Three-Dimensional Chromosome Organization in Eukaryotes: Novel Computational Approaches

BY

GAMZE GÜRSOY
B.S., Bogazici University, Istanbul, Turkey, 2008

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Thesis Committee:

Jie Liang, Chair and Advisor
Yang Dai
Ao Ma
John L Nitiss, Biopharmaceutical Sciences
Bradley Merrill, Biochemistry and Molecular Genetics

This thesis is dedicated to mom and dad, who sacrificed a lot for me, and to my beautiful sisters...

ACKNOWLEDGMENTS

The work in this thesis would not be possible without the encouragement, support and guidance of many wonderful people. I want to start conveying my sincere appreciation by thanking my advisor Prof. Jie Liang for his limitless support, patience, mentorship, and many lessons he taught over the years. I will always be inspired by his ability to think critically and creatively, and his far-reaching intellectual curiosity. I am also extremely thankful to the members of my thesis committee: Drs Yang Dai, Ao Ma, John Nitiss, and Bradley Merill for taking the time to read my hundreds of pages of dissertation and for their insightful feedback.

I am incredibly lucky to have parents who created an environment that nurtured creativity and scientific curiosity in me. They answered my never-ending first scientific questions and encouraged me to ask for more. With all the love and the fun they offered at home, they also gave me something to look forward to every holiday. I love you Mom, Dad, Dilos, and Fatos...

I came to Chicago with a sole purpose of earning my Ph.D. degree. Little that I know I would meet my life partner in this journey. I am completely out of words to describe the immense gratitude and appreciation I feel for my partner in crime, my teammate, my science buddy, my art partner, my mentor and mentee, my best friend: Matthew Vincent "Trikey" Macellaio. Without his generosity, understanding, patience and support, this thesis would never happen.

When I arrived at the Liang lab, I had the privilege of crossing paths with many insightful researchers, Drs. Yun Xu, Joe Dundas, Youfang Cao, Hsiao-Mei Lu, Hammad Naveed, David Jimenez-Morales, Larissa Adamian, and Ke Tang. I am very thankful for their scientific

comments, discussions and suggestions. I specifically thank Dr. Yun Xu for being my project and office-mate and helping me make this thesis happen. We are together proud founding members of the Liang Lab chromatin project. I am also very thankful to my fellow colleagues Jieling Zhao, Meishan Lin, Will Tian, Anna Terebus, and Alan Perez-Rathke for their scientific comments, discussions, and suggestions, but also for the priceless daily help and all the fun moments.

I also want to thank my colleagues from the remaining groups in the Bioinformatics program, especially to Drs. Matt Carson and Georgi Genchev. I greatly appreciate their friendship, guidance, and help. I also thank Drs. Damian Roqueiro and Morten Kallberg, whose scientific advice and support helped me a lot.

Over the many collaborations in graduate school, Dr. Amy Kenter served as a second mentor. I greatly appreciate all the help she has offered. I benefited from her extensive knowledge of biology, enthusiasm for science and patience while teaching me experimental aspects of chromatin organization. I am also very thankful to Dr. Bhaskar DasGupta, a scientist full of wisdom, for inviting me to collaborate with his lab. I learned how to express myself to people from different backgrounds from this collaboration.

I was fortunate to find environments for my scientific training long before graduate school started. When I was an undergraduate student at Bogazici University, I had the privilege of having a caring advisor, Dr. Mehmet Cihan Camurdan, who took the time to teach me how to formulate chemistry and biology problems in the forms of differential equations. I was also lucky to be an undergraduate member of the Polymer Research Center, where I was trained by

Dr. Turkan Haliloglu, who opened the doors of amazing computational biology for me. I am incredibly grateful to her for planting the seeds of this thesis years ago and helping me to get where I am right now.

I would like to thank my many ultimate Frisbee friends, especially every past and present members of [Moose], who have helped keep me sane during graduate school. They provided an outlet for competitiveness on the field, and a fun environment with countless slices of pizza and bottles of beers off the field. They were also unlucky enough to listen to my limitless rants about graduate school. I want to thank my gym coaches Franco and Drew for keeping me healthy and sane and creating an intellectual and clean environment for us to work out. I thank all my gym buddies, proud members of Crossfit Commitment, for throwing hundreds of pounds of weights around with me that greatly helped to burn frustrations of graduate school. I specifically would like to thank Matthew MacDougall, a fellow UIC graduate student and CFC member, for insightful scientific discussions about chromatin and stem cells during our lifting sessions.

I'm forever in debt to Matt's family, who gave me a home away from home. I can not imagine a Christmas without any of the Millers. I thank them all and their dear friends the Meisels for providing the warmth of a family, when I was away from mine.

During my last summer in Chicago, I was a part of not-so-secret, not-so-underground bike gang called Viceroys! I would like to thank Asta, Chris, Matt, Anne and Buckley for the miles of rides, bottles of beers, and unnecessary amounts of carbs.

Last but not the least, I would like to thank all my Turkish friends, who held my hands when I first move to Chicago and showed me how to survive in cold. I specifically thank Dr. Atilla Soner Balkir, who taught me how to write codes and introduced me to the magical world of computer science. He was a big support since the second year of college, and I still use the tricks and the libraries he taught me, when I code.

Contribution of Authors

Chapter 1 is a literature review that places my dissertation question in the context of the larger field and highlights the significance of my research question. **Chapter 2** represents a published manuscript for which I was the primary author and major driver of the research. This chapter has been done in collaboration with former student of Liang lab, Dr. Yun Xu. He helped with the development of the chain growth software described in Section 2.2.2. Drs. Amy L Kenter and Jie Liang played large roles in the writing of the manuscript. This chapter is partially based on the publications (Please see appendices for necessary permissions):

- Gürsoy, G., Xun, Y., Kenter, A., Liang, J.: Spatial confinement is a major determinant of folding landscape of human chromosomes. In Nucleic Acids Research, 42(13):8223-30, 2014.
- Gürsoy, G., Xu, Y., Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. In Conf Proc IEEE Eng Med Biol Soc. 3945-8, 2014.

Chapter 3 is partially based on a published manuscript and largely based on a submitted manuscript for which I was the primary author and major driver of the research. This chapter

has been done in collaboration with former student of Liang lab, Dr. Yun Xu. He helped with the development of the software described in Section 3.2. Dr. Jie Liang played a large role in the writing of the manuscript. This chapter is partially based on the publications (Please see appendices for necessary permissions):

- Gürsoy, G., Xu, Y., Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. In Conf Proc IEEE Eng Med Biol Soc. 3945-8, 2014.
- Gürsoy, G., Xu, Y., Liang, J.: Spatial organization of budding yeast genome in cell nucleus and identification of specific chromatin interactions from multi-chromosome constrained chromatin model. Submitted.

Chapter 4 is based on an unpublished manuscript for which I equally contributed with Dr. Yun Xu. Dr. Yun Xu. He helped with the development of the software described in Section 4.2 as well as the part of the Results section. I generated all the figures of this chapter. Drs. Amy L Kenter and Jie Liang played large roles in the writing of the manuscript. This chapter is partially based on the publication:

- Xu, Y., Gürsoy, G., Kenter, A., Liang, J.: Constructing 3D chromatin ensembles and predicting functional interactions of -globin locus from 5C data. In preparation.

Chapter 5 is based on an unpublished manuscript for which I was the primary author and major driver of the research. Arianna Girardi collected the necessary Hi-C data, as well as the epigenetics data required for this Chapter, and helped with the null model. This chapter is partially based on the manuscript:

- Gürsoy, G., Girardi, A., Liang, J.: Computational prediction of chromatin hotspots using n-Constrained Self-Avoiding Chromatin model. In preparation.

In **Chapter 6**, I review the main topics of this dissertation, highlight the novel contribution of my method, strength and weaknesses of my modeling approach. I also provide perspectives on potential future developments.

TABLE OF CONTENTS

CHAPTER **PAGE**

CHAPTER **PAGE**

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

FISH	Fluorescent *in situ* hybridization
3C	Chromosome Conformation Capture
4C	Circular Chromosome Conformation Capture
5C	Chromosome Conformation Capture Carbon Copy
ChIP	Chromatin immunoprecipitation
g-SIS	Geometrical sequential importance sampling
RMSD	Root mean square deviation
NE	Nuclear Envelope
SPB	Spindle Pole Body
C-SAC	Constrained-Self-Avoiding Chromatin
mC-SAC	multi-chromosome Constrained-Self-Avoiding Chromatin
nC-SAC	n Constrained-Self-Avoiding Chromatin
LAD	Lamin associated domain
TAD	Topologically associated domain
FDR	False Discovery Rate
CTCF	CCCTC factor

LIST OF ABBREVIATIONS (Continued)

RNAPII	RNA Polymerase II
RNAPIII	RNA Polymerase III

SUMMARY

Human Genome Project provided new avenues for advances in medicine and biotechnology by revealing the genetic blueprint of human cells. Our understanding on how biological functions are encoded in genome has therefore been largely based on analysis of one-dimensional DNA sequences. This understanding is now rapidly shifting towards more complex, multi-dimensional models of genome upon the discovery of epigenomic profiles of cells through many high-throughput experimental techniques. Important cellular functions such as gene expression and regulation are largely due to the epigenetic changes in our genome. With the success of The Encyclopedia of DNA Elements (ENCODE) project, there is a tremendous amount of information available on the epigenetic properties of human genome across different cell types. While the quantity of this information is increasing rapidly, the overwhelming question becomes what can be learned from this vast trove. Computational 3D models can provide great help in understanding the mechanisms of gene regulations through construction of three-dimensional chromosome structures using this data. However, it is challenging to computationally construct chromosome structures due geometrical difficulties to satisfy constraints derived from experiments. My research has been focusing on applying principles in physics, biology and computer science to address the important question of how chromosomes are confined in the severely small nuclear environment and how cellular functions are influenced by this organization.

During my PhD training, I have developed novel sampling tools to construct spatial structures of chromosomes, to understand fundamental mechanisms of regulation of gene expression

involving the effect of nuclear space in maintaining the epigenetic state of the cell, long-distance DNA loopings that promote cell-specific gene expression, and the main physical factors and mechanisms that determine the genome organization.

Investigation of genome studies must be carried out with the consideration of its physical environment, namely cell nucleus. However, it is unclear how the architecture of cell nucleus and its small dimension affect the organization of chromosomes. This is largely due to the challenges in computational techniques to properly sample three-dimensional structures in severely confined spaces. We developed a geometrical algorithm based on the Importance Sampling technique to generate ensembles of three-dimensional chromosome chains in the severe confinement of cell nucleus. Our model, named Constrained Self-Avoiding Chromatin (C-SAC), showed how experimentally observed physical behavior of human chromosome folding emerges from the confinement of cell nucleus. Our findings further highlight the importance of nuclear size as a potential regulator of epigenetic programming of cells as in the case of transitioning from stem cells to differentiated cells.

Detailed understanding of different cellular states in mammalian cell differentiation requires comprehensive analysis of the interactions in the whole genome. Well-studied budding yeast is an excellent starting system for genome-wide chromatin construction, as its transcription machineries have been shown to be largely influenced by nuclear architecture. We developed a model that uses constraints derived from microscopy experiments to mimic the nuclear environment and its effects on the folding of yeast genome. We showed that the organization of individual chromosomes of yeast genome is dictated by the confinement of the cell nucleus.

The relative organization of chromosomes in 3D space, however, is largely determined by the centromere clustering in nuclear substructures. We also used these computationally captured interactions arising from the polymer effects under the constraints of nuclear architecture to extract biologically specific interactions from experimental data for further investigation of transcription regulatory mechanism in budding yeast.

Chromosome Conformation Capture (3C) and related techniques are remarkable sources for capturing the pairwise chromatin interactions. Experimentally captured interactions are often sparse and incomplete due to the locations of restriction enzyme sites or sequence mappability issues. Random interactions arising from non-specific formaldehyde fixation introduce further complexity. We further improved our method to remove non-specific spatial interactions from the experimental measurements and incorporate remaining specific interactions in our polymer model to study the differential levels of gene expression in different cell lines. We applied our method on folding of α-globin locus in different cells with different expression levels to understand how spatial organization of genome and epigenetic profiles of highly interacting genomic elements affect the expression level of important genes. Our computational modeling revealed insights that there might be a relationship between levels of expression of α-globin genes in different cell lines and the folding landscape of chromatin. We showed that some of the predicted interactions that were not in the original data are shown to have biological importance by other independent studies.

Finally, we studied the minimum required chromatin interactions in a locus to achieve the transcriptional functions of the cell. Identifying interactions that facilitate the formation of

promoter-enhancer interactions is important to understand the role non-coding DNA. Such pairs involved in these interactions are denoted hotspots. Constructing 3D structures and perturbing these structures with virtual mutations can help to identify hotspots that are not obvious from the experimental measurements. I further developed a computational method that enables virtual mutations. I introduce a repulsion between sites of interest and compare the resulting ensemble with the wild type ensemble obtained by using Hi-C interactions as constraints. This helps to identify the hotspots whose interactions facilitate the interactions between promoters and enhancers. Combined with epigenetic profiling data of the CCL locus, my method revealed that structural hotspots of chromatin do not correlate with their interaction frequencies measured by 3C studies. I further showed the importance of CTCF binding on the regulation of interactions between promoters and enhancers and demonstrated that evolutionary conservation of these bindings sites is a major determinant of the importance of chromatin interactions.

Overall, in this thesis, I use computational methods to construct three-dimensional structures of chromatin, both in genome and locus level. I, first, demonstrated the effects of nuclear space on the organization of chromosome and how it dictates the overall scaling properties of genome. I, then, studied the effects and contribution of nuclear landmarks, confinement as well as the biochemical factors on the folding of budding yeast genome. I, furthermore, focused on the detailed spatial structures of α-globin locus in different gene expression levels to study the effect of three-dimensional organization of chromatin on level of gene expression. I, lastly,

defined the minimum structural units of CCL locus in order to achieve necessary promoter-enhancer interactions.

CHAPTER 1

INTRODUCTION

As we continue to seek to understand the origin of life, we now know that DNA (deoxyribonucleic acid) is the essential unit of every living system. It caries the genetic information and passes it down from one generation to the next. DNA encodes the physical characteristics of every living organism such as height, skin color, eye color, etc. Consequently, a detailed understanding of organization of DNA and how it is assembled is essential to gain insights into how living organisms function. In this thesis, you may not be able to find the answers related to origin of life, but, to the most of my ability, I have demonstrated the large-scale organization of this essential unit of life (DNA) and inferred some understanding of the mechanism behind this organization. These findings are obtained through development of novel computational approaches.

This introductory chapter is organized as follows. We begin reviewing the structural organization and fundamental folding principles of DNA. Through a brief historic review, we visit recent achievements in experimental techniques that built the foundations of our current knowledge of organization of this complex macromolecule. We will then revisit the hierarchical organization of DNA in cell nucleus with the aid of available experimental measurements. Finally, we introduce the aims of the research presented here and its significance, concluding with the outline of the remaining chapters of the dissertation.

1.1 From DNA to Chromosome

In a haploid human cell, there are 23 chromosomes, which are about 3.2 billion base pairs long in total. The fully extended length of about 2 m of human DNA is confined into a cell nucleus, which is only 5-20 μm in diameter (Alberts, 2002). DNA is found to be in the form of chromosome in cell nucleus. Chromosome is the higher-order organization of DNA that provides relevant genetic information readily accessible and transmits that information to the subsequent generations. As cycle of cell life is divided into periods, the organization of chromosome in different periods helps with the function of DNA. For example, the organization of chromosome in interphase and metaphase prevents DNA becoming an unmanagable tangle. The organization of chromosome during mitosis further helps transmitting the genetic information to subsequent generations. Most of the metabolic functions of a cell take place during the interphase and a typical cell spends most of its life in interphase. Chromosome organization in interphase also makes relevant genetic information readily accessible. In this dissertation, my studies are related to the organization of chromosome in interphase, which is also called "chromatin".

How is chromosome formed from DNA? There are certain proteins called histones in the eukaryotic nucleus that compact DNA. Histones fold DNA using the energy from electrostatic interactions (Youngson, 2006; Lehninger et al., 2005). The positively charged histones (H1, H2A, H2B, H3, and H4) interact with negatively charged DNA, in such a way that DNA molecule wraps around the histones and form an octomer. The resulting DNA-histones complex is called nucleosome (Youngson, 2006; Lehninger et al., 2005). The eight histone molecules along with 146 basepairs of tightly wrapped DNA form a nucleosome. In detail, H2A, H2B, H3, and

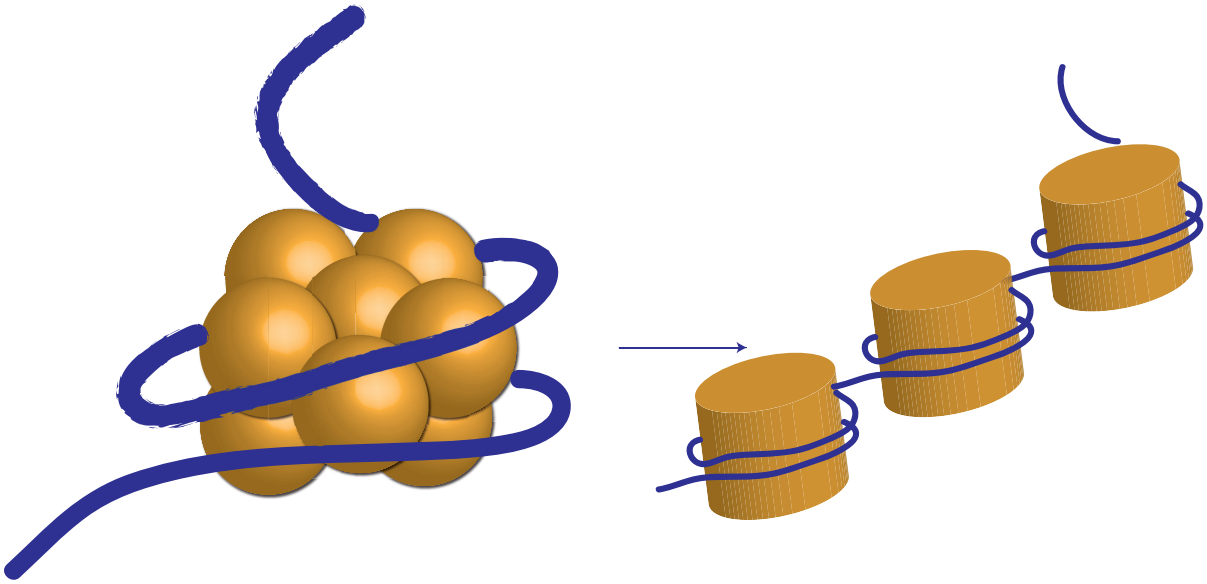

Figure 1. **The formation of nucleosome** DNA (blue ribbon) wraps around eight histone molecules twice and forms a nucleosome. String of beads appearance is also depicted.

H4 histones form an octamer (Lehninger et al., 2005; Hartl et al., 1988). DNA binds to this octomer and wraps 1.7 turns, which is equivalent to 146 basepairs (Figure 1). Each chromosome is formed from thousands of nucleosomes, and these nucleosomes are linked through 20 base pairs long linker DNA. Each chromosome is therefore a collection of nucleosomes with linker DNA in between. This yields the appearance of a string of beads (Figure 1) when viewed using an electron microscope (Farkas and American Association for Clinical Chemistry, 1996). Nucleosomes then fold to form a dense, tightly packed structure, called chromatin fiber (Hartl et al., 1988).

With the completion of human genome project (Lander, 2011), extensive understanding of sequence of human DNA was gained. However, how this information directs gene expression programs and how spatio-temporal changes in the genome take place remain unanswered. Understanding the mechanism behind the nuclear activities is important not only for understanding the small regulatory units but also essential for studying phenotypic variations between cells and the underlying mechanisms behind many human diseases. Growing evidence suggest that the three-dimensional organization of genome along with its sequence orchestrate the unique gene expression machinery in cells.

1.2 Genome in three-dimensional space

The genome is confined within a cell nucleus in eukaryotes. Nucleus is an organelle that separates the transcriptional machinery of cell and cytoplasm. Genome organization is compertmentalized because of the attachment of nuclear substructures such as nuclear envelope (NE) and its lamina. NE is enriched with transmembrane proteins that have affinity to bind to the lamin proteins. binding between these transmembrane proteins and lamin proteins forms the nuclear lamina. The interactions between the chromatin and the nuclear lamina determine the nuclear positioning of the chromosomes (Fawcett, 1966; Amendola and van Steensel, 2014; Zuleger et al., 2011). These interactions are between lamin-associated domains (LADs) that are transcriptionally inactive and NE (Prokocimer et al., 2009). LADs are found to be conserved among different cell types (Meuleman et al., 2013), however can also be specific to lineage and cell type. Another substructure is nuclear pore complexes (NPCs) that play key roles in determining the organization of chromosomes. They are responsible for the transportation

between the cytoplasm and the nucleus and are important for regulation of gene expression. The chromatin around NPCs are different from the chromatin interacting with the NE. There are gaps in the inactive chromatin (also called heterochromatin) around NPCs and these gaps are filled with transcriptionally active chromatin (also called eurochromatin) (Ptak et al., 2014). Nucleolus is another landmark of cell nucleus that is involved in organization of genome in 3D space. Nucleolus is responsible for RNA polymerase I (RNAPI) transcription, and several rDNA genes are clustered in this substructure (Nmeth et al., 2010). We will first review the experimental tools used to analyze and visualize the organization of genome in nucleus, and then will describe the hierarchical genome organization inferred from the measurements of experimental tools in some detail.

1.2.1 Experimental tools to analyze genome organization

At a first level, nuclear landmarks dictate genome organization. But, at a finer resolution, specific chromatin interactions within and between chromosomes mediated by biochemical factors guide the genome organization. These interactions often facilitate regulating the expression of genes and may be specific to different cell types and developmental stages. Current understanding of genome organization is based on the data obtained using two different experimental approaches: Flourescence *in situ* hyridization (FISH) and chromosome conformation capture (3C). The former is a visualization technique and often uses super-resolution light microscopy. It measures the spatial distances between different loci. The latter quantifies frequency of interactions between different loci of a cell population.

1.2.1.1 Flourescence *in situ* hybidridization

FISH is the predominant technique to infer chromatin organization that precedes the development of 3C technique and its derivatives. This technique is based on preparation of fluorescent probes, which are complementary nucleotide sequences for chromatin regions of interest. The spatial distances between the fluorescent probes are then measured under the microscope (Beliveau et al., 2014). Recently, FISH with live-cell imaging enabled observation that transcriptionally active regions of chromosome move from the periphery of nucleus to interior nuclear positions (Chuang et al., 2006; Wang et al., 2016). FISH with live imaging also revealed the compartmentalization of chromatin and large conformational changes in the chromatin that occur during different differentiation stages (Jhunjhunwala et al., 2008).

1.2.1.2 Chromosome conformation capture

The 3C Method. The frequency at which chromatin segments interact with each other in cell populations can now be measured using the technique of chromosome conformation capture (Dekker et al., 2002). This technique is based on chromatin fragmentation and proximity cross-linking. Several of derivatives of 3C are available to quantify chromatin interactions. These techniques always measure the interactions in a population of cells. This is contrary to how FISH is performed, which is at single-cell level. after the collection of a population of cells, they are treated with formaldehyde, which creates covalent bonds between chromatin segments. This enables the fixation of the chromatin interactions. These cross-linked segments are then cut using a restriction enzyme. This is followed by the dilution and ligation of digested chromatin, which produce unique junctions that can be measured by various methods such as PCR.

The 3C technique is used for small scale analysis, as it measures the interaction frequencies between two single loci (Dekker et al., 2002).

The 4C Method. Chromosome conformation capture-on-chip technique is a more developed version of 3C in terms of the throughput and resolution (Simonis et al., 2006). This technique can capture interactions between a single locus, called the anchor site, and all the loci on the entire genome. The interaction partner of the anchor are usually quantified using microarrays or sequencing techniques. After the ligation process of 3C, the final products are further cut with a restriction enzyme. Re-ligation of digested ligation products into circular DNA is then applied. Finally, PCR technique is used to amplify the resulting circular DNA and amplified products are located using deep sequencing (Simonis et al., 2006).

The 5C Method. Chromosome conformation capture carbon copy (5C) technique is designed to capture interactions between many restriction fragment pairs simultaneously. In this variation of 3C technique, primers are computationally constructed for the region of interest that contains restriction enzyme sites. They are then annealed to targeted 3C fragment ends. If the primers are next to each other on a 3C junction, they can be ligated together. The final products are called the 5C library. The frequency of interactions between restriction fragments are then quantified using PCR amplification, followed by high-throughput sequencing. This process allows quantification of all chromatin interactions in the region as long as primers are designed for them. 5C can capture chromatin interactions only for the restriction fragments that are covered by the primer library (Fraser et al., 2009).

Hi-C method or genome-wide chromosome conformation capture can directly quantify the chromatin interaction frequencies of an entire genome. The main advantage of Hi-C technique is usage of high-throughput sequencing for proximity ligation products. Hi-C libraries are produced in a very similar fashion to the 3C libraries, where cross-linking with formaldehyde, and chromatin digestion with restriction enzymes are performed. The restriction fragments are then tagged with biotin. The cross-linked fragments are joined by blunt-end ligation. A purification step follows the reverse cross-linking, which produces the final Hi-C products. These products are sonicated and sequenced to capture their locations in the genome. The organization of human genome have been quantified at resolutions from 1 Mb to 1 kb using this technique (Lieberman-Aiden et al., 2009; Rao et al., 2014).

ChIA-PET Method captures interactions mediated by a transcription factor or an architectural protein by incorporating chromatin immunoprecipitation (ChIP) technique with 3C across the entire genome. As in the previously described techniques, cells are fixed with formaldehyde. Sonication are then used on fixed cells for ChIP of the transcription factor or a protein of interest. Biotinylation of coimmunoprecipitated DNA segments are then performed, followed by ligation. These ChIA-PET products are then digested with restriction enzymes. This is followed by purication and paired-end sequencing. With ChiA-PET, interactions of all the loci on the genome that also contain binding affinity for transcription factors, RNA polymerases, as well as architectural proteins can all be captured (Fullwood et al., 2009).

1.2.2 Inferring genome organization

With the advances in these experimental techniques, the chromosomes are found to be organized hierarchically at different length scales. Chromosomes fold within themselves to form distinct chromosome territories (CTs). Within the territories, chromosomes are composed of compartments A and B, which are active and silent, respectively. There are preferential interactions within each compartment. The higher level of organization after the compartments is topologically associated domains (TADs). They are mostly conserved at different developmental stages and cell types. Below I provide a brief summary of how genome is organized from smaller scales based on the current understanding, from chromosome territories to promoter-enhancer interactions.

1.2.2.1 Chromosome territories

Early light microscopy studies showed that chromosomes form distinct territories in nucleus, with inter-chromosomal interactions minimized. The techniques based on irradiation also showed the emergence of CTs (Cremer et al., 1996). FISH was also used to visualize chromosomes and demonstrated the formation of CTs (Bolzer et al., 2005). Although inter-chromosomal interactions are minimized in the nucleus, interactions between chromosomes in the neighboring territories do exist. Several loci that are on open chromatin and found to be active in expression were shown to loop out of their territories to make interactions with other chromosomes. This intermingling between different CTs were shown to be cell specific (Branco and Pombo, 2006). Since locations of CTs are cell type specific, it suggests that the posi-

tions of chromosomes and the boundaries shared between chromosomes in the nucleus might be functionally relevant (Roix et al., 2003).

1.2.2.2 Chromosome compartments

The first Hi-C study was carried on using two human cell lines in 1 Mb resolution. The heaptmaps of interaction frequencies displayed a checkerboard-like pattern, exhibiting interactions between megabase long regions across large genomic distances (Lieberman-Aiden et al., 2009). Principal component analysis (PCA) was then performed to aggregate the interaction frequencies into principle axes and the formation of two types of compartments were discovered. Compartment A is enriched in genes, transcriptional activity, and is made of open chromatin. In contrast, compartment B is transcriptionally inactive and is made of closed chromatin (Lieberman-Aiden et al., 2009).

