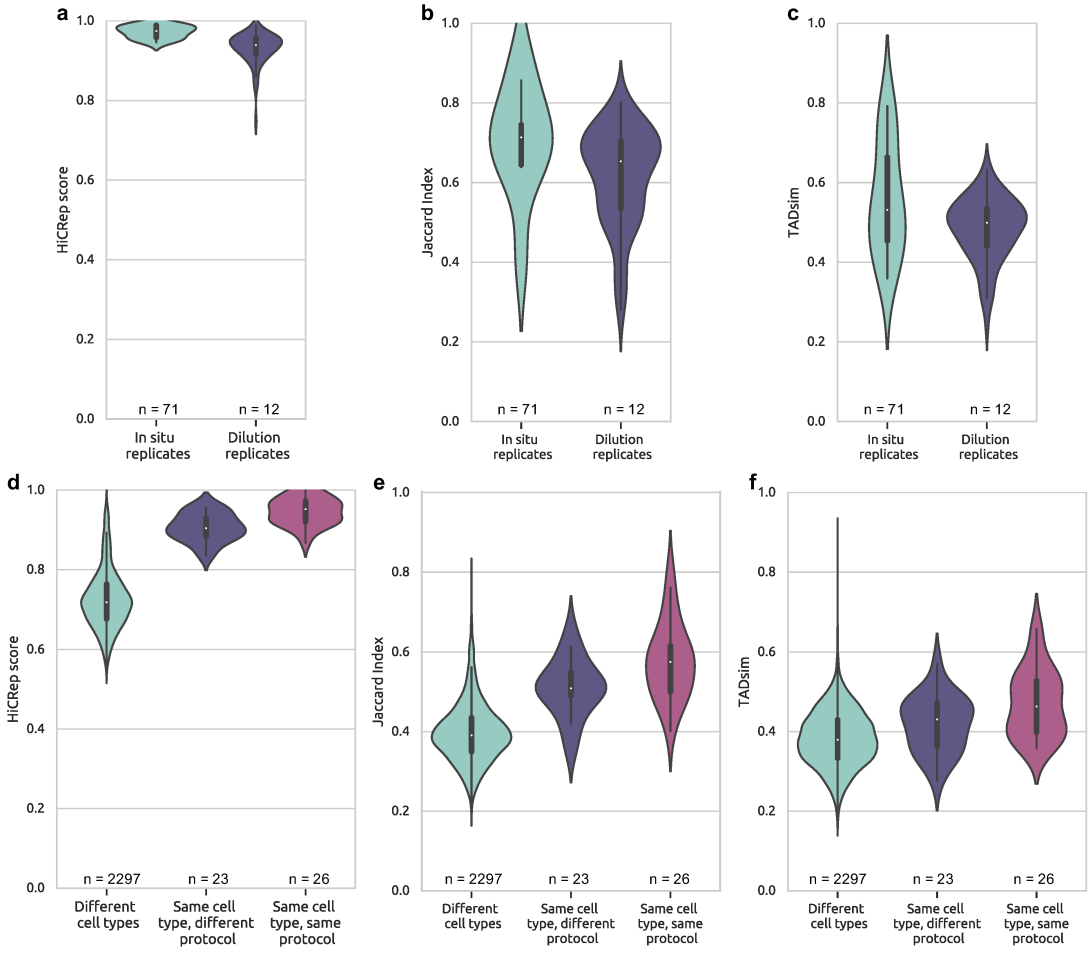


Figure 3.4: **Comparing Hi-C samples generated from the *in situ* and dilution protocols. a:** HiCRep shows that *in situ* Hi-C matrices are more reproducible than dilution matrices ($p < 0.0005$). **b,c:** TAD set reproducibility according to JI ($p = 0.02703$) and TADsim ($p = 0.1547$) shows that protocol choice has less of an impact on reproducibility of TAD sets than full Hi-C matrices. **d,e,f:** Comparisons of same cell type pairs generated by the same and different protocols. The background distribution is all comparisons of different cell types. Under all measures, there is a clear and statistically significant ($p < 0.05$) drop in similarity values of samples generated by different protocols compared to samples generated by the same protocol.

tirely consistent, although pairs from the same cell type still showed more similarity than pairs of different cell types, even among mixed protocol pairs (HiCRep: $p = 3.436 \times 10^{-15}$, JI: $p = 1.2645 \times 10^{-9}$, TADsim: $p = 0.006010$). We observe a similar trend across all non-replicate pairs as well. Overall, we observe some structural differences induced by the protocol variations, but not enough to obscure the general similarities expected from samples of the same cell type.

Restriction enzyme choice has minimal impact on TAD sets

By comparing samples from the same lab of the same cell type, generated with a different restriction enzyme, we see no significant variation in similarity measures induced by restriction enzymes, as shown in Figure 3.5. As expected, the pairs of the same cell type with a different restriction enzyme tend to be more structurally similar than the background distribution, which includes all 2333 other pairwise comparisons. The choice of restriction enzyme does not appear to be a significant source of technical variation in measurements of chromosome structure, as both Hi-C matrices and TAD sets appear robust to this experimental variable.

3.3.5 TAD variation induced by lab-specific differences

Across all of our data, we see no pattern of elevated structural similarity among samples from the same lab (Figure 3.6**a**, JI and TADsim heat maps can be seen in Figures 3.7 and 3.8). A comparison of pairs of the same cell type from different labs shows that these pairs are generally more similar than non-replicate pairs, with similarity values near those of replicate pairs (Figures 3.6**b,c,d**). Consistent with the protocol-driven variation described above, the three lowest pairwise scores for IMR90 in both JI and TADsim are the three mixed protocol comparisons; all other points represent pairs generated by the same protocol. Chromosome structure seems to be robust to the variability across experimental labs.

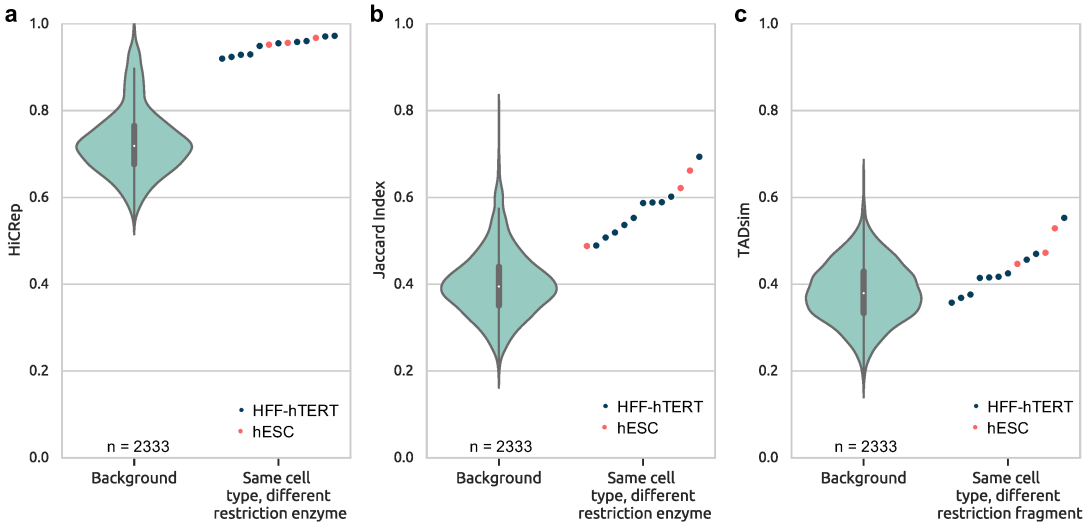


Figure 3.5: **Measurements of structural similarity across the use of different restriction enzymes.** The scattered points represent the similarity of a pair of samples of the same cell type (grey is hESC, red is HFF-hTERT), generated by using different restriction enzymes. The violin plot shows the distribution of all other non-replicate comparisons. As expected, the points that differ only in restriction enzyme are largely more similar than the background, suggesting that this choice is not a significant source of technical variability.

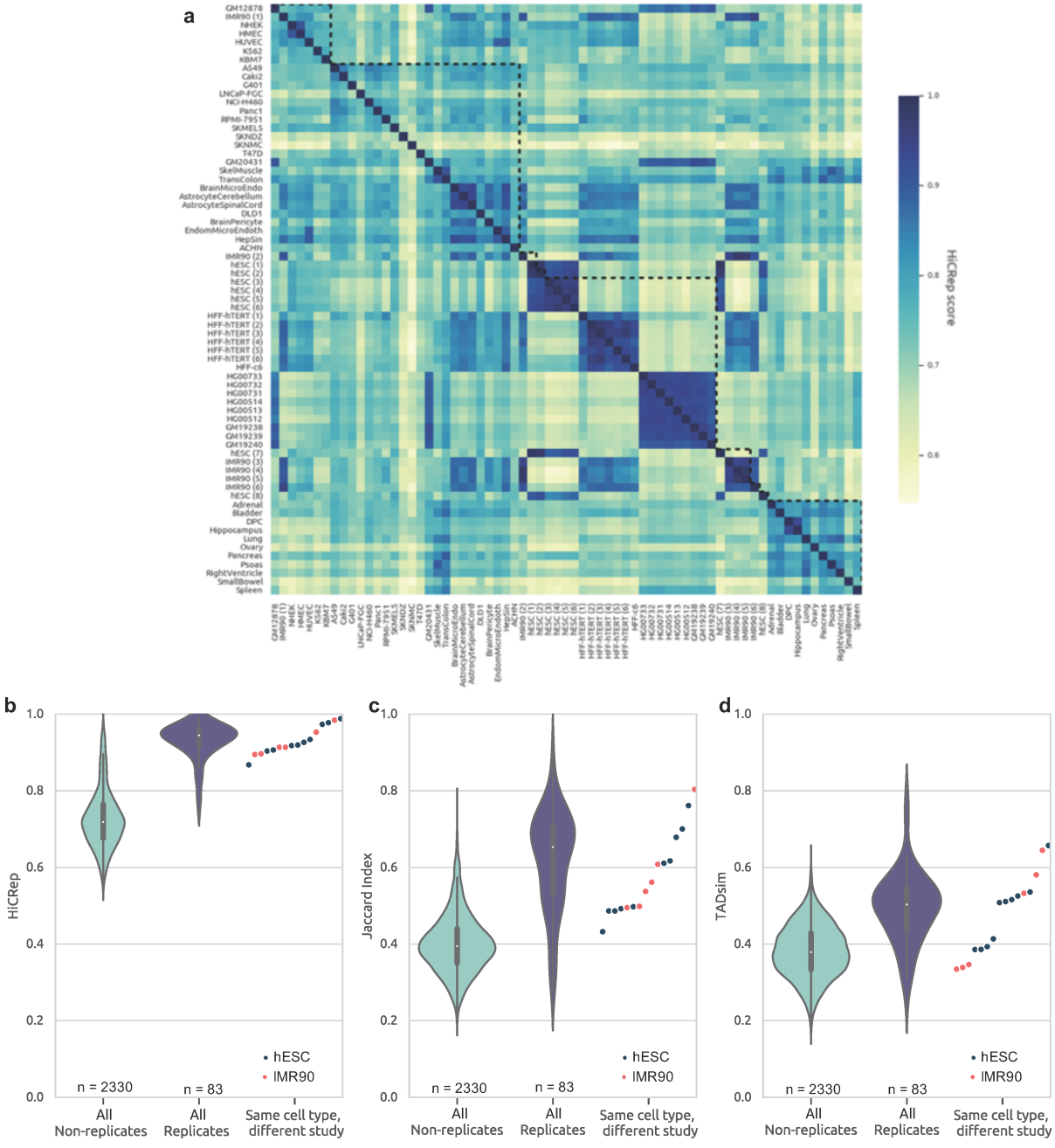

a
HiCRep score
b
HiCRep
c
Jaccard Index
d
TADsim
hESC
IMR90
n = 2330
n = 83
All
Non-replicates
All
Replicates
Same cell type,
different study

Figure 3.6: **Quantifying variation across samples from different labs. a** Summary of all 2346 pairwise sample comparisons as a heatmap of HiCRep scores. The dotted lines outline groups of samples from the same study. **b,c,d**: The effects of lab-specific variation on chromosome structure measurements. The points represent similarity scores of the same cell type (red for IMR90, grey for hESC) from different studies. These can be compared to the distribution of non-replicate pairs and that of replicate pairs, showing that samples from different labs achieve similarity values near those of replicate pairs.

3.3.6 Robustness to TAD size

While the exact similarity values differ somewhat, all trends observed in this work are consistent across TAD sets selected for median TAD sizes of 500kb, 700kb, 880kb, and 1Mb. The true expected size of TADs is fairly unclear, and likely to span a wide range due to their hierarchical nature [20, 42]. Though Armatus does not optimize for a specific TAD length, its resolution parameter γ adjusts a preference for larger or smaller TADs. Throughout this work the γ value was selected by choosing the set with median TAD length closest to 880kb. In order to assess robustness to this parameter, we additionally ran all analyses for TAD sets with median lengths of 500kb, 700kb, and 1Mb. Because HiCRep is performed on the full Hi-C matrix rather than TADs, only JI and TADsim were compared for robustness here. While the similarity values are generally lower for TADs of larger size (Figure 3.9), the trends across conditions compared here were robust (Figures 3.10, 3.11, 3.12, and 3.13).

3.4 Discussion

We have demonstrated that cell or tissue type, rather than individual or genetic difference, appears to be the greatest driver of biological variation in TAD structures and Hi-C matrices, confirming and quantifying the likely biological importance of TADs. However, between replicates,

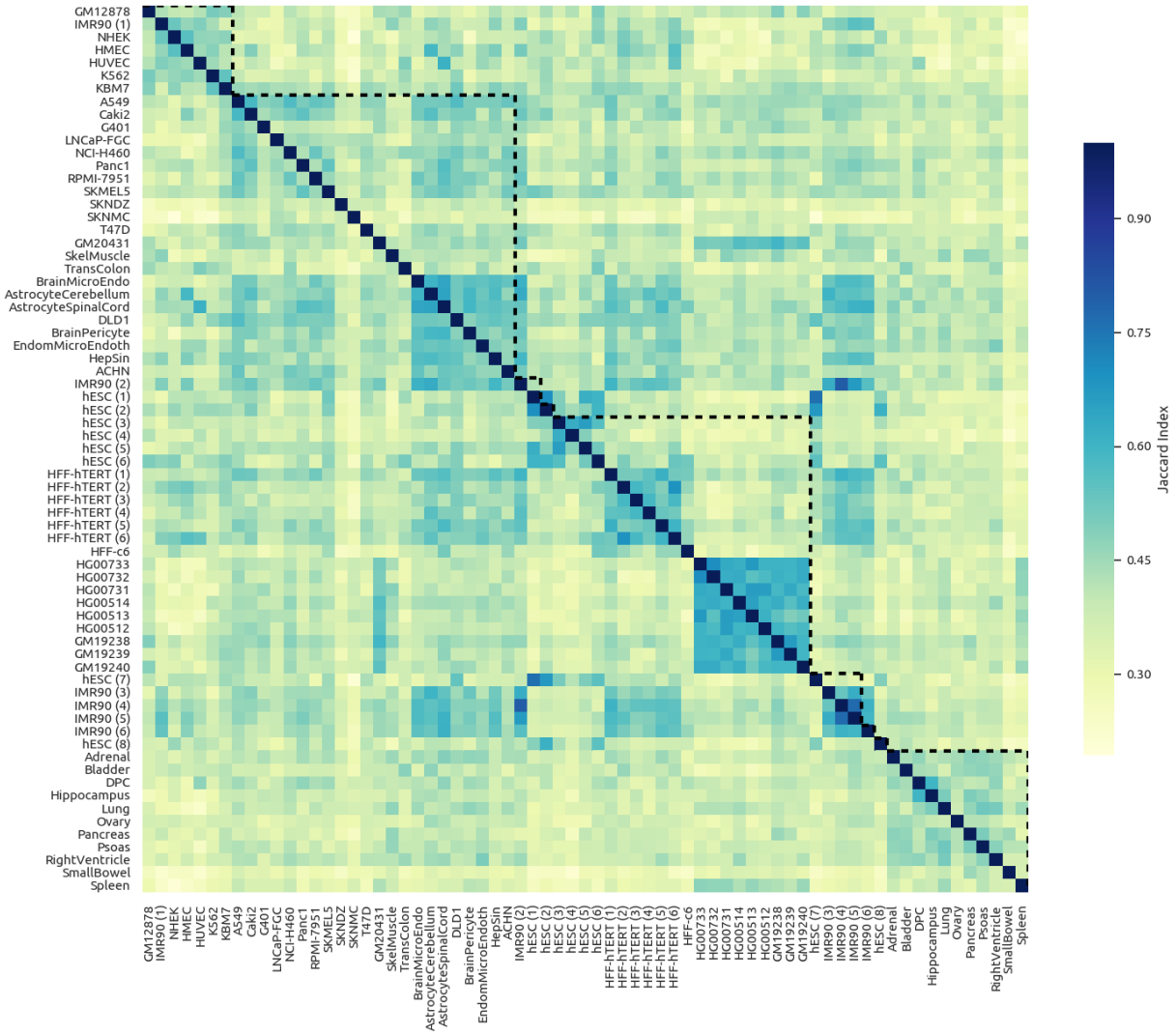


Figure 3.7: Summary of all 2346 pairwise sample comparisons as a heatmap of JI values. Dotted lines mark the samples that came from the same study. We see no systematic elevation in similarity values of intra-lab comparisons. This suggests that lab-specific variations do not significantly impact the similarities of TAD sets.

TAD structures are shown to share only 60% of their boundaries, suggesting that chromosome structure is not a static feature, but remains variable even in identical cell populations. Contrary to previous claims that TADs are highly conserved, we note significant TAD variability across human samples. We observe elevated similarities between samples of the same cell type, sug-

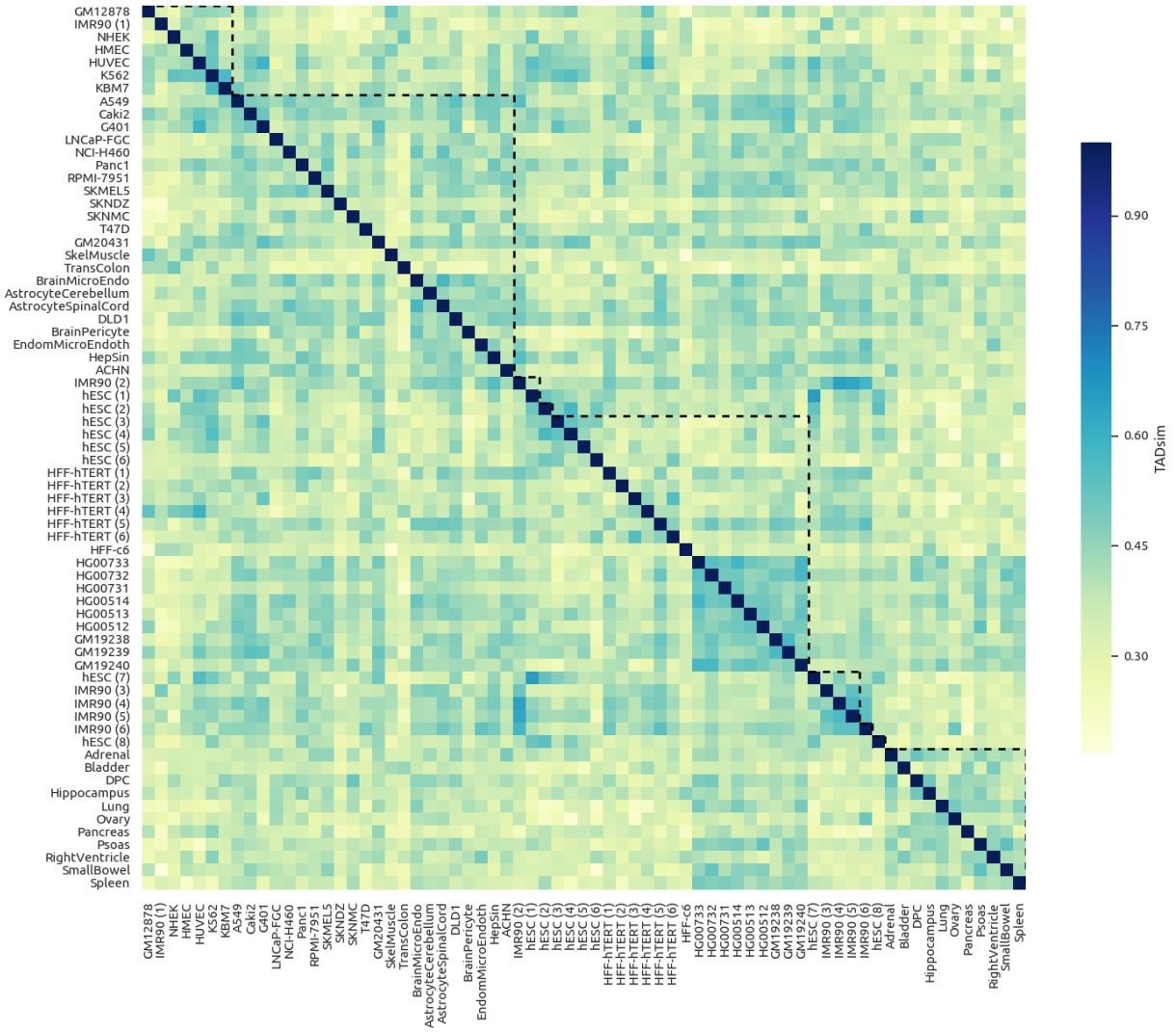


Figure 3.8: Summary of all 2346 pairwise sample comparisons as a heatmap of TADsim values. Dotted lines mark the samples that came from the same study. We see no systematic elevation in similarity values of intra-lab comparisons. This suggests that lab-specific variations do not significantly impact the similarities of TAD sets.

gesting that TAD structures are likely correlated with cellular function, rather than individual genetics. The largest differences due to technical variations appeared in comparing structures generated through *in situ* or dilution protocols, while lab-specific differences and restriction enzyme choices had a smaller impact on the resulting similarities of Hi-C measurements.