1.2.2.3 Topologically Associated Domains

The analysis of the chromatin organization at finer scales based on the data obtained using 5C or Hi-C techniques lead to the discovery of blocks of dense chromatin that interacts more frequently with itself than the neighboring regions (Sexton et al., 2012). These are called Topologically Associated Domains (TADs). This finding is also supported by FISH experiments (Nora et al., 2012). TADs are visible in the heatmaps of frequency of interactions measured from 5C and Hi-C techniques and their average size is between 0.5 and 1 Mb (Dixon et al., 2012). Genes that are located within single TAD are often found to be coexpressed (Dixon et al., 2012). In addition, the distribution of TADs correlate with the position of epigenetic marks related to

activation or repression of gene activities. Factors such as CCCTC (CTCF) along with cohesin are found to be largely bound at boundaries of TADs (Nora et al., 2012).

1.2.2.4 Chromatin loops

While researchers investigate the existence of large scale structural units such as CTs, compartments or TADs, chromatin interactions behind the transcriptional activation span a few kilobases and often involve binding of several transcription factors and RNA polymerases. These interactions are often between promoter of genes and their enhancers. Chromatin looping is one of the most studied structural unit of genome organization, as it directly relates to transcriptional activities, thus cellular functions (Sanyal et al., 2012). As promoters and their enhancers usually are not in sequential proximity, it is important to measure the physical interactions at this finer scale to understand the organizational principles behind transcriptional regulation (Fudenberg and Mirny, 2012). It is also known that a single enhancer can have more than one target genes, and a single gene can be targeted by multiple enhancers. These looping interactions of enhancers and promoters are stabilized by the binding of several factors (such as CTCF and cohesin), transcription factor complexes, or RNA polymerases (Farrell et al., 2002).

1.3 Thesis outline and Project Overview

My research focuses on developing novel computational tools to construct spatial structures of chromosomes for understanding of mechanisms of gene regulation. Specifically, I study the effects of nuclear space in maintaining the epigenetic state of the cell, long-distance DNA loopings that promote cell-specific gene expression, and the main physical factors and mechanisms that determine genome organization. The research described in this dissertation is organized as

follows: **In chapter 2**, I study the effect of nuclear confinement on the folding properties of human genome to asses how severe nuclear confinement plays important roles on chromatin folding and compaction. A fundamental challenge in studying genome organization is the attrition in sampling of three-dimensional structures in severely confined spaces. My Ph.D. work overcame this problem with the development of the technique of geometric sequential importance sampling (g-SIS), with which self-avoiding chromatin chains are grown sequentially. I developed a detailed computational model, named Constrained Self Avoiding Chromatin (C-SAC), for deciphering the folding properties of chromosomes. With C-SAC, nuclear confinement is explicitly modeled and ensemble of chromatin chains are generated inside the cell nucleus. Analysis of C-SAC ensemble shows that the spatial confinement is one of the major determinant of chromatin architecture in cell nucleus. This research further highlights the importance of nuclear size as a potential regulator of epigenetic programming of cells as in the case of transitioning from stem cells to differentiated cells. This chapter has been done in collaboration with former student of Liang lab, Dr. Yun Xu. He helped with the development of the chain growth software described in Section 2.2.2. This chapter is partially based on the publications (Please see appendices for necessary permissions):

- Gürsoy, G., Xun, Y., Kenter, A., Liang, J.: Spatial confinement is a major determinant of folding landscape of human chromosomes. In Nucleic Acids Research, 42(13):8223-30, 2014.
- Gürsoy, G., Xu, Y., Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. In Conf Proc IEEE Eng Med Biol Soc. 3945-8, 2014.

Chapter 3 is devoted understanding the effects of nuclear landmarks and nuclear confinement on the genome organization of budding yeast. Detailed understanding of different cellular states in mammalian cell differentiation and cancer cell formation require comprehensive analysis of the interactions in the whole genome. Well-studied budding yeast is an excellent starting point for genome-wide chromatin construction, as its transcription machineries are shown to be largely dictated by genome organization. With further improvement of the multi-chromosome Constrained Self-Avoiding Chromatin (mC-SAC) model, I studied the effects of nuclear environment on the folding of multi-chromosome yeast genome. Comparison of ensembles of folded chromosomes from mC-SAC model with those from 3C-based studies shows that the majority of measured interactions regulating important cellular functions are captured (at an accuracy of 92%). Further analysis of the folded model genomes shows a high propensity of double stranded DNA breaks to cluster in three-dimensional space. This finding likely has implications in cancer biology and can potentially aid in understanding cancer-promoting translocations due to DNA breaks observed in human genome. This chapter is partially based on the publications (Please see appendices for necessary permissions):

- Gürsoy, G., Xu, Y., Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. In Conf Proc IEEE Eng Med Biol Soc. 3945-8, 2014.
- Gürsoy, G., Xu, Y., Liang, J.: Spatial organization of budding yeast genome in cell nucleus and identification of specific chromatin interactions from multi-chromosome constrained chromatin model. Submitted.

Chapter 4 focuses on understanding the differences in the folding of a chromatin locus between different expression levels. Identifying the difference in chromatin interactions between cell types is key to understanding the phenotypical differences arising from cell-specific gene expression. Constructing 3D structures of a gene locus can help to obtain detailed structural understanding of promoter-enhancer interactions and how they may affect transcriptional machineries and regulate cellular epigenetic states. However, experimental data from Chromosome Conformation Capture (3C) and related techniques are often sparse and incomplete due to systematic biases inevitable in experimental designs and other challenges. It is also challenging to distinguish biologically relevant interactions from non-specific collision of genomic elements in nucleus. I further improved the sampling method so non-specific spatial interactions from the experimental measurements can be removed and remaining specific interactions can be incorporated in a polymer model to study the underlying causation behind how gene expression levels change at different cell states. My computational modeling combined with analysis of epigenetic profiling data provides insights into the understanding of differential expression of important genes. These results show that gene expression is highly influenced by the folding landscape of chromatin. I have further identified novel chromatin interactions that were not captured by 3C data but were shown to have biological importance by other independent studies. This chapter has been studied with equal contribution from Dr. Yun Xu, who helped with the method and pipeline development described in Section 4.2. This chapter is partially based on the publication:

- Xu, Y., Gürsoy, G., Kenter, A., Liang, J.: Constructing 3D chromatin ensembles and predicting functional interactions of -globin locus from 5C data. In preparation.

Chapter 5 answers the question of minimum required chromatin interactions in a locus to achieve the transcriptional functions of the cell. Identifying interactions that facilitate the formation of promoter-enhancer interactions is important to understand the role non-coding DNA. I call such interaction pairs hotspots. Constructing 3D structures and perturbing these structures with virtual mutations can help to identify hotspots that are not obvious from the experimental measurements. I further developed a computational method that enables virtual mutations. I introduce a repulsion between sites of interest and compare the resulting ensemble with the wild type ensemble obtained by using Hi-C interactions as constraints. This helps to identify the hotspots whose interactions facilitate the interactions between promoters and enhancers. Combined with epigenetic profiling data of the CCL locus, my method revealed that structural hotspots of chromatin do not correlate with their interaction frequencies measured by 3C studies. I further showed the importance of CTCF binding on the regulation of interactions between promoters and enhancers and demonstrated that evolutionary conservation of these bindings sites is a major determinant of the importance of chromatin interactions. Arianna Girardi collected the necessary Hi-C data, as well as the epigenetics data required for this Chapter, and helped with the null model. This chapter is partially based on the manuscript:

- Gürsoy, G., Girardi, A., Liang, J.: Computational prediction of chromatin hotspots using n-Constrained Self-Avoiding Chromatin model. In preparation.

Finally, **in chapter 6**, I review the main topics of this dissertation, highlight the novel contribution of my method, strength and weaknesses of my modeling approach. I also provide perspectives on potential future developments.

CHAPTER 2

SPATIAL CONFINEMENT IS A MAJOR DETERMINANT OF THE FOLDING LANDSCAPE OF HUMAN CHROMOSOMES

2.1 Introduction

Human cells must accommodate approximately 6 billion base pairs of DNA in a small nucleus of a diameter of 6 to 20 μm (Alberts, 2002). Comprehensive understanding of chromosome organization is important for studying cellular functions (Fraser and Bickmore, 2007). A major task is to understand the rules that govern the regulation of long-range chromatin interactions (Fraser and Bickmore, 2007; Lieberman-Aiden et al., 2009; Dostie et al., 2006; Helmink and Sleckman, 2012).

FISH and 3C related techniques revealed a wealth of information about spatial chromatin structures across different genomic regions for different cell types (Lieberman-Aiden et al., 2009; Goetze et al., 2007; Mateos-Langerak et al., 2009; Zhao et al., 2006; Gavrilov et al., 2009; Jhunjhunwala et al., 2008; Sexton et al., 2012). A key outcome of FISH experiments is the relationship between the mean-square spatial distance, R^2, and genomic distance, s, of two chromosome loci (Goetze et al., 2007; Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). The folded structures of chromatin fibers follow a scaling relationship of $R^2(s) \sim s^{2\nu}$. In human Chr 1 and 11, the exponent ν is ~ 0.33 at smaller genomic (0.4–2 Mbp) distances, but levels off ($\nu \sim 0$) at larger genomic distances (>10 Mbp) (Mateos-Langerak et al., 2009). In

mouse Chr 12, ν is found to be $\sim$0.25 and $\sim$0.37 for two different cell types at smaller genomic distances ($<$ 0.5 Mbp), and levels off at larger genomic distances ($>$ 0.5 Mbp) (Jhunjhunwala et al., 2008). The leveling-off effects indicate that chromosomes are trapped to a space that is much tinier than volume of nucleus (Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). This reflects the requirement that chromosomes must fit into localized territories (Cremer and Cremer, 2001).

Results from genome-wide 3C (Hi-C) experiments showed that the contact probability ($P_c(s)$) between two sites that are a genomic distance s away from each other follows a power law of $P_c(s) \sim 1/s^{\alpha}$. The exponent α is ~ 1.08 at genomic distances between $0.5-7$ Mbp, when averaged across all chromosomes in a human cell line (Lieberman-Aiden et al., 2009). Further analyses showed that chromosome 11 and 12 exhibit the average human genome scaling behavior, with an exponent $\alpha \sim 1.08$ (Lieberman-Aiden et al., 2009; Barbieri et al., 2012), while exponents of chromosomes X and 19 have values significantly deviating from the average, with $\alpha \sim 0.93$ and $\sim$1.30, respectively (Barbieri et al., 2012). Similar results were obtained from a different Hi-C study (Kalhor et al., 2012) (see also (Barbieri et al., 2012) for analyses).

In order to gain understanding of the principles of spatial organization of chromatin, several polymer models have been developed (Sachs, 1995; Bohn et al., 2007; Tark-Dame et al., 2011). The fractal globule (FG) model (Lieberman-Aiden et al., 2009; Mirny, 2011) offers an explanation of the scaling of $P_c(s)$ and $R^2(s)$ with s at short genomic distances, although it does not account for the leveling-off effects observed in FISH studies (Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). The FG model also does not explain the observed variation in

α among different chromosomes. By attaching diffusible binders to chromatin, the Strings and Binders Switch (SBS) model can account for both the leveling-off effects and the heterogeneous scaling of α (Barbieri et al., 2012). However, individual scaling properties can exist only under carefully tuned conditions of binder concentrations and binding site distributions, which are unknown *a priori*. In addition, the SBS model does not exhibit multiple scaling exponents occuring simultaneously under one set of conditions.

The most important factor that determines how chromosomes fold in the cell nucleus is the amount of available space. In a recent study, spatial constraints were shown to be sufficient to produce the overall structural architecture of budding yeast genome (Tjong et al., 2012), although the general effects of spatial confinement on chromatin folding of human genome are unknown. Polymer models can provide important insights into chromatin compaction in cell nucleus (Hahnfeldt et al., 1993; Heermann et al., 2012; Iyer and Arya, 2012). However, a major obstacle in studying chromatin fibers confined in a small volume is the difficulty in generating a large number of unbiased model chromatin fibers in the form of self-avoiding chains with appropriate physical and spatial properties (Liu and Chen, 1998; Zhang et al., 2003; Lin et al., 2011).

In this chapter, we examined the effects of spatial confinement on chromosome folding using the constrained self-avoiding chromatin (C-SAC) model. We developed a novel algorithm that can generate large ensembles of diverse model chromatin chains in severe spatial confinement, with full excluded volume effect incorporated. We find that spatial confinement is plausibly responsible for much of the observed overall scaling behavior of human chromosome folding.

The heterogeneous ensemble of folded model chromatin chains under spatial confinement also predicts chromosome-specific scaling relationships, as well as formation of highly interactive substructures that might give rise to the formation of topological domains. Our findings highlight the importance of nucleus size in regulating the folding landscape of chromosomes.

2.2 Materials and Methods

2.2.1 Model and parameters

In our C-SAC model, a chromatin fiber is a collection of beads that make up a self-avoiding polymer chain. Each bead has a diameter of 30 nm and is 3,000 base pairs long (Wedemann and Langowski, 2002). Every 5 beads form a persistence unit, which corresponds to a persistence length of 150 nm (Figure 2A) (Wedemann and Langowski, 2002). Our model chain is 4,996 beads long, equivalent to about 15 Mb of DNA.

Each chromatin chain is generated in a confined space of nucleus, which we modeled as a sphere. The sphere diameter, D, is selected to be proportional to the size of the human cell nucleus. We assumed an average nucleus size of a diameter of ~ 11 μm for 6 billion base pairs of human DNA (Alberts, 2002). The diameter of the nucleus for a 15 Mb long chromatin chain is therefore about 1.5 μm. With this model, we grow our chromatin chains sequentially in a sphere of a diameter of $D = 1.5$ μm (Figure 2A). We overcame the difficulties of generating folded chromatin chains inside a small volume by sequentially growing self-avoiding chains one persistence unit at a time using the technique of geometric sequential importance sampling (Liu and Chen, 1998; Zhang et al., 2003; Lin et al., 2011). Subsequently, D was changed to $D = 2.5$,

$D = 5.0$, $D = 7.5$, $D = 10.0$, $D = 30.0$, and $D = 500.0$ μm to explore the effects of size of confined space on the spatial organization of chromatin.

2.2.2 Growing chromatin chains using geometric sequential importance sampling

The chromatin chains in three-dimensional space are generated following a chain growth approach (Liu and Chen, 1998; Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b; Lin et al., 2008a; Zhang et al., 2009; Lin et al., 2011). A chromatin chain contains n persistence units, with the location of the i-th persistence unit denoted as $x_i = (a_i, b_i, c_i) \in \mathbb{R}^3$. The configuration $\boldsymbol{x}$ of a full chromatin chain with n persistence units is:

$$\boldsymbol{x} = (x_1, \cdots, x_n).$$

The target distribution $\pi(\boldsymbol{x})$ is a uniform distribution, in which all chromatin chains within the given confinement can be sampled. To generate a chromatin chain, we grow the chain one persistence unit at a time, ensuring the self avoiding property along the way, namely, $x_i \neq x_j$ for all $i \neq j$. We use a $k = 100$-state off-lattice discrete model (see (Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b; Lin et al., 2008a; Zhang et al., 2009; Lin et al., 2011) for more details). The new persistence unit added to a growing chain with the current persistence unit located at x_t is placed at x_{t+1}, which is a persistence length L_p distance away from x_t. x_{t+1} is randomly taken from one of the unoccupied k-sites neighboring x_t. As random selection from available empty neighboring sites introduce bias for sampling from $\pi(\boldsymbol{x})$, we keep track of the

bias and assign each successfully generated chain a proper weight $w(\boldsymbol{x})$. Details can be found in references (Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b).

Each persistence unit further contains $[(L_p/d_f) - 1]$ number of monomer beads, where $L_p = 150$ nm, and the fiber diameter d_f is 30 nm. These monomers are connected by a chain, and their positions are interpolated as if they are on a rigid rod (Figure 2A). This is to mimic the persistence behavior of the chromatin fiber. We again enforce the self-avoiding property, such that these beads will not intersect with any other beads in the partial chain that has already been grown. All together, there are $N' = N + (N-1) \cdot [(L_p/d_f) - 1] = 1,000 + 999 \cdot [(150/30) - 1] = 4,996$ monomer beads for a $N = 1,000L_p$ unit long chain. For larger confinement space, we generated chains up to $N = 8,100L_p$.

2.2.3 Model Validation:Scaling of C-SAC chains without confinement

We first used our geometric sequential importance sampling technique to generate free space self-avoiding C-SAC chains without confinement, as their scaling behavior is well understood (de Gennnes, 1979). We generated 10,000 C-SAC chains of different length N, for $N \in \{100, 200, \cdots, 1000\}$. Figure 2B and C show the scaling relationship $R(N) \sim N^{\nu}$ and $P_c \sim N^{\alpha}$. The scaling exponents are found to be $\nu \sim 0.59$ and $\alpha \sim -1.88$, which are very close to the expected values of $\nu \sim \frac{3}{5}$ and $\alpha \sim -3\nu$ (de Gennnes, 1979).

2.2.4 Resampling.

To improve the success rate of generating full length chromatin chains, we employ the technique of resampling (Liu and Chen, 1998; Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b; Lin et al., 2008a; Zhang et al., 2009; Lin et al., 2011). When there is no unoccupied

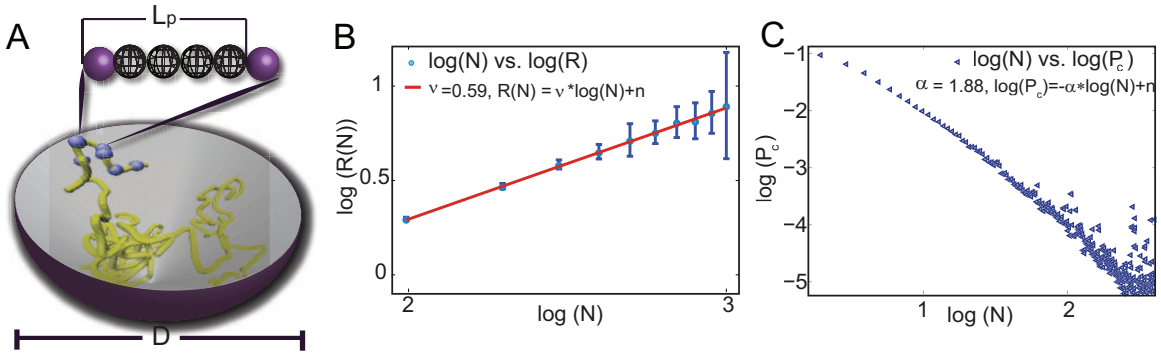

Figure 2. **The physical model of C-SAC chains and scaling properties of chains without confinement. (A)** Cartoon representation of the C-SAC model. Fiber of chromatin is modeled as collection of beads with a persistence length L_p. Purple spheres are the beads at the boundaries of a persistence unit. Spheres in-between are the interpolated beads inside a unit of L_p. Polymers are grown as chains inside a spherical confined space of a diameter D. Beads are not allowed to cross each other or grow beyond the boundary of the spherical volume. **(B)** $R(N)$ vs. N relationship in log-scale. Each data point is derived from an ensemble of 10,000 chains of length N. The exponent ν is found to be 0.59, similar to the theoretical exponent for three-dimensional SAWs in good solvent (de Gennnes, 1979). **(C)** P_c vs. N relationship in log-scale. The scaling expoent α is approximately -1.88, similar to the theoretical exponent of $-3\nu = -1.8$ for three-dimensional SAWs in good solvent (de Gennnes, 1979).

neighboring sites inside the confined space for x_t of a partially grown chain, there is no place to grow x_{t+1} the next persistence unit. In this case, we go back one step and re-grow the chain at x_{t-1} from the $(t-1)^{th}$ monomer. There are also chains with small weights due to biased sampling. They contribute little to the estimation of properties of the population of chromatin chains.

We employ a simple resampling scheme to address these issues. At each t-th step of the chain growth process, we sort all chains by their weights, and divide them into k fractions. The fraction of chains with the lowest weights are then replaced by copies of chains in the fraction of the highest weights. Weights of these chains are then adjusted accordingly. We then continue to grow chains of this new population. This is repeated until all chains reach full length. We use t=10 and k=3. Details of the resampling strategies can be found in references (Liu and Chen, 1998; Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b; Lin et al., 2008a; Zhang et al., 2009; Lin et al., 2011).

We also employ a dynamic resampling scheme (Liu and Chen, 1998) when the chain length N is large and/or the sphere diameter D is small. We calculated effective sample size for every step of the growth process (Liu and Chen, 1998) as:

$$ESS = \frac{\left(\sum_{j=1}^{M} w_i\right)^2}{\sum_{j=1}^{M} w_i^2},$$

where M is the total number of chains. If $ESS < 0.3M$, we assign a probability $p(i)$ to each partial chain i as $p(i) = \exp(w_i - \max_{1 \le i \le M} w_i)$ and sample M chains with replacement according

to $p(i)$ and adjust the weights of each selected chain k as $w_k^* = \frac{w_k}{p(k)}$. We then continue to grow chains of this new population. This is repeated until all chains reach full length.

2.2.5 Chromatin properties.

With m successfully generated chromatin chains, we can calculate the physical properties of the population of chromatin fibers. Denote the configurations of the j-th successfully generated chromatin chain as $\boldsymbol{x}^{(j)} = (x_1^{(j)}, \cdots, x_n^{(j)})$, and its associated weight $w^{(j)}$. To calculate the mean value of a physical property $\bar{h}(\boldsymbol{x})$, we have:

$$\bar{h}(\boldsymbol{x}) = \mathbb{E}_{\pi(\boldsymbol{x})}[h(\boldsymbol{x})] = \frac{\sum_{j=1}^{m} h(\boldsymbol{x}^{(j)}) \cdot w^{(j)}}{\sum_{j=1}^{m} w^{(j)}}.$$

2.2.6 Mean end-to-end distance.

The mean end-to-end distance $R(N)$ is the mean Euclidean distance between the beginning and the end of the chain of a length N. For the j^{th} chromatin chain, we have:

$$R(N)^{(j)} = \|x_1^{(j)} - x_N^{(j)}\|.$$

The mean end-to-end distance is then calculated for the set of m chromatin chains as:

$$\bar{R}(N) = \frac{\sum_{j=1}^{m} R_{(j)}(N) \cdot w^{(j)}}{\sum_{j=1}^{m} w^{(j)}}.$$

2.2.7 Mean-square spatial distance.

The mean-square spatial distance $R^2(s)$ is the mean-square Euclidean distance between genomic regions with a genomic separation s, here in units of persistence length. For the j^{th} chromatin chain, we have:

$$R^2(s)^{(j)} = \frac{\sum\limits_{i=1,\, j=i+s} \|x_i^{(j)} - x_j^{(j)}\|^2}{N - s},$$

in which the denominator $N - s$ is total number of all possible such interactions with s-separations. The mean-square spatial distance is then calculated for the set of m chromatin chains as:

$$\bar{R}^2(s) = \frac{\sum\limits_{j=1}^{m} R^2_{(j)}(s) \cdot w^{(j)}}{\sum\limits_{j=1}^{m} w^{(j)}}.$$

2.2.8 Contact probability.

The contact probability $P_c(s)$ is the probability of two genomic regions separated by genomic distance s to be in spatial proximity of each other for chain of length N. Following Lieberman-Aiden et al. (Lieberman-Aiden et al., 2009), it is calculated by counting the number of times that the Euclidean distance between two regions separated by genomic distance s is smaller than a distance threshold d_θ, divided by the number of all such candidate contacts. Let $\mathbb{I}_c^{(k)}(s)$ be the observed number of i and j contacts that satisfies the condition $\|x_i - x_j\| \leq d_\theta$, with $j - i = s$ in chain k. Let $\mathbb{I}_{all}(s)$ be the number of all possible contacts of two regions separated

by a genomic length of s and is equal to $N - s$. An estimate of contact probability $P_c^{(k)}(s)$ for chain k of length N is:

$$\hat{P}_c^{(k)}(s) = \frac{\mathbb{I}_c^{(k)}(s)}{\mathbb{I}_{all}(s)}.$$

The mean value from the weighted ensemble average is then calculated as:

$$\bar{P}_c(s) = \frac{\sum_k P_c^{(k)}(s) \cdot w^{(k)}}{\sum_k w^{(k)}}.$$

2.2.9 Reweighting.

As chromatin chains are generated following the uniform distribution $\pi(\boldsymbol{x})$ of all geometrically realizable chains, these samples need to be reweighted in order to calculate ensemble properties of chromatin chains following a different distribution $\pi'(\boldsymbol{x})$.

To asses the effect of specific binding on the population of chromatin chains, we recalculate the associated weights of each chain for chromatin following the new distribution $\pi'(\boldsymbol{x})$, which is the Boltzmann distribution after incorporating energies of binding interactions. For a chromatin chain with interactions mediated through protein binder, each interaction between any (i, j) pairs of sites contributes to the weight of the chain by the Boltzmann factor of $\exp(E^{(k)}(i, j)/k_BT)$. Here $E^{(k)}(i, j)$ is the binding energy if both i and j contain binding sites

and are mediated by the binder protein, otherwise $E^{(k)}(i,\, j) = 0$. The total weight of the k^{th} chain previously sampled from the uniform distribution is then re-calculated as:

$$w^{(k)} = \prod_{(i,\, j)} \exp(E^{(k)}(i,\, j)/k_B T).$$

2.2.10 Clustering.

We clustered the generated chromatin chain conformations according to their pairwise distances between persistence units using a k-means clustering algorithm (Hartigan and Wong, 1976). For k-means clustering, we need to calculate the Euclidean distances between persistence units. As we have a population of m=10,000 chains, each with $N = 1,000$ persistence units, this amounts to $n = N \times (N-1)/2 = 499,500$ number of pairwise distances to be calculated. Since the algorithm is of $O(m^{nk+1} \log m)$-complexity, we coarse-grained each chain to speed up the computation. We take sequentially every 33 persistence units as our new unit. This gives 30 connected units, where the number of pairwise distances is now $n = 30 \times 29/2 = 435$. We set the number of clusters k to 20.

2.3 Results

2.3.1 C-SAC model gives observed scaling behavior of human chromosomes

We generated ensembles of 10,000 independent self-avoiding model chromatin chains for different chain length N of 50, 100, 200, and then up to 1,000, with increments of 100 confined to a region of $D = 1.5$ μm. Our C-SAC model chromatin chains exhibit experimentally observed scaling properties. The mean-square spatial distance $R^2(s)$ of partial chains of length s from

10,000 chains of length $N = 1,000L_p$ follows the relationship of $R^2(s) \sim s^{2\nu}$, with an exponent ν of ~ 0.34 at shorter genomic distances, but levels off with $\nu = 0$ at larger genomic distances. The experimentally observed $\nu = 0.33$ was derived from FISH data between 0.4 Mb and 2.0 Mb. In our C-SAC model, $\nu = 0.34$ is derived accordingly between 5 and $25L_p$, by matching the onset points of the leveling-off effect (10 Mb in the FISH study, and $125L_p$ in C-SAC chains) (Mateos-Langerak et al., 2009) (Figure 3A). Since the mass density of chromatin and how it varies in different loci and different chromosomes are unknown, the regime that the exponents are extracted are not directly comparable by genomic distance to the experimental data. The mass density used in this study is an average property, and it may differ from the actual mass density at the loci measured in the FISH experiments (Mateos-Langerak et al., 2009). In the FISH study of ref. (Jhunjhunwala et al., 2008), spatial distances between different loci with different genomic separation s were measured on two different subchromosomal regions of Chr 12 of mouse pre-pro-B cells and pro-B cells. Our non-linear fit (Figure 3) gives a scaling exponent of $\nu \sim 0.37$ when $s < 0.5$ Mb for pre-pro-B cells, and $\nu \sim 0.27$ when $s < 0.5$ Mb for pro-B cells. The leveling-off effects takes place at $s = 0.5$ Mb in mouse Chr 12 of both pre-pro-B cells and pro-B cells.

In the FISH study of human Chr 11 and Chr 1 (Mateos-Langerak et al., 2009), ν was reported to be ~ 0.33 in both human Chr 11 Chr 1 when $0.4 < s < 2$ Mb. The leveling-off effects were reported to takes place at $s \geq 10$ Mb in Chr 11 and $s \geq 3$ Mb in Chr 1 (Mateos-Langerak et al., 2009).

It was also reported in ref. (Barbieri et al., 2012) that the FISH study of mouse Chr 14 in ref. (Langmead and Salzberg, 2009) exhibits a $\nu \sim 0.5$ when $s < 3.5$ Mb, beyond which the leveling-off effects may take place.

In C-SAC chains of length $N = 1,000$ with the confinement of $D = 1.5\ \mu m$, the leveling-off effects are found to take place at around $s = 125L_p$. We calculated the scaling exponent ν of $R(s) \sim s^{\nu}$ between $s = 5L_p$ and $s = 25L_p$. This choice of $25L_p$ is based on the ratio of 25/125, which is the same as the ratio of 2Mb/10Mb between the distance threshold where ν was fitted and the distance threshold beyond which the leveling -off effects occurred in human Chr. 11 (Mateos-Langerak et al., 2009), which was also used in the study of refs. (Lieberman-Aiden et al., 2009; Mirny, 2011). We found $\nu \sim 0.34$ when $5L_p \leq s \leq 25L_p$.

As discussed above, there are some variations in the reported values of the scaling exponent ν from existing FISH studies. Similarly, we found that ν also varies depending on the regime where the exponents were fitted. If $s \leq 60L_p$, ν is found to be ~ 0.25, and $\nu \sim 0.5$ if $s \leq 15L_p$.

To characterize the scaling relationship of contact probability $P_c(s)$ and contour length s between two loci, we harvested partial chains of length s from independent ensembles of different full chain lengths and estimated $P_c(s)$. As contact probability $P_c(s)$ between loci of s genomic distance were derived from fragments from different chromosomes in Hi-C studies (Lieberman-Aiden et al., 2009), partial chains from independent ensembles of varying full lengths are necessary to remove self correlations, which may occur when subchains are taken from the same ensemble of chains with a fixed full length as in (Lieberman-Aiden et al., 2009).

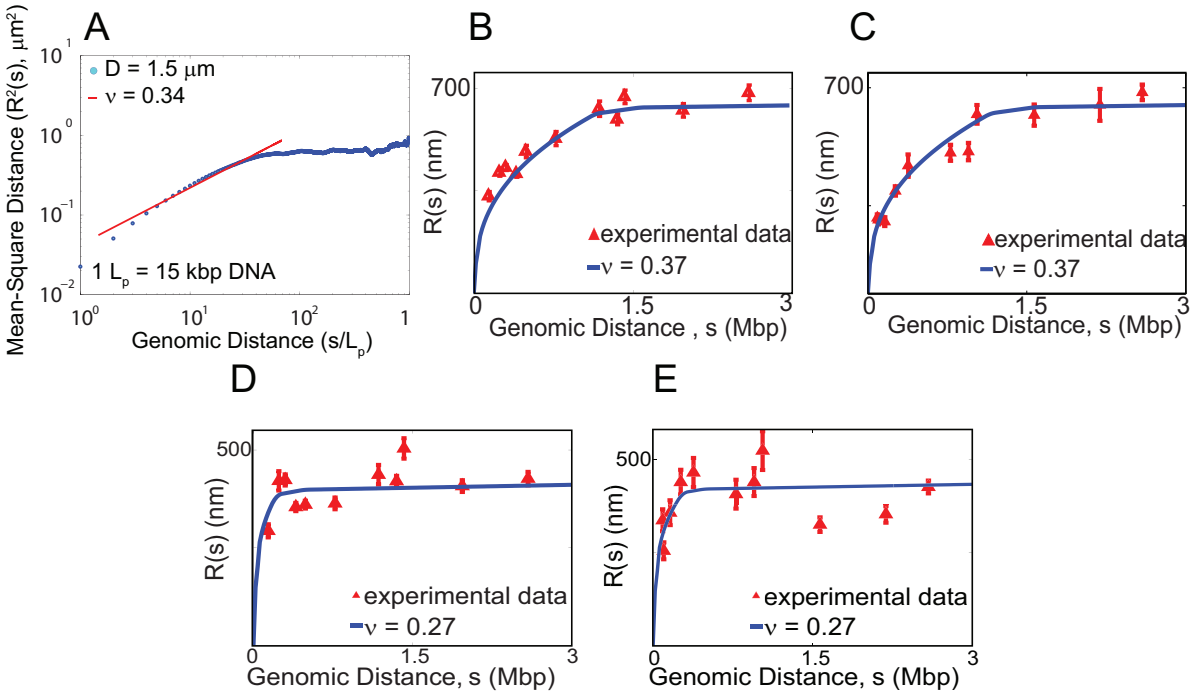

Figure 3. **The mean-square spatial distance *vs.* genomic distance relationship of C-SAC chains. (A)** The scaling of mean-square spatial distance $R^2(s)$ from 10,000 chains of length $1{,}000L_p$ in $\log_{10}$ scale. $R^2(s)$ follows a power law of $\sim s^{2\nu}$, with $\nu \sim 0.34$ (95% confidence interval: $[0.30,\ 0.38]$), similar to measured ν of ~ 0.33 (Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). **(B)** The scaling behavior of $R(s)$ of FISH data spanning 3 Mb when the probe is anchored in genomic element BAC (Jhunjhunwala et al., 2008) (red triangle) in pre-pro-B cells chromosome 12 of mouse genome. **(C)** The scaling behavior of $R(s)$ of FISH data when the probe is anchored in genomic element h1 (Jhunjhunwala et al., 2008) (red triangle) in pre-pro-B cells chromosome 12 of mouse genome. **(D)** The scaling behavior of $R(s)$ of FISH data when the probe is anchored in genomic element BAC (Jhunjhunwala et al., 2008) (red triangle) in pro-B cells chromosome 12 of mouse genome. **(E)** The scaling behavior of $R(s)$ of FISH data when the probe is anchored in genomic element h1 (Jhunjhunwala et al., 2008) (red triangle) in pro-B cells chromosome 12 of mouse genome.

Our C-SAC model can reproduce the scaling relationship of contact probability $P_c(s) \sim 1/s^{\alpha}$, with an exponent α of ~ 1.05 (Figure 4A), which is in excellent agreement with $\alpha \sim 1.08$ measured in Hi-C studies (Lieberman-Aiden et al., 2009).

Our C-SAC model also captures observed deviations in α from the average in individual chromosomes (Barbieri et al., 2012). After clustering the 10,000 C-SAC chains of length $N = 1,000L_p$ according to their spatial similarities measured in pairwise bead distances, the resulting 20 clusters have exponent α ranging from 0.79 to 1.3 (Table 1). The exponents of these clusters give the full range of α observed experimentally. For example, exponents of cluster 10, 15 and 17 agree well with those of Chr 19, X and 11/12, respectively (Figure 4B) (Barbieri et al., 2012; Lieberman-Aiden et al., 2009). These results are obtained without using any characteristics specific to Chromosome X, 19, 11 or 12. Complex scaling property of human genome arises fully from structural clusters resulting from the spatial confinement. Overall, our results indicate that the restriction of volume imposes strong constraints, and chromatin chains under such confinement exhibit experimentally observed scaling behavior of human chromosomes.

2.3.2 Nuclear size determines chromosomal scaling behaviour

To examine the effects of the spatial confinement, we generated independent ensembles of 10,000 C-SAC chains of length N inside a sphere D. Here N is varied from 50, 100, and then up to 1,000, with increments of 100. The sphere diameter D takes the value of 2.5, 5.0, and 7.5 μm, in addition to 1.5 μm. We independently generated 3 different ensembles of 10,000 C-SAC chains at each of the combination of N and D values. Altogether, we have $4 \times 11 = 44$ independent ensembles of 10,000 C-SAC chains for calculating contact probability.

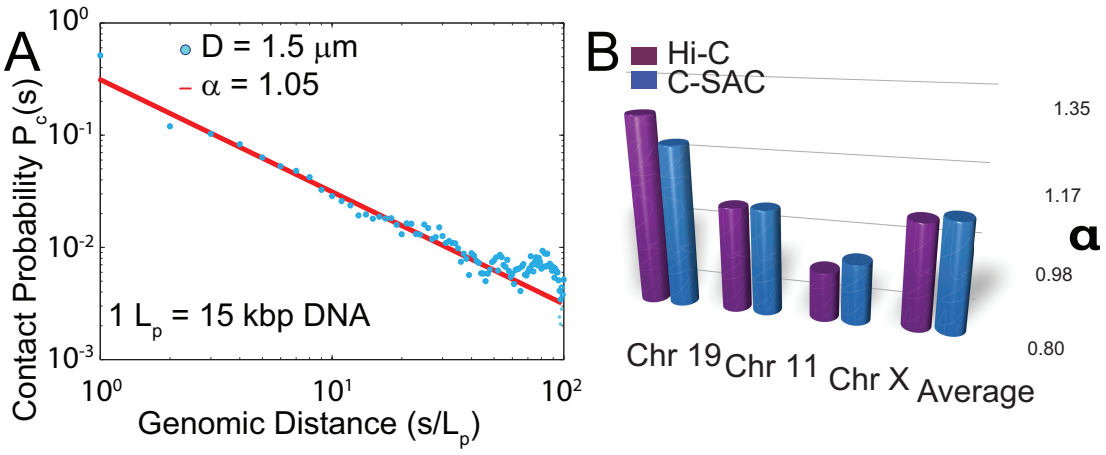

Figure 4. **The contact probability *vs.* genomic distance relationship of C-SAC chains. (A)** The scaling of contact probability $P_c(s)$. $P_c(s)$ follows a power law of $\sim 1/s^{\alpha}$, with $\alpha \sim 1.05$ (95% confidence interval: [1.15, 0.95]), similar to the measured α of 1.08 (Lieberman-Aiden et al., 2009). **(B)** Comparison of exponent α of contact probability, $P_c(s)$, between C-SAC and Hi-C data (Lieberman-Aiden et al., 2009). Values of α for different chromosomes from refs. (Barbieri et al., 2012; Lieberman-Aiden et al., 2009) were compared to those calculated separately for different clusters in the C-SAC population of chromatin chains. Purple bars denote the experimentally observed exponent α for Chr 19, Chr 11, Chr X and the average α across all chromosomes in human genome. Blue bars are αs from the corresponding C-SAC clusters and the average α of entire population.

TABLE I

SCALING EXPONENTS OF THE 20 CLUSTERS OF CHROMATIN CHAINS

Cluster	Size	α	ν
1	458	0.99	0.37
2	122	NA	NA
3	70	NA	NA
4	816	1.02	0.37
5	806	1.13	0.37
6	436	1.13	0.38
7	232	0.87	0.36
8	86	NA	NA
9	560	1.17	0.37
10	560	0.96	0.37
11	666	1.16	0.37
12	388	1.14	0.37
13	496	1.05	0.38
14	332	0.79	0.37
15	374	1.08	0.38
16	1078	1.14	0.37
17	534	1.28	0.37
18	882	1.03	0.37
19	284	1.12	0.40
20	820	1.14	0.38

The average scaling exponents α and ν of each cluster, along with the size of the cluster are listed

We used partial chains of length s from the ensemble of 10,000 chains of $N = 1,000$ of different D for the calculation of mean-square spatial distance $R^2(s)$, following the approach used in the FISH studies (Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). We found that both exponents α and ν increase with D (Figure 5). Furthermore, chromatin chains tend to adopt more open conformation as D increases. At the same time, the leveling-off effect at larger genomic distances disappears (Figure 5B). Further clustering of chromatin structures at different nuclear sizes showed that even with the smallest nucleus size of $D = 1.5$ μm, there exists a substantial amount of open chromatin structures (10.9%), while the compact structures and in-between structures are 18.6% and 70.5% of the population, respectively. As the size of the nucleus increases, the percentage of open-like structures in the population increases. These results therefore suggest that nuclear size is a major factor in influencing the overall folding landscape of chromatin, via modulation of the spatial confinement scale D.

2.3.3 Formation of highly interactive substructures upon confinement and topological domains

We used the C-SAC model to further explore structural properties of chromatin fibers. Topological domains were previously observed in electron microscopy studies (Cremer and Cremer, 2001; Cook, 1999; Kreth et al., 2004) and in recent 3C-based studies (Hi-C) (Dixon et al., 2012; Nora et al., 2012). Such domains are distinctive regions along the chromatin chain with significantly elevated interactions within region (Dixon et al., 2012). Their DNA content range from a few kbp to 1 Mbp, and they occupy a volume of 300 to 800 nm in diameter (Cremer et al., 2000). To examine whether C-SAC chains contain domain-like substructures, we calculated

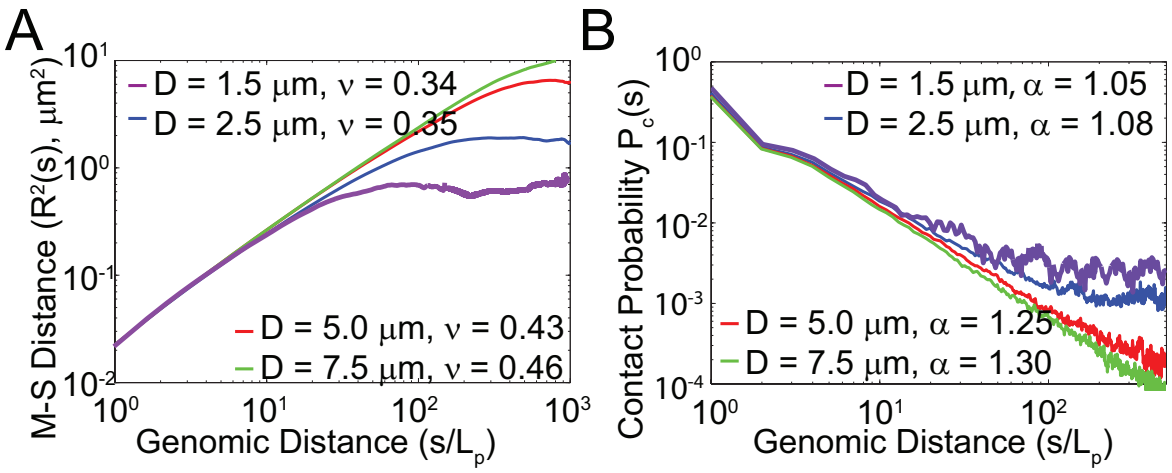

Figure 5. **Size of confinement affects the scaling behavior of human chromosomes.** **(A)** Mean-square spatial distance $R^2(s)$ *vs.* genomic distance s for different confinement sizes D. For larger nuclei, chromatin chains have larger exponent ν. The leveling-off effects disappears as the confinement size increases. **(B)** Contact probability $P_c(s)$ *vs.* chain length s for different confinement sizes D. For larger nuclei, chromatin chains have larger exponent α, indicating more open conformations.

the number of consecutive persistence units in spheres of 800 nm diameter along the chromatin chain. If a sphere contains chromatin fragments that contain more than 400 kbp DNA, it is regarded as a highly interactive substructure. We further define two types of substructures: 1) Interactive substructures in which more than 20% of their persistence units are in spatial proximity with those of other interactive substructures, and 2) independent substructures in which none of their units are in spatial proximity with any units of other substructures (Figure 6A).

On average, there are about ∼6.5 substructures per chain, which occupies around 21% of the entire 15 Mbp C-SAC chain. 41% of these substructures are interactive, whereas the rest of them are independent substructures (Figure 6B). Existence of these highly interactive substructures are also observed from interaction matrices and three-dimensional conformations of individual chains (Figure 6C-D).

In summary, there exist distinct substructures in C-SAC chromatin chains with elevated interactions. These results are observed without requiring special simulation conditions or specific binding sites as in other chromatin models (Barbieri et al., 2012; Lieberman-Aiden et al., 2009; Mirny, 2011). Their existence suggests that the confinement of the cell nucleus is sufficient to induce tentative formation of highly interactive substructures along the chromatin chains, which could further give rise to the formation of topological domains.

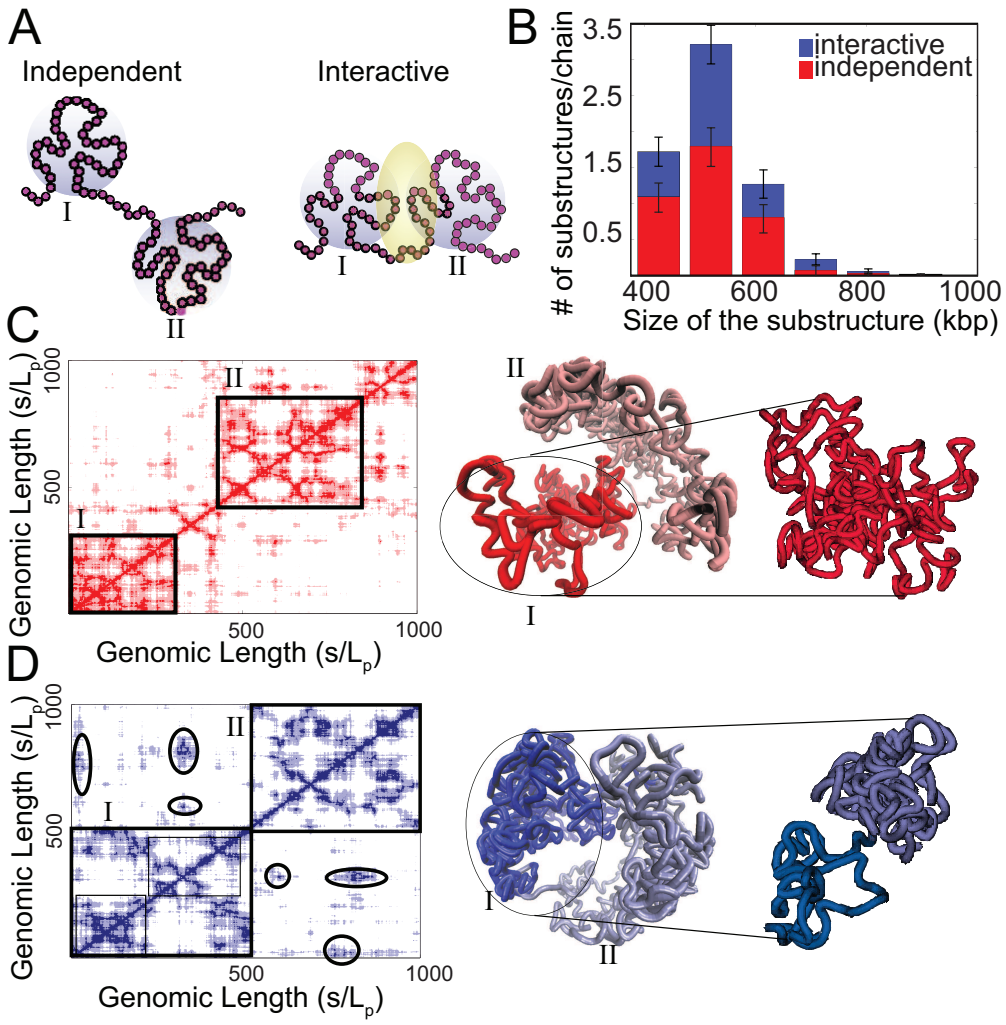

Figure 6. **Formation of domain-like substructures upon confinement.** **(A)** Illustration of substructures on C-SAC chains, where consecutive monomers are contained in spheres of 800 nm diameter (gray circles). Two independent substructures with no interaction between them, as well as two interactive substructures with more than 20% of their monomers participating interaction (interface shaded in yellow) are shown. **(B)** The distribution of number of substructures per chain containing different amount of DNA for both independent (red) and interactive (blue) categories are shown. **(C)** A random C-SAC chromatin chain with independent substructures. The rotated and zoomed-in substructure shows a singular domain-like conformation. The domain-like substructures can also be seen in the corresponding distance matrices, where the spatial distances between different loci of the C-SAC model chains are color coded in red, with darker red representing interactions between chromatin beads. The chromatin chain contains two highly interactive substructures that do not interact with each other. **(D)** A random C-SAC chromatin chain with substructures are shown. There are two small interactive substructures, as can be seen in the rotated and zoomed-in conformation. The domain-like substructures can also be seen in the corresponding distance matrices, where the spatial distances between different loci of the C-SAC model chains are color coded in blue, with darker blue representing interactions between chromatin beads. The chromatin chain contains two highly interactive substructures that interact with each other. Circles highlight regions of interactions.

2.3.4 Scaling behavior of human chromosomes is not altered by random binder-mediated looping interactions

Several polymer models of long-range chromatin organization are based on the introduction of explicit looping probability or looping through binder-mediated interactions (Barbieri et al., 2012; Bohn and Heermann, 2010). To assess how chromatin looping in addition to confinement would affect the scaling behavior of chromatin chains, we distribute different numbers of binding sites randomly along the chromatin chains, which cover from 10 to 50% of the total number of persistence units in the chromatin fiber. Chromatin structures with a large number of binding sites in spatial proximity are subject to binder-mediated interactions. These structures will then have lower energy and therefore higher probability of presence in the chromatin population. We calculated the distribution of the chromatin chains with such binder interactions, in which the binding energy of connecting two interacting sites is assigned to be $6k_BT$ (Barbieri et al., 2012; Renda and Pedone, 2007). This allows us to assess the scaling properties of different populations of chromatin chains under different looping conditions.

Our results showed that there is virtually no change in the scaling exponents α and ν in C-SAC chains after introducing binders compared to the original C-SAC chains, where the only constraint is the spatial confinement of the cell nucleus (Figure 7). These results indicate that random self-avoiding chromatin chains folded inside a confined space have an intrinsic propensity to form loops, without the explicit introduction of additional binders. Overall, our results indicate that the confinement at the scale D is the dominant factor in determining the average scaling behavior of chromatin structures.

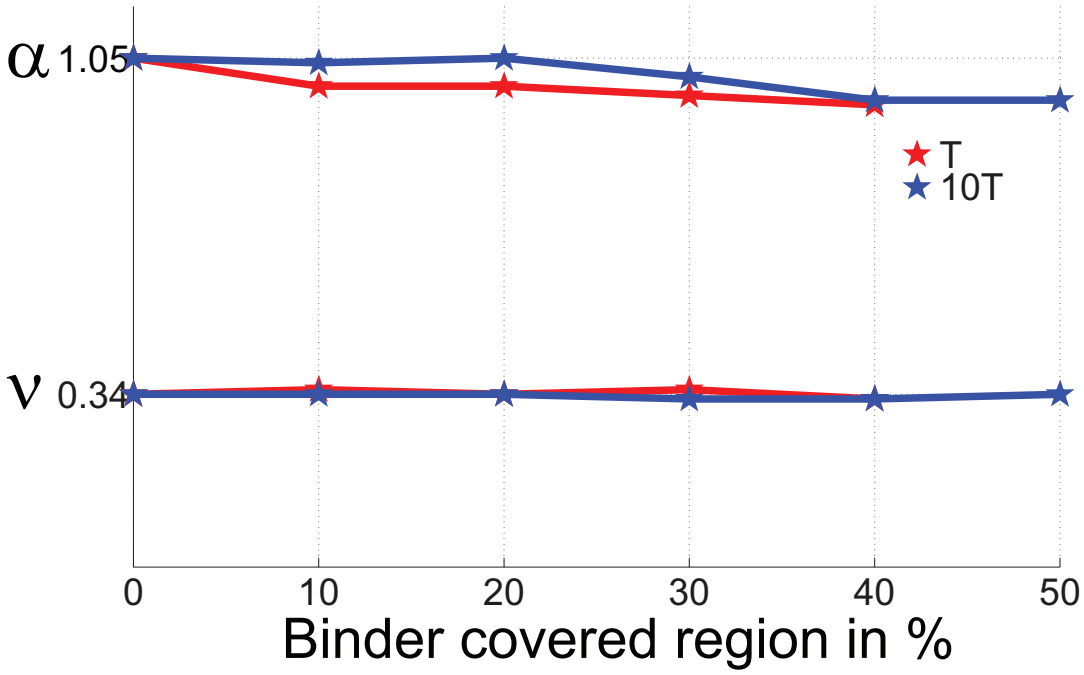

Figure 7. **Scaling exponents α and ν when different fractions of C-SAC chromatin chains are covered by binding sites.** The binding energy is assigned to $6k_BT$ (Barbieri et al., 2012) for two relative temperatures (T in red, 10T in blue). Neither scaling exponents experiences significant changes as the binder coverage increases.

2.3.5 Relevant size regime of chromatin confinement for the spatial organization of chromosomes

Chromosomes are found to occupy localized territories of the size of ~ 2 μm diameter (Cremer and Cremer, 2001). To span a chromosome territory, only a short fragment of chromatin fiber (*ca.* 150 kb, assuming 30 nm diameter and 150 nm persistence length) is required, which is well below the range of 0.5–7.0 Mb measured in Hi-C studies (Lieberman-Aiden et al., 2009). Studying the scaling behavior of the self-avoiding polymer chains in the correct confinement is therefore key to construct the relevant model for understanding chromosome folding. We used C-SAC model to explore the relationship between the size of confinement and the scaling properties of confined chromatin chains, as the calculated scaling exponents of C-SAC chains in the relevant size regime ($\alpha \sim 1.05$) is well below the theoretical scaling exponent of self-avoiding polymer chains in confinement ($\alpha \sim 1.50$) (de Gennnes, 1979).

We generated independent ensembles of C-SAC chains in equilibrium and calculated the relationship between the chain length and the end-to-end distance, as well as the relationship between contact distance and the contact probability. In total, we generated chains with different length N of 50 , 100 , $\cdots$, 1,000 at increment of 100 each with 5 different confinement D (D = 1.5, 2.5, 10, 30 and 500 μm). We also generated C-SAC chains of length N from 50 to 8,100 at different increments for two different confinement D (D = 5.0 and 7.5 μm) (Figure 8). We found that chromatin chains exhibit confinement-dependent scaling behaviour, with ν ranging from 0.30 to 0.60 (Figure 8A). That is, the mean end-to-end distance of self-avoiding C-SAC chains in a spherical confinement is a function of both the chain length N and the

confinement diameter D, when the length of N is larger than D. This confinement-dependent regime is illustrated in Figure 8 for both mean end-to-end distance and contact probability. Figure 6B includes data presented in Figure 5, with additional data for D of 10, 30 and 100 μm to depict comprehensively the relationship between α and the genomic distance. This helps to illustrate the important issue of the cross-over regime for self-avoiding chromatin and the convergence of the scaling exponent α. The asymptotic relationship of $\alpha = 3\nu$ (de Gennnes, 1979) is well-satisfied at larger D value, but less so at smaller D, as the leveling-off effects take place at shorter chain lengths with more severe confinement at smaller D.

Severe spatial confinement has pronounced effects on the conformations of self-avoiding polymers. Overall, we find that the effective scaling exponent slowly changes with increasing D, reflecting a rather slow convergence to the asymptotic behavior expected from simple polymer scaling theory (de Gennnes, 1979).

2.4 Discussion

Chromosomes reside within the severely confined space of the cell nucleus. However, the direct effects of nuclear confinement on chromatin folding and compaction are unknown. A major challenge is the extreme difficulty in adequate sampling of long self-avoiding chromatin chains in the confinement of the cell nucleus. Our C-SAC model enabled us to generate a large number of chromatin conformations in confinement.

Our results showed that the spatial confinement of $\sim$ 15 Mb chromatin within regions of diameter D of 1.5μm gives rise to the chromosomal scaling relationships of the average $\alpha \sim 1.05$ and the average $\nu \sim 0.34$, as well as the leveling-off effects observed experimentally (Mateos-

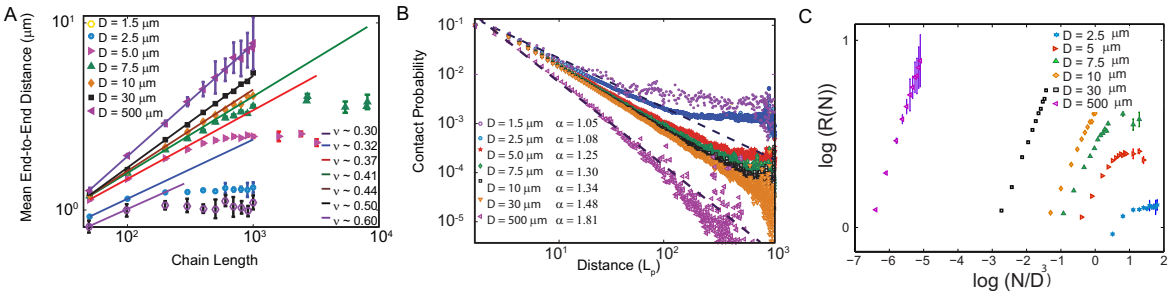

Figure 8. **Scaling properties of self-avoiding C-SAC chains in confinement. (A)** Relationship between the mean end-to-end distance and the chain length. Each data point is an average of 10,000 chains of different length under specific confinement of diameter D. As D increases, the scaling behaviour of self-avoiding walks converges to that of ideal SAW (ν=0.6). **(B)** Relationship between mean contact probability and partial chain length s. Each data point is an average of 10,000 chains of different length N under different confinement of diameter D. **(C)** Relationship between $R(N)$ and N/D^3 of C-SAC chains in confinement. Each data point is an average of 10,000 chains of length N under a specific confinement of D. Under severe confinement when D is small, $R(N)$ is influenced not only by the length of the polymer but also the size D of the confinement.