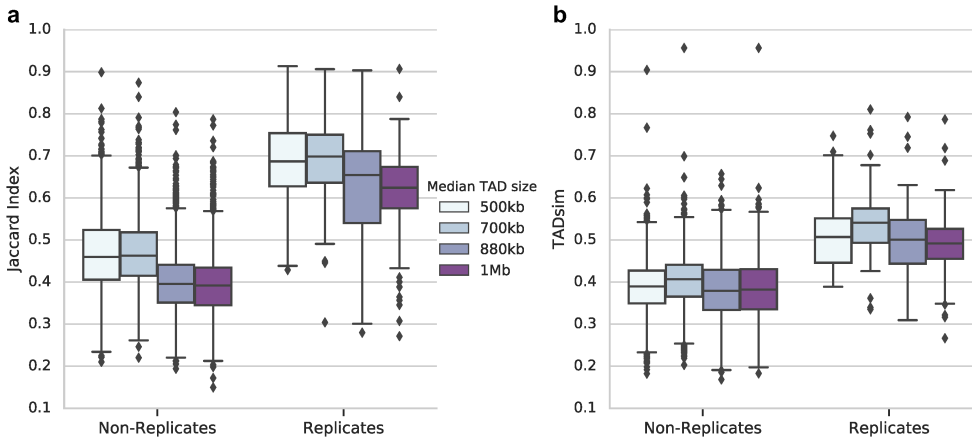


Figure 3.9: **Measuring robustness to TAD size parameter.** Boxplots, with midline representing the median, represent the distributions of JI (**a**) and TADsim (**b**) values for TAD sets selected for median TAD sizes of 500kb, 700kb, 880kb, and 1Mb. The 880kb distributions are also shown in Figure 3.1, and are included here for comparison.

In order to maximize the number of samples analyzed in this work, all comparisons were performed at a fairly low resolution of 100kb, so structural features that would be clearer at higher resolution may have been overlooked. A few observations noted in this work are consistent with previous smaller-scale studies of higher-resolution matrices. In particular, the similarities between replicates that we observed were consistent with those found in Dixon *et al*. [36] and Rao *et al*. [98], though much higher than those reported in Forcato *et al*. [44] due to the different pre-processing methods used in each study. The higher-than-random similarity between pairs of different tissue types was also found by Schmitt *et al*. [109], but our quantification of this similarity suggests significant variability rather than extensive conservation between tissue types.

There are still relatively few available Hi-C data sets compared with other genomic analyses, and many of the observations made here would be strengthened with more samples or with confirmation through single-cell Hi-C. In particular, more trios from other cell types would help to confirm whether there is truly no elevated similarity in genetically related individuals, or

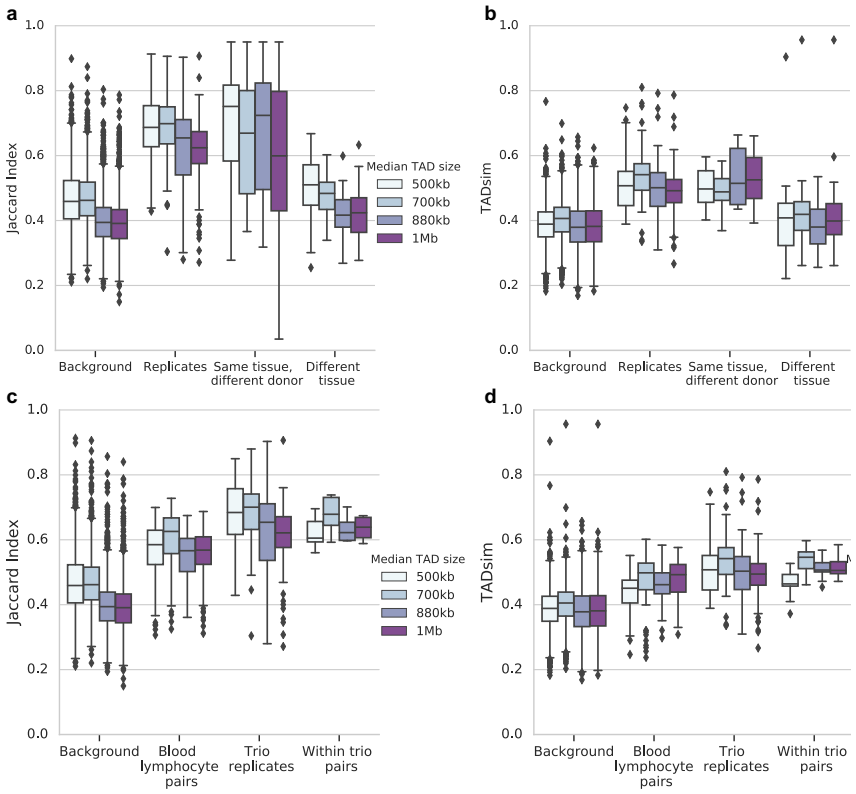


Figure 3.10: Robustness on tissue samples and parent-parent-child trios. These boxplots represent the distributions of JI (**a, c**) and TADsim (**b, d**) values for various median TAD sizes (500kb, 700kb, 880kb, and 1Mb). The 880kb distributions are also shown in Figure 3.3, and are included here for comparison. (**a,b**) The "Background" represents all pairs from our data with at least one cell line (as opposed to a tissue sample), and "Replicates" distributions are the values from all replicate pairs. The "Same tissue, different donor" values come from pairs of samples of the same tissue type collected from different individuals, and "Different tissue" represents all pairs of tissue comparisons from two different tissue types. (**c,d**) The "Background" distributions include all pairs with at least one non-blood lymphocyte sample. "Blood lymphocyte pairs" include all comparisons of two different blood lymphocyte samples that do not come from the trio data. "Trio replicates" are the replicate comparisons from each of the trio samples ($n = 9$). "Within trio pairs" represent comparisons of samples from the same trio (either two parents, or a parent and their child, $n = 9$).

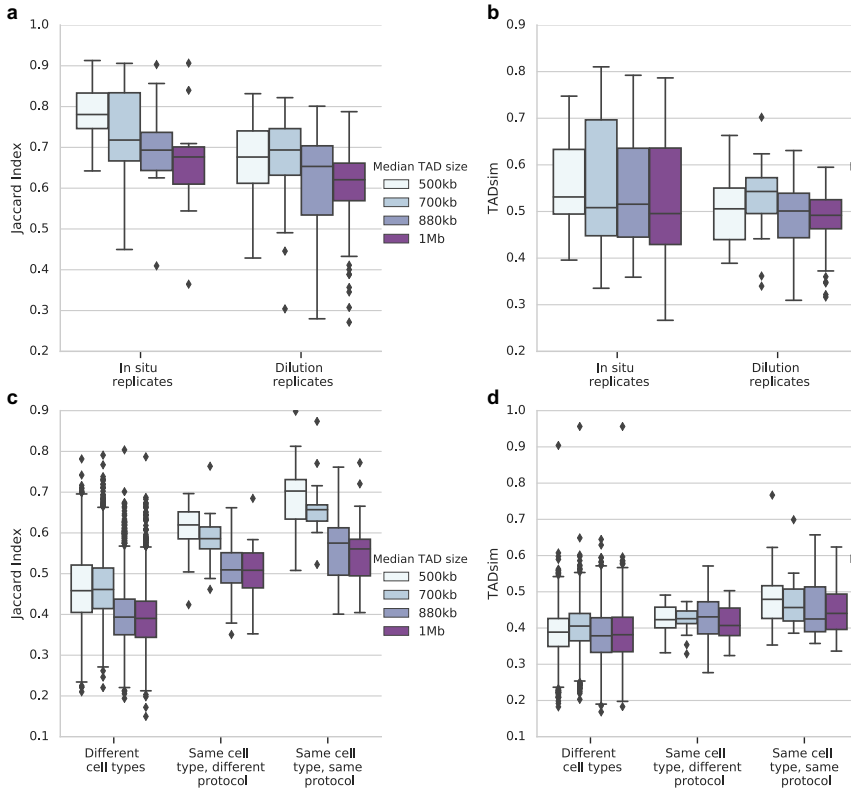


Figure 3.11: Robustness on comparing in situ and dilution Hi-C. These boxplots represent the distributions of JI (**a, c**) and TADsim (**b, d**) values for various median TAD sizes. The 880kb distributions are also shown in Figure 3.4, and are included here for comparison. (**a,b**) "In situ replicates" represents all replicate comparisons of samples collected by an *in situ* protocol, and "Dilution replicates" shows all replicate comparisons collected with a dilution protocol. (**c,d**) "Different cell types" represents all comparisons where the two samples compared are not of the same cell or tissue type. "Same cell type, different protocol" shows the similarity values of pairs in which both samples are the same cell type but one was generated with an *in situ* protocol, and one came from a dilution protocol. "Same cell type, same protocol" shows the similarities of pairs in which both samples come from the same cell type and both were generated by the same protocol (either both *in situ* or both dilution).

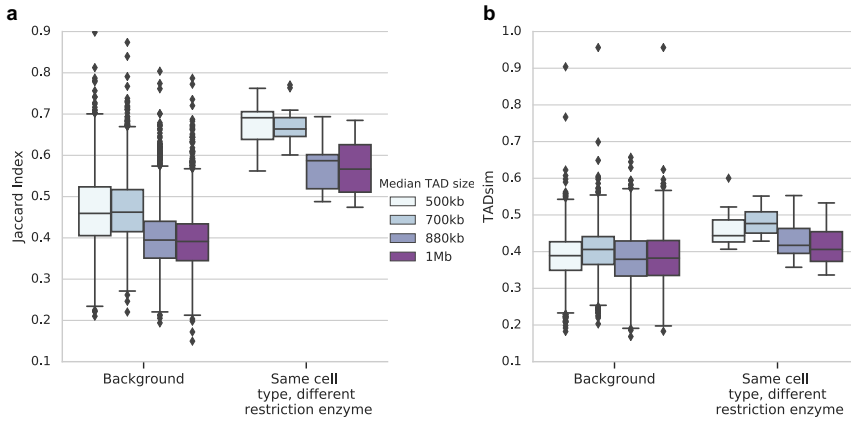


Figure 3.12: Robustness across restriction enzymes. The boxes outline the first to third quartiles of data, the midline represents the median, and outliers are shown as diamonds. These boxplots represent the distributions of JI (**a**) and TADsim (**b**) values for various median TAD sizes. The 880kb distributions are also shown in Figure 3.5, and are included here for comparison. The "Same cell type, different restriction enzyme" values represent similarities of pairs of samples from the same cell line (either HFF-hTERT or hESC), where different restriction enzymes were used during the Hi-C protocol ($n = 13$). The "Background" here represents all other non-replicate pairs.

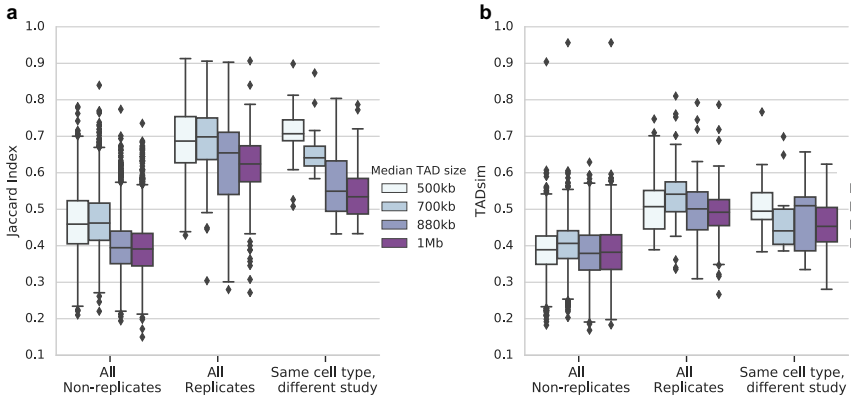


Figure 3.13: Robustness on lab-specific variation. The boxes outline the first to third quartiles of data, the midline represents the median, and outliers are shown as diamonds. These boxplots represent the distributions of JI (**a**) and TADsim (**b**) values for various median TAD sizes. The 880kb distributions are also shown in Figure 3.6, and are included here for comparison."Same cell type, different study" distributions show the similarities of pairs of samples of the same cell line (either IMR90 or hESC), which came from different studies ($n = 16$). "All replicates" distributions show the values from all replicate pairs in the data collected, and "All non-replicates" shows all non-replicate pairs which are not in the "Same cell type, different study" category.

whether this conclusion was specific to the blood lymphocyte samples studied in this work. As more single-cell Hi-C data becomes available through studies such as Flyamer et al. [43], Nagano et al. [80], Stevens et al. [121] and analysis methods improve, cell-to-cell variations in chromosome structure will be easier to assess, and we will be able to determine whether these population trends hold within individual cells.

All samples studied here were processed from sequencing reads to Hi-C matrices through the same pre-processing pipeline, and all TADs were computed using Armatus [42]. These choices may have influenced the trends we observed in this work, because different pre-processors, Hi-C normalization methods, or TAD callers could result in different patterns in the resulting structural measurements. All TAD sets represent some amount of uncertainty given the limited power of TAD callers, and our conclusions on TADs are only as strong as the reliability of the TAD sets they are based on. The consistencies with previous work using different methods for each of these steps suggests that they did not have a major effect, but more study is needed to assess the overall robustness of Hi-C measurements to these processing choices. Additionally, there may be other possible comparison methods for Hi-C matrices and TAD sets, which may or may not agree with the three measures used here.

Further study of the structural differences across cell types may lead to insights into the mechanisms of chromosome structure. These comparison techniques could also be used to determine the differences between chromosome structures in healthy and diseased cells and could point to the locations of structural changes that are present across diseased cells. There is already significant evidence of structural abnormalities in many diseases (review, [71]). Additional systematic, genome-wide analyses of TAD structures could increase our understanding of a range of human diseases. Here, we have taken the first step towards systematically quantifying, at a large scale, the extent of TAD structure variability.

This work compares Hi-C data and TAD structures from nine studies using three different measures, in order to identify trends in the variables controlling chromosomal structural similarity.

We observe that even replicates display a certain amount of variability in chromosome structure. Chromosome structure appears most conserved within cell types and tissue types and not influenced more strongly by genetic similarity or differences across individuals. Differences in the cross-linkage step of the Hi-C protocol can induce variation in the resulting Hi-C and TAD measurements, but they seem robust to both lab-specific differences and choice of restriction enzyme.

Cell type	Description	Replicates	Res frag	Protocol	Accession(s)	Citation
IMR90	lung fibroblast	2	MboI	*in situ*	SRR1658672, SRR1658673, SRR1658674, SRR1658675, SRR1658676, SRR1658677, SRR1658678	[98]
GM12878	blood lymphocyte	2	MboI	*in situ*	SRR1658570, SRR1658571, SRR1658572, SRR1658573, SRR1658574, SRR1658575, SRR1658576, SRR1658577, SRR1658578, SRR1658579, SRR1658580, SRR1658581, SRR1658582, SRR1658583, SRR1658584, SRR1658585, SRR1658586, SRR1658587, SRR1658588, SRR1658589, SRR1658590, SRR1658591, SRR1658592, SRR1658593, SRR1658594, SRR1658595, SRR1658596, SRR1658597, SRR1658598, SRR1658599, SRR1658600, SRR1658601, SRR1658602, SRR1658603	[98]
HMEC	mammary epithelial	2	MboI	*in situ*	SRR1658680, SRR1658681, SRR1658682, SRR1658683, SRR1658684, SRR1658685	[98]
HUVEC	umbilical vein endothelial	1	MboI	*in situ*	SRR1658709, SRR1658710, SRR1658711, SRR1658712, SRR1658713, SRR1658714	[98]
K562	chronic myeloid leukemia	2	MboI	*in situ*	SRR1658693, SRR1658694, SRR1658695, SRR1658696, SRR1658697, SRR1658698, SRR1658699, SRR1658700, SRR1658701, SRR1658702	[98]
KBM7	chronic myeloid leukemia	2	MboI	*in situ*	SRR1658703, SRR1658704, SRR1658705, SRR1658706, SRR1658707, SRR1658708	[98]
NHEK	epidermal keratinocyte	1	MboI	*in situ*	SRR1658689, SRR1658690, SRR1658691	[98]
A549	adenocarcinomic alveolar basal epithelial	2	HindIII	dilution	ENCLB571HTP, ENCLB222WYT	[125]
Caki2	clear cell renal cell carcinoma (epithelial)	2	HindIII	dilution	ENCLB555CZE, ENCLB858SVS	[125]
G401	rhabdoid tumor kidney epithelial	2	HindIII	dilution	ENCLB506SDM, ENCLB589RBY	[125]
LNCaP-FGC	prostate carcinoma epithelial-like	2	HindIII	dilution	ENCLB191OGC, ENCLB473XWD	[125]
NCI-H460	large cell lung cancer	2	HindIII	dilution	ENCLB118KAE, ENCLB104ZTM	[125]
Panc1	pancreas ductal adenocarcinoma	2	HindIII	dilution	ENCLB951HSJ, ENCLB134IVX	[125]
RPMI-7951	malignant melanoma	2	HindIII	dilution	ENCLB210AAY, ENCLB016TGU	[125]
SKMEL5	malignant melanoma	2	HindIII	dilution	ENCLB296ZFT, ENCLB462TWE	[125]
SKNDZ	neuroblastoma	2	HindIII	dilution	ENCLB524GGK, ENCLB952BSP	[125]
SKNMC	neuroepithelioma	2	HindIII	dilution	ENCLB215KZO, ENCLB914GYK	[125]
T47D	ductal carcinoma	2	HindIII	dilution	ENCLB758KFU, ENCLB183QHG	[125]
IMR90	lung fibroblast	2	HindIII	dilution	SRX116345, SRX128222	[36]
hESC	human embryonic stem cell	2	HindIII	dilution	SRX116344, SRX128221	[36]
H1-hESC	human embryonic stem cell	1	NcoI	*in situ*	4DNES4DGHDMX	[32]
H1-hESC	human embryonic stem cell	3	DpnII	*in situ*	4DNESRJ8KV4Q	[32]
H1-hESC	human embryonic stem cell	1	HindIII	dilution	4DNES78Y8Y5K	[32]
H1-hESC	human embryonic stem cell	2	DpnII	*in situ*	4DNES2M5JIGV	[32]
HFF-hTERT	foreskin fibroblast	4	HindIII	dilution	4DNES9L4AK6Q	[32]
HFF-hTERT	foreskin fibroblast	2	DpnII	*in situ*	4DNESVUMGLG2	[32]
HFF-hTERT	foreskin fibroblast	1	NcoI	*in situ*	4DNESY859VLG	[32]
HFF-hTERT	foreskin fibroblast	2	HindIII	*in situ*	4DNESB6MNCFE	[32]
HFF-hTERT	foreskin fibroblast	1	HindIII	*in situ*	4DNES8J78WV2	[32]
HFF-hTERT	foreskin fibroblast	1	MboI	*in situ*	4DNESAPF27TG	[32]
HFFc6	subclone of HFF-hTERT	2	DpnII	*in situ*	4DNES2R6PUEK	[32]
HG00733	blood lymphocyte	2	HindIII	dilution	4DNESTAPSPUC	[32]
HG00732	blood lymphocyte	2	HindIII	dilution	4DNESI2UKI7P	[32]
HG00731	blood lymphocyte	2	HindIII	dilution	4DNESJ1VX52C	[32]
HG00514	blood lymphocyte	2	HindIII	dilution	4DNESE3ICNE1	[32]
HG00513	blood lymphocyte	2	HindIII	dilution	4DNESJIYRA44	[32]
HG00512	blood lymphocyte	2	HindIII	dilution	4DNES4GSP9S4	[32]
GM19238	blood lymphocyte	2	HindIII	dilution	4DNESYUYFD6H	[32]
GM19239	blood lymphocyte	2	HindIII	dilution	4DNESVKLYDOH	[32]
GM19240	blood lymphocyte	2	HindIII	dilution	4DNESHGL976U	[32]
hESC	human embryonic stem cell	1	HindIII	dilution	SRR639047, SRR639048, SRR639049	[60]
IMR90	lung fibroblast	6	HindIII	dilution	SRX212172, SRX212173, SRX294948, SRX294949, SRX294950, SRX294951	[60]
IMR90	lung fibroblast	6	HindIII	dilution	SRX212174, SRX212175, SRX294952, SRX294953, SRX294954, SRX294955	[60]
IMR90	lung fibroblast	1	HindIII	dilution	SRR639045, SRR639046	[60]
hESC	human embryonic stem cell	2	HindIII	dilution	SRX378271, SRX378272	[37]
GM20431	blood lymphocyte	3	HindIII	dilution	ENCLB097VEW, ENCLB167NGL, ENCLB938LSX	[125]
skeletal muscle tissue	gastrocnemius medialis, 4 donors	4	MboI	*in situ*	ENCLB925XYW, ENCLB361HQM, ENCLB966EDS, ENCLB645GUM	[125]
transverse colon	from 4 donors	4	MboI	*in situ*	ENCLB584CUK, ENCLB920LTI, ENCLB724QSQ, ENCLB527HSP	[125]
brain microvascular	endotheilial	2	HindIII	dilution	SRX3322341, SRX3322340	[125]
astrocyte	cerebellum	2	HindIII	dilution	ENCLB672PAB, ENCLB174TEA	[125]
astrocyte	spinal cord	2	HindIII	dilution	SRX3322978, SRX3322979	[125]
DLD1	colon adenocarcinoma epithelial	2	HindIII	dilution	SRX3321987, SRX3321988	[125]
pericyte	brain	2	HindIII	dilution	SRX3322286, SRX3322287	[125]
HEMEC	endometrial microvascular endothelial	2	HindIII	dilution	SRX3322599, SRX3322600	[125]
hepatic sinusoid	endothelial	2	HindIII	dilution	ENCLB284TIY, ENCLB618NVM	[125]
ACHN	renal cell adenocarcinoma epithelial	2	HindIII	dilution	SRX3322373, SRX3322374	[125]
IMR90	lung fibroblast	2	HindIII	dilution	GSM2595584, GSM2595585	[138]
hESC (H9)	human embryonic stem cell	1	HindIII	dilution	GSM2309023	[47]
adrenal gland	tissue	1	HindIII	dilution	SRX2179246	[109]
bladder	tissue	2	HindIII	dilution	SRX2179247, SRX2179248	[109]
DPC	dorsolateral prefrontal cortex tissue	1	HindIII	dilution	SRX2179249	[109]
hippocampus	tissue	1	HindIII	dilution	SRX2179250	[109]
lung	tissue from 2 donors	2	HindIII	dilution	SRX2179252, SRX2179251	[109]
ovary	tissue	1	HindIII	dilution	SRX2179253	[109]
pancreas	tissue from 2 donors	2	HindIII	dilution	SRX2179254, SRX2179255, SRX2179256, SRX2179257	[109]
psoas muscle	tissue from 2 donors	2	HindIII	dilution	SRX2179260, SRX2179258, SRX2179259	[109]
right ventricle	tissue	1	HindIII	dilution	SRX2179261	[109]
small bowel	tissue	1	HindIII	dilution	SRX2179262	[109]
spleen	tissue from 2 donors	2	HindIII	dilution	SRX2179264, SRX2179263	[109]

Table 3.1: All Hi-C data used in this study

Cell type	Description	Replicates	Res frag	Protocol	Accession(s)	Citation
SJCRH30	rhabdomyosarcoma fibroblast	2	HindIII	dilution	ENCLB379VAF, ENCLB821TDJ	[125]
GM06990	blood lymphocyte	1	HindIII	dilution	SRR027956, SRR027957, SRR027958, SRR027959	[69]
K562	chronic myeloid leukemia	1	HindIII	dilution	SRR027962, SRR027963	[69]
HeLa-S3	cervix adenocarcinoma epithelial	2	HindIII	dilution	ENCLB693EVR, ENCLB696DUT	[125]
HepG2	hepatocellular carcinoma epithelial	2	HindIII	dilution	ENCLB022KPF, ENCLB625TGE	[125]
hESC	human embryonic stem cell	2	HindIII	*in situ*	GSE70181	[81]
hESC	human embryonic stem cell	2	HindIII	dilution	GSE70181	[81]

Table 3.2: Hi-C samples that could not be analyzed at 100kb resolution

Chapter 4

Chromosome dynamics revealed by an elastic network model

Hi-C technology has permitted the study of 3D genome organization, but provides only a static picture of a dynamic system, and we still lack an understanding of the structural dynamics of chromosomes. The dynamic couplings between regions separated by large genomic distances ($>$ 50 megabases) have yet to be characterized. This chapter describes an adaptation of a well-established protein-modeling framework, the Gaussian Network Model (GNM), to model chromatin dynamics using Hi-C data. We show that the GNM can identify spatial couplings at multiple scales: it can quantify the correlated fluctuations in the positions of gene loci, find large genomic compartments and smaller topologically-associating domains (TADs) that undergo en-bloc movements, and identify dynamically coupled distal regions along the chromosomes. We show that the predictions of the GNM correlate well with genome-wide experimental measurements. We use the GNM to identify novel cross-correlated distal domains (CCDDs) representing pairs of regions distinguished by their long-range dynamic coupling and show that CCDDs are associated with increased gene co-expression. Together, these results show that GNM provides a mathematically well-founded unified framework for modeling chromatin dynamics and assessing the structural basis of genome-wide observations.