Langerak et al., 2009; Jhunjhunwala et al., 2008; Lieberman-Aiden et al., 2009). Our model also captured the complex folding behavior of the chromosome-specific variation in scaling (Barbieri et al., 2012). In addition, the tentative formation of domains (Cremer and Cremer, 2001; Cook, 1999; Kreth et al., 2004; Dixon et al., 2012; Nora et al., 2012) also emerged in C-SAC model as highly interactive substructures, without the need of introducing additional binder molecules and fine tuning of their concentrations. These interactive substructures could be stabilized by introducing more specific interactions through evolutionary selection pressure to form functional topological domains.

We found that D, and therefore nuclear size, is a major factor in influencing the overall folding landscape of chromatin. As nuclear size changes, there are significant differences in the chromosome architecture, which are reflected in variations in the scaling exponents. These conclusions are in good agreement with results from Hi-C studies using different cell lines (Barbieri et al., 2012). For example, lines of differentiated cells (GM06990 (Lieberman-Aiden et al., 2009), GM12878 (Kalhor et al., 2012), IMR90 (Dixon et al., 2012)) have similar overall average scaling behavior, with $\alpha \sim 1.08$, while embryonic stem cells (hESC) (Dixon et al., 2012) behave differently, with an α close to 1.6. A characteristic property of an hESC nucleus is that it occupies almost the entire cell volume (Zwaka and Thomson, 2003; Dahl et al., 2008) and is plastic and deformable (Pajerowski et al., 2007). This provides an enlarged space for chromosome organization. As a result, hESC chromatin is largely diffuse (Zwaka and Thomson, 2003). Our calculation also showed an increase in α when the confined space is enlarged. This observed variation in scaling corresponds well with the confinement of different nucleus sizes.

The average compactness of the chromatin chains and the fractions of open, compact, and in-between chromatin structures are all different when the nuclear size is changed. Nuclear size likely alters the overall structural organization of chromosomes, allowing previously unlikely long-range interactions to occur, at the same time prohibiting certain other genomic interactions present at a different nucleus size. Thus, nuclear confinement may bias distant sites towards spatial proximity.

Our results showed that randomly placed binders do not affect directly the scaling behaviour. Biological binders such as CTCF may play more specific roles of modifying or biasing chromosomes towards formation of specific domains required for cell function. Future work on the selection of properly placed CTCF binding and its effects will likely be fruitful for understanding the effects of biochemical binding on spatial organization of chromosome

We compared predictions from C-SAC models with those from other chromatin models. As experimentally observed $\nu \sim 0.3$ deviates significantly from the expected ν of 0.5 for sub-chains in equilibrium globule, chromatin fibers were conjectured to be in non-equilibrium fractal globule (FG) state, in which the exponent of $\nu \sim 0.3$ would be retained at every scale (Lieberman-Aiden et al., 2009; Mirny, 2011). The lack of leveling-off effects in $P_c(s)$ with s observed in (Lieberman-Aiden et al., 2009) is consistent with the prediction of the FG model. However, leveling-off effects are observed in FISH studies on different chromosomes at several different length scales (Mateos-Langerak et al., 2009; Jhunjhunwala et al., 2008). These leveling-off effects are not accounted for by the FG model. In addition, the significant variation of exponent

α among different chromosomes of human cells (Lieberman-Aiden et al., 2009; Barbieri et al., 2012) is not explained by the FG model.

An important consideration in studying the scaling relationship of chromosomes is the relevant size regime dictated by experimental observations. An average of 50–100 Mbp chromosome occupies a territory of size ~ 2 μm (Cremer and Cremer, 2001). As a result, a chromatin must traverse back and forth many times in the chromosome territory, and severe spatial confinement is at play and will have pronounced effects on the folding and scaling of chromatin fibers. General asymptotic scaling analysis of polymers is overly simplistic to offer much insight under such strong effects of finite sizes. Conventional simulation studies based on Metropolis Monte Carlo are also challenged to generate adequate samples to study the equilibrium ensemble of severely confined chromatin chains in the relevant regime.

Simulation using the novel technique of geometric sequential importance sampling allows the effects of finite size of confinement to be examined in detail. Our results offer an alternative explanation on the scaling relationship of chromosomes to the existing FG and SBS models. Overall, our results show that equilibrium ensemble of C-SAC chromatin chains under severe confinement of scale $D = 1.5$ μm exhibit scaling behavior consistent with known experimental data, which are different from that of asymptotic random chains in the relevant biological scale.

A useful result that can be inferred from our analysis is that chromosomes are restricted via confinement of sub-chromosomal regions of size about 15 Mb, each within a D of about 1.5 μm-diameter region. Therefore, it may be useful to consider the nucleus to be made up of close-packed regions of size D, each containing ~ 15 Mb of DNA. For example, one can consider

whole chromosomes to be made up of individual units of 15 Mb of DNA, confined to spherical regions of diameter ~ 1.5 μm. The whole human nucleus containing ~ 6 Gbp of DNA can be considered to be a collection of $\sim 6,000/15 = 400$ such units, which can be fit into a nucleus of diameter $\sim 400^{1/3}D \approx 7.4D \sim 11\mu$m, compatible with the observed size of human cell nuclei (Alberts, 2002). Therefore, the subchromosome confinement parameter D, namely, the size of a region containing 15 Mb of DNA, is an important parameter in our structural model.

As spatial confinement is a dominant factor in determining chromosome folding, the specific epigenetic state of genes and transcription activities in different cell types are likely influenced by the degree of nuclear confinement. Cell nucleus size at different developmental stages or physiological states may be altered to induce different chromosome folding landscape, enabling different genetic programming to be activated. Overall, how nuclear size and shape relate to cell size and shape, and how their relative ratio or pattern regulate the epigenetic programs of the cells at different developmental stages are important problems requiring further investigations.

Although our approach can generate a large ensemble of chromatin chains under spatial confinement, there still exists uncertainty in the physical parameters used in the current C-SAC model, including the persistence length, the chromatin fiber diameter, and the mass density (Fussner et al., 2012). In addition, current chromatin models are based on growing a single chromosome chain, and cannot be used to study inter-chromosomal interactions. Another question is how the 15 Mb sequence scale, and the parameter D are controlled in the cell. These issues will likely be resolved when chromosomal properties are better understood and the C-SAC algorithm is further improved. It is interesting that our simplistic model can

capture complex folding characteristic of human genome. The current study highlighted the importance of spatial confinement in dictating the chromatin folding landscape. With the accumulation of high resolution chromosome conformation capture data, it is envisioned that more specific spatial information inferred from 3C-based studies can be incorporated into the C-SAC model, and realistic ensemble of chromatin conformations reflecting 3C-based information can be reconstructed to gain insight into the structural basis of gene regulation and expression.

CHAPTER 3

SPATIAL ORGANIZATION OF BUDDING YEAST GENOME FROM LANDMARK CONSTRAINTS AND IDENTIFICATION OF BIOLOGICAL CHROMATIN INTERACTIONS

3.1 Introduction

Genome organization largely determines important nuclear activities such as repair, recombination, and replication of DNA, as well as the control of transcriptional status of genes (Fraser and Bickmore, 2007; Taddei and Gasser, 2012). The overall organization of genome has been shown to be compartmentalized in the form of chromosome territories (Cremer and Cremer, 2001), topologically associated domains (Nora et al., 2012; Dixon et al., 2012), and spatial localization of individual gene loci (Berger et al., 2008). Such compartmentalization affects the expression levels of genes in eukaryotes such as yeast (Taddei and Gasser, 2012) and mammalians (Fraser and Bickmore, 2007). With the well understood nuclear architecture and transcriptional machineries (Taddei and Gasser, 2012), budding yeast provides an excellent model system for investigating cellular activities related to genome organization. Furthermore, there is now clear evidence that important nuclear events such as cancer-promoting chromosomal translocations observed in human nuclei and relocation of genomic elements upon breaks of double stranded DNA observed in budding yeast originate from analogous underlying cellular machineries (Taddei and Gasser, 2012).

Studies using electron microscopy techniques revealed detailed structures of architectural landmarks of budding yeast nucleus. These include the spindle pole body (SPB), the nucleolus, and the nuclear envelope (NE) (Hediger et al., 2002; Taddei et al., 2004; Taddei et al., 2009; Mekhail and Moazed, 2010; O'Toole et al., 1999; Yang et al., 1989; Dvorkin et al., 1991; Bystricky et al., 2005; Berger et al., 2008). SPB is functionally equivalent to centrosome in mammalian nuclei, where all heterochromatic centromeres are attached to throughout interphase (O'Toole et al., 1999). Nucleolus, where ribosome synthesis and assembly take place, contain clusters of ribosomal DNA (rDNA) repeats (Yang et al., 1989; Dvorkin et al., 1991; Bystricky et al., 2005; Berger et al., 2008; Mekhail and Moazed, 2010; Taddei et al., 2010). NE, where telomeric regions of yeast chromosomes are anchored, facilitates silencing of telomeric genes (Hediger et al., 2002; Taddei et al., 2004; Taddei et al., 2009; Mekhail and Moazed, 2010). In addition, microscopy experiments further revealed the dynamics behavior of important genes of budding yeast (Berger et al., 2008).

With genome-wide studies using Chromosome Conformation Capture (3C) technique (Duan et al., 2010; Lieberman-Aiden et al., 2009), large-scale long-range chromatin looping interactions across the budding yeast genome have been revealed (Duan et al., 2010). Studies of polymer modeling of both human (Sanborn et al., 2015; Chiariello et al., 2016; Trieu and Cheng, 2014; Zhang and Wolynes, 2015; Meluzzi and Arya, 2013) and yeast (Tjong et al., 2012; Wong et al., 2012; Wang et al., 2015; Tokuda et al., 2012) genomes have revealed important information on the folding principle of genome. For example, recent computational studies demonstrated that chromosomes of budding yeast behave as randomly folded flexible self-avoiding polymer

chains that are subject to the constraints of nuclear landmarks and nuclear confinement (Tjong et al., 2012; Wong et al., 2012). It was shown that tethering of genomic elements such as centromeres and telemores to the nuclear landmarks gives rise to the preferential localization of functional loci in nucleus (Tjong et al., 2012; Wong et al., 2012). However, the correlations of modeled inter-chromosomal interactions with experimentally captured interactions are modest at best (Tjong et al., 2012; Wong et al., 2012). In addition, these volume exclusion models (Tjong et al., 2012; Wong et al., 2012) may be able to capture only interactions arising from generic polymer effects. After correction of measured interaction frequencies using a statistical null model, budding yeast genome no longer exhibit properties of randomly folded polymer chains under constraints (Ay et al., 2014). The question whether the organization of yeast genome is dictated by physical tethering of landmarks and the excluded-volume effects as argued in (Tjong et al., 2012; Wong et al., 2012), with specific protein-mediated interactions playing negligible roles, remains unanswered. Overall, the exact roles of nuclear landmarks, volume confinement, biochemically mediated interactions, as well as their relative contributions to the overall organization of yeast genome are unclear.

In this study, we explored computationally the structural properties of budding yeast genome under different combinations of landmark constraints and nuclear confinement. Our goal is to answer these questions: (1) how does the confinement of cell nucleus affect the organization of yeast genome, (2) to what extent genome organization is determined by the physical architecture of the nucleus through landmarks, (3) what are the contributions of the individual nuclear landmarks on overall genome organization, (4) how can we distinguish chromatin looping in-

teractions arising from biochemical factors from those arising from generic polymer properties. Our study is based on the multi-chromosome Constrained Self-Avoiding Chromatin (mC-SAC) method and the generation of ensembles of ∼150,000 model genomes using the geometrical Sequential Importance Sampling technique (g-SIS) (Gürsoy et al., 2014a; Gürsoy et al., 2014b).

Our results showed that indeed the overall patterns of chromatin interactions of budding yeast genome are well captured when only polymer effects under the spatial confinement of cell nucleus and landmark constraints are considered (row-based Pearson correlation coefficient R of 0.95). We found that the size of the nuclear confinement is the key determinant of intra-chromosomal interactions, while centromere tethering is responsible for much of the observed inter-chromosomal interactions and correlation of pairwise telomere distances to chromosomal arm lengths. Furthermore, novel chromatin interactions undetected in experimental studies (Duan et al., 2010) can be uncovered from the ensemble of model genomes generated with nuclear confinement and landmark constraints, and are found to be stabilized by binding of a transcription factor and RNA polymerase. In addition, we found there are important specific genomic elements enriched with tRNA genes that were not captured by polymer properties under landmark constraints, but are detected in experimental studies (Duan et al., 2010). Overall, our findings define the specific roles of confinement and individual landmarks, and can uncover likely biologically relevant interactions from genome-wide 3C measurements that are beyond polymer effects.

3.2 Materials and Method

3.2.1 Model and parameters

Budding yeast nuclear architecture is composed of NE, SPB, nucleolus and 16 chromosomes. The locations of SPB, NE and nucleolus are fixed according to the imaging experiments (Figure 9A) (Hediger et al., 2002; Taddei et al., 2004; Taddei et al., 2009; Mekhail and Moazed, 2010; O'Toole et al., 1999; Yang et al., 1989; Dvorkin et al., 1991; Bystricky et al., 2005; Berger et al., 2008) and the locations of the 16 chromosomes are modeled as independent but interacting polymer chains (Gürsoy et al., 2014b).

In our mC-SAC model, we used 30 nm chromatin fiber model (Bystricky et al., 2004; Wedemann and Langowski, 2002; Gürsoy et al., 2014a; Gürsoy et al., 2014b), in which each monomer of the polymer chain is modeled as spheres with 30 nm diameter and corresponds to a 3 kb of DNA (Bystricky et al., 2004; Wedemann and Langowski, 2002). Every 5 monomers form a persistence unit that corresponds to a persistence length L_p of 150 nm (Bystricky et al., 2004; Wedemann and Langowski, 2002). The entire budding yeast genome is modeled a total of 796 L_p (3990 monomers) divided into 16 chromosomes.

3.2.2 Chain growth by geometrical Sequential Importance Sampling (g-SIS)

The mC-SAC model is developed based on our single C-SAC chain growth model (Gürsoy et al., 2014a; Gürsoy et al., 2014b). First, we mapped the locations of centromeres, telemores and rDNA repeats onto the polymer chains that corresponds to each chromosome. Each chromosome is the divided into right and left arms from their centromeres, except Chr 12 (Figure 9D). The

polymer chain representing Chr 12 is divided into three segments to accommodate for the nucleolus constraint (Figure 9D).

The budding yeast genome is therefore composed of 33 chromosomal arms, each represented by a polymer chains. The genome $\gamma = (x^1, x^2, ..., x^{33})$ is a collection of chromosomal arms, where each arm x^k consists of n units as $x_k = (x_1^k, x_2^k, ..., x_n^k)$. The three-dimensional location of the i-th unit of the k-th chromosome arm is denoted as $x_i^k = (a_i^k, b_i^k, c_i^k) \in \mathbb{R}^3$.

To generate a chromosomal arm, we grow the mC-SAC chain one unit at a time, ensuring the self avoiding property along the way, namely, $x_i^k \neq x_j^l$ for all $i \neq j$. We use a $s = 1640$-state off-lattice discrete model (see (Liang et al., 2002; Zhang et al., 2003; Gürsoy et al., 2014a; Gürsoy et al., 2014b) for more details). The new unit added to a partial chain is placed at x_{t+1}^k, taken from one of the unoccupied s-sites neighboring x_t^k, with a probability of growth $g(\boldsymbol{x})$, which is the trial distribution. This selection introduce bias away from the target distribution $\pi(\boldsymbol{x})$, therefore the bias is corrected by assigning each successfully generated genome a proper weight $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/g(\boldsymbol{x})$. Details can be found in references (Liang et al., 2002; Zhang et al., 2003; Gürsoy et al., 2014a; Gürsoy et al., 2014b).

Multiple chain growth process starts with a random selection of a chromosomal arm and placement of its corresponding centromere at a random location in the SPB. We then employ the chain growth strategy to grow chromosomal arms until the telomere of the corresponding arm reaches to the target location, *i.e.* NE. In the case of Chr 12, we select a random location on the nucleolus to place the rDNA repeats and grow the chain towards to the target location(*i.e.* NE or SPB). We repeat this process until all 33 chromosomal arms are completely generated.

3.2.3 Target distribution

The target distribution $\pi(\boldsymbol{x})$ is Boltzmann distribution, in which all chains are self avoiding, their centromeres are attached to the SBP, the rDNA repeats are in the nucleolus and telomeres can be attached to NE at any point of growth. The first persistence unit of each chromosomal arms (except for Chr12) are randomly attached to any location in the SPB. Each partial chromosomal arm x_t^k is grown from centromeres according to the target distribution $\pi(x_t)$ based on the geometrical constraints derived from experimental data by conserving the self-avoiding property and confinement of cell nucleus.

The target distribution $\pi(x_t^k)$ of a partial chain follows Boltzmann distribution as

$$\pi(x_t^k) = \exp(-E(x_t^k)/k_B T),$$

where $E(x_t^k)$ is an energy like term that is derived from the landmark constraints.

(1) Potential from telomere closing constraints

This potential is designed to obtain model genomes where the telemores are either attached to the NE when the full arm length is reached or can be attached to the NE at any point of chain growth,

Let $H_1(x_t^k)$ be the potential from telomere closing probability constraints. For each candidate node x_{t_m} that does not violate the self-avoiding property and inside the nuclear confinemet, we calculate the energy-like term as

$$H_1(x_t^k) = |||\ x_{t_m}\ \| - R| - (N-t) \times L_p - d_{thres}|, \tag{3.1}$$

where L_p is the persistence length, N is the total number of nodes in a chromosomal arm, R is the nuclear radius and d_{thres} is the threshold distance which was taken as 50 nm.

(2) Potential from centromere tethering constraints.

This potential is used only for the Chr12 chromosomal arms where the rDNA repeats are sampled from nucleolus and designed to obtain model genomes where the centromere are either in the SPB when the full arm length is reached or can be in the SPB at any point of chain growth,

Let $H_2(x_t^k)$ be the potential from centromere tethering constraints. For each candidate node x_{t_m} that does not violate the self-avoiding property and inside the nuclear confinemet, we calculate the energy-like term as

$$H_2(x_t^k) = |||\ x_{t_m} - x_{SPB}\ \| - R_{SPB}| - (N-t) \times L_p|, \tag{3.2}$$

where x_{SPB} is the center coordinates of SPB, L_p is the persistence length, N is the total number of nodes in a chromosomal arm, and R_{SPB} is the radius of SPB as we modeled as a sphere.

3.2.4 Trial distribution

Trial distribution is designed to introduce a bias to chose the highest probability partial chain x_t^k with respect to the target distribution $\pi(x_t^k)$. The trial distribution $g(x_t^k)$ of a partial chain x_t^k is

$$g(x_t^k) = \exp(\pi(x_t^k) - \max_{t=1,\ldots,1640} \pi(x_t^k))$$

.

3.2.5 Random model

An ensemble of 150,000 model genomes with only excluded volume constraint and nuclear confinement are generated. To improve the sampling efficiency, we employ a dynamic resampling technique that is described in Chapter 2 (Gürsoy et al., 2014a).

3.2.6 Statistical properties of model genomes

With m successfully generated model genomes, the physical properties of the ensembles of model genomes are calculated. If the configurations of the j-th successfully generated model genome as $\boldsymbol{x}^{(j)} = (x_1^{(j)}, \cdots, x_n^{(j)})$, and its associated weight $w^{(j)}$. To calculate the mean value of a physical property $\bar{h}(\boldsymbol{x})$ such as the spatial distance between genomic elements, we have:

$$\bar{h}(\boldsymbol{x}) = \mathbb{E}_{\pi(\boldsymbol{x})}[h(\boldsymbol{x})] = \frac{\sum_{j=1}^{m} h(\boldsymbol{x}^{(j)}) \cdot w^{(j)}}{\sum_{j=1}^{m} w^{(j)}}.$$

3.2.7 Normalization and calculation of propensity

We normalized the interaction frequencies of experiments and the model ensembles following the previous work (Tjong et al., 2012; Lieberman-Aiden et al., 2009; Duan et al., 2010). Let f_{ij} be the interaction frequencies between the genomic elements i and j. We obtained the normalized interaction frequency as

$$f_{ij}^{n} = f_{ij} \times \frac{\sum_{k=1}^{N} \sum_{l=k+1}^{N} f_{kl}}{\sum_{k=1}^{N} f_{ik} \sum_{k=1}^{N} f_{kj}},$$

where N is the total number of the genomic elements. All the calculations in this paper are employed after normalization of experimental and model ensembles.

The propensity of an interaction is the observed/expected for the experiment and the modeled ensembles. First, we calculated the probability of an interaction in the experimental interaction matrix, interaction matrix of modeled ensemble and random model as following,

$$q^{exp}(ij) = \frac{f^{exp}(ij)}{\sum_{i=1}^{N} \sum_{j=1}^{N} f^{exp}(ij)}$$

$$q^{model}(ij) = \frac{\sum_{k} w_k I(i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k} w_k I(i,j)},$$

$$q^{random}(ij) = \frac{\sum_{k} w_k I(i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k} w_k I(i,j)},$$

where N is the total number of genomic elements, w_k is the weight of the $k^t h$ chain in the ensemble and $I(i,j)$ is an indicator function, which equals to 1 when elements i and j interacts, equals to 0 otherwise. We calculated the propensity of each interaction as,

$$propen^{exp}(ij) = \frac{q^{exp}(ij)}{q^{random}(ij)}$$

$$propen^{model}(ij) = \frac{q^{model}(ij)}{q^{random}(ij)}$$

3.2.8 Calculation of p-value for the correlation between experimental matrix and model ensemble matrix

We shuffled the each row of the experimental interaction matrix for 1000 times and generated 1000 shuffled interaction matrices. We calculated the mean row-based Pearson correlation coefficient between each shuffled matrix and the modeled ensemble and calculated the probability of obtaining mean row-based Pearson correlation coefficient of 0.95 as the p-value.

3.2.9 Mean combined occupancy enrichment

We mapped the genome-wide occupancy enrichment of RNAPIII and TFIIS on to beads. We used a geometrical mean approach for the coupled enrichment of pairs. Mean enrichment value for each pair can be measured as

$$en_{mean}(i) = \sqrt{\sqrt{(en_{RNAPIII}(i) * en_{RNAPIII}(j)} * \sqrt{en_{TFIIS}(i) * en_{TFIIS}(j)}}$$

3.3 Results

3.3.1 mC-SAC model of budding yeast genome

We model the chromatin fiber of budding yeast as chained beads, where each bead corresponds to 3 kb of DNA (Bystricky et al., 2004; Gerchman and Ramakrishnan, 1987; Wedemann

and Langowski, 2002). Following previous studies (Tjong et al., 2012; Wong et al., 2012; Gürsoy et al., 2014b), we used light microscopy data to model the architecture of yeast nucleus. The nucleus is modeled as a sphere of a diameter of 2 μm and contains the Spindle Pole Body (SPB), the Nuclear Envelope (NE, modeled as a shell of thickness of 0.5 nm), the nucleolus, and 16 chromosomes (Figure 9A,B and D) (Gürsoy et al., 2014b). Chromosomes all reside inside the nucleus as independent but interacting self-avoiding chromatin fibers. The entire budding yeast genome is modeled as a total of 3,990 beads divided into 16 different chromosomes (Figure 9B).

An ensemble of ∼150,000 independent model genomes are generated that are subject to the nuclear confinement, centromere clustering at SPB, telomere attachment at the NE, and rDNA repeat clustering at the nucleolus. This is achieved by sequentially growing self-avoiding chromatin chains one unit at a time using the technique of geometrical Sequential Importance Sampling (g-SIS) (Gürsoy et al., 2014a; Gürsoy et al., 2014b; Liang et al., 2002; Lin et al., 2008b). We call this *fully-constrained ensemble* of mC-SAC chains. In addition, we examined the effect of individual landmark constraints and generated four separate ensembles of ∼150,000 independent model genomes, after turning off each of the three separate constraints and with only the constraint of centromere tethering imposed (see Table II). As an overall control, we also generated a *random ensemble* of ∼150,000 model genomes, which is subject only to the constraint of nuclear confinement (see Table II).

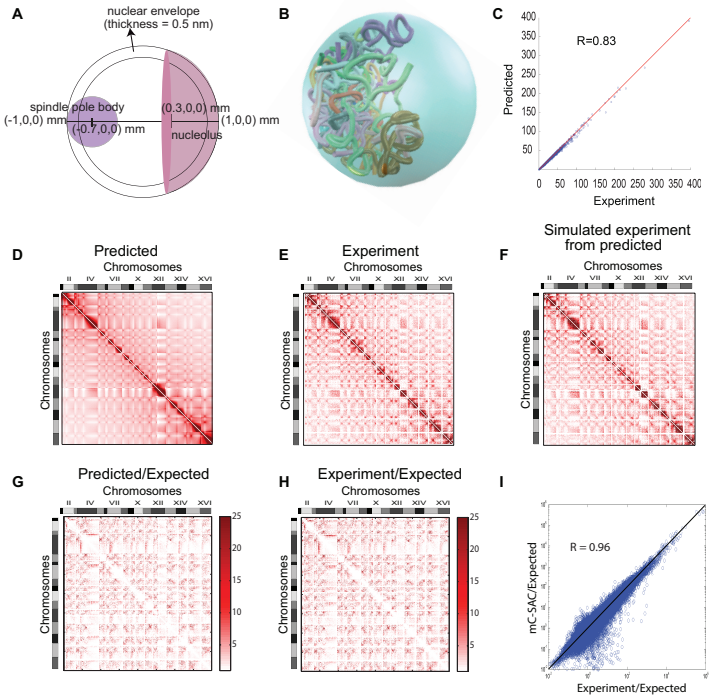

Figure 9. **Nuclear architecture of budding yeast and the mC-SAC model of budding yeast genome.** **(A)** Schematic representation of the nucleus and nuclear landmarks of budding yeast and their corresponding coordinates and the dimensions. **(B)** An example 3D structure of mC-SAC genome confined in the cell nucleus. **(C)** Correlation between genome-wide chromatin conformation capture interaction frequencies and the interaction frequencies measured from the fully-constrained mC-SAC ensemble of model yeast genomes. **(D)** Yellow chromosome represents the Chr12 where the rDNA elements are highlighted as blue spheres and the centromere is highlighted as red sphere. Purple chromosome represents the rest of the chromosomes where centromere is highlighted as red sphere. The direction of chain growth is shown with the arrows. **(E)** Schematic representation of the chromosomes and the special case of Chr 12 where we used 3 chromosomal arms for chain growth process. **(F)** Histogram of the mean row-based correlation coefficients between shuffled experimental data and the model ensemble. **(G)** Heatmap of the interaction frequencies measured in the fully-constrained mC-SAC ensemble. Darker color indicates higher interaction frequency. **(H)** Heatmap of the interaction frequencies from the experiment. **(I)** Heatmap of interactions in the fully-constrained mC-SAC ensemble The interactions between restriction fragments of the genome-wide 3C experiment (Duan et al., 2010) are shown for direct comparison between the predicted model and experiment. **(J)** Heatmap of the interaction frequencies of the fully-constrained mC-SAC ensemble that are corrected after removal of non-specific interaction frequencies. **(K)** Heatmap of the interaction frequencies of the genome-wide 3C experiments that are corrected by expected interaction frequencies. **(L)** Correlation between genome-wide chromatin conformation capture interaction frequencies and interaction frequencies from the fully-constrained mC-SAC ensemble after removal of expected interactions.

TABLE II

THE EFFECTS OF DIFFERENT CONSTRAINTS ON THE FOLDING OF BUDDING YEAST GENOME

	mC-SAC	without telomere	without nucleolus	without centromere	with only centromere	Random
Overall	0.82	0.83	0.90	0.90	0.81	0.77
Inter	0.91	0.90	0.92	0.76	0.90	0.54
Intra	0.82	0.84	0.90	0.92	0.82	0.80

Correlation coefficients between the interactions of different ensembles and the genome-wide chromosome conformation capture experiments at 15 kb resolution. Spatial confinement of a nucleus of diameter 2 μm is imposed for all cases.

3.3.2 mC-SAC model with nuclear confinement and landmark constraints recapitulates long-range chromatin interactions of budding yeast genome

Recent genome-wide Chromosome Conformation Capture (3C) studies have quantified the frequency of chromatin looping interactions of budding yeast genome that can be summarized by an interaction frequency matrix (Duan et al., 2010). To examine how well our model can capture the overall genome organization, we calculated the correlation between interaction frequency matrices from the fully-constrained ensemble and from genome-wide 3C experiment (Duan et al., 2010) following previous studies (Tjong et al., 2012; Wong et al., 2012). Interaction frequency matrices obtained from our predicted ensemble (Figure 9G and I) and from genome-wide 3C experiments (Figure 9H) are highly correlated, with an R of 0.83 at 15 kb resolution (as calculated in (Wong et al., 2012), with R of ~ 0.60 reported) and a row-based R of 0.94 at 15 kb (as calculated in (Tjong et al., 2012), with R of 0.94 at 32 kb reported, Figure 9C, p-value ¡ 0.001, Figure 9F). Furthermore, the calculated inter-chromosomal interaction frequencies in the

fully-constrained ensemble and those observed in genome-wide 3C experiments are in excellent agreement, with an R of 0.91 at 15 kb resolution, compared to previously reported R of 0.54 at 32 kb resolution (Tjong et al., 2012). The heatmaps obtained from experiments (Duan et al., 2010) and from mC-SAC ensemble have nearly identical patterns (Figure 9G–I).

To eliminate the effect of proximity interactions and non-specific interactions arising from nuclear confinement of self-avoiding chromatin chains, we used our random ensemble as the null model to calculate the propensity (observed/expected) of each interaction in both fully-constrained ensemble (Figure 9J) and the genome-wide 3C data (Figure 9K). After removal of non-specific interactions, the propensities from the fully-constrained ensemble and propensities from genome-wide 3C measurements has strong correlation, with an R of 0.96 at 15 kb resolution (Figure 9I).

Overall, our fully-constrained models of budding yeast genome showed that model genomes generated under the constraints of nuclear confinement and all three nuclear landmarks can capture much of the experimentally measured intra- and inter-chromosomal interactions. These results suggest that nuclear confinement and nuclear landmarks play key roles in determining the overall organization of yeast genome.

3.3.3 Nuclear size is a major determinant of overall spatial chromatin interactions in the budding yeast genome

Effects of confinement on patterns of genome-wide interactions.

To understand the effects of the nuclear confinement on chromatin interactions, we examined the frequency of interactions of model yeast genome with different degrees of confinement in

nuclei of diameters of 2, 4 and 16 μm, respectively, each with and without landmark constraints. A total of 6 ensembles, each with ∼150,000 model genomes are generated. As the nuclear diameter increases, the correlation between the interaction frequencies of fully-constrained ensemble and those of genome-wide 3C experiments decreases from R of 0.83 to 0.55 (Figure 10A). When the landmark constraints are removed, the interaction frequencies of random ensemble and frequencies of genome-wide 3C experiments decreases from R of 0.77 to 0.25 as the nuclear diameter increases from 2 μm to 16μm (Figure 10A). These results showed that the degree of confinement is a major source of the organization of budding yeast genome, as when only nuclear confinement constraint is employed, the correlation R is still quite strong at $R = 0.77$, so long as the appropriate confinement size is imposed.

Effects of confinement on pairwise distances between telomeres.

Fluorescence imaging data suggested that telomeres are positioned on the nuclear periphery according to their armlengths (Therizols et al., 2010). The distance between two telomeres increases as the length of the chromosomal arms increase (Therizols et al., 2010). The function that represents this relationship can be represented by two linear regimes, with a change in the slope at around the arm length of 310–326 kb (Therizols et al., 2010).

We examined the origin of this correlation. In the fully-constrained ensemble at a nuclear diameter of 2μm, the median telomere-telomere distances and chromosome arm lengths are linearly correlated in two regimes, with a change in the slope at around 356 kb for small chromosomes, and 396 kb for longer chromosomes (Figure 10B). This behavior is fully consistent with experimental findings (Therizols et al., 2010) as well as results from a previous polymer

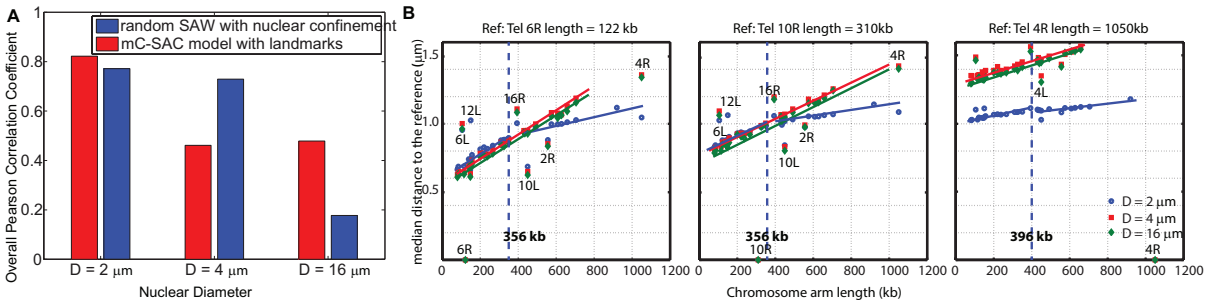

Figure 10. **Effects of confinement on the overall folding behavior of budding yeast genome. (A)** Overall correlation coefficient of the frequencies between genome-wide 3C measurements and modeled ensemble. As the nuclear size increases, correlation decreases. **(B)** Effects of nuclear size and chromosomal arm length on the median distances between telomeres. Relationships between arm length and median telomere distances at different nuclear sizes for the fully-constrained ensemble, with different telomeres as references are shown. Two linear regimens becomes one linear regime as D increases from 2 μm to 4 and 16 μm.

model (Tjong et al., 2012). However, we found that only long chromosomal arms can be mapped to two linear regimes when cell nuclei is enlarged to D=4 and D=16 μm (Figure 10B). For shorter arm lengths, telomere distances and chromosomal arm lengths have a single linear regime (Figure 10B). In contrast to suggestions from a previous study (Therizols et al., 2010) that the telomeres of budding yeast genome are not positioned randomly, our results suggest that telomeres indeed positioned randomly on the nuclear envelope. The size of the nucleus and the chromosomal arm lengths largely determine the random positioning of telomeres.

3.3.4 Attachment of centromeres to SPB is a major determinant of inter-chromosomal interactions

After turning off individual constraints one or more at a time, we generated ensembles of ∼150,000 model genomes to understand the specific roles of each landmark on the folding pattern of budding yeast genome. In total, we have 4 ensembles for the conditions of "without telomere", "without nucleolus", "without centromere", and "with only centromere" constraints, and the nuclear confinement and self-avoiding property of chromatin are imposed in each case. The overall correlation between the interaction frequencies from each ensemble and frequencies from experimental measurements is strong (R ¿ 0.80, Table II), suggesting again nuclear confinement and excluded-volume effects that are common to all four ensembles are the dominant factors in determining the overall interaction patterns of the budding yeast genome.

Inter-chromosomal interactions in most of ensembles, except the ones in which centromere tethering is off, are also highly correlated with experimentally captured inter-chromosomal interactions. These findings suggest that imposing the constraint of centromere tethering to the SPB in addition to the volume confinement is sufficient to capture inter-chromosomal interactions observed in genome-wide 3C experiments. We further examined the importance of centromere tethering on the pairwise distances between telomeres. When the centromeres are not attached to the SPB, the linear relationship between pairwise telomere distances and chromosomal arm lengths that was observed in fluorescence imaging experiments disappears (Figure 11).

Overall, our results showed that centromere attachment to the SPB largely determines the chromosome-chromosome interactions, hence the chromosomal positioning in the nucleus.

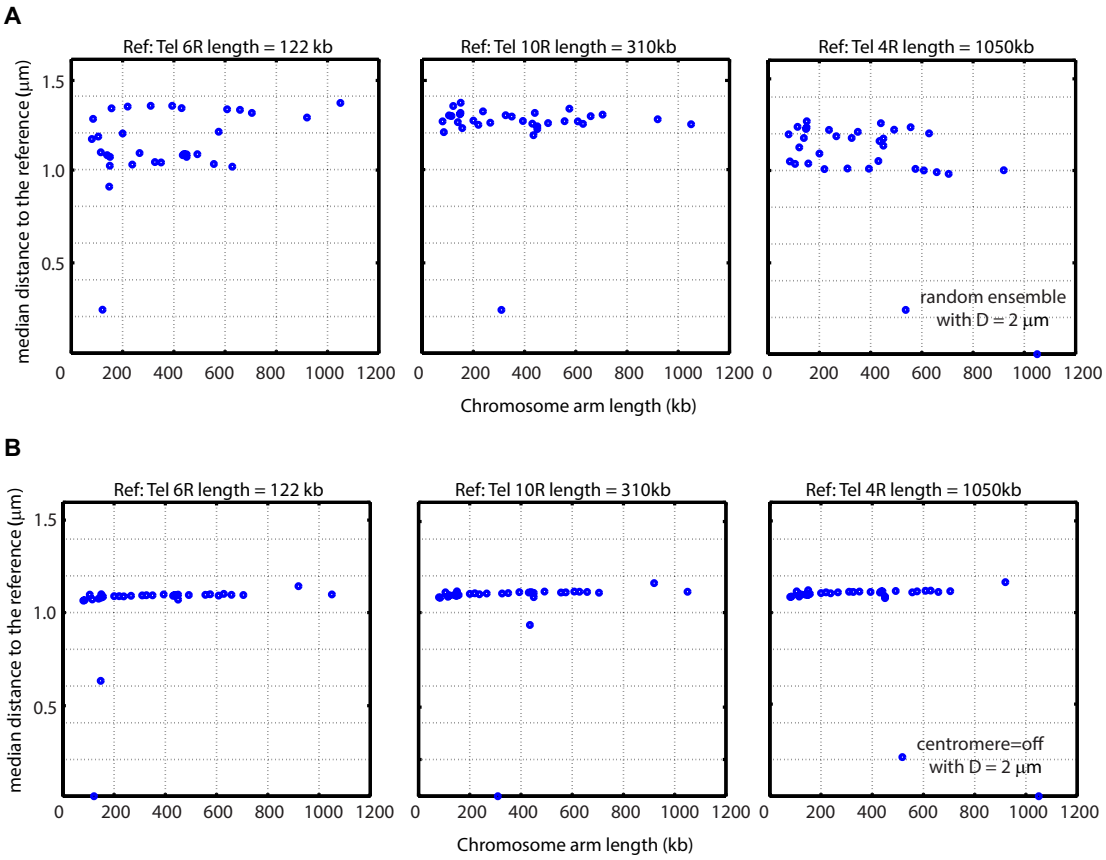

Figure 11. **The effect of centromere tethering on the median distances between telomeres. (A)**Relationship between chromosome armlength and median telomere distances for the random model. No correlation between armlength and the median telomeric distances was observed. **(B)** Relationship between chromosome armlength and median telomere distances for the "centromere=off" ensemble. No correlation between armlength and the median telomeric distances was observed.

The folding landscape of individual chromosomes, on the other hand, is largely determined by the crowding effects due to nuclear confinement and volume exclusion. Furthermore, telemore attachment to the NE has insignificant effects on the overall structural properties of the genome. In addition, nucleolus constraint has effects only on the folding of Chromosome 12 (Figure 12).

3.3.5 Spatial location of eight important genes are determined by their genomic distances to the centromeres.

The spatial locations of genes affect their transcriptional status (Fraser and Bickmore, 2007). The relative locations of eight important budding yeast genes with respect to the SPB were measured in a fluorescent imaging study (Berger et al., 2008). We compared experimentally observed relative positions of these genes with positions measured from the fully-constrained ensemble. Overall, they are in excellent agreement (R^2 =0.95, Figure 13A).

The relative position of these genes were found to be inversely correlated with their genomic distances to corresponding centromeres in a previous study (Berger et al., 2008). The same relationship is also observed in our model (Figure 13B and Figure 13C). We hypothesise that the relative locations of these genes are determined by their genomic distances to centromeres. To test this hypothesis, we generated two artificial genomes that have the same overall genome size and architecture as the budding yeast nucleus. Artificial Genome 1 (AG1) has the same number and lengths of chromosomes as the budding yeast genome, but with randomized locations of the centromeres. Artificial Genome 2 (AG2) has only 12 chromosomes, with the locations of centromeres also randomized. We found the same cross-like pattern in the interaction frequency heatmap as the budding yeast genome for AG1 and AG2 (Figure 13D and E), suggesting that

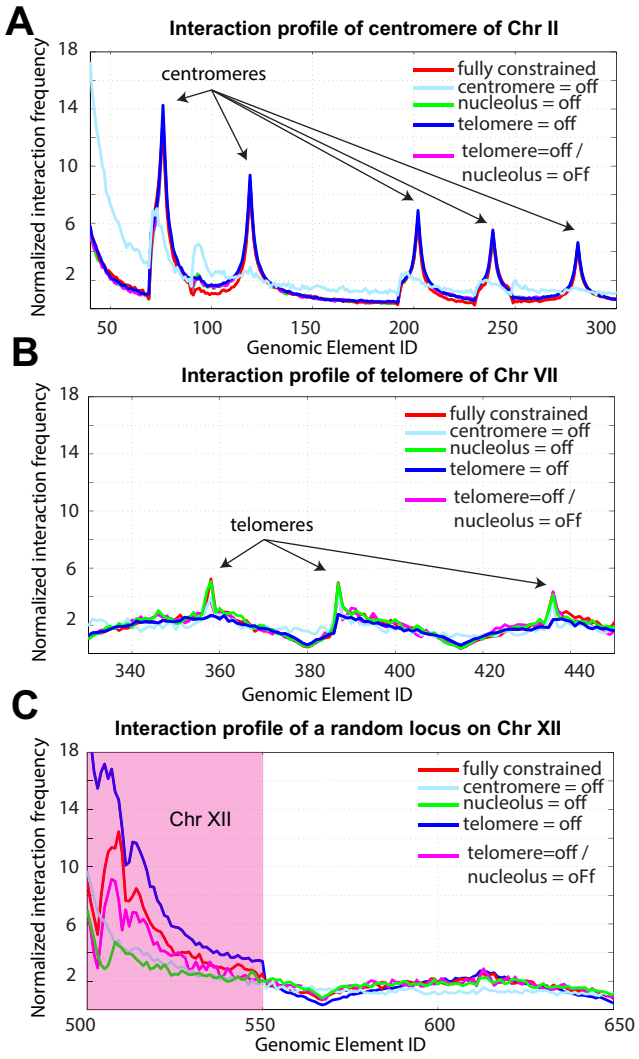

Figure 12. **Effects of different constraints on the interaction profiles of different genomic elements.** **(A)** Interactions between the centromere of Chr II and the other genomic elements in the yeast genome. The interactions between the centromere of Chr II and the other centromeres are the same for the ensembles, in which centromere constraint is on. Despite the high-correlation coefficient between the experiment and the ensemble of centromere =off, this ensemble fails to capture the centromere-centromere interactions. **(B)** Interactions between the telomere of Chr VII and the other genomic elements in the yeast genome. The interactions between the telomere of Chr VII and the other telomeres are the same for the ensembles, in which telomere constraint is on. However, since the frequency of telomere-telomere interaction is low, the ensembles, in which telomere constraint is off still have high correlation with Hi-C data. **(C)** Interactions between a random locus on Chr XII and the other genomic elements in the yeast genome. The intra-chromosomal interactions within Chr XII differs for all the ensembles.

the number and the length of the chromosomes have little effects on the overall pattern of yeast genome organization.

However, when the genomic locations of the eight genes were mapped to the artificial genomes, their relative positions deviate significantly from the experimentally measured positions (R^2=0.16 and R^2=0.11 for AG1 and AG2, respectively, Figure 13F). Surprisingly, the inverse relationship between the genomic distance to the corresponding centromere and the relative positions of these genes observed in wild type yeast is preserved ($R^2 = -0.87$ and $R^2 = -0.91$ for both artificial genomes, respectively, Figure 13G).

We further compared experimentally measured relative positions of these genes with their positions obtained from the ensembles of "with only centromere" and "without centromere" to explore the roles of centromere tethering on genome organization. The ensemble of "with only centromere" captured the relative spatial positions of these genes quite well ($R^2 = 0.88$, Figure 13H), whereas the relative positions in the ensemble of "without centromere" do not correlate well with experimental measurements (R^2=0.11, Figure 13I).

Overall, these results further suggest that centromere tethering is a key determinant of the folding of yeast genome and the positions of several important genomic elements are largely determined by their genomic distances to their corresponding centromeres.

3.3.6 Chromosomal fragile sites are clustered in three-dimensional space

In eukaryotes, chromosome can break at specific locations when DNA replication is perturbed (Song et al., 2014). These specific locations are called fragile sites. A recent genome-

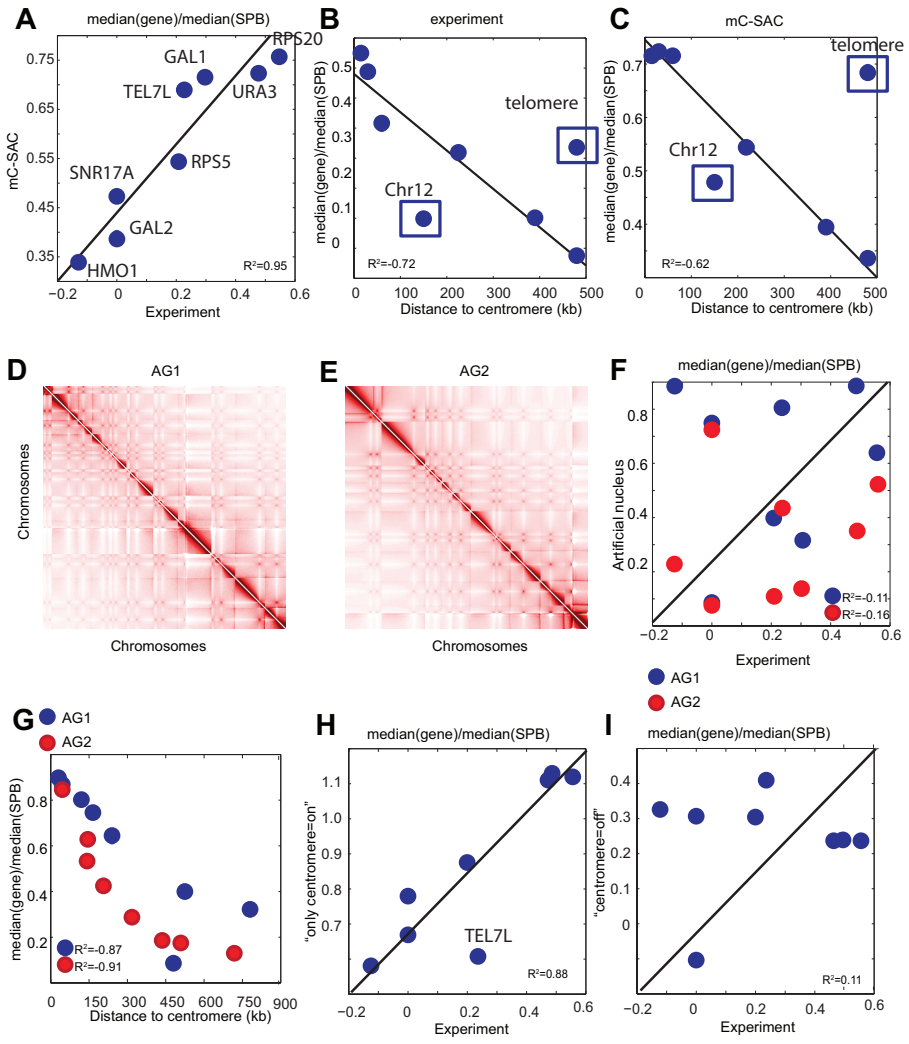

Figure 13. **Relationship between sequence and spatial positions of eight genes.** **(A)** The correlation between the relative positions of these genes by electron microscopy (Berger et al., 2008) (x-axis) and by fully-constrained ensemble (y-axis). **(B)** The relationship between the experimentally measured relative spatial positions of the important genes and their distance to the corresponding centromeres. The two locations of genes that correlate poorly are on Chr12 and telomere, which are subject to nucleolus and telomere attachment constraints. **(C)** The same relationship can be seen from computationally generated fully-constrained ensemble. **(D)** Heatmap of interaction frequencies of Artificial Genome 1 (AG1) with 16 total chromosomes. **(E)** Heatmap of interaction frequencies Artificial Genome 2 (AG2) with 12 total chromosomes. **(F)** The correlation between the relative position of the genes measured experimentally and measured from AG1 (blue) and AG2(red) ensembles. **(G)** The relationship between the relative positions of the genes measured from AG1 (blue) and AG2 (red) ensembles and their distances to the corresponding centromeres. **(H)** The correlation between the relative positions of the genes measured by electron microscopy (Berger et al., 2008) and by "with only centromere" ensemble. **(I)** The same correlation between the positions measured by electron microscopy (Berger et al., 2008) and by "without centromere" ensemble.

wide study of fragile site mapping revealed all breakable sites of budding yeast genome with high-resolution (Song et al., 2014).

We mapped all 201 experimentally identified fragile sites to beads in our polymer model of yeast genome and calculated the mean interaction frequencies among them. Only non-local interactions between fragile sites that are more than 45 kb apart are considered, so proximity effects are eliminated in our consideration. Overall, the mean interaction frequency between the 95 mapped beads containing fragile sites is 35.9. The random probability of observing similar or higher frequency is $p < 0.001$ (Fig Figure 14A), which is estimated through bootstrapping of 10,000 sets of 95 random beads that are at least 45 kb apart. These results showed that fragile sites have high propensity of clustering spatially together in the nucleus (Figure 14B and C), indicating that the underlying mechanism of double-stranded DNA breaks coming together in 3D space to create a repair foci (Lisby et al., 2003) may be facilitated by the centromere tethering and the confinement of the cell nucleus.

3.3.7 Predicting novel long-range chromatin interactions of budding yeast genome

While genome-wide 3C technique has identified many long-range pairwise chromatin interactions in budding yeast (Duan et al., 2010), these interactions are incomplete due to the distribution of restriction enzyme sites and lack of full mappability of the fragments. Our fully-constrained ensemble can be used to predict novel interactions that are not captured by genome-wide 3C experiments. In addition, as much of the spatial organization of budding yeast genome is likely dictated by the landmark constraints and nuclear confinement, it would be

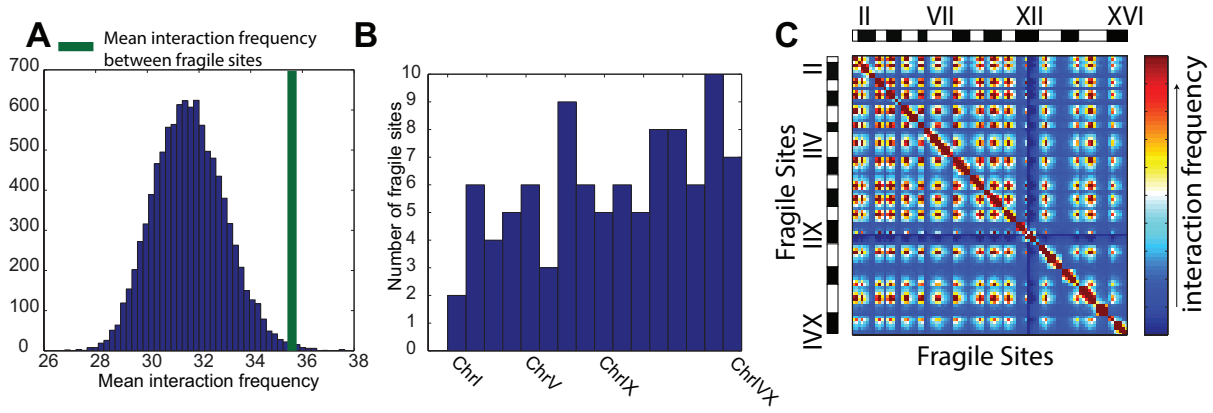

Figure 14. **Interactions among fragile sites and their distribution in the budding yeast genome. (A)** Mean interaction frequency between fragile sites (shown as thick green line) and the histogram of mean interaction frequencies between 10,000 sets of 95 random sites. **(B)** The distribution of fragile sites in the 16 chromosomes. **(C)** Heatmap of interaction frequencies between fragile sites as computed from the fully-constrained ensemble. The length of each chromosome is proportional to the number of fragile sites it contains. All high frequency interactions (red) are predicted to occur between different chromosomes, except those on the diagonal.

important to identify biologically specific interactions captured in genome-wide 3C studies but are unaccounted for by landmark constraints and nuclear confinement.

Predicted genomic interactions involving RNAPIII and TFIIS.

There are 14 interactions occurring between 10 loci that appear in more than 15% of the chains in the fully-constrained ensemble but are absent in the genome-wide 3C data (Figure 15A). We examined the available ChIP-chip study of RNAPIII and TFIIS binding ((Ghavi-Helm et al., 2008), see SI Methods) and found that there is an enrichment of 182.10 on average in binding of these factors to the 10 loci. This is higher than the enrichment of 112.25 at a significance level $p < 10^{-2}$ (Figure 15B), which is estimated from 10,000 sets of 14 random

TABLE III

LANDMARK GENES THAT ARE SPECIFIED IN THE YEAST DATABASE.

Chr1	FLO9, CLN3, MAK16, CYS3, ADE1, PHO1
Chr2	ILS1, MCM2, RAD16, SUP45, MET8
Chr3	HMLALPHA1, MATALPHA1, HMRA1
Chr4	CDC9, CDC2, SIR4, XRS2, TRP1
Chr5	CAN1, CUP5, FCY2, MET6, RAD3
Chr6	YPT1, SMC1, HIS2, HXK1
Chr7	ADH4, CUP2, TRP5, GCD2, PFK1
Chr8	SPOII, ARD1, CUP11, FUR1, ERG9
Chr9	SUC2, HIS5, BCY1, LYS1
Chr10	TPK1, ARG3, CYR1, CYC1, ECM17
Chr11	URA1, APE2, ELM1, VPS1, SIR1
Chr12	CDC25, LEU23
Chr13	HMG1, NDC1, MCM1, PFK2, ADE4
Chr14	DAL82, KEX2, RPC31, TOP2, LYS9
Chr15	HXT11, TOP1, DED1, PPO2, RAD17
Chr16	GAL4, TPK2, PEP4, ERG10, HTS1, RPC40

interactions of loci pairs. In addition, all 14 interactions are between centromeres and contain at least one tRNA gene (Table V). Only 3 out of 14 interactions have enrichment of RNAPIII and TFIIS lower than the mean enrichment of random interactions (112 ± 21). These findings are consistent with the observation of the tRNA gene localization at centromeres (Iwasaki et al., 2010), as well as the association of elongation factor TFIIS with RNAPII that are important for tRNA gene expression (Ghavi-Helm et al., 2008). These findings showed that a subset of computationally predicted interactions may have originally arisen from confinement and landmark constraints, but were subsequently stabilized through evolution with binding of RNAPIII and binding of TFIIS. The abundance of tRNA genes involved points to likely biological roles of these genomic interactions.

TABLE IV

PREDICTED 14 INTERACTIONS BETWEEN CENTROMERES OF CHROMOSOMES, WHETHER THEY CONTAIN TRNA GENE, AND THEIR COMBINED ENRICHMENT VALUE OF RNAPIII AND TFIIS

Chr	tRNA gene	Chr	tRNA gene	enrichment
II	yes	XIII	yes	56.98
II	yes	XV	yes	127.24
II	yes	XVI	no	61.53
IV	yes	XVI	no	163.63
V	yes	XVI	no	204.64
VII	no	XV	yes	183.53
X	yes	XIV	yes	372.52
X	yes	XV	yes	373.15
X	yes	XVI	no	180.46
XI	no	XV	yes	134.95
XIII	yes	XV	yes	146.34
XIII	yes	XVI	no	70.77
XIV	yes	XV	yes	326.23
XV	yes	XVI	no	158.03

Origin of tRNA-tRNA gene interactions.

Genome-wide 3C experiments and polymer modeling strongly suggests that tRNA genes cluster together in 3D space (Duan et al., 2010; Tjong et al., 2012; Wong et al., 2012). However, the origin of this clustering is unclear, as clustering could arise from the landmark constraints, or alternatively, from biological factors such as cohesin (Mizuguchi et al., 2014) and/or condensin (Haeusler et al., 2008). We sort all possible tRNA gene interactions according to their average separation distance from the corresponding centromeres, and find that mean spatial distances between tRNA genes are smaller when the their average genomic distances from their corresponding centromeres are within 30 kb (Figure 15C). While association of tRNS genes

with condensin is suggested to mediate the tRNA gene clustering in yeast nucleus (Haeusler et al., 2008), our results indicate that to a large extent, the clustering of tRNA genes is likely a consequence of the spatial clustering of centromeres to the SPB.