A version of this chapter was published in *Nucleic Acids Research* and is joint work with She Zhang (co-first author), Carl Kingsford, and Ivet Bahar [107].

4.1 Background

The spatial arrangement of chromosomes within the nucleus plays a crucial role in gene regulation, cell replication and mutations [18, 23, 46, 55, 114]. Recent experimental methods such as Hi-C [69] derived from chromosome conformation capture (3C) [31] have made it possible to characterize the physical contacts between gene loci on a genome-wide scale. These studies revealed hierarchical levels of organization, from large (so called "A" and "B") compartments corresponding to active and inactive chromatin respectively [69], to smaller compact regions called topologically associated domains (TADs) [36]. Hi-C-measured spatial relationships have been related to chromosomal alterations in cancer [48] and TADs have been pointed out to contain clusters of genes that are co-regulated [83]. Interactions between sequentially (but not necessarily spatially) distant genes along the DNA 1-dimensional (1D) structure, termed long-range interactions, have been implicated in gene regulation – for example, distal expression quantitative trait loci (eQTLs) tend to be much closer in 3D space [40] to their target genes than expected by chance.

Several computational methods have contributed to these and other characterizations of chromosomal architecture [36, 42, 67, 98, 102, 115, 129, 134]. However, chromosome structure is dynamic and complex, and its exact nature and influence on gene expression and regulation remain unclear. The scale, complexity, and noise inherent in the available data make it challenging to determine exact spatial relationships and underlying chromatin architecture, and its structure-based dynamics. In particular, long-range spatial interactions have proven difficult to characterize with Hi-C data, and most computational analyses attempt to identify a static chromosomal architecture despite its known dynamic nature. There have also been efforts to mathematically characterize the dynamics of the genome separate from its structure, particularly through describing the emer-

gence of cell types during development as bifurcations from a stable equilibrium [95].

Chromatin structure is often described in terms of TADs, whose identification is a 1D problem: it involves searching for sequentially contiguous groups of highly interconnected loci along the diagonal of the Hi-C matrix of intra-chromosomal contacts. Spatial couplings between sequentially distant genomic regions, on the other hand, represent a new dimension to search and the identification of such long-range couplings is a more challenging problem. Several methods have sought to identify long-range interactions from 3C-based data [60, 98, 103, 105, 130], but the scale of these interactions is still small compared to that of the full chromosome. Most methods detect interactions within 1-2 Mbp, or up to 10Mbp [7], so extending the span of predicted long-range couplings to the order of tens of millions of base pairs may yield further insights into regulatory actions. Such long-range correlations may originate from physical proximity in space, or other indirect effects similar to those in allosteric structures. Assessment of such long-range correlations is important for gaining a better understanding of the physical basis of gene expression and regulation.

We adopt here the Gaussian Network Model (GNM), a highly robust and widely tested framework developed for modeling the intrinsic dynamics of biomolecular systems [9, 11, 51], and we adapt it to the topology-based modeling of chromosomal dynamics. Chromosomal dynamics refers to the coupled spatial movements of loci under equilibrium conditions, as uniquely defined by the topology of an elastic network representative of the chromosome architecture. The only input GNM requires is a map of 3D contacts. Here, this information is provided by Hi-C data, which gives contact frequencies between genomic loci. The Hi-C matrix is used for constructing the Kirchhoff (or Laplacian) matrix Γ which uniquely defines the equilibrium dynamics of the network nodes (genomic loci) as well as their spatial cross-correlations. Notably, the use of Laplacian-based graph segmentation has been recently shown to help identify topological domains from Hi-C data [25, 26]. Our approach differs in the method of construction of the network topology embodied in Γ, the inclusion of the complete spectrum of motions, and the application

to a broad range of observables. We show, and verify upon comparison with an array of experimental data and genome-wide statistical analyses, that the GNM provides a robust description of accessibility to the nuclear environment as well as co-expression patterns between gene-loci pairs separated by tens of megabases. The analysis is mathematically rigorous, efficient, and extensible, and may serve as a framework for drawing inferences from Hi-C and other advanced genome-wide studies toward establishing the structural and dynamic bases of regulation.

4.2 Materials and methods

4.2.1 Extension of the Gaussian Network Model to modeling chromatin dynamics

The GNM has proven to be a powerful tool for efficiently predicting the equilibrium dynamics of almost all proteins and their complexes/assemblies which can be accessed in the Protein Data Bank (PDB) [68], and has been incorporated into widely used molecular simulation tools such as CHARMM [21]. It is particularly adept at predicting topology-dependent dynamics and identifying long-range correlations – the type of modeling that has been a challenge in chromatin 3D modeling studies. Hi-C matrices, in which each entry represents the frequency of contacts between pairs of genomic loci, can be interpreted as chromosomal contact maps similar to those between residues adopted in the GNM representation of proteins.

There are several differences between the Hi-C and GNM Γ matrices. The first is the size: human chromosomes range from $\approx$ 50 to 250 million base pairs. When binned at 5kb resolution, this leads to $10,000 - 50,000$ bins per chromosome. GNM provides a scalable framework, where the collective dynamics of supramolecular systems represented by 104-105 nodes (such as the ribosome or viruses) can be efficiently characterized. GNM may therefore be readily used for analyzing intrachromosomal contact maps at high resolution. The second is the precision of the data. Experimental methods for resolving biomolecular structures such as X-ray crystallography,

NMR, and even cryo-electron microscopy yield structural data at a much higher resolution than current genome-wide studies. The Hi-C method is population-based (derived from hundreds of thousands to millions of cells) and noisy. Hi-C matrices furthermore contain unmapped regions. However, the GNM results are usually robust to variations in the precision/resolution of the data on a local scale, and require information on only the overall contact topology rather than detailed spatial coordinates, which supports the utility of Hi-C data and applicability of the GNM. Third, the chromatin is likely to be less "structured" than the structures at the molecular level, and it is likely to sample an ensemble of conformations that may be cell- or context-dependent. Single-cell Hi-C experiments have indicated cell-cell variability in chromosome structure on a global scale, though the domain organization at the megabase scale is largely conserved [80]. Therefore, structure-based dynamic features may be assessed at best at a probabilistic level. With these approximations in mind, we now proceed to the extension of GNM to characterize chromosomal dynamics (see Figure 4.1).

The GNM describes the structure as a network of beads/nodes connected by elastic springs. The network topology is defined by the Kirchhoff matrix Γ, whose elements are

$$\Gamma_{ij} = \begin{cases} -\gamma_{ij} & \text{for } r_{ij} < r_{cut} \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

$$\Gamma_{ii} = -\sum_{j, j \neq i} \gamma_{ij}.$$

Here γ_{ij} represents the strength or stiffness of interaction between beads i and j (or the force constant associated with the spring that connects them), r_{ij} is their separation in the 3D structure, and r_{cut} is the distance limit for making contacts (or for being connected by a spring). In the application to proteins, the beads represent the individual amino acids (n of them), their positions are identified with those of the α-carbons, and a uniform force-constant $\gamma_{ij} = \gamma$ is adopted for all pairs ($1 \leq i, j \leq n$), with a cutoff distance of $r_{cut} \approx 7$Å. In the extension to human chromosomes, we redefine the network nodes and springs such that beads represent genomic

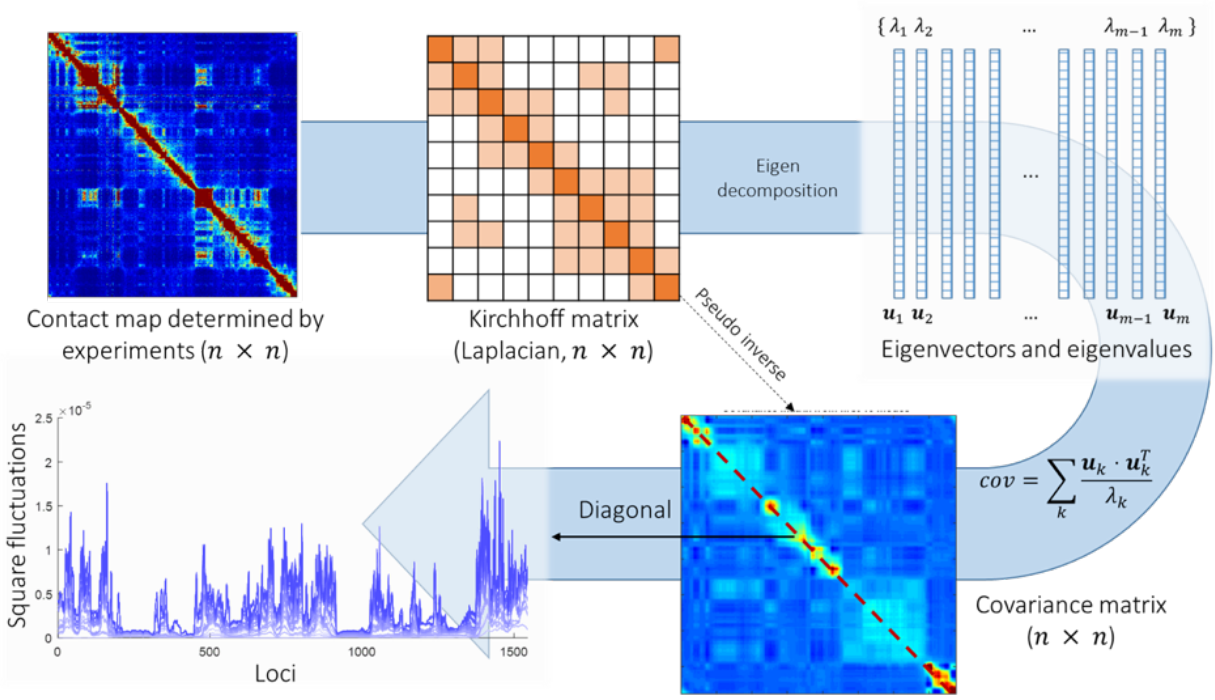


Figure 4.1: **Schematic description of GNM methodology applied to Hi-C data.** The inter-loci contact data represented by the Hi-C map (upper left, for n genomic bins (loci)) is used to construct the GNM Kirchhoff matrix, Γ (top, middle). Eigenvalue decomposition of Γ yields a series of eigenmodes which are used for computing the covariance matrix (lower, right), the diagonal elements of which reflect the mobility profile of the loci (bottom, left), and the off-diagonal elements provide information on locus-locus spatial cross-correlations. $\vec{u}_k$, k^{th} eigenvector; γ_k, k^{th} eigenvalue; m, number of nonzero modes, starting from the lowest-frequency mode, included in the GNM analysis ($m \leq n-1$). In the present application to the chromosomes, n varies in the range $10,248 \leq n \leq 49,850$, the lower and upper limits corresponding respectively to the respective chromosomes 22 and 1.

loci consistent with the resolution of the Hi-C data. We set γ_{ij} equal to γz_{ij} where z_{ij} is the Hi-C contact counts reported for the pair of genomic bins i and j after normalization by vanilla coverage (VC) method [98], and γ is taken as unity. The element Γ_{ij} is thus taken to be directly proportional to the actual number of physical contacts between the loci i and j, which permits us to directly incorporate the strength of interactions in the network model. The parameter γ uniformly scales all elements, physically representing the strength (or spring constant) of individual contacts. A recent study normalized the diagonal elements of the Laplacian matrix (constructed using Hi-C contact counts, similarly to Γ) after construction [25, 26], but we choose not to, because it removes the information on packing density of nodes, renders the calculation of square fluctuations meaningless, and disables the comparison with chromatin accessibility.

The cross-correlation between the spatial displacements of loci i and j is obtained from the pseudoinverse of Γ, as

$$< \Delta r_i \cdot \Delta r_j > \approx [\Gamma^{-1}]_{ij} = \sum_{k=1}^{n-1} \frac{1}{\lambda_k} [\vec{u}_k \vec{u}_k^T]_{ij}, \tag{4.2}$$

where the summation is performed over all modes of motion intrinsically accessible to the network, obtained by eigenvalue decomposition of Γ. The respective frequencies and shapes of these modes are given by the $n-1$ non-zero eigenvalues (λ_k) and corresponding eigenvectors ($\vec{u}_k$) of Γ, and $[\vec{u}_k \vec{u}_k^T]_{ij}$ designates the ij^{th} element of the matrix enclosed in square brackets. The eigenvector $\vec{u}_k$ is an n-dimensional vector representing the normalized displacements of the n loci along the k^{th} mode axis, and $1/\lambda_k$ rescales the amplitude of the motion along this mode. Lower frequency modes (smaller λ_k) make higher contributions to observed fluctuations and correlations; they usually embody large substructures, if not the entire structure, hence their designation as global modes. In contrast, high frequency modes are highly localized, and often filtered out to better visualize cooperative events represented by global modes.

Cross-correlations are organized in the nn covariance matrix, C (and displayed by an $n \times n$ map). The i^{th} diagonal element of C, $\langle (\Delta r_i)^2 \rangle$, is the predicted mean-square fluctuation (MSF) in the positions of the i^{th} loci under physiological conditions; and $\langle (\Delta r_i)^2 \rangle$ plotted as a function of

locus index i is called the mobility profile. The MSFs are inversely proportional to the elastic spring constant γ. While their absolute values uniformly depend on this parameter, their relative magnitudes do not; the MSF profiles thus provide a measure of the relative size of motions of the different gene loci (irrespective of γ), exclusively defined by the particular loci-loci contact topology. They represent ensemble averages over all accessible motions to a given locus.

4.2.2 Removal of unmapped regions

In the Hi-C map there are regions where no cross-linked DNA fragments can be mapped. These unmapped regions are isolated from the system, and their existence may lead to multiple zero-eigenvalue modes. These unmapped regions are not constrained by other loci, so they may cause large fluctuations that obscure the signal from other regions. These extra zero-eigenvalue modes and unphysically large fluctuations were removed by discarding the unmapped regions. The removal of the unmapped regions will not cause disconnections because the chromosomes are highly compact, so the loci next to the unmapped regions remained connected to the loci located at the other end of the region.

4.2.3 Data

We used high-resolution Hi-C data from Rao et al. [98] (GEO accession GSE63525), pre-processed using vanilla coverage (VC) normalization [98]. We used Hi-C data at 5 kb resolution unless otherwise noted. DNase-seq data were collected as part of the ENCODE project (ENCFF000SKV for GM12878 cells, ENCFF740JVK for IMR90 cells) [93]. The ATAC-seq measurements [22] were also obtained for GM12878 and IMR90 cells (GEO accessions GSM1155959 and GSM1418975, respectively). For both of these experimental datasets, bed-formatted peak files were downloaded from the study authors and the data was binned to the same resolution as the Hi-C data by adding all peak values within each bin. The binned data were then smoothed using moving average with a window size of 200kb. The long-range inter-

actions from ChIA-PET were from ENCODE (ENCFF002EMO) [53]. We used a two-sample t-test assuming unequal variances to quantify the difference between the covariance distributions of ChIA-PET and background interactions.

4.2.4 Hi-C data normalization

We tested three types of normalization methods applied to the Hi-C contact map: Vanilla-Coverage normalization (referred to as VCnorm), square-root Vanilla-Coverage normalization (referred to as sqrtVC) [69] and Knight-Ruiz normalization (referred to as KRnorm) [63]. All three methods aim to eliminate the so-called "one-dimension bias" [98]. We found that the GNM performed best on Hi-C maps normalized by VCnorm when benchmarked against experimental data (Figure 4.2). Not only are the correlations with the chromatin accessibility lower, but also the square fluctuations become flatter and flatter by adding more modes in the calculation when KRnorm or sqrtVC has been applied on the contact map. In the extreme case, when all the modes are used, the square fluctuations become almost completely flat along the chromosome using KRnorm. This is because KRnorm ensures that every row and column sums to 1. As a consequence, all loci become almost equally constrained and the differences in their square fluctuations are suppressed. In addition, computations with the three normalization methods were repeated at different resolutions, and VCnorm yielded the most robust agreement between theoretically predicted MSFs and experimentally observed accessibilities across all resolutions. Both KRnorm and sqrtVC showed poor correlations at high resolution (5kb) (Figure 4.3). Furthermore, VCnorm showed the expected improvement in correlation using increasing number of modes included in the analysis, while KRnorm or sqrtVC led to inconsistent results, even at 50kb resolution (Figure 4.3). Due to the better performance across resolutions and numbers of modes, shown by agreement with experimental data, we chose VC normalized contact maps to perform further analyses.

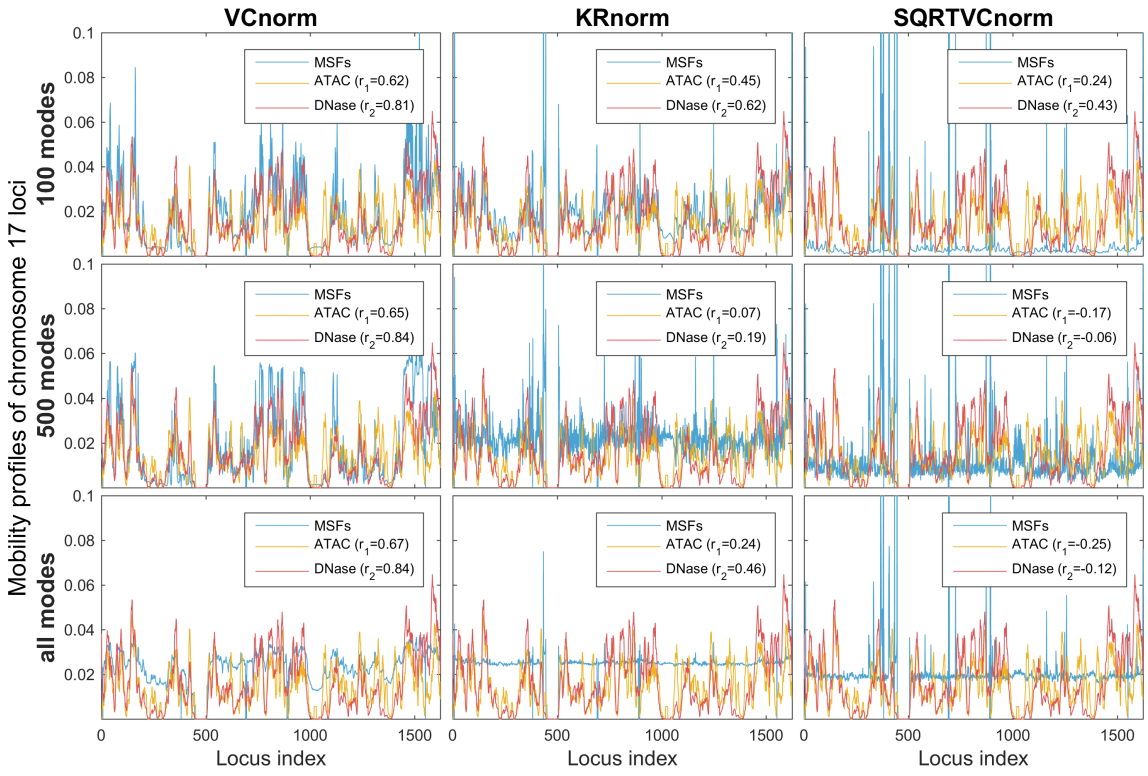


Figure 4.2: Comparison of the MSFs obtained from different number of GNM modes (rows), and three different normalization methods (columns): Vanilla Coverage normalization (left), Knight-Ruiz normalization (middle), and square root Vanilla Coverage normalization (right). MSFs in this figure are calculated from Hi-C data at 50kb resolution for GM1287 chromosome 17.

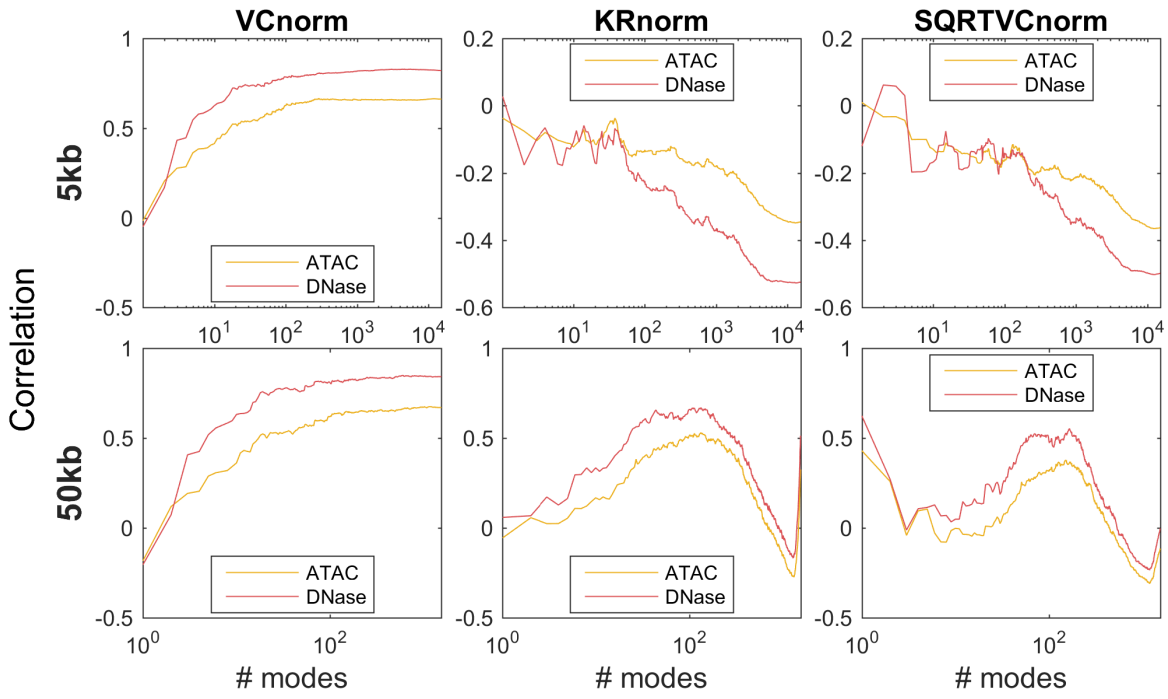


Figure 4.3: The scanning of correlations between chromatin accessibility data from experiments and square fluctuations from theory calculated as a function of the number of modes included in the GNM analysis. The rows compare the correlations at different resolutions, and the columns compare those computed from three different normalization methods. Note the poor performance of KRnorm and SQRTVCnorm.

4.2.5 GNM domain identification

GNM domain boundaries for a given mode k are identified by plotting the elements of $\vec{u}_k$ as a function of loci index, and identifying the crossover region (also referred to as hinge regions), from positive to negative direction motion (or vice versa), along the mode axis. To reduce minor fluctuations in the eigenvectors that could lead to spurious small domains, we first smoothed the eigenvectors with local regression using weighted linear least squares and a first-degree polynomial model. The smoothing window was chosen to be the smallest value that minimizes the number of domains of length one, where a domain of length one is defined as a domain that begins and ends in the same bin. In general, the size of the domains decreases with increasing mode number. The domains resulting from the superposition of multiple modes were delimited by the union of hinge sites.

4.2.6 Variation of Information metric

As a quantitative measure of agreement between GNM-predicted domains, TADs, and compartments, the variation of information (VI) metric was used. This metric is based in information theory, and measures the difference in information contained in two clusterings, or partitions, of a data set. If we consider each domain to be a cluster of nodes/points, this type of comparison becomes very natural. Formally, for two sets of clusters C and C', VI is defined as follows:

$$VI(C, C') = H(C) + H(C') - 2I(C, C'),$$

where $H(C)$ represents the entropy of a set of clusters C, and $I(C, C')$ is the mutual information between the two partitions, given by

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k),$$

$$I(C, C') = -\sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P(k')},$$

where the probability of picking a node in cluster C_k, $P(k)$, is simply the number of points in that cluster divided by the total number of points in the data set. In this work, a "cluster" is the set of loci placed into the same domain or compartment. This is a true metric on the space of clusterings; VI is commutative, satisfies the triangle inequality, and is always non-negative and equal to zero if and only if the two clusterings are identical. More intuitively, VI is a measure of the amount of information that is lost and gained by changing from one clustering to another, without any assumptions placed on the clusterings themselves or how they were generated. More information can be found in Meilă [76].

4.2.7 Co-expression calculation

In order to calculate co-expression values for genes in this cell type, we downloaded every publicly available RNA-seq experiment on GM12878 cells from the Sequence Read Archive [64], which gave 212 data sets. These raw read data were quantified using Salmon [90], resulting in 212 transcripts per kilobase million (TPM) values for every gene. Quantification was performed with and without bias correction, with qualitatively similar results. Co-expression was then measured as the Pearson correlation of the two vectors of TPM values for a given gene pair.

4.3 Results

4.3.1 Loci dynamics correlate well with experimental measures of chromatin accessibility

We first evaluated the mobility profiles of the chromosomes for GM12878 cells, a human lympho-blastoid cell line with relatively normal karyotype, and IMR90 cells, a human lung fibroblast cell line. Figures 4.4 and 4.5 illustrate the MSFs obtained with the GNM (blue curves) for the loci on three chromosomes (1, 15 and 17, in respective panels A, B and C) of the two different cell lines. GNM application to H/D exchange data has shown that the MSFs of network

nodes can be directly related to the accessibility of the corresponding sites: exposed sites enjoy higher mobility, and those buried have suppressed mobilities [10]. The entropic cost of exposure to the environment for a given site can be shown to be inversely proportional to its MSFs based on simple thermodynamic arguments applied to macromolecules subject to Gaussian fluctuations (such as those represented by the GNM) [10]. We examined whether GNM-predicted mobility profiles were also consistent with data from chromatin accessibility experiments. We compared our predictions with two measures of chromatin accessibility, DNase-seq [118] and ATAC-seq [22], shown respectively by the yellow and red curves in Figures 4.4 A-C and 4.5 A-C.

Figures 4.4 and 4.5 show that the MSFs of chromosomal loci, predicted by the GNM, are in very good agreement with the accessibility of loci as measured by DNase-seq. For GM12878, the corresponding Spearman correlations for the three chromosomes illustrated in panels A-C vary in the range 0.78-0.85 (see inset), and the computations for all 23 chromosomes (panel D, yellow bars) yield an average Spearman correlation of 0.800 (standard deviation of 0.044). The average Spearman correlation between GNM MSFs and ATAC-seq data is somewhat lower: 0.552 ± 0.112. Interestingly, the average Spearman correlation between the two sets of experimental data was 0.741 ± 0.089, suggesting that the accuracy of computational predictions is comparable to that of experiments, and that the DNase-seq provides data more consistent with computational predictions. ATAC-seq maps not only the open chromatin, but also transcription factors and nucleosome occupancy [126], which may help explain the observed difference. The same analysis on IMR90 cells demonstrated even better agreement with experiments (Figure 4.5). The Spearman correlation between the computed MSFs and experimental ATAC-seq data averaged over all chromosomes was 0.63 ± 0.08 IMR90 cells, and that between MSFs and DNase-seq data was 0.82 ± 0.03. Consistently, the two sets of experiments also exhibit a higher correlation (0.81 ± 0.06) in this case.

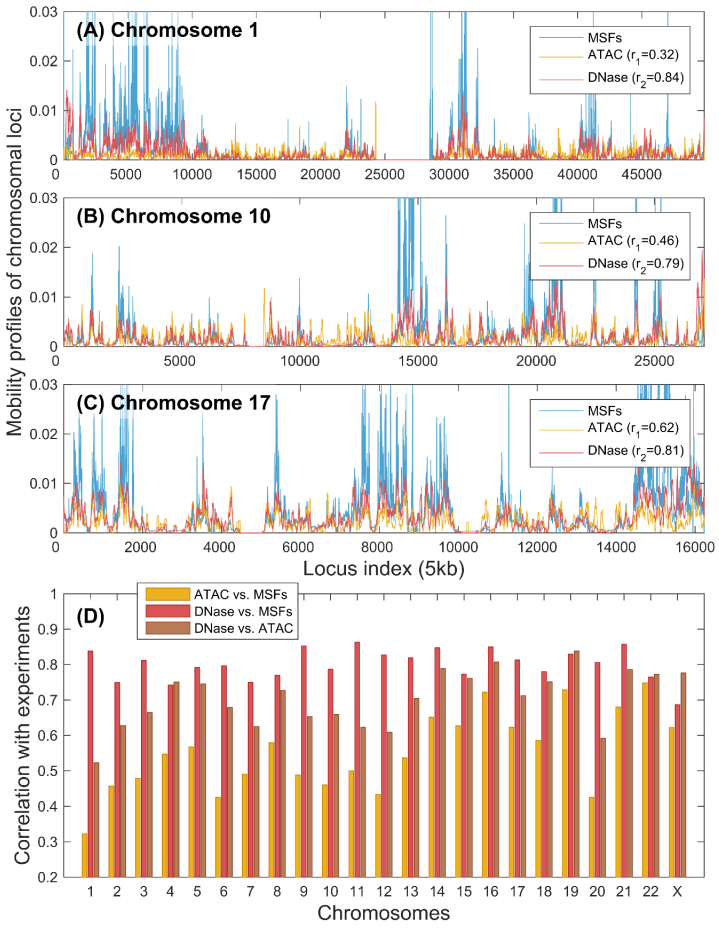


Figure 4.4: GNM-predicted mobilities of chromosomal loci in GM12878 show good agreement with data from chromatin accessibility experiments. (A) – (C) Mobility profiles (MSFs of loci) obtained from GNM analysis of the equilibrium dynamics of chromosomes 1, 17, and X, respectively, shown in blue, are compared to the DNA accessibilities probed by ATAC-seq (yellow) and DNA-seq (red) experiments. GNM results are based on 500 slowest modes. r_1 is the Spearman correlations between GNM predictions and DNase-seq experiments; and r_2 is that between GNM and ATAC-seq. (D) Spearman correlations between theory and experiments for all chromosomes (red and yellow bars, as labeled). The Spearman correlation between the computed MSFs and experimental ATAC-seq data averaged over all chromosomes is 0.55 ± 0.11, and that between MSFs and DNase-seq data is 0.80 ± 0.04. For comparison, we also display the Spearman correlation between the two sets of experimental data (brown bars); the average in this case is 0.70 ± 0.08.

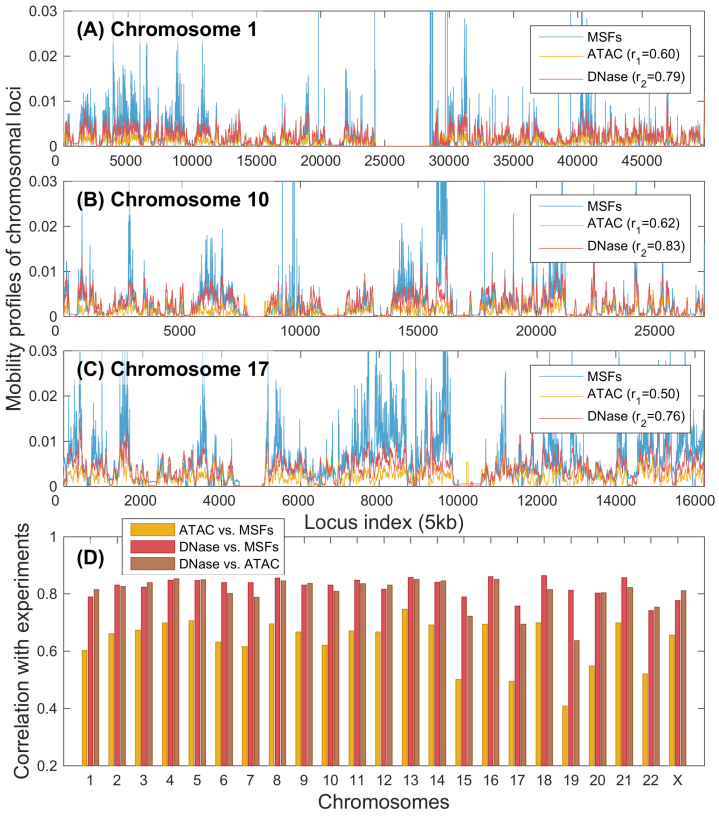


Figure 4.5: GNM-predicted mobilities of chromosomal loci in IMR90 similarly show good agreement with data from chromatin accessibility experiments. (A) – (C) Mobility profiles (MSFs of loci) obtained from GNM analysis of the equilibrium dynamics of chromosomes 1, 17, and X, respectively, shown in blue, are compared to the DNA accessibilities probed by ATAC-seq (yellow) and DNA-seq (red) experiments. GNM results are based on 500 slowest modes. r_1 is the Spearman correlations between GNM predictions and DNase-seq experiments; and r_2 is that between GNM and ATAC-seq. (D) Spearman correlations between theory and experiments for all chromosomes (red and yellow bars, as labeled). The Spearman correlation between the computed MSFs and experimental ATAC-seq data averaged over all chromosomes is 0.63 ± 0.08, and that between MSFs and DNase-seq data is 0.82 ± 0.03. For comparison, we also display the Spearman correlation between the two sets of experimental data (brown bars); the average in this case is 0.81 ± 0.06.

4.3.2 GNM results are robust to changes in the resolution of Hi-C data and can be efficiently reproduced with a representative subset of global modes

These results for two different types of cell lines show that the mobility profiles predicted by the GNM for the 23 chromosomes accurately capture the accessibility of gene loci. The agreement with experimental data lends support to the applicability and utility of the GNM for making predictions on chromatin dynamics. The current results were obtained by using subsets of $m = 500$ GNM modes for each chromosome, which essentially yield the same profiles and the same level of agreement with experiments as those obtained with all modes (see Figure 4.6). The use of a subset of modes at the low frequency end of the spectrum improves the efficiency of computations, without compromising the accuracy of the results. Computations repeated for different levels of resolution (from 5kb to 50kb per bin) also showed that the results are insensitive to the level of coarse-graining (Figure 4.6) which further supports the robustness of GNM results. All results are obtained by adopting the VC normalization for Hi-C data. Computations repeated with two alternative normalization schema, square-root VC [69] and Knight-Ruiz [63] normalization, showed a significant decrease in the level of agreement with experimental data regardless of the number of modes included in the GNM computations (Figures 4.2 and 4.3), and the underperformance of these schema became particularly pronounced in the case of high resolution data (Figure 4.3), in support of the VC normalization adopted here.

4.3.3 Domains identified by GNM at different granularities correlate with known structural features

Compartments, first identified by Lieberman-Aiden et al. [69], are multi-megabase-sized regions in the genome that correspond to known genomic features such as gene presence, levels of gene expression, chromatin accessibility, and histone markers. Hi-C experiments have revealed two

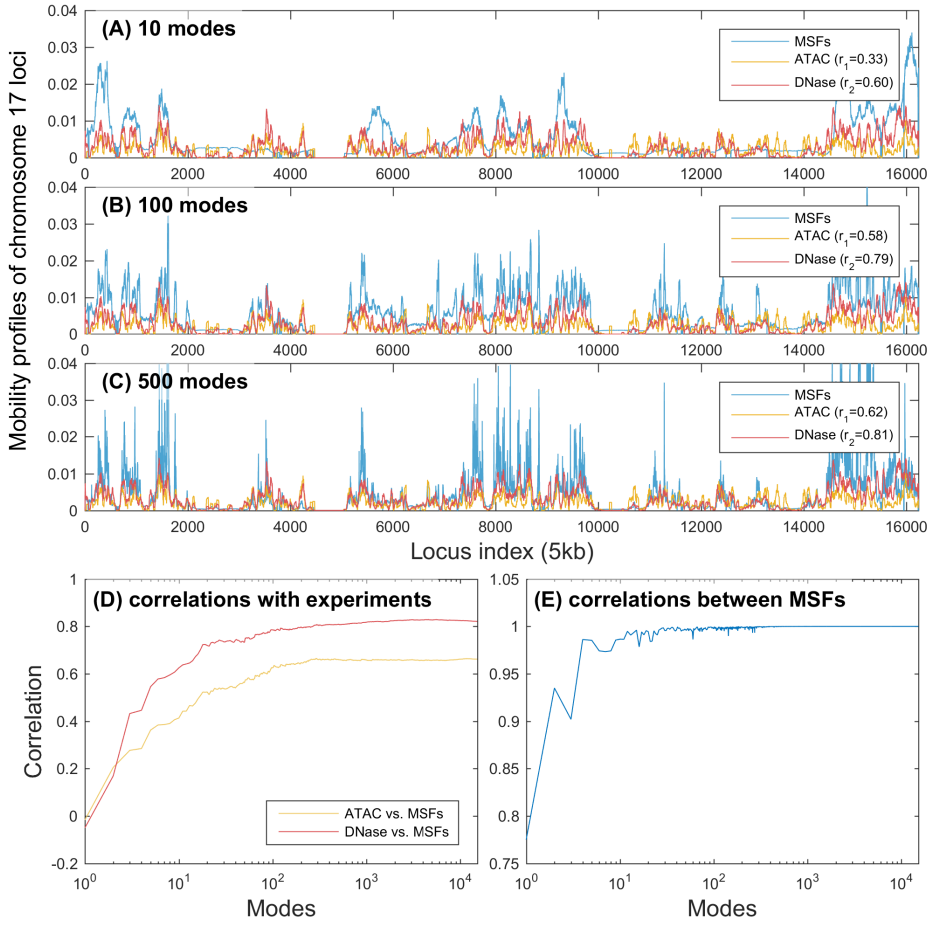


Figure 4.6: GNM computations of mobility profiles using different subsets of modes show the robust convergence of results with a small subset of modes. Results are presented here for GM12878 chromosome 17, at 5kb resolution. (A) – (C) Comparisons between experimental data and computed MSF profiles obtained using 10, 100, and 500 GNM modes. (D) Spearman correlations between experimental and computationally predicted fluctuation/accessibility profiles obtained with different numbers of modes. (E) Spearman correlations between MSFs computed from slowest i modes and $i + 1$ modes. The abscissa is in logarithmic scale in panels D and E. The correlation levels off at around a few hundreds of modes, showing that the addition of higher modes does not practically change the predicted MSF profile, and a small subset of < 500 modes can be efficiently used for evaluating the MSFs.

broad classes of compartments: "A" compartments generally associated with active chromatin, containing more genes, fewer repressive histone markers, and more highly expressed genes; and "B" compartments, for less accessible DNA, sparser genes, and higher occurrence of repressive histone marks. TADs [36] are finer resolution groupings of chromatin distinguished by denser self-interactions and associated with characteristic patterns of histone markers and CTCF binding sites near their boundaries. The multiscale nature of GNM spectral analysis allows hierarchical levels of organization to be identified computationally, and it is of interest to examine to what extent these two levels can be detected. As presented above, the GNM low frequency modes reflect the global dynamics of the 3D structure, and increasingly more localized motions are represented by higher frequency modes. We identified domains from subsets of GNM modes that group regions of similar dynamics (see Methods). In order to verify whether these dynamical domains correspond to TADs at various resolutions, we used the TAD-finder Armatus [42], varying its γ parameter that controls resolution. We refer to this latter parameter as the Armatus γ, to distinguish it from the force constant in the GNM. We measure the agreement between GNM domains and TADs using the variation of information (VI) distance, which computes the agreement between two partitions, and where a lower value indicates greater agreement [76]. For each choice k of number of modes, the Armatus γ_k that minimizes the VI distance between the GNM domains and the Armatus domains was selected. This resulted in a mean VI value for optimal parameters of 1.251, significantly lower than the VI distance of 1.946 obtained when the GNM domains were randomly re-ordered along the chromosome and compared back to the original TADs (empirical p-value < 0.01 for all chromosomes). Figures 4.7A and 4.8 show the comparisons for each chromosome between the VI value for the optimally matched TAD boundaries with the GNM domains and the distribution of VI values from the randomly shuffled domains. As the number of included GNM modes is increased, γ_k monotonically increases as well, showing that the number of GNM modes is a good proxy for the scale of chromatin structures sought. Furthermore, GNM predicts large-scale global motions using a relatively low

number of modes, so we compared these to larger-scale compartments. We found that the first 5-20 non-zero modes correspond fairly well to compartments. For each chromosome, we selected the number of modes that produced the smallest VI distance between Lieberman-Aiden compartments and GNM domains. This yielded a mean optimal VI distance of 1.771 (using an average of 13 modes). This is significantly lower than the mean optimal VI distance of 2.088 when the locations of Lieberman-Aiden compartments are randomly shuffled along the chromosome, though the difference is only statistically significant for 16 of the 23 chromosomes, with p-value equal to 0.05. The comparisons of GNM domains with compartments for each chromosome in GM1287 cells can be seen in Figure 4.7B. The same calculations were performed on IMR90 cells, with qualitatively similar results. For the comparisons with randomly shuffled domains on IMR90 cells, only 1 chromosome for TADs and 3 for compartments were statistically insignificant (see Figure 4.8). The ability of GNM to recapitulate both TADs and compartments – two organizational levels of wildly different scales – shows the flexibility and generality of the GNM approach. A TAD-finding method using only the second eigenpair (Fiedler value/vector) of the Laplacian has also been developed [26] and tested on 100kb resolution data. By including more eigenvectors, we are able to identify TADs closer to Armatus on all chromosomes (as measured by lower VI) at 5kb and for 18/23 chromosomes at 100kb resolution (see Figure 4.9A and C). Though the Fiedler vector-based method identifies compartments better at low resolution, that method performs poorly at finer resolution, while GNM remains robust to resolution changes. We are also able to identify compartment sets with lower VI on all chromosomes at 5kb (Figure 4.9B and D). Further corroborating the benefit of using multiple modes, it has been shown in early studies that spectral clustering by using more eigenvectors can outperform partitioning methods which only use one eigenvector [2, 3].