Biologically specific interactions beyond polymer effects.

We further identify chromatin interactions that are unaccounted for by random polymer interactions and are likely biologically significant. We computed propensities of interactions in the fully-constrained ensemble and in the genome-wide 3C experimental measurements using the random ensemble under the constraint of confinement only as the null model. There are 19 experimentally captured interactions with a propensity $\geq$3.5 in genome-wide 3C data but $<$1 in the fully-constrained ensemble (Figure 15D, see also SI Methods). Among the 19 interaction pairs, 4 are between tRNA genes. To further confirm that these interactions are not due to polymer effects, we calculated the correlation of the frequencies of these 19 interactions between fully-constrained ensemble and genome-wide 3C data, and found a small R value of 0.11. Furthermore, there are 70 important genes considered to be landmark genes in the budding yeast genome according to literature (Cherry et al., 2012) (for a list see Table III). We found that 8 of the identified 19 specific interactions are between these landmark genes (see Table V). Among these pairs, the genetic interaction between genes CYS3 and ADE4 has already been recently reported (Chen and Gartenberg, 2011), although the genetic relationship of the rest of the interacting landmark genes require further experimental investigations.

TABLE V

PREDICTED INTERACTING LANDMARK GENES. EACH ROW CONTAINS A PAIR OF INTERACTING GENES, IDENTIFIED FROM GENOME-WIDE 3C MEASUREMENTS USING FULLY-CONSTRAINED ENSEMBLE AS NULL MODEL.

gene	gene
CYS3	ADE4
TRP1	TOP2
SMC1	CYC1
SPOII	CYC1
FUR1	PEP4
ARG3	RAD17
MCM1	PEP4
PFK2	TOP1

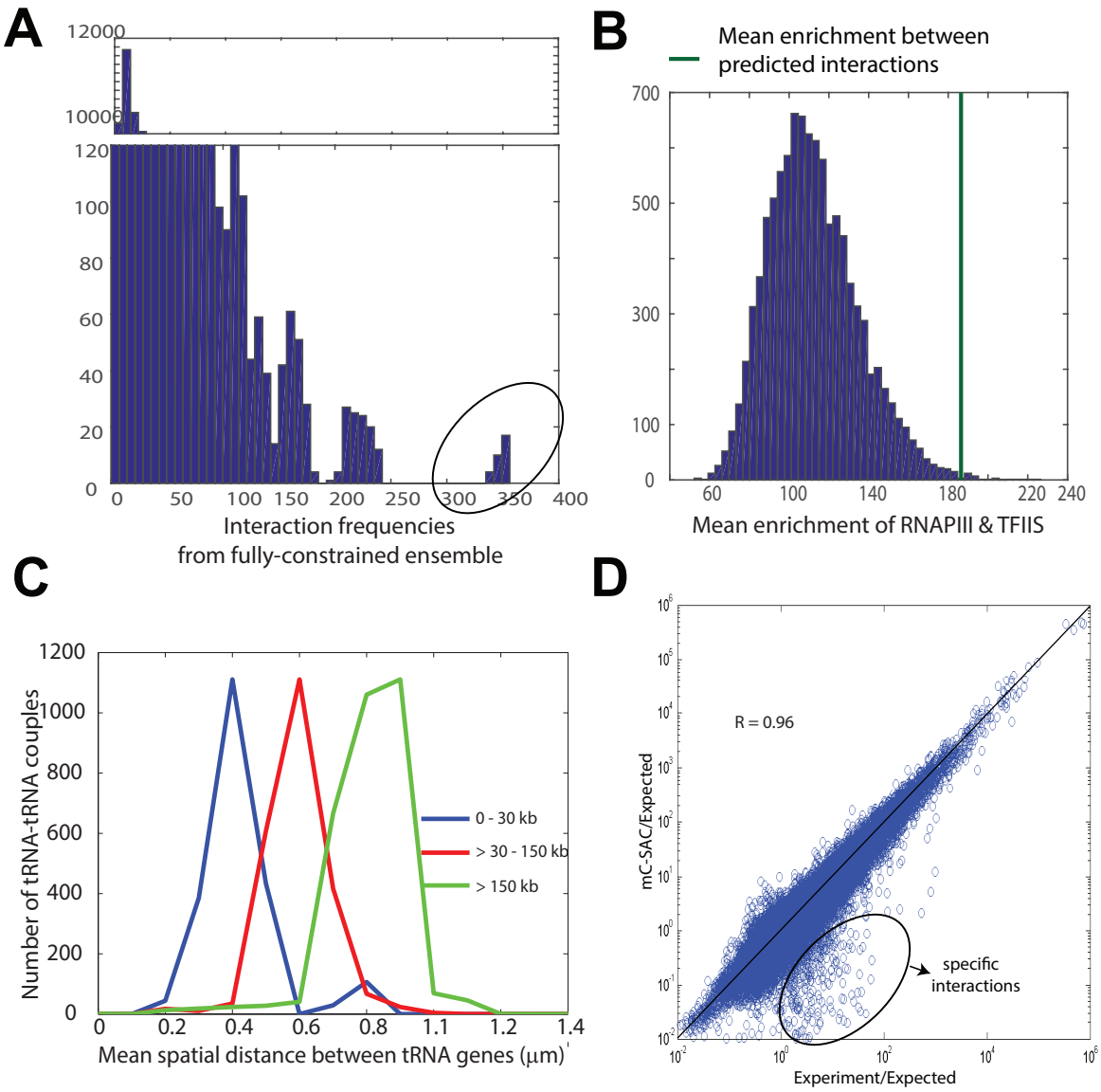

Figure 15. **tRNA gene interactions and differentiating biologically specific interactions from non-specific interactions arising from polymer effects. (A)** Distribution of frequencies of interactions enriched in fully-constrained ensemble, but absent in genome-wide 3C data. The 14 novel interactions with significant amount of interaction frequencies are encircled. The x-axis is the interaction frequencies and the y-axis is the number of interactions that these frequencies are observed. **(B)** Histogram of enrichment of RNAPIII and TFIIS binding. Mean enrichment of predicted interactions are shown with solid green line, along with the histogram of enrichment of 10,000 random sets of 14 interactions. **(C)** Distribution of mean-spatial distances between tRNA genes grouped according to their genomic distances to centromeres. **(D)** Interaction propensities of genome-wide 3C data (x-axis) and the fully-constrained ensemble (y-axis) calculated using random ensemble as null model. Interactions enriched in genome-wide 3C data over the fully-constrained ensemble are enclosed in the black circle.

3.4 Discussion

Genome in eukaryotes reside within the confined space of cell nucleus. Genome organization is directed by interactions with substructures called nuclear landmarks. With the advent of genome-wide chromosome conformation capture techniques (Duan et al., 2010), previous computational studies (Tjong et al., 2012; Wong et al., 2012) have already shown that random configurations of tethered chromosomes can reproduce measured interaction patterns (Duan et al., 2010) of budding yeast genome, although the correlation between modeled and measured inter-chromosomal interactions is not strong. The direct effects of individual nuclear landmarks on genome folding, as well as the origin of inter-chromosomal interactions are unknown. A major technical challenge is the extreme difficulty in adequate sampling of multiple chromatin chains subject to landmark constraints and the confinement of the cell nucleus. The mC-SAC model, developed in this study, which is based on a novel sampling technique (Gürsoy et al., 2014a; Gürsoy et al., 2014b) enabled us to generate large ensembles of model genomes with different combinations of landmark constraints in the nuclear confinement.

Our results showed that nuclear confinement and excluded-volume effects alone largely determine intra-chromosomal interaction patterns of individual yeast chromosomes, without the requirement of centromere tethering to the SPB and attachment of telomeres to the NE. This is in agreement with the results from the polymer-diffusion studies (Rosa and Everaers, 2008). Our results also highlight the importance of nuclear size on the patterns of interactions of genomic elements. When the nuclear size is enlarged, the experimentally captured interaction patterns disappeared. Our results further demonstrated that centromere tethering to the SPB,

along with the nuclear confinement and excluded-volume effect, are sufficient to capture the patterns of inter-chromosomal interactions. Measured inter-chromosomal interactions are enriched with interactions between pericentromeric regions, hence a cross-like pattern originating from centromeres is observed. Our results also showed that, with only the landmark constraint of centromere tethering to the SPB is introduced, measured patterns of inter-chromosomal interactions can be reproduced. Our results suggest that gene-regulatory systems originating from long-range chromatin interactions might have been inherited from the telephase of budding yeast, and the key difference in the regulatory machineries between the telophase and the interphase cells might be the silencing of telemoric genes through attachment to the NE. Such attachment, however, has no significant effects on the overall genome organization of budding yeast (Figure 12).

Previous studies showed the presence of co-localization and clustering of important genomic elements such as early replicating sites or tRNA genes (Duan et al., 2010; Tjong et al., 2012). However, the origin of such clustering remained unclear. Here, we showed that this clustering is largely due to the attachment of centromeres to the SPB. Except genes on Chr 12 and telomeres, positions of genomic elements on the chromosomes relative to the SPB are strongly correlated with their genomic distances to their corresponding centromeres. We also showed that the relative positions of genes can be reproduced, as long as their genomic distances to the corresponding centromeres are given. This finding may be useful for predicting spatial positions of important genes from their genomic locations. For example, the spatial distances between tRNA genes decrease as their genomic distances to the centromeres decrease (Figure 15C).

Our results are consistent with the suggestion that genomic locations of important elements in budding yeast were selected by evolutionary pressure (Tjong et al., 2012)

Our model of budding yeast can be used to make biological inference of the organization of yeast genome. Fully constrained ensemble not only can reproduce the pattern of spatial interactions from genome-wide 3C studies, but can also provide additional details by filling the gaps in the sparse interaction matrices. We found that there are interactions arising from landmark constraints, but are absent in the genome-wide 3C data. These interactions are enriched with transcription factor TFIIS as well as RNAPIII. They are located in pericentromeric regions of chromosomes, and contain significant amount of tRNA genes. These results suggest that chromatin interactions arising from landmark constraints may be subsequently stabilized by biological factors through evolution. We also found that chromosomal fragile sites are clustered together in three–dimensional space, most likely as a result of their location on pericentromeric sites and a consequence of centromere clustering at the SPB. As SPB functionally corresponds to centrosome in mammalian cell nuclei, where the centromeres are attached during metaphase, our results raised the question whether fragile sites of human genome form spatial clusters and are also in genomic proximity to the centromeres. It is possible that translocations due to the errors in mitosis in human genome that may be cancer promoting may also be related to centromere clustering.

Because of the dominant effects of landmark constraints and confinement on the folding patterns of budding yeast genome, it is challenging to uncover the specific spatial interactions that are due to biological factors. One approach to identify such interactions is to generate

ensembles of model genomes that are subject to landmark constraints. Taking this ensemble as a null model, one could in principle subtract polymer effects from interactions captured in genome-wide chromosome conformation capture study. However, current polymer models are inadequate for such a task, as they failed to reproduce the inter-chromosomal interaction patterns (Tjong et al., 2012). Previous studies also suggested that volume exclusion models capture only expected interactions when such expected interactions were removed, there was no strong correlation between model genomes and experimental measurements (Ay et al., 2014). Our findings reveal that this correlation can be improved significantly with better sampling techniques. To further understand whether the budding yeast genome organization are dictated by landmark constraints, we removed the interactions arising from excluded-volume effect, chain connectivity and nuclear confinement from both experimental measurements and fully-constrained ensemble, and compared the remaining interaction frequencies. Our results suggest that remaining experimentally measured interactions are in excellent agreement with the remaining interactions of fully-constrained ensemble of modeled genomes. Nevertheless, there exist a small set of interactions that are high frequency in genome-wide 3C data but almost absent in the fully-constrained ensemble. These interactions turn out to involve several important genes. That is, we were able to extract interactions of potential biological interest from the interaction frequencies of genome-wide 3C data, a very challenging task due to the dominance of polymer effects in experimental measurements.

In summary, we showed that much of the folding patterns of budding yeast genome can be recapitulated by our constrained mC-SAC polymer model. Our model can also be used to

identify novel interactions that are not measured experimentally. It can also extract biologically specific interactions that are unaccounted by polymer effects. Because of the coarse-grained nature of both current polymer models and genome-wide 3C techniques, our model does not contain detailed spatial information of yeast genome. Inferring structural units of gene regulatory machineries that span just a few kilo bases requires chromatin models of much finer resolution. As the advances in theory, model, and experimental measurements continues, it is envisioned that high resolution models of yeast genome can be computed.

CHAPTER 4

CONSTRUCTING 3D CHROMATIN ENSEMBLES AND PREDICTING FUNCTIONAL INTERACTIONS OF α-GLOBIN LOCUS FROM 5C DATA

4.1 Introduction

Understanding the spatial organization of the genome inside a cell nucleus and how 3D genome folding dictates important cell activities such as gene expression are important problems to adress in biology (Fraser and Bickmore, 2007). Recent advent of chromosome conformation capture (3C) and related techniques (4C, 5C, Hi-C) enabled large-scale discovery of long-range chromatin looping interactions among distant chromosomal elements (Dekker et al., 2002; Hagge et al., 2007; Lieberman-Aiden et al., 2009; Duan et al., 2010; Montefiori et al., 2016). The discovery of topologically associated domains (TADs) with elevated chromatin interactions (Nora et al., 2012; Dixon et al., 2012) suggests a detailed structural network involving binding of architectural proteins (Phillips-Cremins et al., 2013). These findings point to likely 3D structural units of chromatin that accommodate spatial clustering of different regulatory elements and transcription factors important for cell activities.

Chromatin is highly dynamic and experiences significant conformational changes (Lucas et al., 2014). As 3C data are from collection of cell populations and may reflect a mixture of different conformations at a particular moment, it is important to uncover an ensemble of

3D structures of a gene locus that collectively best describe the bulk measurements (Ay and Noble, 2015). This would enable precise structural measurements and identification of spatial organizational units of genomic elements. However, it is difficult to generate well-sampled ensembles of detailed chromatin chains using many constraints from 3C-related data.

To overcome the limitations of the pairwise nature of Chromosome Conformation Capture data and to gain detailed mechanistic understanding of gene regulation, there have been significant efforts in constructing 3D structures of chromatin. 3D polymer models based on minimal physical assumptions revealed important information on general rules genome organization (Lieberman-Aiden et al., 2009; Tokuda et al., 2012; Barbieri et al., 2012; Tjong et al., 2012; Wong et al., 2012; Gürsoy et al., 2014a; Kang et al., 2015; Goloborodko et al., 2016; Fudenberg et al., 2016; Sanborn et al., 2015; Chiariello et al., 2016). Modeling of 3D ensemble of chromatin chains using 3C-related (4C/5C/Hi-C) data (Giorgetti et al., 2014; Rousseau et al., 2011; Ay et al., 2014; Bau et al., 2011; Wang et al., 2015; Trieu and Cheng, 2014; Zhang and Wolynes, 2015; Meluzzi and Arya, 2013; Tjong et al., 2016), transcription factor binding information (Junier et al., 2012; Brackley et al., 2016), as well as epigenomic states of the chromosomes (Jost et al., 2014) provided rich information on biological properties of genomic elements. Specifically, experimentally obtained interaction patterns can be reproduced computationally with simple assumptions (Tjong et al., 2012; Wong et al., 2012; Zhang and Wolynes, 2015; Meluzzi and Arya, 2013; Giorgetti et al., 2014; Ay et al., 2014; Tjong et al., 2016; Kalhor et al., 2012; Goloborodko et al., 2016; Fudenberg et al., 2016; Sanborn et al., 2015; Chiariello et al., 2016), with co-localization of co-expressed genes uncovered (Ay et al., 2014). The for-

mation of TADs (Brackley et al., 2016; Jost et al., 2014) and their boundaries (Tjong et al., 2016) can also be predicted. Nevertheless, current methods based on the general folding principles of genome do not generate detailed spatial structures for understanding the underlying mechanism of differential gene expression (Ay et al., 2014; Zhang and Wolynes, 2015; Meluzzi and Arya, 2013; Brackley et al., 2016; Jost et al., 2014; Goloborodko et al., 2016; Fudenberg et al., 2016; Sanborn et al., 2015). Other methods have limited resolution for capturing structural differences of a small locus (Tjong et al., 2016; Kalhor et al., 2012; Chiariello et al., 2016), as they are designed to study overall genome organization or a larger genomic region. In addition, difficulties in sampling of chromatin conformations poses additional challenges for unbiased assessment of populations of chromatin structures (Rousseau et al., 2011; Giorgetti et al., 2014; Bau et al., 2011).

In this study, we describe the n Constrained-Self-Avoiding Chromatin (nC-SAC) computational method for predicting configurations of ensembles of chromatin chains with spatial details based on the interaction frequencies of the α-globin locus measured by the 5C technique (Bau et al., 2011). While Hi-C measurement is emerging as a method of choice in studying chromatin structures at high-resolution (Rao et al., 2014), currently it comes at a great cost and is not yet widely accessible. With focus on a specific genomic region of interest, 5C technique can provide valuable information on biologically important interactions through specifically designed primer for a particular locus. However, computationally reconstructing 3D structures of chromatin from 5C study is challenging, as experimentally captured interactions are often sparse and incomplete due to the locations of restriction enzyme sites and uneven distribution

of primers. Another significant problem in using 5C as well as other 3C related data for constructing 3D chromatin ensembles is the added complexity due to random interactions arising from non-specific collision of chromosomal regions to one another (Belmont, 2014).

Our nC-SAC model first generates 100,000 random self-avoiding chromatin chain conformations inside the crowded cell nucleus, which are used as the null model to distinguish the most significant 5C interactions from non-specific interactions in the α-globin locus. Overcoming severe sampling problems using the geometrical Sequential Importance Sampling (g-SIS) technique, the nC-SAC model then generates two large ensembles of 3D chromatin chains of the α-globin locus for two cell lines with different expression levels. These ensembles satisfy $\sim 90\%$ of the imposed constraints of significant 5C interactions. Our model predicts a large number of novel looping interactions with spatial details that were not captured by the original 5C experiment due to lack of primer coverage. A subset of our predicted interactions were shown in two independent ChIA-PET studies to be mediated by proteins such as CTCF, RNAPII, RAD21 and are associated with concurrent histone modifications (Li et al., 2012; Heidari et al., 2014). Our model further suggests the existence of a many-body structural unit involving α-globin gene, enhancers HS40/46/48, and POL3RK gene for regulating α-globin expression in the silent cell line. Furthermore, our models uncover global differences in the spatial structures of the α-globin in cells with high and low expression. Our findings suggest that a homogeneous and dominant structural population of the locus may be associated with the high expression level of the α-globin.

4.2 Materials and Methods

The overall computational pipeline of nC-SAC is illustrated in Figure 16. The interaction frequencies obtained from the 5C study (Bau et al., 2011) (Figure 16A) are compared to the interaction frequencies of random C-SAC ensemble. For this purpose, an ensemble of 100,000 C-SAC chains (Gürsoy et al., 2014a) confined to a spherical confinement is generated (Figure 16B). This ensemble is then bootstrapped for 1,000 steps. p-value of observation of the experimentally captured interaction frequency in the random ensemble is calculated. The correction for multiple hypothesis testing is then employed (Figure 16C). The significant 5C interaction frequencies at the False Discovery Rate of $\alpha < 5\%$ (Figure 16D) are converted into spatial distances using a half-Gaussian model (Figure 16E) and an ensemble of 10,000 3D chromatin chains that satisfy these spatial distance constraints are generated using the technique of g-SIS (Figure 16F). A full resolution interaction map is computed from the generated ensemble of structures (Figure 16G).

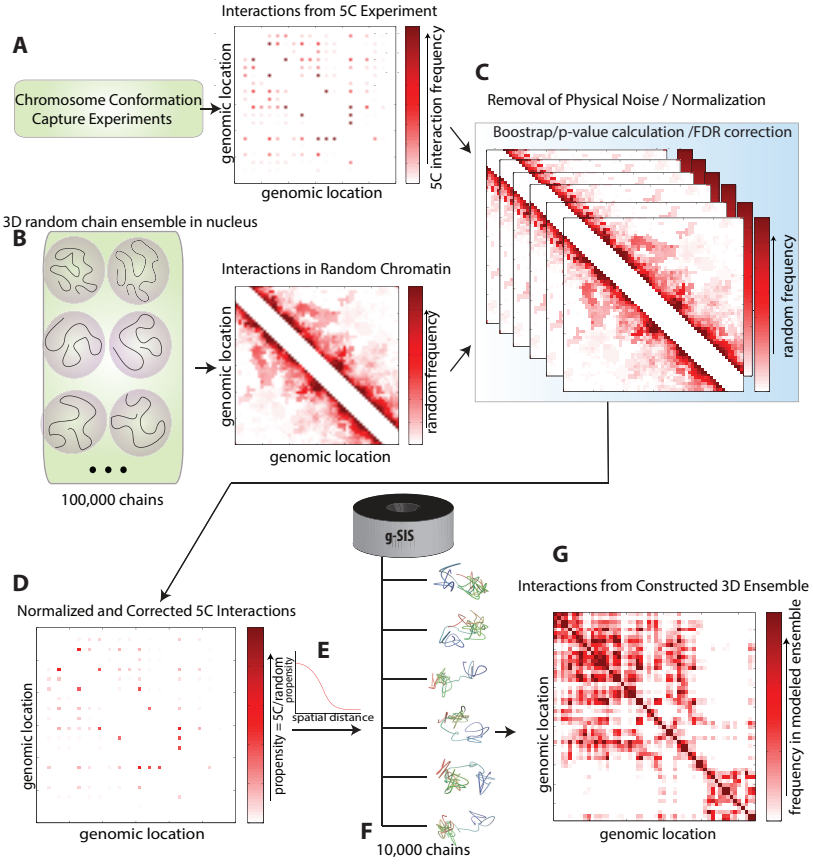

Figure 16. **The nC-SAC computational pipeline to predict structural ensembles of chromatin chains from 5C data.** **(A)** 5C interactions are compared with **(B)** the interactions of a random 3D ensemble of 10^5 C-SAC chains in cell nucleus that is generated to obtain a contact matrix of random interactions. **(C)** 1,000 bootstrapped random contact matrices to calculate the p-value of each 5C interaction. **(D)** After FDR adjustment for multiple hypothesis testing, non-specific 3D random interactions are excluded. **(E)** Remaining significant 5C interactions are normalized and converted into distances using a half-Gaussian model. **(F)** An ensemble of $> 10^4$ 3D-chains of the locus is then generated and **(G)** a full resolution contact map is computed.

4.2.1 Mapping 5C data on to a polymer chain.

Our polymer model consists of monomers that are modeled as spheres with 30 nm diameter and a genome density of 2,727 bp. Each HindIII fragments that are used in the original 5C study (Bau et al., 2011) corresponds to a fragment unit, which is modeled as a non-bendable collection of monomers. The last monomer of each fragment unit corresponds to a primer site. The maximum fragment length is used as 150 nm (Bystricky et al., 2004), which corresponds to 5 monomers (13,635 bp of DNA). The HindIII fragments that are larger than 13,635 bp are divided into multiple fragment units that are not larger than 5 monomers and a virtual node is placed onto the last monomer of artificially divided unit. This is done to give the bending property to the chromatin (Figure 21A). In total, α-globin locus polymer chain contains 53 nodes and $\sim$183 monomers. The interaction frequency $f_{cell}^{5C}(i,j)$ between each nodes i and j is mapped and a total of 367 and 425 unique pairwise chromatin interactions are obtained for the K562 and GM12878 cell lines using the 5C data (Bau et al., 2011).

4.2.2 Exclusion of non-specific physical interactions.

An ensemble of 100,000 randomly folded polymer chains that have the same physical properties as α-globin locus (53 nodes, 183 monomers, 150 nm maximum fragment length and 2,727 bp/30 nm DNA density) are generated inside a confined space of nucleus using previously described C-SAC model using a chain growth method (Gürsoy et al., 2014a) to asses the statistical significance of each 5C interaction. The size of the confined space is the volume that 500 kb DNA occupies, which is calculated to be proportional to a diploid human nucleus. This allows

us the determine the interaction frequency $f^{null}(i,j)$ between each nodes i and j that occur due to the available space in cell nucleus, excluded volume effect and the chain connectivity.

First, we normalized the interaction frequencies by total number of 5C interaction frequencies for each cell line and random model as following,

$$q^{5C}_{cell}(i,j) = \frac{f^{5C}_{cell}(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} f^{5C}_{cell}(i,j)}$$

$$q^{null}(i,j) = \frac{\sum_k w_k I(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_k w_k I(i,j)},$$

where N is the total number of nodes, w_k is the weight of the k^{th} chain in the ensemble and $I(i,j)$ is an indicator function, which equals to 1 when nodes i and j interacts, equals to 0 otherwise. We calculated the propensity of each 5C interaction as,

$$prop^{5C}_{cell}(i,j) = \frac{q^{5C}_{cell}(i,j)}{q^{null}(i,j)}$$

and by bootstrapping the random ensemble for 1,000 time, we calculated a p-value for each interaction that satisfies $prop^{5C}_{cell}(i,j) > 1$. After a multiple hypothesis testing through FDR (Benjamini and Hochberg, 1995) with an $\alpha = 0.05$, any interaction that cannot pass FDR are excluded from the original 5C data (Figure 16A–C). Assuming an inverse relationship between propensity and spatial distance, we applied a Gaussian model to convert $prop^{5C}_{cell}(i,j)$ to $d^{5C}_{cell}(i,j)$, where $d^{5C}_{cell}(i,j)$ stands for the spatial distance between node i and j.

4.2.3 nC-SAC model: Incorporation of significant 5C interactions.

nC-SAC model is an extension of C-SAC model where n spatial distances ($d_{cell}^{5C}(i,j)$) used as constraints during the chain growth process with geometrical Sequential Importance Sampling (Gürsoy et al., 2014a; Liang et al., 2002; Zhang et al., 2003; Lin et al., 2008b; Lin et al., 2008a; Zhang et al., 2009). The configuration $\boldsymbol{x}$ of a full chromatin chain with N nodes, with the location of the i-th node denoted as $x_i = (a_i, b_i, c_i) \in \mathbb{R}^3$, is:

$$\boldsymbol{x} = (x_1, \cdots, x_N).$$

The target distribution $\pi(\boldsymbol{x})$ is a Boltzmann distribution of chromatin chains that the spatial distances between nodes, $d_{cell}^{pred}(i,j)$, is equal to spatial distances derived from 5C interaction frequencies, $d_{cell}^{5C}(i,j)$, while ensuring the self avoiding property. To generate a chromatin chain, we grow the chain one node at a time, by using a $k = 640$-state off-lattice discrete model. The new node added to a growing chain with the current node located at x_t is placed at x_{t+1}, which is a persistence unit L_p distance away from x_t. x_{t+1} is taken from one of the unoccupied k-sites neighboring x_t according to a probability distribution that favors the $d_{cell}^{pred}(i,j) = d_{cell}^{5C}(i,j)$ and self-avoiding property, namely, $x_i \neq x_j$ for all $i \neq j$. As satisfying the $d_{cell}^{5C}(i,j)$ where (i,j) pair is far away from each other on genome is extremely challenging, we introduce a look-ahead bias to our selection from available empty neighboring sites for (i,j) pairs that do not have constraints. We keep track of the bias and assign each successfully generated chain a proper weight $w(\boldsymbol{x})$.

4.2.4 Details of exclusion of non-specific physical interactions from 5C Data

4.2.4.1 Bootstrap and False Discovery Rate

Calculation of p-value.

To test if the interaction between nodes i and j is significant, we compare how many times $q^{5C}_{cell}(i,j)$ (the normalized interaction frequency between nodes i and j in 5C data) is less than $q^{null}(i,j)$ by bootstrapping 1000 times of 100,000 random C-SAC chains with replacement. $q^{null}(i,j)_m = \frac{\sum_{k'} w_{k'} I(i,j)}{\sum_i \sum_j \sum_{k'} w_{k'} I(i,j)}$ is the normalized interactions frequency of nodes in the random C-SAC ensemble, and $w_{k'}$ is the weight of k'th random chain from the mth bootstrapped 100,000 samples with replacement. The p-values p_{ij} of interaction is:

$$p_{ij} = \frac{\sum_{m=1}^{M} I(q^{5C}_{cell}(i,j) < q^{null}(i,j)_m)}{M},$$

where $M = 1000$, and $I(\cdot)$ is a indicator function, which equals to 1 when condition is satisfied, equals to 0 otherwise. FDR correction is employed for each interactions with a genomic separation $s = |i-j|$. For each constant s, we sorted p_{ij} ascendantly to get new p-value set $\{p^{(m)}_{ij}\}$, such that $p^{(1)}_{ij} \leqslant p^{(2)}_{ij} \leqslant \cdots \leqslant p^{(m)}_{ij}$ are ordered, where m is the total number of the p-values in the set $\{K|K = j-i\}$. We then used Hochberg adjustment method (Benjamini and Hochberg, 1995) to adjust p-values $p^{(m)}_{ij}$,

$$\tilde{p}^{(l)}_{ij} = \begin{cases} p^{(m)}_{ij} & \text{for } l = m, \\ \min(\tilde{p}^{(l+1)}_{ij}, \frac{m}{l} p^{(l)}_{ij}) & \text{for } l = m-1, \ldots, 1. \end{cases}$$

After the FDR adjustment, the null hyphothesis is rejected with significance level of $\alpha = 5\%$.

4.2.5 Details of nC-SAC Model

4.2.5.1 Obtaining Distance Constraints from Significant 5C Interaction Frequencies

After the calculation of propensities described in the Methods section, we assume that the relationship between the propensity $prop_{ij}$ and the distance constraint d_{ij} between node i and j follows half Gaussian distribution,

$$\frac{prop_{ij}}{\max prop_{ij}} = \exp \frac{-(d_{ij} - \mu)^2}{2\sigma^2}, \text{ and } d_{ij} > \mu.$$

where $\sigma = \frac{d_c - \mu}{\sqrt{2 \log \frac{\max prop_{ij}}{\min prop_{ij}}}}$. In equation 4.2.5.1, d_{ij} is an entry of matrix D, corresponds to a spatial distance constraint of interaction of i and j and can be calculated as

$$d_{ij} = \mu + \sqrt{2\sigma^2 \log \frac{\max prop_{ij}}{prop_{ij}}}, \text{ and } prop_{ij} > 0.$$

where μ is 30 nm which is the minimum possible distance between any node i and j and $\max prop_{ij}$ is the maximum propensity. The maximum possible distance between any node i and j is taken as 80 nm following experimentally determined threshold (Giorgetti et al., 2014).

4.2.5.2 Geometrical Sequential Importance Sampling Algorithm for nC-SAC Model

Conformations that satisfy the spatial distance constraints can be generated by minimizing an error function. This function measures the deviations from the desired spatial distances, i.e. distance constraints derived from 5C frequencies.

$$\mathcal{E}(\boldsymbol{x}_n^{(k)}) = \frac{\sum\limits_{(i,j)\in P_{\boldsymbol{x}_n}} |\| x_i - x_j \| - d_{i,j}|}{\sum\limits_{(i,j)\in P_{\boldsymbol{x}_n}} d_{i,j}},$$

in which $P_{\boldsymbol{x}_n}$ is list of i-j interactions with distance constraint d_{ij} and i,j = 1,...,n. Our objective is to generate chromatin chains that satisfy distance constraints, hence follow target distribution $\pi(\boldsymbol{x}_n)$,

$$\pi(\boldsymbol{x}_n) = \exp(-\mathcal{E}(\boldsymbol{x}_n))$$

$\boldsymbol{x}_t = (x_1, \ldots, x_t)$ is a vector defining the three-dimensional coordinates of nodes. We place a node t at coordinate x_t that follows a growth function $g_t(x_t|\boldsymbol{x}_{t-1})$. Growth function is designed in a way that candidate positions have different probabilities, which helps the sampling efficiency, as well as approximation of target distribution. The growth function of a conformation with t nodes at coordinate $x_1, \ldots, x_t$ is

$$g_t(\boldsymbol{x}_t) = g_1(\boldsymbol{x}_1) g_2(x_2|\boldsymbol{x}_1) \ldots g_t(x_t|\boldsymbol{x}_{t-1}).$$

Final sample chain $\boldsymbol{x}_n$ is weighted to remove the biases originating from the design of the trial distribution, so that target distribution $\pi(\boldsymbol{x}_n)$ can be recovered (Liang et al., 2002; Liu and Chen, 1998). The assigned weight is

$$w(\boldsymbol{x}_n) = \pi(\boldsymbol{x}_n)/g_n(\boldsymbol{x}_n)$$

The statistical mean of physical properties such as interaction probabilities can be represented by $h(\boldsymbol{x}_n)$ of chain $\boldsymbol{x}_n$ that follows the target distribution $\pi(\boldsymbol{x}_n)$ that follows as

$$E_\pi(h(\boldsymbol{x}_n)) \simeq \frac{\sum_{k=1}^{m} w(\boldsymbol{x}_n^{(k)}) \cdot h(\boldsymbol{x}_n^{(k)})}{\sum_{k=1}^{m} w(\boldsymbol{x}_n^{(k)})},$$

where k=1,..,m is the number of chains in the ensemble. The algorithm that is used in this work is adopted from (Zhang et al., 2004; Lin et al., 2011; Lin et al., 2008b; Lin et al., 2008a)

Trial Distribution.

The growth function $g_t(x_t|\boldsymbol{x}_{t-1})$ (also called trial distribution) for the partial chain $\boldsymbol{x}_{t-1}$ takes the form of priority score $\beta_t^{(l)}$ in the g-SIS algorithm. It biases the chain to grow towards to the regions that will potentially satisfy the target distribution. A growth function is required since the distance constraints of the future nodes can only be used when all the participating nodes are being generated. The priority score is calculated from three components: growth functions of excluded volume constraints, distance constraints and loop consideration.

The priority score $\beta_t^{(l)}$ for chain $\tilde{\boldsymbol{x}}_t^{(l)}$ is set as

$$\beta_t^{(l)} = \exp\left[-\frac{\lambda_1 f_1(\tilde{x_t}^{(l)}) + \lambda_2 f_2(\tilde{x_t}^{(l)}) + \lambda_3 f_3(\tilde{x_t}^{(l)})}{T}\right]$$

where λ_1, λ_2, and λ_3 are coefficients and T is temperature (we used $\lambda_1 = \lambda_2 = \lambda_3 = T = 1$ in this study).

(1) Growth function of excluded volume constraints.

This function is designed to maintain the self-avoiding property of a 30 nm chromatin fiber. If $f_1(x_t)$ is the growth function of excluded volume constraint, then

$$f_1(x_t) = \sum_{B_{\boldsymbol{x}_{t-1}}} h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0),$$

where h_1 is the loss function to quantify the violation of excluded volume of x_t with its previous partial chain $B_{\boldsymbol{x}_{t-1}}$.

$$h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = I(||x_t - \tilde{x}_i|| < r_0), \text{any } \tilde{x}_i \in B_{\boldsymbol{x}_{t-1}},$$

where $I(\cdot)$ is an indicator function, such that $h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = 0$, when $||x_t - \tilde{x}_i|| \geqslant r_0$, and $h_1(x_t, B_{\boldsymbol{x}_{t-1}}, r_0) = 1$, when $||x_t - \tilde{x}_i|| \leqslant r_0$, and $r_0 = 30$ nm, adapted from experimentally verified threshold (Giorgetti et al., 2014).

(2) Growth function of distance constraints.

A partial chain $\boldsymbol{x}_{t-1}$, the coordinates of the current node t ($x_t \notin \boldsymbol{x}_{t-1}$) is determined according to the distance constraints derived from 5C interactions. If $f_2(x_t)$ is the growth function of distance constraints, then

$$f_2(x_t) = h_2((\parallel x_{i_1} - x_t \parallel, \ldots, \parallel x_{i_K} - x_t \parallel), (d_{i_1,t}, \ldots, d_{i_K,t})),$$

where i_k is the k^{th} node that has distance constraint $d_{i_k t}$ with the node t and K is the total number of nodes that have distance constraints $d_{i_k t}$ with the current node t in the partial chain x_{t-1}. h_2 is the loss function to quantify the error between distances between the nodes in the chain and their corresponding distance constraints.

$$h_2((\parallel x_{i_1} - x_t \parallel, \ldots, \parallel x_{i_K} - x_t \parallel), (d_{i_1,t}, \ldots, d_{i_K,t})) = \frac{\sum\limits_{i_k \in P_t} |\parallel x_{i_k} - x_t \parallel - d_{i_k,t}|}{\sum\limits_{i_k \in P_t} d_{i_k,t}},$$

(3) Growth function of loop constraints.

Due to the sparseness of 5C interactions, there are several nodes that do not have any distance constraints. For a node t with no distance constraints from 5C data in the partial chain $\boldsymbol{x}_{t-1}$, we employ a loop constraint to enforce node t to follow triangle inequality. If $f_3(x_t)$ is the function of loop constraints, then

$$f_3(x_t) = h_3(x_t, O_t),$$

where $O_t = \{(t_{i_k}, t_{j_k}) \mid \text{interaction pair } t_{i_k} \text{ and } t_{j_k} \text{ and } t_{i_k} < t < t_{j_k}\}$, and, h_3 is a loss function to quantify the triangle inequality,

$$h_3(x_t, O_t) = \sum_{(t_{i_k}, t_{j_k}) \in O_t} I(\Big| \parallel x_t - x_{t_{i_k}} \parallel - d_{t_{j_k}, t_{i_k}} \Big| > \sum_{l=t}^{t_{j_k}-1} d_{l,l+1}),$$

$d_{l,l+1}$ is the length of segment between node l and $l+1$, l from t to $t_{j_k}-1$, and $I(\cdot)$ is an indicator function such that it is equal to 1, when the distance between node t and node t_{i_k} is greater than the sum of the rest of the segment length between the node t and node t_{j_k}, it is equal to 0 otherwise.

Target distribution.

The target score $\gamma_t^{(l)}$ that represents the target distribution for chain $\tilde{\boldsymbol{x}}_t^{(l)}$ is

$$\beta_t^{(l)} = \exp\left[-\frac{\lambda_1 f_1(\tilde{x_t}^{(l)}) + \lambda_2 f_2(\tilde{x_t}^{(l)})}{T'}\right],$$

where λ_1, λ_2 are coefficients. T' is temperature, and $T' = \frac{1}{2}T$.

4.2.5.3 Density-based algorithm for clustering

A density based algorith is adopted to cluster the chromatin conformations according to their similarities for each cell line. The details of this algorithm can be found in (Ester et al., 1996).

The RMSD between pair of conformations are calculated and used as a similarity measure for clustering. The minimum RMSD required to cluster two conformations together is taken as 34 nm. A conformation i belongs to the cluster c_k, if i has more than 5 similar conformations in the cluster c_k.

RMSD is calculated as following,

$$RMSD(\boldsymbol{c}_m, \boldsymbol{c}_n) = \sqrt{\frac{\sum_{k=1}^{l} ||c_m^k - c_n^k||^2}{l}}, \tag{4.1}$$

in which $\boldsymbol{c}_m = \{d(i,j)|(i,j)$ are nodes$\}$ is the set of spatial distances between nodes of mth predicted conformation, l is number of interactions in total and c_m^k is the kth distance element in the set $\boldsymbol{c}_m$.

4.3 Results

4.3.1 Identifying the most significant 5C interactions by the nC-SAC method

To build three-dimensional chromatin chains of the α-globin locus, we use published 5C data from the K562 (α-globin expressing) and GM12878 (silent) cell lines where HindIII restriction digestion was used for preparation of 5C libraries (Bau et al., 2011). We remove 5C interactions associated with short (<2.7 kb) fragments as they are considered to be unreliable (Naumova et al., 2012). We also remove interactions between consecutive fragments as they are likely due to proximity effects (Dekker et al., 2002).

We use the C-SAC polymer model (Gürsoy et al., 2014a) to model the α-globin chromatin chain. In this model, we represent the chromatin as a collection of beads and the chain is constrained by 5C interaction frequencies and the crowding effects in the cell nucleus. We divide the 500 kb locus into 184 beads, each corresponds to 2.7 kb DNA. We used fragment units to mimic the HindIII restriction fragmentation. Each fragment unit is equal to the size of the HindIII fragments in the 5C study and is modeled as a rigid body consisting of a maximum of 5 beads. To incorporate the bending properties of chromatin, we divide the fragment units that are longer than 5 beads ($\sim$ 13.5 kb) into 13.5 kb or shorter units. We call the last bead in each unit a node, and the node number is used as the identifier of the unit (Figure 21A). In total, we have 54 fragment units for the 500 kb α-globin locus.

During the construction of 3C chromatin libraries, formaldehyde treatment can covalently link genomic elements within certain spatial distances, regardless whether specific interactions exist (Belmont, 2014). We hypothesize that a significant number of such interactions arise from

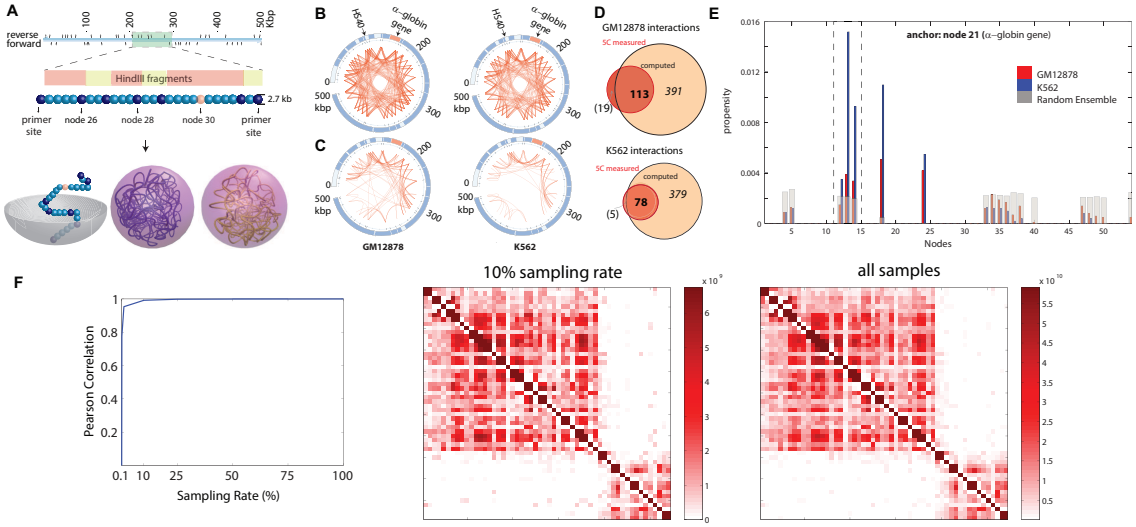

Figure 17. **Mapping 5C interactions onto C-SAC model chromatin chains, identifying non-specific interactions, and predicting novel interactions between genomic elements in α-globin locus.** (**A**) Mapping the 500 kb α-globin locus and 5C interactions onto the C-SAC polymer model of chromatin chain. Up and down arrows in the linear diagram of the α-globin locus represent reverse and forward primers at ends of HindIII fragments (Bau et al., 2011), respectively. Fragments between primer sites 25 and 32 are enlarged to demonstrate details of the C-SAC model. Alternating fragments are shown in pink and yellow. HindIII fragments are mapped onto fragment units, with which bending can occur at the primer sites (darker blue) or every 6-th bead (pink) for fragments mapped onto one or more units. Ensemble of random C-SAC chromatin chains are generated through chain-growth one bead at a time in a confined sphere representing the cell nucleus (Gürsoy et al., 2014a). Representative partial and full C-SAC chains in spherical confinement are shown. (**B**) All reported 5C interactions between elements in the α-globin locus (Bau et al., 2011). (**C**) Significant 5C interactions remaining after non-specific interactions were identified. (**D**) Comparison of statistically significant 5C interactions (red circle) and interactions predicted by the nC-SAC model (beige circles). 5C interactions that were not captured by the nC-SAC model are indicated (numbers in parenthesis). Among predicted interactions (beige circles), many are novel predictions (in italic) that are not measured in the original 5C study. (**E**) Interaction profiles of selected nodes in both cell lines and in the random ensemble. x-axis denotes the nodes and y-axis is the propensity of interaction between the anchor node and the rest of the locus. The interactions between α-globin gene and the nodes 12,13 and 14 (enhancers HS40/46/48) are highlighted in the dotted box. (**F**) Conversion of ensemble by the sampled number of chains. The Pearson Correlation between the interaction frequencies of full ensemble and partial ensembles with different sampling rates. The heatmaps of interaction frequencies of a partial ensemble where only 10% of 10,000 chains were sampled and the heatmap of interaction frequencies of the full ensemble.

self-avoiding chromatin chains confined in the crowded cell nucleus (Kang et al., 2015; Gürsoy et al., 2014a). As loci from different chromosomes can coexist inside the cell nucleus, the nuclear confinement as well as the crowding effects result in limited available space for individual locus. We hypothesized that many 5C interactions are of non-specific nature. To identify such non-specific interactions, we generated an ensemble of 100,000 random C-SAC chains using a chain growth strategy (Gürsoy et al., 2014a) (Figure 21A). Without *a prior* information, we assume the 500 kb (5×10^5 bp) α-globin locus occupies $(5 \times 10^5) \ / \ (6 \times 10^9)$ of the available space of $\sim 7.5^3$ μm^3 inside average cell nucleus of $\sim 10^3$ μm^3 with a diploid human genome size of $2 \times 3 \times 10^9$ bp. This corresponds to a sphere of a diameter of 330nm.

We bootstrap the chains in the random ensemble to generate 1,000 ensembles of 100,000 C-SAC chains and calculate the probabilities of observing normalized 5C interaction frequencies in these random ensembles and used these probabilities as p-values subject to correction of multiple hypothesis testing at the False Discovery Rate (FDR) of $\alpha < 5\%$ (Materials and Methods). A total of 293 of 425 experimentally captured 5C-interactions (77%) in the GM12878 cell line and 284 of 367 5C-interactions (87%) in the K562 cell line are not statistically significant and are therefore not used as spatial constraints (Figure 21B and C). Recognizing that the available space may not be perfectly spherical, we use a stringent FDR criteria to ensure that we only identify the most significant interactions, which will be present even if there are deviations from the ideal spherical shape.

We then asked whether the 5C interactions identified as significant are generally in line with the general findings of distant chromatin interactions of the locus. Interactions of α-globin gene

with its enhancer HS40, as well as interactions with neighboring hypersensitive sites HS48 and HS46 were identified as key factors determining the expression levels of α-globin gene in globin expressing mouse cells in prior knock-out and 3C studies (Zhou et al., 2011; Vernimmen et al., 2009). We found that indeed pairwise interactions involving the α-globin gene and the enhancers HS40, HS46 and HS48 are all preserved after excluding non-specific interactions (Figure 21E).

4.3.1.1 nC-SAC can generate large ensemble of chromatin chains of α-globin locus

To build structural models of the α-globin locus, we developed the nC-SAC algorithm to generate 3D chromatin chains that satisfy 5C interactions identified as significant. Our goal is to generate conformations from a Boltzmann distribution, which all geometrically possible chromatin chains that satisfy the significant 5C interactions are properly sampled. Following previous studies (Bau et al., 2011; Duan et al., 2010; Rousseau et al., 2011; Ay et al., 2014), we assume an inverse relationship between 5C frequencies and spatial distances, and employ a simple half-Gaussian model to map frequencies of significant 5C interactions to spatial distances between nodes (detailed in Materials and Methods). These spatial distances are then regarded as physical constraints that the 3D chromatin chains need to satisfy. Two separate ensembles of 10,000 chromatin chains of the α-globin locus are then generated for the GM12878 and the K562 cell lines (Figure 21F).

We first assessed the statistical significance of the interactions in these ensembles by using random C-SAC ensemble as our null model. After temporarily excluding interactions that do not pass FDR test, we captured a total of 78 out of 83 (94 %) significant 5C interactions for K562

cells and 113 out of 132 (86 %) for GM12878 cells (Figure 21D). 113 and 78 temporarily excluded 5C interactions for the GM12878 and K562 cell lines also re-appeared in predicted structural configurations, respectively. These observations suggest that while our conservative approach for excluding non-significant interactions are stringent and only the most significant interactions are used as constraints, the resulting 3D chromatin chains contain many moderately strong 5C interactions, which may be biologically relevant. Furthermore, the predicted chromatin chains of the α-globin locus exhibit many novel interactions not present in the original 5C data (278 and 301 interactions in the GM12878 and K562 cell lines, respectively).

4.3.1.2 nC-SAC uncovers structural differences of α-globin locus

There are global structural differences in the organization of the α-globin gene locus between K562 and GM12878 cells, as seen in heatmaps of spatial interactions from predicted α-globin chains (Figure 18A). Overall, α-globin locus of the silent GM12878 cell line forms a single compact chromatin globule. In contrast, chains of the active K562 cell line are extended, forming two non-interacting globules, which exhibit two separate domains in the heatmap (Figure 18A). These findings are consistent with previous results (Bau et al., 2011).

Our model predicts additional global structural differences in chromatin chains between the two cell lines. Using a density based clustering algorithm (Ester et al., 1996), which does not require specification of the number of clusters *a priori*, we partitioned the ensemble of 3D chromatin chains of each cell line into clusters based on their pairwise structural similarity (Materials and Methods). Chromatin chains with structural similarity above a threshold are grouped into the same cluster. The ensemble of the α-globin expressing K562 cell line is remarkably homogeneous. There is overall a small number of clusters (a total of $\sim$13), with the most populated cluster accounting for $\sim$97% of the chromatin chains in the ensemble. In contrast, the ensemble of the non-expressing GM12878 cell line is structurally diverse, with many different clusters (a total of $\sim$148). The most prominent cluster accounts for only $\sim$24% of the whole ensemble. Figure 18b depicts representative 3D chromatin chains of the 1st and the 2nd most populated clusters in both cell lines. While the exact number of the clusters are subject to choices of the clustering technique used, our results indicate that there are significant differences in subpopulation heterogeneity of chromatin chains in the two cell lines. While a

folding landscape of diverse chromatin chains with many subpopulations is evident for the nonexpressing cell line, a major structural subpopulation dominates the α-globin chains in the active cell line.

4.3.1.3 nC-SAC predicts novel interactions that are mediated by proteins and associated with concurrent histone modifications

To determine the biological relevance of the predicted long range α-globin interactions absent in 5C measurements, we examined results from two independent ChIA-PET studies of K562 cells (Li et al., 2012; Heidari et al., 2014) and an independent ChIA-PET study of GM12878 cells (Heidari et al., 2014). ChIA-PET is a 3C-based technique that incorporates chromatin immunoprecipitation analysis to capture looping interactions mediated by proteins or interactions that are associated with histone modifications (Fullwood et al., 2009). Recent ChIA-PET studies revealed looping interactions in the α-globin locus mediated through RNAPII, CTCF and RAD21 binding, as well as interactions associated with histone modifications (Li et al., 2012; Heidari et al., 2014).

Among the 68 RNAPII-mediated interactions in K562 cells detected by ChIA-PET (Li et al., 2012) (Figure 19A, blue circle in the Venn diagram and Figure 19B1, blue and grey arcs), 33 are also predicted by nC-SAC (Figure 19A, orange circle). Notably, 21 of the 33 predicted interactions are novel interactions absent in the 5C measurements (Figure 19B2, red arcs) and 12 are interactions captured by 5C measurements (Figure 19B3, green arcs). Among the 35 RNAPII-mediated interactions undetected by nC-SAC (Figure 19B1, gray arcs), 26 have no primer coverage and therefore are not reflected in the 5C data. The remaining 9 RNAPII-

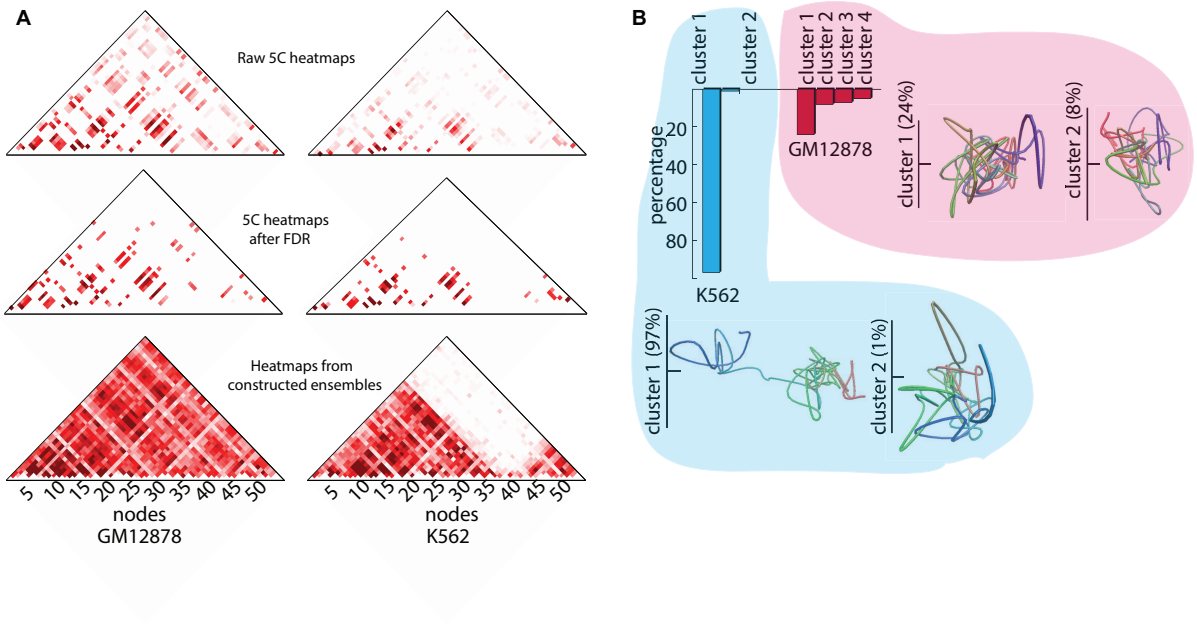

Figure 18. **Ensembles of predicted 3D chromatin chains of the α-globin locus.** Interactions between genomic elements of the α-globin locus from predicted structural ensembles of 10,000 chromatin chains in the silent GM12878 and the active K562 cell lines. (**A**) Heatmaps of spatial interactions of α-globin locus including raw 5C counts, most significant 5C counts after exclusion of non-specific interactions, and counts from the modeled structural ensembles. The normalized frequency of $i - j$ interactions after exclusion of non-specific interactions is color coded. Red intensity indicates increased frequency. (**B**) The histogram (top) shows the proportion of structures associated with different structural clusters (K562, blue; GM12878, red). The predominant three dimensional structures associated with structural clusters 1 and 2 are also shown for both cell lines.

mediated interactions have low or no 5C interactions, imposing very weak constraints for our model. A separate ChIA-PET study (Heidari et al., 2014) revealed two more RNAPII mediated interactions between α-globin gene (node 21) and HS40 (node 12), which is also predicted by nC-SAC and is present in 5C measurements (Table VI).

TABLE VI

THE INTERACTIONS THAT ARE CAPTURED BY CHIA-PET STUDY IN K562 CELL LINE. GENOMIC LOCATIONS ARE IN BASEPAIRS AND REFERENCE GENOME IS HG18.

Genomic Location 1	Genomic Location 2	Node 1	Node 2	Factor	Status
55,200-56,599	169,200-171,999	7	21	H4K4me3	New prediction
55,000-56,999	351,800-354,999	7	39	H4K4me1	Not predicted / not in 5C
96,800-98,199	124,000-130,599	12	16	H4K4me3	New prediction
96,600-98,599	132,000-135,599	12	18	H4K4me2	New prediction
96,600-98,599	167,200-168,799	12	21	H4K427ac	Predicted / in 5C
94,200-96,199	169,200-172,199	12	21	H4K4me1	Predicted / in 5C
96,800-98,199	169,200-171,999	12	21	H4K4me3	Predicted / in 5C
94,897-95,586	170,079-171,864	12	21	PolII	Predicted / in 5C
96,798-97,306	170,079-171,864	12	21	PolII	Predicted / in 5C
96,600-98,599	169,200-172,199	12	22	H4K4me1	Predicted / in 5C
96,600-98,599	169,200-173,399	12	22	H4K427ac	Predicted / in 5C
96,600-98,599	167,200-172,199	12	22	H4K4me2	Predicted / in 5C
100,400-101,199	167,200-168,799	13	21	H4K427ac	Predicted / in 5C

We also examined CTCF-mediated interactions in K562 cells detected by ChIA-PET (Li et al., 2012). Among the 11 reported interactions (Figure 19C, blue circle in the Venn diagram and Figure 19D1, blue and grey arcs) (Li et al., 2012), 8 are predicted by the nC-SAC model (Figure 19C, orange circle). Of those, 6 are absent in the 5C study (Figure 19D2, red

arcs) and 2 interactions are captured by 5C (Figure 19D3, green arcs). The 3 CTCF-mediated interactions detected by ChiA-PET but undetected by nC-SAC (Figure 19D1, gray arcs) either have no 5C frequency or have no primer coverage, hence impose no constraints for our model.