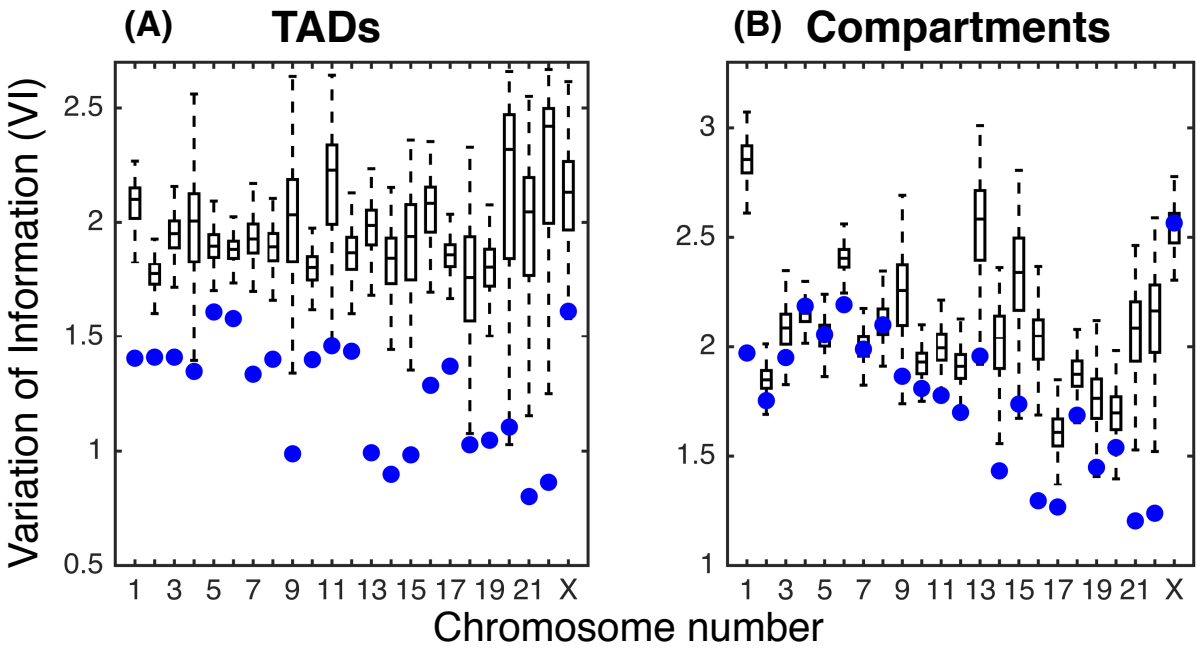


Figure 4.7: **Comparison of GNM domains with TADs and Compartments in GM12878.** Variation of information (VI) measures for comparing GNM domains with (A) TADs and (B) compartments (lower VI indicates greater agreement). Box plots show the distribution of VI values obtained by randomly shuffling GNM domains and comparing to original TAD and compartment boundaries. Blue dots represent the VI value of the true GNM domains with TADs and compartments, respectively.

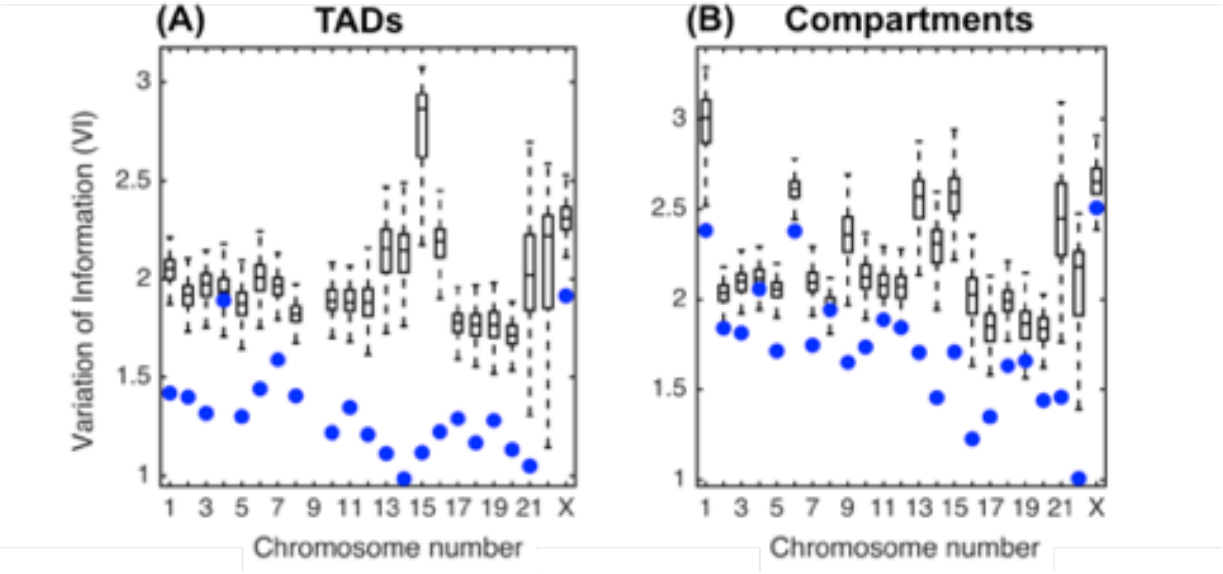


Figure 4.8: **Comparison of GNM domains with TADs and Compartments in IMR90.** Variation of information (VI) measures for comparing GNM domains with (A) TADs and (B) compartments (lower VI indicates greater agreement). Box plots show the distribution of VI values obtained by randomly shuffling GNM domains and comparing to original TAD and compartment boundaries. Blue dots represent the VI value of the true GNM domains with TADs and compartments, respectively.

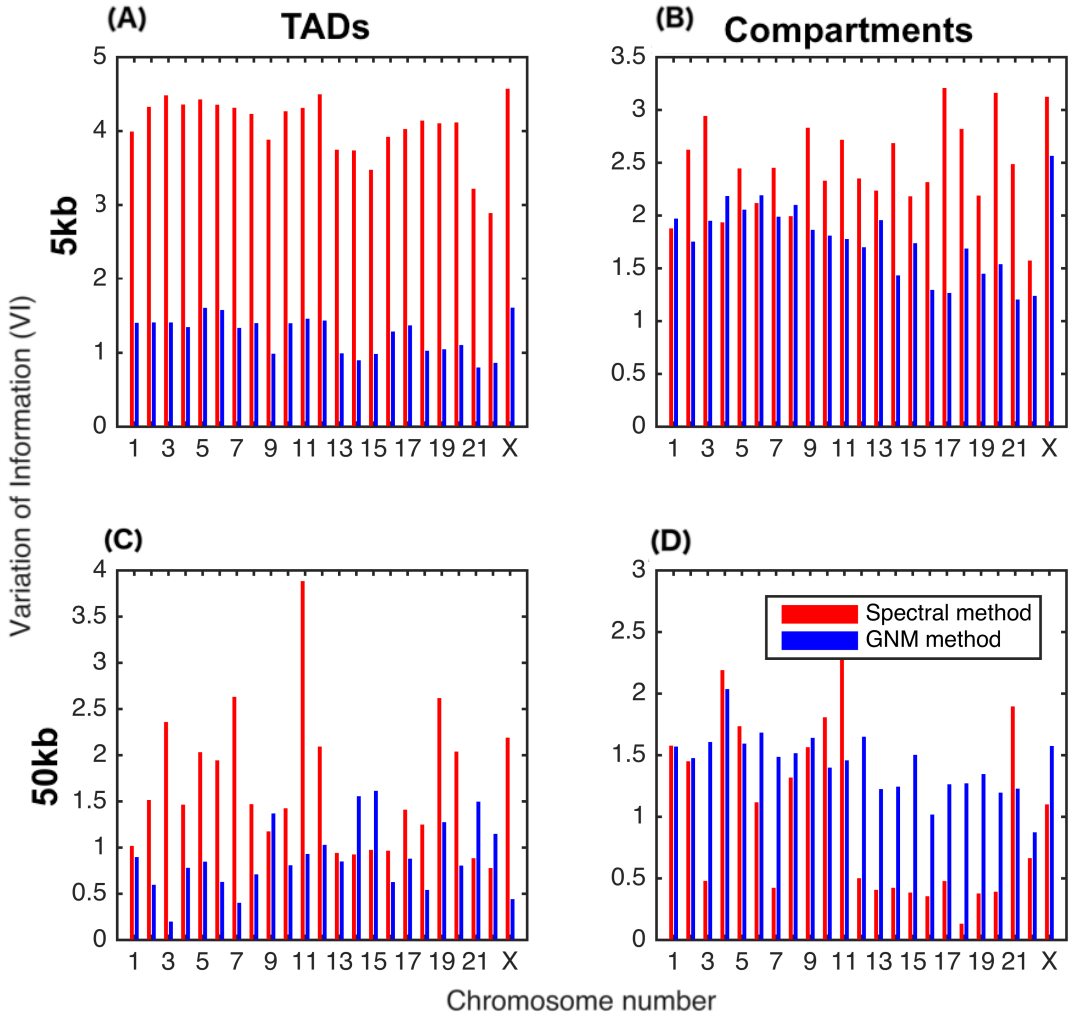
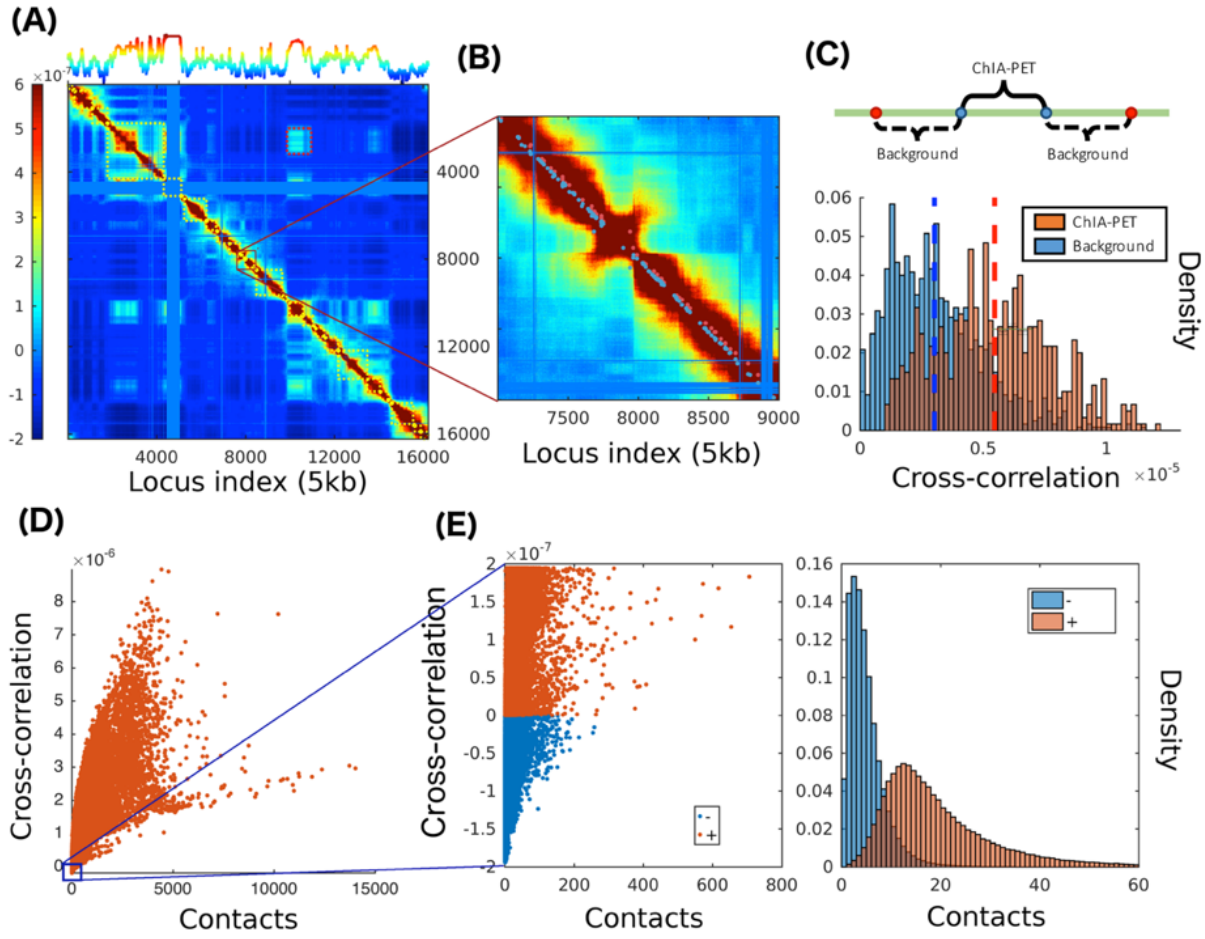


Figure 4.9: Comparison of GNM-based method for finding TADs and compartments to the spectral method from [26] for GM12878. In all panels, blue bars represent the results from the GNM method and red bars, those from the spectral method. A lower variation of information (VI) value demonstrates better agreement. Compartments were calculated based on the method described in [69]. TADs were computed using Armatus, which requires a resolution parameter γ. The VI value shown for every comparison with TADs represents the lowest VI from comparing to TAD sets obtained from γ ranging from 0 to either 0.5 (for 5kb resolution) or 1 (for 100kb resolution), with a step size of 0.05. (A) Comparison to TADs at 5kb resolution. (B) Comparison to compartments at 5kb resolution. (C) Comparison to TADs at 100kb resolution. (D) Comparison to compartments at 100kb resolution.

4.3.4 Loci pairs separated by similar 1D distance exhibit differential levels of dynamic coupling, consistent with ChIA-PET data

Figure 4.10 displays the covariance map generated for the coupled movements of the loci on chromosome 17 (of GM12878 cells), based on Hi-C data at 5 kb resolution. Panel A displays the cross-correlations (see equation 2) between all loci-pairs as a heat map. Diagonal elements are the MSFs (presented in Figures 4.4C and 4.5). The curve along the upper abscissa in Figure 4.10A shows the average cross-correlation of each locus with respect to all others; the peaks indicate the regions tightly coupled to all others, probably occupying central positions in the 3D architecture. The covariance map is highly robust and insensitive to the resolution of the Hi-C data. The results in Figure 4.10A were obtained using all the $m = 15,218$ nonzero modes corresponding to 5kb resolution representation of chromosome 17. Calculations repeated with lower resolution data (50kb) and fewer modes (500 modes) yielded covariance maps that maintained the same features. Owing to their genomic sequence proximity, the entries near the main diagonal of the covariance map tend to show relatively high covariance values (colored yellow-to-brown; Figure 4.10A). Even the close vicinity of the diagonals (e.g. loci intervals of ≥ 200) represents (at 5 kb resolution) genomic loci separated by more than 1 megabase. The covariance map clearly shows that there are strong couplings between loci separated by a few megabases. We show an example of such regions in Figure 4.10B. While the loci pairs located in the dark red band along the diagonal appear all to exhibit strong couplings, a closer examination reveals differential levels of cross-correlations that are in good agreement with the data from Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) experiments [135]. The 'long-range' interactions identified by ChIA-PET [53] are indicated in panel B by red dots (close to the diagonal). These are interacting loci separated by several hundreds of kb. We selected background pairs separated by the same 1D distance, on both sides of the ChIA-PET pair, and compared the cross-correlations predicted for the two sets along each chromosome (Figure 4.10C). The background pairs (blue bars) show weaker GNM cross-correlations compared to the ChIA-PET pairs (red



bars) although they are separated by the same genomic distance along the chromosome. Similar statistical analysis repeated for all 23 chromosomes showed that the cross-correlations of ChIA-PET pairs were greater than those of background pairs of the same genomic distance on every chromosome, with all p-values less than 10^{-19} (two-sided t-test).

Figure 4.10: **Covariance map computed for chromosome 17 and comparison with ChIA-PET data and contacts from Hi-C experiments in GM12878.** (A) Covariance matrix computed for chromosome 17, color-coded by the strength and type of cross-correlation between loci pairs ranged from 5th to 95th percentile of all cross-correlation values (see the color bar on the left). The curve on the upper abscissa shows the average overall off-diagonal elements in each column, which provides a metric of the coupling of individual loci to all others. The blocks along the diagonal indicate loci clusters of different sizes that form strongly coupled clusters. The red dashed boxes indicate the pairs of regions exhibiting weak correlations despite genomic distances of several megabases. The blue bands correspond to the centromere, where there are no mapped interactions. (B) Close-up view of a region along the diagonal. Red dots near the diagonal indicate pairs (separated by $\approx$100 kb) identified by ChIA-PET to interact with each other; nearby blue points are control/background pairs. (C) Stronger cross-correlations of ChIA-PET pairs compared to the background pairs. (D) Dependence of cross-correlations on the number of contacts observed in Hi-C experiments. A broad distribution is observed, indicating the effect of the overall network topology (beyond local contacts) on the observed cross-correlations. (E) Loci pairs exhibiting anti-correlated (same direction, opposite sense) movements usually have fewer contacts, compared to those exhibiting correlated (same direction, same sense) pairs of the same strength.

4.3.5 Cross-correlations between loci motions are global properties that result from the overall chromosomal network topology

In general, loci-loci cross-correlations become weaker with increasing distance along the chromosome, and some pairs show anticorrelations (i.e. move in opposite directions; see scale bar in Figure 4.10A). Yet, we can distinguish distal regions that exhibit notable cross-correlations in the spatial movements (off-diagonal lighter-colored blocks). The levels of cross-correlations do not necessarily need to scale with the interaction strengths between the correlated loci (or number of

contacts detected by Hi-C). On the contrary, a broad range of cross-correlations is observed for a given number of contacts, indicating that the observed correlations are global properties defined by the entire network topology and reflect the collective behavior of the entire structure. Figure 4.10D displays the computed cross-correlations as a function the number of contacts, showing that some pairs of loci display much stronger correlations revealed by the GNM than others that make more Hi-C contacts. Figure 4.10E shows that the anticorrelated pairs of loci (blue) usually have fewer contacts than those (red) exhibiting positive cross-correlations of the same strength.

4.3.6 Distal regions that are predicted to be strongly correlated in their spatial dynamics exhibit higher co-expression

The GNM covariance map further shows correlations between farther apart (> 10 Mbp) regions. In contrast to the main diagonal, the majority of the off-diagonal space typically shows significantly weaker correlations. Regions in this space with higher than expected covariance values represent dynamically linked windows along the chromosome, which may represent long-range interactions. We call these pairs of windows cross-correlated distal domains (CCDDs). To identify CCDDs, we set a threshold for each covariance matrix equal to the absolute value of the minimum covariance. Treating the remaining adjacent pairs as edges in a graph, we then locate connected components beyond the widest section of the main diagonal and above the threshold that contain more than one bin pair, and find the maximal-area rectangle contained within each connected region of high covariance values (see Figure 4.11). These CCDDs are therefore pairs of regions distant along the chromosome, composed each of highly interconnected loci, which also exhibit relatively high cross-correlations compared to other regions of similar genomic separation. Previous methods for identifying long-range chromatin interactions [98, 105, 130, 135] have focused on locating individual points of interaction within 1-2 Mbp apart, while CCDDs tend to be on the order of tens of Mbp apart and supported by groups of interacting loci.

The covariance matrix results from the overall coupling of the complete network of loci upon

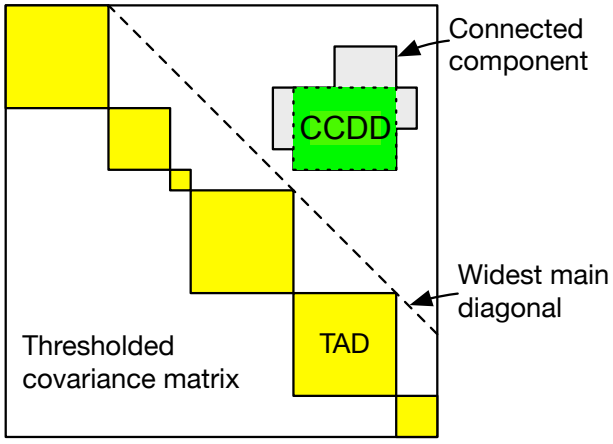


Figure 4.11: Identification of cross-correlated distal domains (CCDDs). CCDDs are found by searching for connected components outside of the widest point of the main diagonal. The CCDD is then the rectangle of maximal area contained entirely within the connected component.

inversion of the connectivity/Kirchhoff matrix for the entire chromosomes. As such, it permits to capture, or better discriminate, the long-range correlations resulting from the complex topology of loci-loci contacts, as opposed to the raw data on local loci-loci contacts described by Hi-C maps. The covariance data also permit the identification of an appropriate threshold value for defining the significant CCDDs, consistent with the cooperative couplings within the entire structure, including distal correlations. There is no correspondingly clear threshold value for raw Hi-C data, which makes identifying these regions difficult without covariance matrices. Highly distant gene pairs within CCDDs show greater co-expression values than gene pairs outside these regions (p-value $< 10^{-7}$ using the background defined below). For each CCDD, we identified the genes contained within the region and measured the co-expression of each gene pair from distant chromosomal segments. The background gene pairs were gathered from outside the CCDDs but with similar genomic separation as the CCDD gene pairs. We computed gene expression correlations from 212 experiments (see Section 4.2.7). As seen in Figure 4.12, the CCDDs containing specifically gene pairs that are between 50 and 100 Mbp apart are much more highly co-expressed than background gene pairs at the same genomic distance (p-value $< 10^{-19}$, see

Section 4.2.7 for details). This indicates that the dynamic coupling of these genes, as revealed by GNM, may often be biologically important. CCDDs at smaller genomic distance (< 50 Mbp) exhibit similar co-expression distributions to the background gene pairs, likely due to the effect of shorter genomic distances including more co-regulated genes within the background. Beyond distances of 100 Mbp, there are not sufficient gene pairs within CCDDs to draw any meaningful conclusions. Dynamically coupled regions that are very distant sequentially but biologically linked through gene expression are therefore identifiable using the GNM covariance matrix.

4.4 Discussion

This work represents the first analysis of chromosome dynamics using an elastic network model, GNM, which has found wide applications in molecular structural biology. Though other models [25, 26] have examined genome structure through graph theoretical methods, the inclusion of the complete spectrum of motions in the analysis provides a more realistic picture of chromosomal dynamics in accord with a wealth of experimental data. The approach brings three key advantages. First, this is a mathematically rigorous, based on first physical principles, with intuitive interpretations and well-established theoretical and physical underpinnings. Second, it enables us to evaluate, compare and consolidate with the help of a unified model a broad range of biologically significant genome-wide properties. These include the evaluation of loci MSFs at 5kb resolution, the discrimination of short-range regulatory interactions among close-neighboring loci, and the identification of TADs and compartments. These respective predictions were shown to satisfactorily compare with data from chromatin accessibility (DNase-seq and ATAC-seq) and ChIA-PET experiments, and predictions from previous computational methods. The agreement with experiments not only validates the applicability of the GNM, but also provides a new set of independent data, which consolidate those from experiments, especially when the experimental data themselves exhibit some differences (see Figure 4.4). The application to two different cell types also showed that GNM data comply with cell-cell variability. This unifying frame-

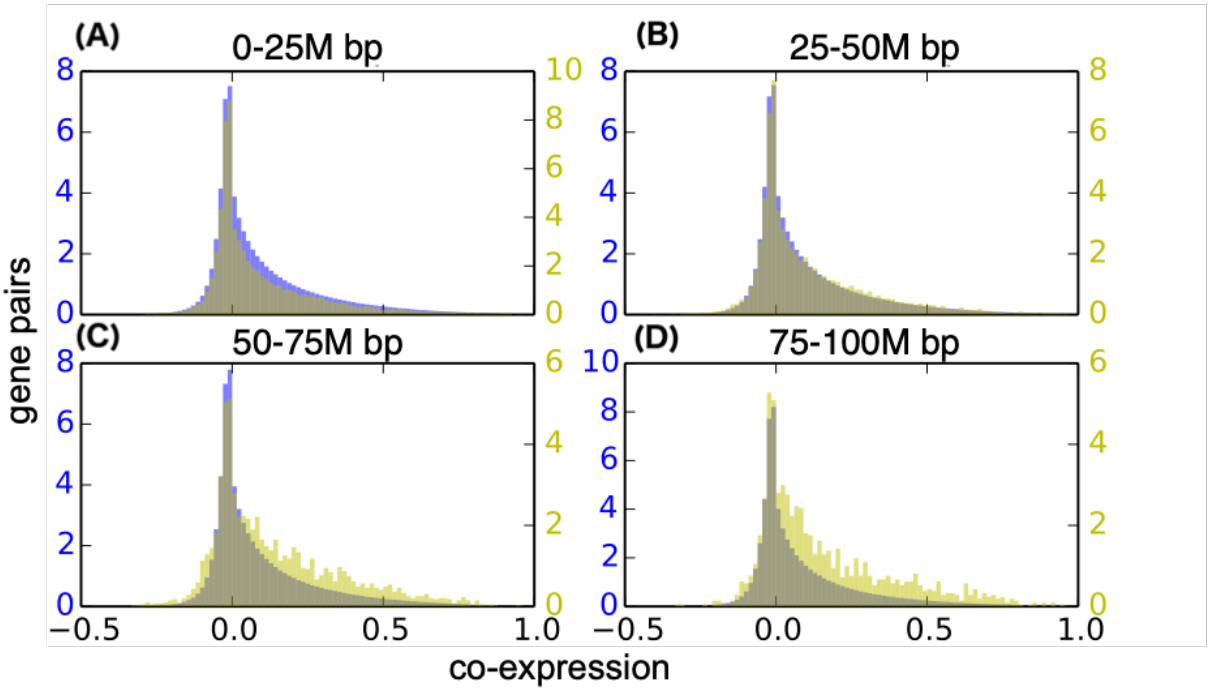


Figure 4.12: **Co-expression is significantly enriched in distant CCDDs in GM12878 cells.** In each histogram, the yellow distribution represents gene pairs from CCDDs and the blue distribution represents background gene pairs. All are showing the normalized number of gene pairs with a particular Pearson expression correlation for gene pairs within a distance of (A) 0-25 million base pairs, (B) 25-50 million base pairs, (C) 50-75 million base pairs, and (D) 75-100 million base pairs. The more distant pairs (50-100 million base pairs apart) within the CCDDs show enriched expression correlations as compared to the background pairs. There were not enough gene pairs within CCDDs more than 100M base pairs apart to draw significant conclusions.

work further led to the discovery of biologically significant, dynamically coupled regions, termed CCDDs. No existing method has located spatially coupled co-expressed regions of the genome which are so distant (over 50 Mbp apart) along the chromosomes, and this information cannot be found from gene expression or other experimental data alone.

Future GNM analyses of chromatin dynamics could focus on the nature of the long-range couplings, analysis of their biological significance, or the meaning of genomic regions that exhibit high covariances. GNM also predicts a measure of overall coupling of each genomic locus to others (see the curve along the upper abscissa in Figure 4.10A), the significance of which requires further investigation. The GNM was shown to capture several biological properties of chromosomes, but further insights on cooperative events, including the interchromosomal (trans) interactions is within reach by focusing on the softest (lowest frequency) modes of motion predicted by the GNM. Finally, advances in 3D embeddings of Hi-C data may open the way to adopting the Anisotropic Network Model (ANM) [5, 41, 124] for efficient modeling and visualization of the whole chromatin dynamics.

Chapter 5

Relationships between computational and biological TAD properties

One of the major challenges for the computational TAD-finding field is the fact that there is significant disagreement between the TADs identified by different methods. Beyond this disagreement, it is unclear which method is "best" because they tend to perform well in certain evaluation metrics but not others, as documented by several recent reviews. We study the relationships between computational TAD definitions and biological TAD properties through a combination of algorithm development and analysis. We design a flexible TAD finder based on several computational TAD definitions, with parameters that can be optimized for any desired property. We then explore the space of computational TAD sets by optimizing for various biological properties, and analyze the relationships between the resulting TAD sets.

5.1 Background

Since the introduction of a sequencing technique to study the genome-wide three-dimensional structure of chromosomes [69], many studies have shown a connection between this architecture and regulatory mechanisms. Genome structure plays a role in gene regulation [58], cell cycle

regulation [35], and many human diseases and disorders (see [119] for review). This structure has been described by topologically associating domains (TADs), the building blocks of genome architecture that are regions of the chromosome interacting more highly within themselves than outside [36]. TADs in particular have been suggested to bring together target genes with their intended regulatory elements and insulate them from interactions with other nearby enhancers and promoters, though their exact role and even definition remains an area of active study [13].

One of the main challenges in describing and identifying TADs is a lack of consensus on how to define these structures, with many different potentially defining features used by different analyses. Computationally, we expect TADs to have dense intra-TAD contacts, sparse inter-TAD contacts, and show a strong shift in contact direction at TAD boundaries, among other features [13, 36]. TADs are generally identified through optimization of one of these computational properties. One TAD finder, Armatus [42], finds the TAD set with the greatest sum of TAD densities, where TAD density is a scaled sum of the Hi-C contact counts within each TAD. TopDom [116] and Insulation Score [28], both successful TAD finders, instead set TAD boundaries at the minima of a function adding up the contacts within a window centered on each bin, thereby identifying TADs as the regions in between highly insulated bins. The first computational method for TAD finding, called Domain Caller [36], was based on quantifying the direction of contact bias, or whether a bin preferentially interacted with other bins upstream of it or downstream, and used a Hidden Markov Model to define TADs where this directionality index switched from upstream to downstream. These concepts are all related and have all been successful in TAD identification, though no one method performs best across all evaluation metrics [30, 44, 139].

Without a gold standard or ground truth TAD set to which we can compare computational predictions, we assess TAD sets based on various properties we expect from them. Biologically, TADs should show increased binding of CTCF, certain histone markers, and cohesin at their boundaries, and remain fairly consistent across replicates [36, 42, 108]. CTCF is a critical structural protein for TAD boundaries, and we therefore expect to see an enrichment of CTCF binding

at these boundary locations [36]. Similarly, the histone marker H3K36me3 has been shown to be enriched at TAD boundaries, to the extent that it can be combined with other histone markers to predict TAD locations [42, 112]. Because of the significant involvement of cohesin in TAD formation, two of its proteins, RAD21 and SMC3, have also been shown to peak around TAD boundaries [116]. While TADs have been shown to vary between single cells [128], on the population level they should remain fairly consistent between replicates [108]. On a fundamental level, TADs are defined by increased contact frequency within their interiors and somewhat depleted contact frequency between different TADs. This can be quantified by looking at the distributions of inter-TAD contacts and intra-TAD contacts, with the expectation that the distribution of intra-TAD contacts is generally higher than that of inter-TAD contacts.

Several reviews have assessed the performance of TAD finders on their ability to fit these biological properties, always finding that no one method captures every one equally well, and there is a significant amount of disagreement in the TADs output by the different methods [30, 44, 139]. It is unclear why these different computational TAD definitions lead to such different TAD sets, and why some perform better on certain metrics than others. For example, while Armatus TADs are fairly reproducible [44] and show higher intra-TAD contacts than inter-TAD contacts [30], the boundaries do not contain as many CTCF, RAD21, or SMC3 peaks as other tools [139]. On the other hand, Domain Caller TADs have very different inter-TAD and intra-TAD contact distributions [30], align extremely well with CTCF, RAD21, and SMC3, but do not have particularly impressive histone marker measures [139], and have very low reproducibility [44]. TopDom and Insulation Score seem to occupy a space of doing fairly well across all metrics and generally agreeing well with other TAD finding results, but rarely coming out as the best TAD finder in any particular measure.

The relationship between the computational definitions each of these methods implements and the resulting TAD sets remains unclear, and it is still an open question whether there exists an algorithm to identify TADs which would be superlative across assessment metrics, or whether

there are inherent tradeoffs between these evaluation criteria. To study these questions and the variability of computational TAD predictions, we have developed a general TAD finder with several tunable parameters to reflect the space of potential TAD sets. This allows exploration of the computational definitions of TADs, and the ability to combine computational components rather than relying on a single TAD definition.

Parameters of the model can be chosen to guide TAD selection towards any desired property, provided the relevant data. This connects the computational definitions with the biological properties we use to assess TAD sets, identifying TADs that fit a combination of computational definitions and optimize a specific biological property such as high CTCF occupancy at boundaries. We choose parameter sets to optimize six different desirable TAD properties: (1) many CTCF binding sites at boundaries, high occupancy (measured by ChIP-seq data) of (2) CTCF, (3) RAD21, or (4) H3K36me3 at TAD boundaries, (5) reproducibility, and (6) large difference in inter- and intra-TAD contact frequency. We compare the resulting TAD sets across 12 different cell and tissue types to quantify variability and study the computational and biological properties that lead to similar TAD outputs. We find that some cell types show extremely high variability in TAD sets, while others are much more consistent, and analyze which properties lead to each of these outcomes.

5.2 Methods

To reflect the space of potential computationally-defined TAD sets, we developed a general algorithm for TAD identification. This algorithm, called FrankenTAD, optimizes a linear combination of three computational TAD properties that have been used in successful TAD finders. Tuning the parameters of this algorithm can lead to very different TAD sets, so we can use various desired properties to select parameters that will guide the resulting TAD set towards the given property. Parameters are optimized for one of several possible objective functions, reflecting various biological and technical TAD properties discussed below.

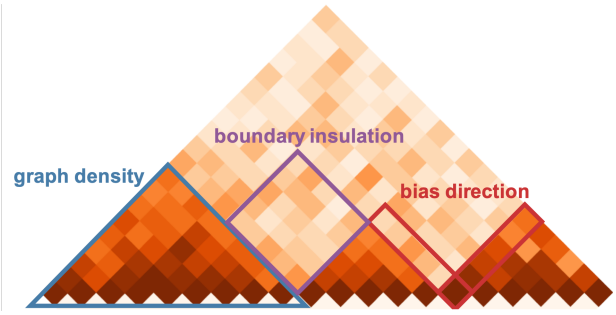


Figure 5.1: Illustration of the three TAD features optimized by FrankenTAD.

5.2.1 TAD-finding algorithm

To explore the space of TADs generated through computational TAD definitions, we developed a flexible algorithm called FrankenTAD based on several features previously used to identify TADs. FrankenTAD optimizes a linear combination of three TAD features: density within TADs, insulation between TADs, and a change in contact bias around TAD boundaries (Figure 5.1). Varying the model parameters leads to different TAD sets that emphasize different computational and definitional aspects of TADs.

Below, $\mathcal{D}$ represents a set of TADs on a given chromosome, the $N \times N$ matrix A is the normalized Hi-C matrix with N bins, each bin representing a segment of k bases where k is the data resolution, $A_{i,j}$ is the normalized Hi-C value between bin i, and bin j, and an individual TAD can be represented by $[a, b]$.

The core of FrankenTAD is the following objective function, combining three different TAD definitions, where $\lambda = \langle \lambda_1, \lambda_2, \lambda_3 \rangle$:

$$F(\mathcal{D}, \lambda, \gamma, \alpha) = \lambda_1 f_{GD}(\mathcal{D}, \gamma) + \lambda_2 f_{BI}(\mathcal{D}, \alpha) + \lambda_3 f_{BD}(\mathcal{D}). \tag{5.1}$$

The three components of this function, f_{GD}, f_{BI}, and f_{BD} will be explained in detail below. Each scores a different aspect of a TAD set: dense intra-TAD contacts, insulation between TADs, or a shift in bias direction at TAD boundaries.

Graph density component

The first property included in FrankenTAD is inspired by Armatus [42], in which TADs are chosen to maximize the sum of scaled subgraph densities of each TAD. Computationally, this defines a TAD as a high density subgraph of the graph induced by the Hi-C matrix. The following objective function $f_{GD}(\mathcal{D}, \gamma)$ is used by FrankenTAD and Armatus:

$$f_{GD}(\mathcal{D}, \gamma) = \sum_{[k,l] \in \mathcal{D}} [s(k, l, \gamma) - \mu_s(l - k)] \tag{5.2}$$

$$s(k, l, \gamma) = \frac{\sum_{g=a}^{b} \sum_{h=g}^{b} A_{gh}}{(b - a)^{\gamma}} \tag{5.3}$$

where γ is a parameter to be optimized, and $\mu_s(l - k)$ is the mean value of $s(k, l, \gamma)$ for all possible TADs of length $l - k$. For details on this function, see Fillipova et al. [42].

Boundary insulation component

The next component of FrankenTAD is inspired by TopDom [116], in which TADs are defined by the strength of their boundaries rather than the density of their contacts. A sliding window is used to quantify the insulation from other nearby regions, and boundaries are identified by finding the local minima of this function. The function quantifying the average contact frequency for each bin is

$$binsignal(i) = \frac{1}{w^2} \sum_{l=i-w-1}^{i} \sum_{m=i+1}^{i+w} A_{l,m}, \tag{5.4}$$

where the parameter w controls the window size. The expectation is for this $binsignal(i)$ to be high if i is near the center of a TAD, and low if i is at or near a TAD boundary. Details on this function can be found at Shin et al. [116]. The following function should be maximum when TAD boundaries are at the local minima of the $binsignal(i)$ function. We therefore score a TAD set based on the distance from its boundaries to the nearest local minima in the $binsignal$

function, and the average slope of $binsignal$ around the boundary:

$$f_{BI}(\mathcal{B}, \alpha) = \sum_{i \in \mathcal{B}} [-dist(i) + \alpha \cdot slope(i)] \tag{5.5}$$

$$slope(i) = \frac{1}{3} \sum_{x=i-3}^{i-1} [binsignal(x) - binsignal(x+1)] + \frac{1}{4} \sum_{x=i}^{i+3} binsignal(x+1) - binsignal(x) \tag{5.6}$$

The set $\mathcal{B}$ contains all TAD boundaries, $dist(i)$ is the number of bins from i to the nearest local minimum, $slope(i)$ is the average slope of $binsignal$ around i, and α is a parameter to be optimized. This should prioritize TAD boundaries not only at local minima, but those with especially high differences in insulation around the boundary.

Bias direction component

The third component of FrankenTAD comes from the Domain Caller TAD finder [36], which is based on the insight that TADs and their boundaries will display particular contact patterns, with bins near the start of a TAD displaying a strong bias towards contacts downstream, while bins near the end of a TAD will show a strong bias towards upstream contacts. At TAD boundaries, we therefore expect to see a switch from downstream to upstream bias. This notion of contact bias was quantified in the following way in Dixon et al. [36]:

$$C(i) = \sum_{g=i-w_d}^{i-1} A_{gi} \tag{5.7}$$

$$B(i) = \sum_{g=i+1}^{i+w_d} A_{ig} \tag{5.8}$$

$$E(i) = (A(i) + B(i))/2 \tag{5.9}$$

$$DI(i) = \left(\frac{B(i) - C(i)}{|B(i) - C(i)|} \right) \left(\frac{(C(i) - E(i))^2}{E(i)} + \frac{(B(i) - E(i))^2}{E(i)} \right). \tag{5.10}$$

In this formula, $C(i)$ represents the number of upstream contacts of bin i, $B(i)$ represents the downstream contacts, $E(i)$ is the expected number of contacts, and w_d is a window size parameter computed based on the resolution to consider contacts within 2Mb up or downstream of

the bin of interest. We expect $DI(i)$ to be strongly positive at the start of a TAD and strongly negative at the end of a TAD, leading to the following scoring function:

$$f_{BD}(\mathcal{D}) = \sum_{[a,b]\in\mathcal{D}} \sum_{g=0}^{w} [DI(a+g) - DI(b-g)]. \tag{5.11}$$

Dynamic program to optimize multi-feature TAD definition

Ideally, TADs should fit all of these definitions: high density within a TAD, strong insulation at boundaries, a strong downstream contact bias near the start, and a strong upstream contact bias near the end of the TAD. A linear combination of these properties provides a flexible framework for identifying TAD sets based on these computational features.

We identify TADs by maximizing Equation 5.1, which can be done efficiently through a dynamic program. For any position b on the chromosome, the optimal TAD set over the interval $[0, b]$ is given by:

$$OPT(b) = \max_{a<b} \big(OPT(a-1) + F([a,b], \lambda, \gamma, \alpha)\big) \tag{5.12}$$

Efficiency is achieved through significant precomputation of the elements of each subfunction, as well as the above dynamic program.

5.2.2 Data-driven objective functions

We define six different data-driven objective functions to select parameters for FrankenTAD, thereby guiding the TAD sets we find towards these properties. Each of these has properties has been used to assess the quality of TAD sets, with existing TAD finders showing clear tradeoffs between them. We formulate each property as an objective function, and parameters are chosen to find TAD sets that optimize this objective function. These objective functions are computationally expensive to evaluate because they involve running FrankenTAD with the test parameter set to find TAD sets, and then assessing the objective on those TADs. We use Bayesian optimization [91] to select parameters according to each of these properties.

CTCF binding sites

While its exact role is unclear, the structural protein CTCF is widely understood to be critical to TAD architecture. Peaks in the number of CTCF binding sites at TAD boundaries have been used to validate the quality of the TADs [42]. While less descriptive than ChIP-seq data, binding site locations are not cell-type specific and are relatively easy to identify. Binding site locations are the best way to optimize for protein locations in the absence of ChIP-seq data which may not exist for a given species, cell, or tissue type. To optimize for CTCF binding sites, we maximize a function that counts the number of binding sites at each TAD boundary in a candidate TAD set. Let $n_{site}(i)$ be the number of CTCF binding sites within bin i, and recall that $\mathcal{B}$ is the set of all TAD boundaries. Then

$$OBJ_{bind} = \sum_{i \in \mathcal{B}} n_{site}(i). \tag{5.13}$$

The parameters of FrankenTAD are chosen to identify TAD sets that maximize OBJ_{bind}.

ChIP-seq peaks: CTCF, RAD21, and H3K36me3

While binding site locations are informative of where proteins could bind, ChIP-seq data reveals the true locations of bound proteins in a specific cell or tissue type. We therefore use a similar objective function to maximize the number of peaks of bound proteins at TAD boundaries for three different structurally associated proteins. RAD21 is a protein component of the cohesin complex, which has been implicated as a key structure in TAD formation, along with CTCF. H3K36me3 is a histone marker that has also been associated with TAD structures [42, 112] and is expected to be enriched at TAD boundaries. We therefore use ChIP-seq peaks of CTCF, RAD21, and H3K36me3 to select parameters that will identify TAD sets with boundaries at the locations of binding peaks, with $n_{bound}(i)$ as a function that counts the number of ChIP-seq peaks with midpoints within the bin i:

$$OBJ_{chip} = \sum_{i \in \mathcal{B}} n_{bound}(i). \tag{5.14}$$

Highlighting the weakness of relying on ChIP-seq data, only 6 of 12 total cell and tissue types studied here had publicly available RAD21 ChIP-seq data, so the analysis of cohesin is more limited than the others.

Reproducibility

There can be significant variability between single cells of the same population [19, 82, 121], but on a population level we broadly expect replicates to show similar TAD sets [108]. The Jaccard Index, $JI(\mathcal{D}_1, \mathcal{D}_2)$, a distance metric quantifying the overlap between two sets, is used to quantify the similarity between replicate TAD sets. In this case, $\mathcal{D}_i$ represents a TAD set, and the Jaccard Index is computed as the ratio of the intersection of the TAD sets to their union. Parameters are chosen here to identify TAD sets on each replicate with maximal Jaccard Index between them:

$$OBJ_{rep} = JI(\mathcal{D}_1, \mathcal{D}_2) \tag{5.15}$$

In this work we use two replicates for each cell type, but if more replicates are available this objective could be expanded as the sum of all pairs of replicate JI values.

Inter- versus intra-TAD contact frequency

Perhaps the most fundamental property of TADs is that there are more interactions within TADs than between them, which can be quantified by comparing the inter-TAD contact frequency with the intra-TAD contact frequency. This property is the closest to the computational features optimized by FrankenTAD, compared to the biological properties reflected by ChIP-seq data. Intra-TAD contact frequency is computed as the mean of all Hi-C values within TADs in the given TAD set (Equation 5.16), where n is the total number of intra-TAD matrix entries. Inter-TAD contact frequency is computed as the mean of all Hi-C values with bins in adjacent TADs. If we describe the t^{th} TAD along the chromosome as $[a_t, b_t]$ (it begins at bin a_t and ends at bin b_t), the next TAD along the chromosome can be described as $[a_{t+1}, b_{t+1}]$. The inter-TAD

contact mean is given by Equation 5.17, where m is the total number of inter-TAD matrix entries considered. To identify TADs with the greatest change in these two values, we maximize their difference with OBJ_{int}:

$$INTRA(\mathcal{D}) = \frac{1}{n} \sum_{[a,b]\in\mathcal{D}} \sum_{i=a}^{b} \sum_{j=i}^{b} A_{ij} \tag{5.16}$$

$$INTER(\mathcal{D}) = \frac{1}{m} \sum_{[a_t,b_t]\in\mathcal{D}} \sum_{i=a_t}^{b_t} \sum_{j=a_{t+1}}^{b_{t+1}} A_{ij} \tag{5.17}$$

$$OBJ_{int} = INTRA(\mathcal{D}) - INTER(\mathcal{D}). \tag{5.18}$$

5.2.3 Data

Parameters optimizing each relevant objective function were used to predict TADs on all autosomal chromosomes of 12 different cell types from a variety of studies (see Supplementary Table 5.1) at 40kb resolution. The testing data was chosen to represent a range of biological conditions (e.g. cancerous and healthy), cell and tissue types, sequencing depths, and availability of relevant ChIP-seq data. All Hi-C data was uniformly processed from sequence data to ICE-normalized Hi-C matrices using HiC-Pro [113] (accession numbers available in Table 5.1). All ChIP-seq data was processed into peak formats (accession numbers available in Table 5.2).

5.3 Results

Each parameter-choice was used to run FrankenTAD on each cell type for which the necessary data was available, thereby guiding TAD identification towards each desirable property. We use the Jaccard Index (JI) to quantify similarity between the TAD sets within each cell type and study the variability of TAD sets under various objectives. One cell type (NHEK) did not have replicate data, so we were unable to optimize it for reproducibility, and six of the cell types did not have publicly available RAD21 ChIP-seq data, so they could not be optimized for cohesin

Cell type	Description	Replicates	Accession(s)	Citation
IMR90	lung fibroblast	2	SRR1658672, SRR1658673, SRR1658674, SRR1658675, SRR1658676, SRR1658677, SRR1658678	[98]
GM12878	blood lymphocyte	2	SRR1658570, SRR1658571, SRR1658572, SRR1658573, SRR1658574, SRR1658575, SRR1658576, SRR1658577, SRR1658578, SRR1658579, SRR1658580, SRR1658581, SRR1658582, SRR1658583, SRR1658584, SRR1658585, SRR1658586, SRR1658587, SRR1658588, SRR1658589, SRR1658590, SRR1658591, SRR1658592, SRR1658593, SRR1658594, SRR1658595, SRR1658596, SRR1658597, SRR1658598, SRR1658599, SRR1658600, SRR1658601, SRR1658602, SRR1658603	[98]
K562	chronic myeloid leukemia	2	SRR1658693, SRR1658694, SRR1658695, SRR1658696, SRR1658697, SRR1658698, SRR1658699, SRR1658700, SRR1658701, SRR1658702	[98]
NHEK	epidermal keratinocyte	1	SRR1658689, SRR1658690, SRR1658691	[98]
A549	adenocarcinomic alveolar basal epithelial	2	ENCLB571HTP, ENCLB222WYT	[125]
LNCaP-FGC	prostate carcinoma epithelial-like	2	ENCLB191OGC, ENCLB473XWD	[125]
T47D	ductal carcinoma	2	ENCLB758KFU, ENCLB183QHG	[125]
hESC	human embryonic stem cell	2	SRX116344, SRX128221	[36]
pancreas	tissue from 2 donors	2	SRX2179254, SRX2179255, SRX2179256, SRX2179257	[109]
spleen	tissue from 2 donors	2	SRX2179264, SRX2179263	[109]
skeletal muscle	gastrocnemius medialis tissue, 4 donors	4	ENCLB925XYW, ENCLB361HQM, ENCLB966EDS, ENCLB645GUM	[125]
HFFc6	subclone of HFF-hTERT	2	4DNES2R6PUEK	[32]

Table 5.1: Hi-C data used to generate all results in this chapter.

Cell/tissue type	CTCF	RAD21	H3K36me3
IMR90	GSM935404	GSM935624	GSM1055820
GM12878	GSM935611	GSM935332	ENCFF479XLN
K562	GSM1817654	GSM935319	ENCFF676RWX
NHEK	GSM822271	–	ENCFF253CMF
A549	ENCFF543VGD	ENCFF958VNQ	GSM1003494
LNCaP-FGC	GSM2827202	–	GSM875814
T47D	ENCFF903ZMF	GSM3415550	GSM1541451
hESC	GSM822297	GSM935379	GSM450268
pancreas	GSM2827540	–	ENCFF928PAZ
spleen	ENCFF459AHK	–	GSM2700497
skeletal muscle	GSM733762	–	GSM733717
HFFc6	GSM1022644	–	GSM817238

Table 5.2: Accessions for all ChIP-seq data used in this chapter.

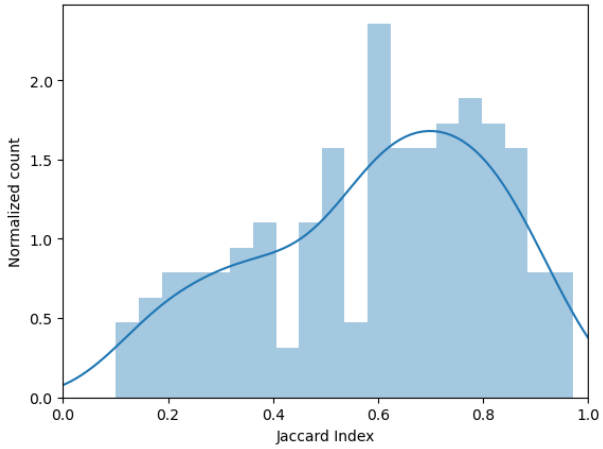


Figure 5.2: Normalized histogram of JI values across all 12 cell types and 6 objective functions.

occupancy. We find that the JI values follow a roughly bimodal distribution (Figure 5.2) with many high values indicating significant agreement between TAD sets, and another cluster of very low values, indicating significantly different TAD sets.

5.3.1 Variability within cell types

The level of variability we observe in TAD sets is highly dependent on the cell type: while several showed consistency, others displayed high levels of variation. Four cell types in particular, A549, HFF-c6, NHEK, and Skeletal Muscle, showed little difference across their TAD sets (Figure 5.3). In each of these cell types, only one JI value lies below 0.5: between the TADs optimized for a large inter- versus intra-TAD difference, and either the TADs optimized for CTCF binding sites (A549, HFF-c6, and NHEK) or for the CTCF ChIP-seq data (Skeletal Muscle). On the other hand, several cell types resulted in much lower JI values across their TAD sets, with IMR90 and LNCaP averaging below 0.5, and hESC and Spleen averaging at or below 0.55 (Figure 5.4).

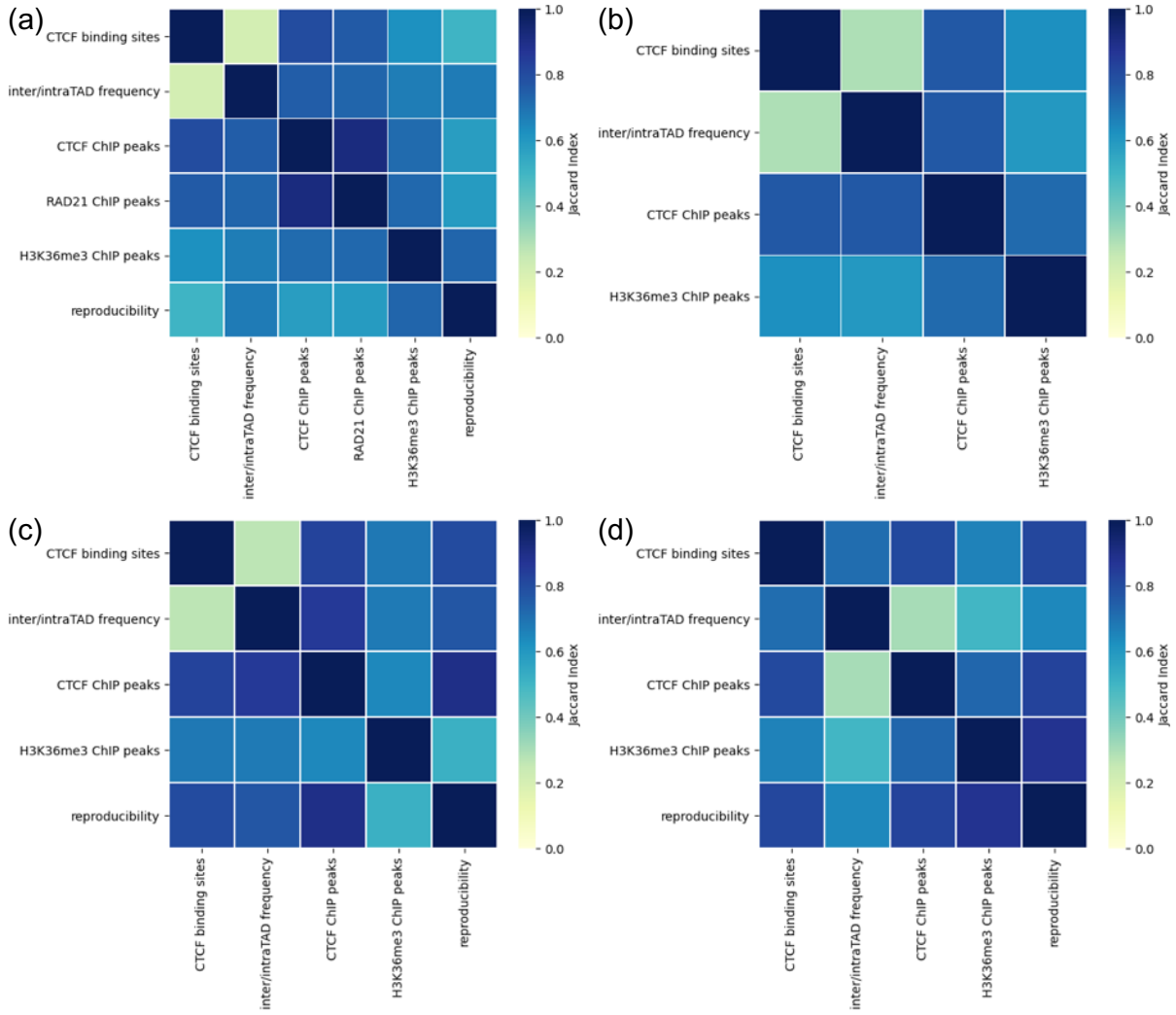


Figure 5.3: Low TAD set variability for various objective functions. Cell types: **(a)** A549 **(b)** NHEK **(c)** HFF-c6 **(d)** Skeletal Muscle tissue.

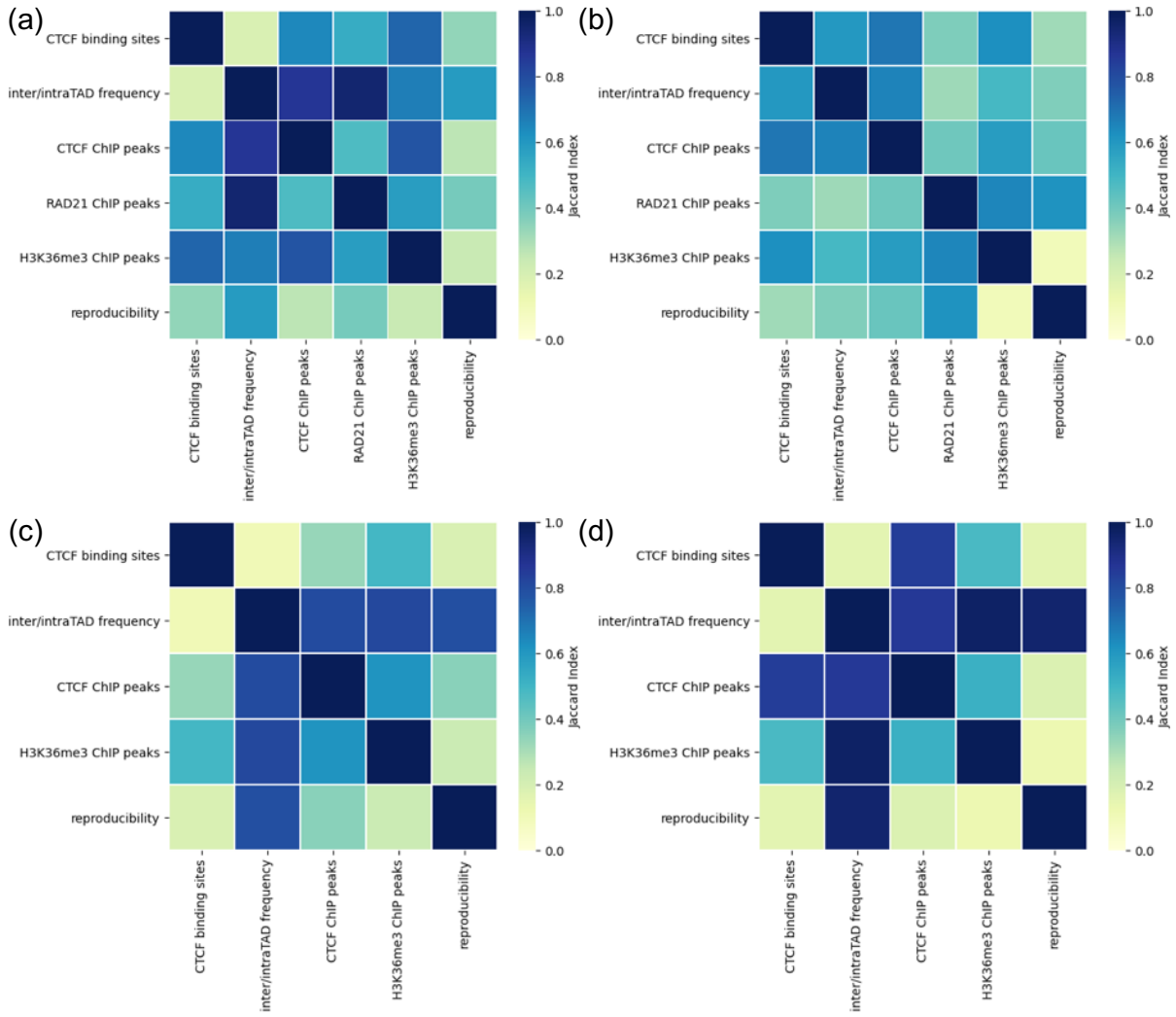


Figure 5.4: High TAD set variability for various objective functions. Cell types: **(a)** hESC **(b)** IMR90 **(c)** LNCaP-FGC **(d)** Spleen tissue.

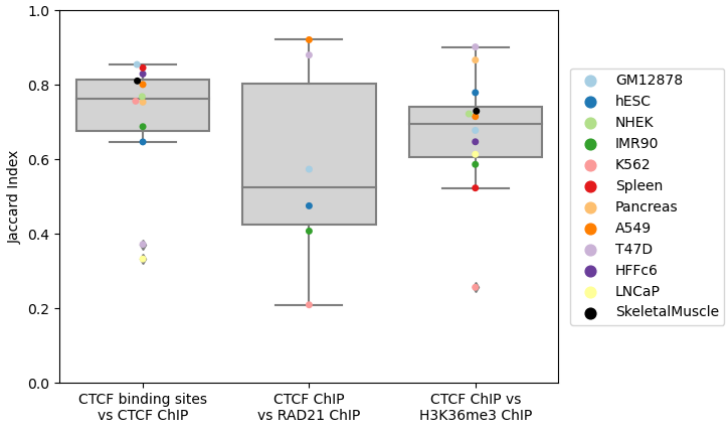


Figure 5.5: Variability of TAD sets identified with ChIP-seq or binding site data.

5.3.2 Relationships between protein-based objective functions

Optimizing for TADs with high peaks of CTCF or H3K36me3 generally leads to very similar TAD sets, but the relationship is muddled for RAD21, the cohesin subunit protein. For all but two cell types (T47D and LNCaP-FGC), the CTCF binding sites appear to be a good proxy for CTCF ChIP-seq data: TAD sets optimized for each tend to be very similar (Figure 5.5). Similarly, choosing parameters with CTCF ChIP-seq and H3K36me3 ChIP-seq result is generally similar TAD sets, with the exception of the Pancreas cells. Though CTCF and cohesin are believed to work together in TAD formation [49], the relationship between TADs identified by CTCF ChIP-seq and RAD21 ChIP-seq is much less clear. While two cell types (A549 and T47D) show very high similarity between TAD sets, the other four cell types have similarity values below 0.6. One possible explanation comes from the loop extrusion model, which suggests that cohesin binds to DNA to create TADs, but falls off when it hits barrier elements, believed to be CTCF. Under this hypothesis, we would not expect CTCF and cohesin to be frequently bound at the same genomic locations.

5.3.3 High variability in TADs from some objective functions

Looking at the variation across cell types of each objective function, we often see more variability than within a single cell type. Comparing TAD sets obtained by optimizing parameters for reproducibility, for example, shows completely different distributions in similarity values for each cell type (Figure 5.6). In some cell types (GM12878, K562, HFF-c6, and Skeletal Muscle), TAD sets optimized for reproducibility are generally similar to those optimized for other biological properties, reflected in the higher JI value distributions. In contrast, hESC, IMR90, and Spleen cells produce low JI values between TADs selected by optimizing for reproducibility and those optimized for other objectives, suggesting that sometimes these objectives are not maximized by the same or even similar TAD sets. While the JI values vary significantly between cell types, in most cases the agreement between TADs identified for high reproducibility and those selected for high inter- versus intra-TAD difference is fairly high (green points in Figure 5.6), suggesting that these two objectives, reproducibility and high inter- versus intra-TAD difference, are generally compatible. On the other hand, the TAD sets given by parameters optimized for high reproducibility seem fairly incompatible with those given by parameters optimized for H3K36me3 peaks, shown by the lower values of the orange points in Figure 5.6.

The range of similarity values between TAD sets optimized for inter- and intra-TAD frequency difference versus other objective functions also reflects significant variability. For most cell types, there was very little overlap in TADs produced by optimizing for CTCF binding sites and optimizing for inter- vs intra-TAD interaction differences (blue points in Figure 5.7). Despite poor agreement with CTCF binding sites, the agreement with CTCF ChIP-seq peaks tends to be much higher, possibly suggesting that binding sites are insufficient to capture both true CTCF occupancy and expected contact distribution in TAD sets.

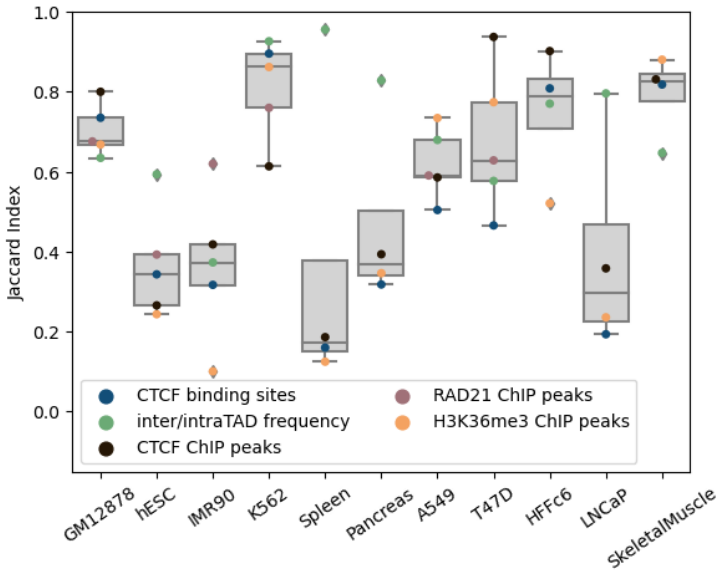


Figure 5.6: Similarity values between TAD sets optimized for reproducibility and all 5 other objective functions across cell types.

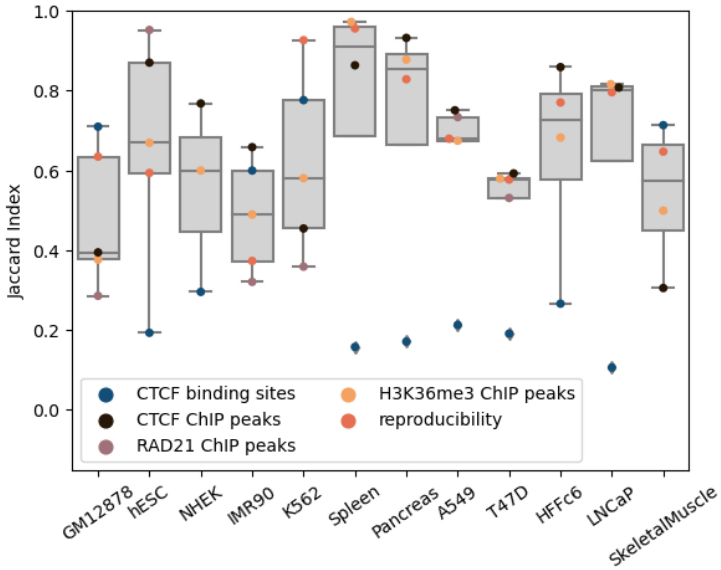


Figure 5.7: Similarity values between TAD sets optimized for high inter- and intra-TAD contact difference and all 5 other objective functions across cell types.

5.4 Discussion

We study the space of computational TAD predictions, and the relationships of TAD properties with each other, by tuning the parameters of a flexible TAD finder model according to a variety of data-driven properties. By choosing TADs specifically to optimize a desired outcome, and comparing them with TADs chosen to optimize other outcomes, we gain insight into the tradeoffs inherent in computational TAD finders. These tradeoffs seem to vary by cell type, with some cell types showing little difference in TAD sets selected for various properties while others show extreme differences. We find that optimizing the most basic TAD property, higher interaction counts within TADs than between TADs, tends to return very different TADs than optimizing for any of the associated biological properties such as high CTCF or RAD21 occupancy at TAD boundaries. This suggests an inherent tradeoff between TAD properties defined solely by the Hi-C data, and those associated with other data types. These results may also suggest different families of TADs based on different properties, rather than a single TAD definition accommodating all expected biological and computational TAD properties at once.

Chapter 6

Discussion and Conclusions

The work presented here has advanced our understanding and ability to study the variability and complexity of three-dimensional chromosome structure through the development of new algorithms, the application of algorithms from other fields, and extensive data analysis.

We opened up the ability to perform rigorous analyses of TAD structure differences between samples by developing the first algorithm to quantify similarity of any two TAD sets, as described in Chapter 2. This method called TADsim not only gives a measure of overall similarity between samples, but identifies regions of significantly similar TAD structures. Unlike general metrics such as the Jaccard Index, TADsim measures TAD similarity by looking at the size of the overlap between TADs, rather than focusing only on boundary locations. We therefore provide a more meaningful measurement of TAD similarity based on the full TADs, as well as finding where in the genome these similar regions exist. Using TADsim, we showed how much TAD structures can vary both within and across chromosomes, and explored more closely the nature of TAD disruption in cancer cell lines. Our results suggest that cancer cell lines do not show widespread TAD structure disruption, but rather localized differences near highly mutated cancer genes. Further study of this result looking at the TAD structures around regions known to be highly mutated, possibly with FISH experiments to validate these predictions, would be necessary to confirm this hypothesis.

In Chapter 3, we applied TADsim, along with another measure of TAD similarity and one of Hi-C similarity, to perform the largest study of chromosomal structural variability across Hi-C samples to date. We showed that despite the claim that TADs are high conserved across cell types and tissues [36, 98, 109], there is in fact significant variability between samples, and even some variability between replicates. This high level of variability has since been supported by single cell studies [19, 122] showing that TADs are not consistent, fixed structures, but vary even within a cell population. We were able to compare TAD sets from tissue samples drawn from different donors, as well as TADs from family members to show that the genetic differences across individuals, and the genetic similarities between parents and children, do not impact the similarities of the resulting TAD sets. Our results suggest that TADs are more similar based on biological function rather than underlying genetics. We also studied some sources of technical variation in Hi-C data and determined that the choice of protocol - *in situ* or dilution Hi-C - seems to have an effect on the resulting TAD sets, so one should be careful comparing samples from differing protocols. On the other hand, lab of origin and restriction enzyme do not seem to induce significant technical variation in TAD sets. All results were shown to be robust to resolution and expected TAD size, though only the Armatus TAD finder was used so it is still unclear whether they hold for all TAD finding tools.

In addition to developing and applying our own algorithms, we also adapted a method from biophysics to study chromosome dynamics in Chapter 4. The Gaussian Network Model (GNM) is a widely used method for inferring protein dynamics from contact maps. Applying the GNM to Hi-C data, we demonstrated that the dynamics predicted by this model correlate highly with two experimental techniques for measuring chromatin accessibility, robust to the data resolution. Dynamic domains identified by GNM additionally matched up well with TADs and compartments. ChIA-PET data, which measures long distance interactions between pairs of loci, also validated high interaction counts between loci predicted by GNM to have high cross-correlation values. GNM additionally permitted us to identify extremely long range dynamically coupled regions

that had not been seen directly from Hi-C matrices. We showed that these regions contain gene pairs with higher co-expression values than expected for gene pairs of similar genomic separation distance, suggesting that the dynamic coupling of these regions could play a regulatory role for the genes they contain.

Finally, using a combination of method development and analysis, we study the variability of computationally-derived TAD sets optimized for various data-driven properties in Chapter 5. We designed an algorithm to be a flexible computational TAD finder combining the primary ideas behind several successful TAD finders to cover much of the space of computational TAD finding. We then tune the parameters to this model according to several properties used to assess the quality of TAD sets, quantifing the variability they induce and compare the resulting TADs. Consistent with the loop extrusion model which suggests that cohesin falls off the DNA strand after TAD formation, using RAD21, a component of the cohesin complex, to optimize TADs leads to very different results than CTCF. We see significant variability of TAD sets within some cell types and across different objective functions, supporting the tradeoffs observed in TAD assessment from various TAD finding methods.

Many of the limitations of this work have been touched on in earlier chapters, but they are important to reiterate. All of the Hi-C data used for this dissertation is bulk Hi-C, meaning it is an aggregation of a large population of cells. Recent work has shown significant TAD variability within cell populations [19, 82, 121], so all of our conclusions hold only on population averages, not individual cells. Because of the relative cost and complexity of the Hi-C experiment, we were restricted to a limited amount of publicly available data. The amount of available data limited the types of methods we could use as well as the types of questions we could ask. For example, in comparing the TAD structures of cancer cell lines to normal cell lines, we did not have enough samples from the same cancer types to be able to identify cancer-specific structural disruptions, but instead had to focus on pan-cancer genes and treat all cancer samples as one group. We also focus largely on TADs here, despite some ongoing debate in the community about their

meaning and importance [13]. TADs are certainly not the only meaningful structural component of chromosomes: compartments, loops, lamina-associating domains (LADs), subTADs, and nuclear speckles have all additionally been shown to play important roles in genomic architecture and its contribution to gene regulation, and there are very likely other important structures both known and unknown.

Perhaps the greatest limitation to all work in the Hi-C field currently is the challenge of validation. Hi-C is based on an indirect measure (proximity ligation counts) of an extremely complex system, and any hypothesis drawn from this data is difficult to independently confirm. The first study to image chromosome structure across the full genome was just published in August 2020 [122], but there is currently no gold standard TAD set to compare a TAD finders' results, or even a clear consensus biological or computational definition of a TAD. We therefore rely on indirect validation measures such as comparisons with other genomic data (ATAC-seq, DNase-seq, ChIP-seq, etc), or quantifying properties we expect to see from TADs, such as higher interaction frequencies within TADs than between them. All of these validation measures are far from perfect, but as experimental techniques for studying chromosome structure improve and diversify, it will become easier to test these computational predictions.

The recent advances in imaging technology are paving the way for exciting new computational methods combining existing Hi-C and scHi-C with imaging data to greatly expand our understanding of chromosome structure. With imaging and Hi-C data from the same cells, models can be trained to infer accurate 3D distances from scHi-C data, and to deconvolve bulk Hi-C into reasonable sub-populations of single cell structures. Imaging data can provide gold standard TAD sets on which to evaluate Hi-C based TAD finding, and perhaps suggest better features to optimize or even provide a more concrete definition of a TAD. While imaging data is a more direct, accurate measure of chromosome structure, it is extremely expensive and much less widely available than Hi-C. Until this changes, Hi-C will still be a critical component in the study of chromosome structure, but we should be able to leverage the strengths of both data types to

develop a greater understanding of this complex system.

Many open questions remain in this field; we have so far only scratched the surface of chromosome structure study. With the development of more experimental techniques for measuring genomic architecture, computational techniques for integrating these results to provide a more complete picture of chromosome structure will need to be developed. We will need new methods to handle the combination of increased sparsity and increased numbers of single cell measurements. Methods, both experimental and computational, will need to be developed to understand the changes over time of TAD structures, which could clarify their role in gene and cell cycle regulation.

As single cell techniques improve and data becomes more widely available, the meaning of the TAD-like structures identified in single cells and their variability will be an important topic to study. Are TAD structures highly variable across the population because their exact locations are not very important to cellular processes, or are they variable precisely because they are important and therefore must be responsive to the exact state of the cell? Integrating this data with bulk Hi-C could lead to better deconvolution methods, allowing inferred subpopulations from bulk Hi-C data to augment the existing single cell data and improve predictive power.

It is critical to reconcile the seemingly contradictory results from studies that show little to no changes in transcription from structural disruptions and those that show dramatic phenotypic changes from relatively minor structural changes. Whether the explanation is that certain genomic regions are less robust than others to these disruptions, or that there is a confounding factor we have yet to consider, it is impossible to understand the true role chromosome structure plays in gene regulation without understanding why it sometimes appears necessary and other times meaningless. This understanding is likely to come from a combination of statistical methods to uncover the underlying patterns of changes that disrupt transcription and those that do not, and experiments to test these computational predictions.

The methods, techniques, and ideas presented in this dissertation will inform the study of chro-

mosome structure as data availability and quality improves. The information we can draw from comparing samples, along with the statistical power to trust this information, is greatly enhanced with larger numbers. The dynamics of chromosome structures will continue to be studied with methods drawn from other fields and new experiments. A more complete TAD definition provides a framework for understanding and studying these structures as our statistical power improves. In order to see the full picture of the genomic architecture and the processes it influences, we will need to continue using all of the approaches mentioned here: new method development, rigorous data analysis, adaptation of existing methods, and careful, quantifiable definitions of the objects of study.

Bibliography

[1] Kadir C Akdemir, Victoria T Le, Sahaana Chandran, Yilong Li, Roel G Verhaak, Rameen Beroukhim, Peter J Campbell, Lynda Chin, Jesse R Dixon, and P Andrew Futreal. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics*, 52(3):294–305, 2020.

[2] Charles J Alpert and So-Zen Yao. Spectral partitioning: the more eigenvectors, the better. In *Proceedings of the 32nd annual ACM/IEEE Design Automation Conference*, pages 195–200, 1995.

[3] Charles J Alpert, Andrew B Kahng, and So-Zen Yao. Spectral partitioning with multiple eigenvectors. *Discrete Applied Mathematics*, 90(1-3):3–26, 1999.

[4] Anthony T. Annunziato. DNA packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008.

[5] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[6] Ferhat Ay and William S Noble. Analysis methods for studying the 3D architecture of the genome. *Genome Biology*, 16(1):183, 2015.

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

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

[9] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[10] Ivet Bahar, Anders Wallqvist, David G. Covell, and Robert L. Jernigan. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry*, 37(4):1067–1075, 1998.

[11] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual Reviews Biophysics*, 2010.

[12] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Research*, 41(D1):D991–D995, 2012.

[13] Jonathan A Beagan and Jennifer E Phillips-Cremins. On the existence and functionality of topologically associating domains. *Nature Genetics*, pages 1–9, 2020.

[14] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee CA Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017.

[15] Paweł Bednarz and Bartek Wilczyński. Supervised learning method for predicting chromatin boundary associated insulator elements. *Journal of Bioinformatics and Computa-*

tional Biology, 12(06):1442006, 2014.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[17] Alessandro Bertero, Paul A Fields, Vijay Ramani, Giancarlo Bonora, Galip G Yardimci, Hans Reinecke, Lil Pabon, William S Noble, Jay Shendure, and Charles E Murry. Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. *Nature Communications*, 10(1):1–19, 2019.

[18] Wendy A Bickmore and Bas van Steensel. Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284, 2013.

[19] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), 2018.

[20] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678, 2016.

[21] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[22] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12): 1213, 2013.

[23] Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nature*

Structural and Molecular Biology, 20(3):290–299, 2013.

[24] Abhijit Chakraborty and Ferhat Ay. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics*, 34(2):338–345, 2017.

[25] Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4D Nucleome. *Proceedings of the National Academy of Sciences*, 112(26): 8002–8007, 2015.

[26] Jie Chen, Alfred O Hero III, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 2016.

[27] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K Brinkman, Stephen A Adam, Robert Goldman, Bas van Steensel, Jian Ma, and Andrew S Belmont. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *The Journal of Cell Biology*, 217(11):4025–4048, 2018.

[28] Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R Lajoie, Bayly S Wheeler, Edward J Ralston, Satoru Uzawa, Job Dekker, and Barbara J Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559): 240–244, 2015.

[29] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301, 2001.

[30] Rola Dali and Mathieu Blanchette. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*, 45(6):2994–3005, 2017.

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

[32] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'Shea, Peter J Park, Bing Ren, et al. The 4D

nucleome project. *Nature*, 549(7671):219–226, 2017.

[33] Alexandra Despang, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing-Lee Chan, Bernd Timmermann, Lars Wittler, Martin Vingron, et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics*, 51(8):1263–1271, 2019.

[34] Michele Di Pierro, Ryan R Cheng, Erez Lieberman Aiden, Peter G Wolynes, and José N Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, 114 (46):12126–12131, 2017.

[35] Vishnu Dileep, Ferhat Ay, Jiao Sima, Daniel L Vera, William S Noble, and David M Gilbert. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Research*, 25:1104–1113, 2015.

[36] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[37] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539): 331–336, 2015.

[38] Jesse R Dixon, David U Gorkin, and Bing Ren. Chromatin domains: the unit of chromosome organization. *Molecular Cell*, 62(5):668–680, 2016.

[39] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for

mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, 2006.

[40] Geet Duggal, Hao Wang, and Carl Kingsford. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Research*, 42(1):87–96, 2014.

[41] Eran Eyal, Gengkon Lum, and Ivet Bahar. The anisotropic network model web server at 2015 (ANM 2.0). *Bioinformatics*, 31(9):1487–1489, 2015.

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

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

[44] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, 14(7):679–685, 2017.

[45] James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L Moore, Dorothee CA Kraemer, Stuart Aitken, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*, 11(12):852, 2015.

[46] James Fraser, Iain Williamson, Wendy A Bickmore, and Josée Dostie. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, 79(3):347–372, 2015.

[47] Paula Freire-Pritchett, Stefan Schoenfelder, Csilla Várnai, Steven W Wingett, Jonathan

Cairns, Amanda J Collier, Raquel García-Vílchez, Mayra Furlan-Magaril, Cameron S Osborne, Peter Fraser, et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife*, 6:e21926, 2017.

[48] Geoff Fudenberg, Gad Getz, Matthew Meyerson, and Leonid A. Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature Biotechnology*, 29(12):1109–1113, 2011.

[49] Geoffrey Fudenberg, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A Mirny. Emerging evidence of chromosome folding by loop extrusion. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 82, pages 45–55. Cold Spring Harbor Laboratory Press, 2017.

[50] Yad Ghavi-Helm, Aleksander Jankowski, Sascha Meiers, Rebecca R Viales, Jan O Korbel, and Eileen EM Furlong. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, 51(8):1272–1282, 2019.

[51] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Physical Review Letters*, 79(16):3090, 1997.

[52] Anders S Hansen, Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, 6: e25776, 2017.

[53] Nastaran Heidari, Douglas H Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q Zhang, and Michael P Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Research*, 24(12):1905–1917, 2014.

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

[55] Chunhui Hou, Li Li, Zhaohui S Qin, and Victor G Corces. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular Cell*, 48(3):471–484, 2012.

[56] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[57] Jialiang Huang, Eugenio Marco, Luca Pinello, and Guo-Cheng Yuan. Predicting chromatin organization using histone marks. *Genome Biology*, 16(1):162, 2015.

[58] Daniel M Ibrahim and Stefan Mundlos. The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. *Current Opinion in Genetics & Development*, 61:1–8, 2020.

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

[60] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013.

[61] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471): 333, 2013.

[62] Rieke Kempfer and Ana Pombo. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, pages 1–20, 2019.

[63] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

[64] Yuichi Kodama, Martin Shumway, and Rasko Leinonen. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):D54–D56, 2012.

[65] François Le Dily, Davide Baù, Andy Pohl, Guillermo P Vicent, François Serra, Daniel Soronellas, Giancarlo Castellano, Roni HG Wright, Cecilia Ballare, Guillaume Filion, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, 28(19):2151–2162, 2014.

[66] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research*, 39: D19–D21, 2010.

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

[68] Hongchun Li, Yuan-Yu Chang, Lee-Wei Yang, and Ivet Bahar. iGNM 2.0: the Gaussian network model database for biomolecular structural dynamics. *Nucleic Acids Research*, 44(D1):D415–D422, 2016.

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

[70] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer

interactions. *Cell*, 161(5):1012–1025, 2015.

[71] Darío G Lupiáñez, Malte Spielmann, and Stefan Mundlos. Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genetics*, 32(4):225–237, 2016.

[72] Jennifer M Luppino and Eric F Joyce. Single cell analysis pushes the boundaries of TAD formation and function. *Current Opinion in Genetics & Development*, 61:25–31, 2020.

[73] Hongqiang Lyu, Erhu Liu, and Zhifang Wu. Comparison of normalization methods for Hi-C data. *BioTechniques*, 68(2):56–64, 2020.

[74] Claire Marchal, Jiao Sima, and David M Gilbert. Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology*, pages 1–17, 2019.

[75] Karen J Meaburn, Prabhakar R Gudla, Sameena Khan, Stephen J Lockett, and Tom Misteli. Disease-specific gene repositioning in breast cancer. *The Journal of Cell Biology*, 187(6):801–812, 2009.

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

[77] Leonid A Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, 2011.

[78] Tom Misteli. Higher-order genome organization in human disease. *Cold Spring Harbor Perspectives in Biology*, 2(8):a000794, 2010.

[79] Benoit Moindrot, Benjamin Audit, Petra Klous, Antoine Baker, Claude Thermes, Wouter de Laat, Philippe Bouvet, Fabien Mongelard, and Alain Arneodo. 3D chromatin conformation correlates with replication timing and is conserved in resting cells. *Nucleic Acids Research*, 40(19):9470–9481, 2012.

[80] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[81] Takashi Nagano, Csilla Várnai, Stefan Schoenfelder, Biola-Maria Javierre, Steven W Wingett, and Peter Fraser. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biology*, 16(1):175, 2015.

[82] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, 2017.

[83] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485 (7398):381–385, 2012.

[84] Elphège P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.

[85] Heidi K Norton, Daniel J Emerson, Harvey Huang, Jesi Kim, Katelyn R Titus, Shi Gu, Danielle S Bassett, and Jennifer E Phillips-Cremins. Detecting hierarchical genome folding with network modularity. *Nature Methods*, 15(2):119–122, 2018.

[86] Johannes Nuebler, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706, 2018.

[87] Donald E Olins and Ada L Olins. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology*, 4(10):809–814, 2003.

[88] Chin-Tong Ong and Victor G Corces. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4):234, 2014.

[89] Horng D Ou, Sébastien Phan, Thomas J Deerinck, Andrea Thor, Mark H Ellisman, and Clodagh C O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349):eaag0025, 2017.

[90] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14 (4):417–419, 2017.

[91] Martin Pelikan, David E Goldberg, Erick Cantú-Paz, et al. BOA: The Bayesian optimization algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99*, volume 1, pages 525–532. Citeseer, 1999.

[92] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527): 402–405, 2014.

[93] ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[94] Sofia A Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M Lai, Alexander A Shishkin, Prashant Bhat, Yodai Takei, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3):744–757, 2018.

[95] Indika Rajapakse and Steve Smale. Mathematics of the genome. *Foundations of Computational Mathematics*, 17(5):1195–1217, 2017.

[96] Vijay Ramani, Darren A Cusanovich, Ronald J Hause, Wenxiu Ma, Ruolan Qiu, Xinxian Deng, C Anthony Blau, Christine M Disteche, William S Noble, Jay Shendure, et al. Mapping 3D genome architecture through in situ DNase Hi-C. *Nature Protocols*, 11(11): 2104–2121, 2016.

[97] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

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

[99] Suhas SP Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M Engreitz, Elizabeth M Perez, Kyong-Rim Kieffer-Kwon, Adrian L Sanborn, Sarah E Johnstone, Gavin D Bascom, Ivan D Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2): 305–320, 2017.

[100] Judhajeet Ray, Paul R Munn, Anniina Vihervaara, James J Lewis, Abdullah Ozer, Charles G Danko, and John T Lis. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proceedings of the National Academy of Sciences*, 116(39):19431–19439, 2019.

[101] Sarah Rennie, Maria Dalby, Lucas van Duin, and Robin Andersson. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nature Communications*, 9(1):487, 2018.

[102] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1): 1–16, 2011.

[103] Sushmita Roy, Alireza Fotuhi Siahpirani, Deborah Chasman, Sara Knaack, Ferhat Ay, Ron Stewart, Michael Wilson, and Rupa Sridharan. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Research*, 43(18):8694–

8712, 2015.

[104] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C Schulz, Allan J Robins, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6):761–770, 2010.

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

[106] Natalie Sauerwald and Carl Kingsford. Quantifying the similarity of topological domains across normal and cancer human cell types. *Bioinformatics*, 34(13):i475–i483, 2018.

[107] Natalie Sauerwald, She Zhang, Carl Kingsford, and Ivet Bahar. Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic Acids Research*, 45(7):3663–3673, 2017.

[108] Natalie Sauerwald, Akshat Singhal, and Carl Kingsford. Analysis of the structural variability of topologically associated domains as revealed by Hi-C. *NAR Genomics and Bioinformatics*, 2(1):lqz008, 2020.

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

[110] Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv*, page 103614, 2017.

[111] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H Haering, Leonid Mirny, et al. Two independent modes of chromatin organization revealed by co-

hesin removal. *Nature*, 551(7678):51–56, 2017.

[112] Emre Sefer and Carl Kingsford. Semi-nonparametric modeling of topological domain formation from epigenetic data. *Algorithms in Bioinformatics*, pages 148–161, 2015.

[113] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, 16(1):259, 2015.

[114] Tom Sexton, Heiko Schober, Peter Fraser, and Susan M Gasser. Gene regulation through nuclear organization. *Nature Structural & Molecular Biology*, 14(11):1049–1055, 2007.

[115] Tom Sexton et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–72, 2012.

[116] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, 44(7):e70–e70, 2016.

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

[118] Lingyun Song and Gregory E Crawford. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):1118–1129, 2010.

[119] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3D genome. *Nature Reviews Genetics*, 19:453–467, 2018.

[120] John C Stansfield, Kellen G Cresswell, Vladimir I Vladimirov, and Mikhail G Dozmorov. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, 19(1):279, 2018.

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

[122] Jun-Han Su, Pu Zheng, Seon S Kinrot, Bogdan Bintu, and Xiaowei Zhuang. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, 2020.

[123] James H Sun, Linda Zhou, Daniel J Emerson, Sai A Phyo, Katelyn R Titus, Wanfeng Gong, Thomas G Gilgenast, Jonathan A Beagan, Beverly L Davidson, Flora Tassone, et al. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell*, 175(1):224–238, 2018.

[124] Florence Tama and Y-H Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14(1):1–6, 2001.

[125] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[126] Maria Tsompana and Michael J Buck. Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, 7(1):1–16, 2014.

[127] Bas van Steensel and Andrew S Belmont. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791, 2017.

[128] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, 2016.

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

[130] Zheng Xu, Guosheng Zhang, Cong Wu, Yun Li, and Ming Hu. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics*,

32(17):2692–2695, 2016.

[131] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43 (11):1059, 2011.

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

[133] Galip Gürkan Yardımcı, Hakan Ozadam, Michael EG Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R Lajoie, Fan Song, et al. Measuring the reproducibility and quality of Hi-C data. *Genome Biology*, 20(1):57, 2019.

[134] Bin Zhang and Peter G Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067, 2015.

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

[136] Jingyu Zhang, Hengyu Chen, Ruoyan Li, David A Taft, Guang Yao, Fan Bai, and Jianhua Xing. Spatial clustering and common regulatory elements correlate with coordinated gene expression. *PLoS Computational Biology*, 15(3):e1006786, 2019.

[137] Ye Zheng and Sündüz Keleş. FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation. *Nature Methods*, 17(1):37–40, 2020.

[138] Anne Zirkel, Milos Nikolic, Konstantinos Sofiadis, Jan-Philipp Mallm, Chris A Brackley, Henrike Gothe, Oliver Drechsel, Christian Becker, Janine Altmüller, Natasa Josipovic,

et al. HMGB2 loss upon senescence entry disrupts genomic organization and induces CTCF clustering across cell types. *Molecular Cell*, 70(4):730–744, 2018.

[139] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19(1):217, 2018.