In addition, we examined RAD21-mediated interactions in K562 cells detected by a recent ChIA-PET study (Heidari et al., 2014). Among the 8 reported interactions (Figure 19E, blue circle in the Venn diagram and Figure 19F1, blue and grey arcs), 5 are predicted by the nC-SAC model (Figure 19E, orange circle), 3 of them are novel interactions that are absent in 5C measurement (Figure 19F2, red arcs) and 2 interactions are captured by 5C (Figure 19F3, green arcs). The 3 RAD21-mediated interactions detected by ChiA-PET but undetected by nC-SAC (Figure 19F1, gray arcs) have no 5C coverage, imposing no constraints for our model.

We then examined the interactions that are found to be associated with histone modifications in K562 cells in a recent ChIA-PET study (Heidari et al., 2014). Among the 7 reported interactions, 6 of them are predicted by nC-SAC model, 3 of them are novel interactions that are absent in 5C measurement, and 3 of them are captured by 5C study. The only undetected interactions has no 5C coverage (Table VI).

We further examined RAD21-mediated interactions in the silent GM12878 cells detected by ChIA-PET study (Heidari et al., 2014). Among the 4 reported interactions (Figure 19G, blue circle in the Venn diagram and Figure 19H1, blue and grey arcs), 3 are predicted by the nC-SAC model (Figure 19G, orange circle), including one novel interaction absent in the 5C study (Figure 19H2, red arcs), as well as 2 interactions captured by 5C (Figure 19H3, green

arcs). The only undetected interaction (Figure 19H1, gray arcs) has no 5C coverage, imposing no constraints for our model.

Overall, our nC-SAC method has predicted 52% of the 68 RNAPII-mediated interactions, 75% of the 11 CTCF-mediated interactions, 62% of the 8 RAD21-mediated interactions in K562 cell line, 86% of the 7 interactions that are associated with histone modifications in K562 cell line, and 80% of the 5 RAD21-mediated interactions in GM12878 cell line (Table VII-Table X). In total, 89 interactions are detected by ChiA-PET in K562 cell line and 52 of them are among 457 predicted significant interactions. A randomization test is then carried out and the probability of finding any 52 or more interactions out of the 89 ChIA-PET detected interactions by random chance is found to be $p < 0.01$ (Figure 19I), indicating that our discovery is highly significant.

TABLE VII

DETAILS OF THE PREDICTIONS AND COMPARISON WITH CHIA-PET RNAPII DATA IN K562 CELL LINE

Node	Node	New Prediction	Already in 5C Data	No Primer Site	No Record in 5C	5C Count
5	11			$		
6	16			$		
6	26	+				
6	54					0
8	22	+				
8	38	+				
8	40					0
8	54					5
9	22				#	
9	23	+				
9	27	+				
9	40					33
9	41				#	
11	22	+				
12	22		-			
13	17		-			
14	21		-			
14	22		-			
15	20	+				
15	22	+				
16	18			$		
17	24	+				
17	44			$		
18	39			$		
20	22			$		
21	32				#	
21	36	+				
22	24					37
22	27				#	
22	54					6
23	26	+				
26	28	+				
26	31			$		
26	38	+				
26	39			$		
27	36	+				
27	39			$		
28	31	+				
29	31	+				
30	39			$		
31	34	+				
31	38			$		
31	39			$		

TABLE VIII

DETAILS OF THE PREDICTIONS AND COMPARISON WITH CHIA-PET RNAPII DATA

Node	Node	New Prediction	Already in 5C Data	No Primer Site	No Record in 5C	5C Count
34	38	+				
35	37				#	
35	39			$		
35	44			$		
36	39			$		
37	41					27
38	40				#	
38	44			$		
39	42			$		
39	44			$		
39	54			$		
40	43					23
40	54				#	
41	44				#	
41	54					2
42	52	+				
42	54	+				
43	54		-			
44	50			$		
44	52			$		
44	54			$		
46	51			$		
46	54	+				
49	54	+				
50	54	+				

TABLE IX

DETAILS OF THE PREDICTIONS AND COMPARISON WITH CHIA-PET CTCF DATA IN K562 CELL LINE

Node	Node	New Prediction	Already in 5C Data	No Primer Site	No Record in 5C	5C Count
8	40					0
8	11	+				
8	22	+				
9	11	+				
9	22				#	
11	40			$		
11	22	+				
12	22		-			
13	22		-			
15	22	+				
43	52	+				

4.3.1.4 nC-SAC predicts detailed 3D structural interactions

Expression of the α-globin gene is thought to be regulated through enhancer-promoter interactions (Bau et al., 2011). Indeed, the interaction between the α-globin gene and enhancers HS40/46/48 are found in 90% of predicted chains of the active K562 cells. However, this represents an increase by a factor of only 1.29 compared to the silent GM12878 cells, as this interaction is also present in 69.8% of predicted chains of the GM12878 cells (Figure 20A). Our finding is consistent with a previous ChIA-PET study, in which interactions between HS40 and α-globin gene is found to be mediated by RAD21 in the silent GM12878 cell line (Figure 19G,H) (Heidari

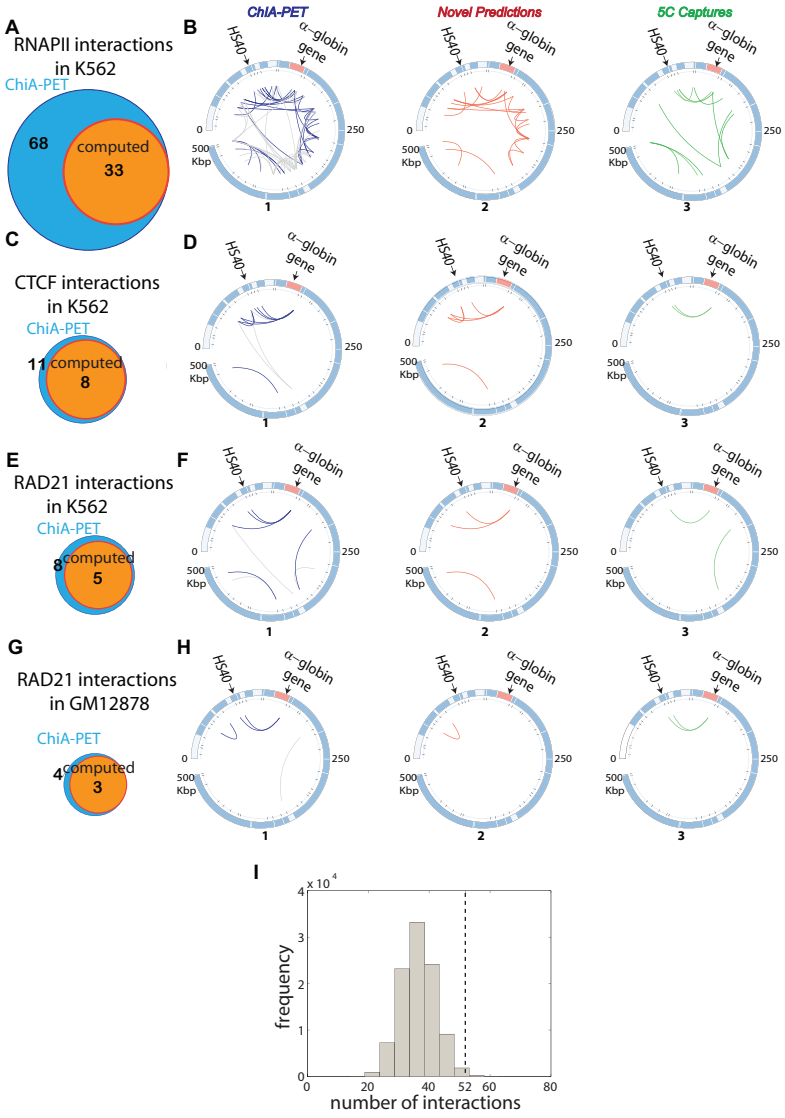

Figure 19. **Predicting novel interactions between genomic elements in α-globin locus and validation of their biological relevance.** **(A,C,E,G)** Comparing looping interactions detected by ChIA-PET (Li et al., 2012) and nC-SAC 3D ensemble predicted interactions in K562 cells. The Venn diagrams show ChIA-PET measured (blue circles) and nC-SAC predicted (orange circles) interactions. **(B1,D1,F1,H1)** The circos diagrams show interactions detected by ChIA-PET (blue arcs for captured interactions by 3D model and gray arcs for interaction that are absent in the 3D model), **(B2,D2,F2,H2)** nC-SAC predicted interactions detected by ChIA-PET but absent in 5C (red arcs), **(B3,D3,F3,H3)** interactions predicted by nC-SAC and captured by the 5C and ChiA-PET techniques (green arcs). **(I)** Histogram of number of interactions among the 89 ChiA-PET interactions that are found in the 100,000 sets of randomly generated 457 interactions. p-value of obtaining 52 out of 89 among 457 interactions is 0.006

TABLE X

DETAILS OF THE PREDICTIONS AND COMPARISON WITH CHIA-PET RAD21 DATA IN K562 CELL LINE

Node	Node	New Prediction	Already in 5C Data	No Primer Site	No Record in 5C	5C Count
7	21	+				0
7	39					
11	21	+				
12	21		-			
26	38		-			
33	38			$		
43	51			$		
43	52	+				

et al., 2014). These observations indicate that α-globin promoter-enhancer interactions alone do not determine the expression level and additional regulatory elements may be at play.

We examined nC-SAC predicted 3D structures for K562 and GM12878 cells to assess the presence of other looping interactions, which may regulate α-globin expression (Figure 20A–D). While it is difficult to compare absolute interaction frequencies between cell lines, we can compare the relative fractions of chromatin chains containing specific interactions in each cell line. We analyze interactions that both α-globin gene and enhancers participate simultaneously for both cell lines and identified the differential interactions (Figure 20E). We then overlapped the epigenetic profiles of each of the differentially interacting nodes and found the ones that are associated with epigenetic marks (Figure 20E and F). As a result, our nC-SAC study predicts that the POL3RK gene engages in a three-way interaction with the α-globin gene

TABLE XI

DETAILS OF THE PREDICTIONS AND COMPARISON WITH CHIA-PET RAD21 DATA IN GM12878 CELL LINE

Node	Node	New Prediction	Already in 5C Data	No Primer Site	No Record in 5C	5C Count
7	10	+				0
12	21		-			
14	21		-			
26	39				#	

and enhancers in 70% of α-globin chromatin structures from GM12878 cells. In contrast, the POL3RK gene has a much lower three-way interaction frequency (18%) with the α-globin gene and enhancers in K562 cells (Figure 20A). The interaction between POL3RK and enhancers was not detected in the original 5C study due to primer design strategy (Bau et al., 2011). With explicitly generated 3D structures, we can measure the exact Euclidean distances between genomic elements in individual chains and can calculate their ensemble averages. We found chains from GM12878 cells with POL3RK:α-globin:enhancers three-body interaction all have average pair-wise distances between elements (50.1 ± 20 nm, 62.4 ± 18 nm, and 80.0 ± 5nm) shorter or near the threshold of interaction ($\sim$80$\pm$5 nm) given in previous studies (Giorgetti et al., 2014) (Table XII, Figure 20D). In contrast, the averaged spatial distances of POL3RK:α-globin ($\sim$135$\pm$20 nm) and POL3RK:enhancers ($\sim$140$\pm$18 nm) are both much longer than this threshold in active K562 cells (Table XII).

TABLE XII

THE AVERAGE SPATIAL DISTANCES BETWEEN NODES IN THE CHAINS WITH THREE-WAY INTERACTION IN GM12878 CELL LINE AND IN THE CHAINS WITHOUT THREE-WAY INTERACTION IN K562 CELL LINE

Node(s)	Node	Distance in GM12878 (nm)	Distance in K562 (nm)
12/13/14 (HS40/46/48)	21 (α-globin gene)	80±5	74.9±5.1
12/13/14 (HS40/46/48)	5 (POL3RK)	50.1±20.0	134.6±30.2
21 (α-globin gene)	5 (POL3RK)	62.4±18.5	140.1±28.2

We speculate that the three-way looping interaction of POL3RK with the α-globin gene and enhancers may occlude access of transcription factors to the α-globin transcriptional elements, thus silencing the α-globin expression (Figure 20B–D). This denial of access could be aggravated when transcription factors bound to the POL3RK gene occupies much of the available space. This scenario is consistent with epigenetic data, in which POLR3K in the silent GM12878 cells is enriched for the binding of transcription factors Pu.1 and Sp1 and for histone modifications H2A.Z and H3Kme2, both of which are related to transcriptional activation (Figure 20F) (ENCODE Project Consortium, 2012). Furthermore, it is also consistent with the observed lack of H3Kme2 modifications on α-globin enhancers in the silent cells, which is related to abundance of transcription factor binding, and with the lack of RNAPII enrichment, which is related to absence of gene expression (Figure 20F).

4.4 Discussion

We describe a method that can transform 2D maps of 5C frequencies of interactions into a population of 3D chromatin chains. Our method identifies the most significant spatial inter-

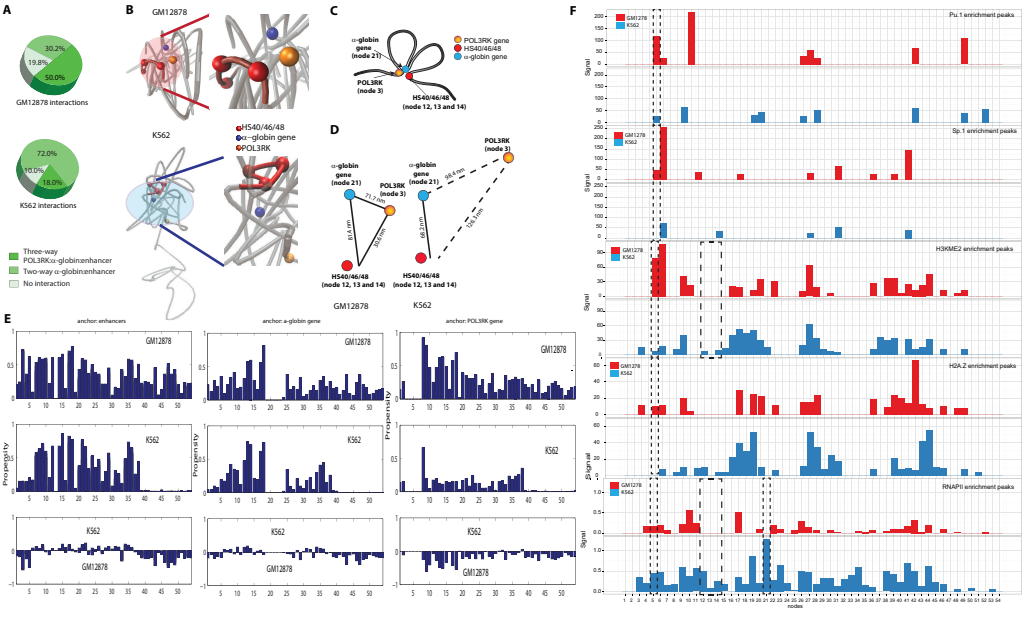

Figure 20. **Three-way interaction of POL3RK:α-globin gene:enhancers is likely a unique feature in the non-expressing GM12878 cells.** (**A**) Pie charts depicting the percentages of the ensembles that have two way (α-globin:enhancer) and three-way (POL3RK:α-globin gene:enhancers) interactions in both GM12878 and K562 cell lines. (**B**) The spatial structures of α-globin locus chromatin were reconstructed from nC-SAC predicted 3D chromatin chains, with the enhancers HS40/46/48 (red), POL3RK (orange), and the α-globin gene (blue) depicted. The depicted structures are drawn from the most populated clusters of GM12878 and K562 cells. (**C**) A schematic representation of the three-body interaction of α-globin gene (blue), enhancer (red), POL3RK (orange) observed in GM12878 cells. (**D**) The spatial distances between the enhancers HS40/46/48 (red), POL3RK (orange), and the α-globin gene (blue) of the three-way interaction unit in the representative structure depicted in (b). (**E**) Interaction profiles of enhancers, α-globin gene and POL3RK gene in the constructed 3D ensembles. x-axis denotes the nodes and y-axis is the propensity of interaction between the anchor node and the rest of the locus. The interactions between enhancers and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively. The interactions between α-globin gene and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively. The interactions between POL3RK gene and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively. (**F**) Chip-Seq enrichment peaks obtained from ENCODE database (ENCODE Project Consortium, 2012). Red bars denote the enrichments in the silent GM12878 cell line and the blue bars denote the enrichments in the active K562 cell line. Enrichment for transcription factor Pu.1 on node 5 (POL3RK) is highlighted in dashed box, where more than 2 fold increase is measured in the silent GM12878 cell line compared to the active K562 cell line.

actions, overcomes the sampling problem, and generates a large number of properly sampled self-avoiding chromatin chains that satisfy constraints imposed by 5C interactions. While its resolution is limited to that of the HindIII fragments in this study (5–13 kb) and no direct information is provided on chromatin dynamics, this method enables us to examine structural properties of the α-globin locus, allowing structural and distance measurements at the population level in a manner consistent with the basic requirement of the physical chromatin chains and the 5C interactions.

Our results show that non-specific spatial interactions arising from nuclear confinement and crowding effect are pronounced in 5C measurements, as up to ∼50–70% of 5C interactions can be accounted for when self-avoiding chromatin chains are confined in the available space of the crowded nucleus. To eliminate false positives, we focus on long-range interactions detected with the strongest statistical confidence. As this strategy is rather conservative, a portion of the 5C interactions that are temporarily excluded do appear subsequently in the constructed 3D ensemble structures, likely resulting from constraints from the stronger interactions and the confinement of the self-avoiding chromatin chains. Biology may take advantage of such interactions and endow them with functional roles. While the reappearance of such interactions in our model does not guarantee that we will detect all functionally relevant moderate or weak interactions, we recognize that their detection with high precision and recall is an overall challenging task in the field. In fact, a randomization test in which 100,000 sets of 457 random i-j interaction pairs are generated shows that the probability of finding any ≥ 52 out of 89 ChIA-PET detected interactions by random chance is $p < 10^{-2}$. Thus, we can conclude that

nC-SAC can predict distant chromatin interactions, which are mediated with proteins or are associated with histone marks. Furthermore, our approach is not overly sensitive to the choice of parameters such as the diameter of the spherical confinement. While recent studies showed that confinement of nucleus is a key determinant of the chromosome organization (Gürsoy et al., 2014a; Kang et al., 2015), as much of the scaling rules including the exponent of looping probability can be explained by the confinement of the self-avoiding chromatin chains, the scaling exponent α changes only slowly as the nuclear diameter in the relevant size regime changes (Gürsoy et al., 2014a). Therefore moderate changes in nuclear diameter and deviations from the spherical shape will likely not affect the identification of the most significant 5C interactions as our criteria are rather stringent.

Our model further predicts global differences in the chromatin chains between cells at different expression levels. It predicts that this locus adopts more homogeneous configurations in the active cells. This finding suggests a common structural scaffold in the active cell line that is required for α-globin expression. The nC-SAC model further allows structural examination of subpopulations of chromatin chains adopting different configurations. As demonstrated by recent single cell studies, cells with identical hormonal stimulation may exhibit diverse levels of gene expression, highly expressed genes at the population level may exhibit bimodal distributions, and epigenetic modifications may be highly heterogeneous (Shalek et al., 2014; Shalek et al., 2013; Rotem et al., 2015). Access to 3D chromatin structures of subpopulations of cells will help to gain understanding of the structural diversity of chromatin chains associated with the heterogeneity of gene expression and epigenetic modifications.

Our method can make detailed predictions of spatial interactions between distant genomic elements, which are validated by available ChIA-PET studies. Excluding locations that lack 5C coverage or locations where ChIA-PET and 5C measurements disagree, our model recovered all remaining RNAPII, CTCF, and RAD21 mediated long-range chromatin interactions, as well as the interactions associated with concurrent histone modifications. While we cannot extrapolate to declare all novel interactions predicted by our model are biologically important, the overall validation by ChIA-PET suggests that our method can make detailed predictions accurately.

Our method can also suggest highly specific and testable mechanistic models of gene regulation. While 5C measurement has identified many important chromatin interactions, details of our predicted chromatin chains suggest a complex many-body mechanism of gene regulation that is beyond a simple gene-enhancer model. Although the α-globin gene and the enhancers HS40/46/48 interact in both cell lines, the enhancers interact strongly with POL3RK in the silent but not in the active cells. As POL3RK is observed to have bound transcription factors, we speculate it may occlude access of enhancers to factors necessary for α-globin activation in the silent cells. This mechanism of gene inactivation through denial-of-access is also consistent with the epigenetic profiles of the enhancers and the POL3RK gene in both cell lines. Analogous to the mechanism of a multi-gene complex for co-transcription, in which the promoter of the first gene acts as an enhancer of the second gene (Li et al., 2012), a multi-gene complex for inactivation may be at play. Since the accessibility of transcription factor binding is a key determinant of gene regulation (Fraser and Bickmore, 2007), the POL3RK gene in this case may act on the enhancers of α-globin gene as a silencer through denial of access of transcription

factors. Although these predictions are rather speculative, they can be tested by genetic perturbation of the identified multi-body structural unit. While recent Hi-C studies (Rao et al., 2014) can identify chromatin interactions at high resolution, the discovery of this many-body mechanism would not be possible without constructing 3D ensemble of structures because of the pairwise capture of Hi-C technique. The importance of 3D model of chromatin interactions was also demonstrated in a recent study, where a many-body interactions between Sox9 and Kcnj2 genes were discovered (Chiariello et al., 2016).

Our study also suggests that integrating 3D models of chromatin chains with epigenetic data can reveal mechanistic insight into the regulation of cell activities. While genome-wide epigenetic studies such as CTCF enrichment and histone modification point to potential regulatory elements and suggest possible long-range interactions along the one-dimensional genome (ENCODE Project Consortium, 2012), it is challenging to interpret and integrate such information. Recent studies showed that important organizational properties of genome such as the formation of TADs can be inferred from the integration of epigenome data with 3D structure construction (Jost et al., 2014; Brackley et al., 2016; Junier et al., 2012). By projecting epigenetic data onto predicted 3D chromatin chains, we showed one can gain better understanding of the complex many-body machineries of gene regulation that involves multiple genomic elements.

nC-SAC method can be used to determine configurations of other gene loci, hence it is general. However, successful predictions are limited by the availability, consistency, and resolution of experimental measurements. In addition, while our method can predict novel interactions, such predictions can only be made in neighborhoods with rich contact information. As the

density of experimentally captured interactions decreases, successful predictions become less likely. In regions devoid of primer coverage, spatial interactions will likely go undetected. For instance, a subset of ChIA-PET identified interactions in regions with no primer coverage or low 5C frequencies are undetected by our method. Regions where no predictions can be made, however, can be identified *a priori* through analysis of primer distribution and 5C frequencies. In principle, any 3C and related data (4C/5C/Hi-C) can be used as spatial constraints to infer 3D chromatin ensembles. Recent high-resolution ($\sim$ 1 kb) Hi-C studies provide great resources of information on 3D genome folding of different cells (Rao et al., 2014). With additional algorithm development, the nC-SAC method can be further improved so it can generate 3D ensembles of chromatins from high resolution Hi-C data. In summary, the nC-SAC method can model chromatin structures of gene loci in cell populations and subpopulations with different expression levels. It can also provide a powerful new approach for identifying spatial structures and interactions and for assessing their roles in regulating gene activities. These results point to exciting opportunities of leveraging limited and pairwise chromosome conformation capture data through modeling of 3D chromatin structures to gain additional knowledge on long-range interactions. Combined with further genetic manipulation, we expect future studies will lead to novel findings on organization of the genome.

CHAPTER 5

COMPUTATIONAL PREDICTIONS OF CHROMATIN HOTSPOTS USING N-CONSTRAINED SELF-AVOIDING CHROMATIN MODEL

5.1 Introduction

Recent development of 3C and related techniques enabled large-scale discovery of distant chromatin contacts among chromosomal locations (Dekker et al., 2002; Lieberman-Aiden et al., 2009; Duan et al., 2010; Montefiori et al., 2016). Understanding the 3D organization of genome using such data is crucial for inferring biological functions such as transcription (Fraser and Bickmore, 2007). The detailed analysis of pairwise interaction frequencies of chromatin revealed the understanding of likely 3D structural units of chromatin that accommodate spatial clustering of different regulatory elements and transcription factors important for cell activities (Phillips-Cremins et al., 2013).

Chromosomes in interphase show a collection of DNA interactions arising from topological constraints, architectural protein binding and significant conformational changes due to the dynamics property of chromatin fiber (Lucas et al., 2014). 3C data are averaged over cell populations and reflect a mixture of different conformations at a particular moment (Ay and Noble, 2015). Therefore, it is challenging to dissect the structural core units of 3D genome organization in the cell nucleus from the analysis of the data. The pairwise nature of 3C poses additional challenges to de-convolute the organization of genome into small units that mediates

the overall folding of chromosomes and are important for transcriptional activation (Dekker et al., 2013).

Current 3D structure modeling approaches that are based on minimal physical assumptions (Gürsoy et al., 2014a; Lieberman-Aiden et al., 2009; Barbieri et al., 2012; Tjong et al., 2012; Wong et al., 2012; Kang et al., 2015; Tokuda et al., 2012; Rousseau et al., 2011; Kalhor et al., 2012; Meluzzi and Arya, 2013; Ay et al., 2014; Trieu and Cheng, 2014; Zhang and Wolynes, 2015; Wang et al., 2015; Tjong et al., 2016), chromosome conformation capture data (Giorgetti et al., 2014), transcription factor binding (Junier et al., 2012; Brackley et al., 2016) and epigenomic states of chromatin (Jost et al., 2014) revealed a wealth of information on the driving forces behind the overall genome organization as well as identification of locus specific interactions that may be important for biological functions. Recent study by Giorgetti et al. (Giorgetti et al., 2014) further established the notion of important loci that determine the structure of a topologically associating domain through virtual mutations of 5C interactions. However, a study that dissects the internal structure of a locus using both epigenetics data and Hi-C measurements in an effort to identify structural hotspots that are responsible for promoter-enhancer interactions is still necessary.

In this chapter, we used n Constrained-Self-Avoiding Chromatin (nC-SAC) computational method for constructing configurations of chromatin chains at the level of large ensembles based on the Hi-C interaction frequencies of the 1 Mb long CCL locus at 3 kb resolution. The interaction frequencies of our nC-SAC ensemble and the interaction frequencies of Hi-C measurements correlate with an R of 0.80 at 10 kb resolution. Our model identifies the

interactions between promoters of CCL genes and distant genomic elements that are subject to histone modifications related to enhancer activity. Majority of identified enhancers are in excellent agreement with experimental studies (Jin et al., 2013; Bonello et al., 2011). We further predicted putative enhancers that have elevated interactions with promoters of genes and are subject to necessary histone modifications, but have not been identified by Hi-C study (Jin et al., 2013). Our findings point to spatially clustered transcriptional units that are composed of many active genomic elements and further show highly variable conformations of these units in the cell population. We further integrated epigenomic profiles of the genomic elements to hypothesize putative structural hotspots that determine the internal structure of CCL locus. Using the nC-SAC method, we created virtual mutations at hypothesized hotspots and measured the resulting changes in the chromatin structure. We proposed that a small number of genomic elements that are highly conserved and enriched with CTCF and cohesin determine the internal structure of the locus as well as are responsible for the interactions between the promoters and enhancers.

5.2 Materials and Methods

5.2.1 Model and Parameters

The overall computational pipeline of nC-SAC model is described and illustrated in Chapter 4. The generation of null model, the calculation of p-values for Hi-C interactions as well as the FDR procedure are done following Chapter 4 Materials and Methods section. The chain growth algorithm is also in Chapter 4 except instead of enforcing distances between monomers,

we enforced an interaction (any spatial distance $\leq$ 850 nm between the monomers that are selected according to interaction probability of their corresponding HindIII fragments.

5.2.1.1 Mapping Hi-C data on to a polymer model

Following Chapter 4, we model CCL lcus chromatin as a polymer chain consisting of monomers that are spheres with 30 nm diameter and 3 kbp genome density. Each HindIII fragment of Hi-C study (Jin et al., 2013) is mapped to several monomers according to their lengths (Figure 21). In total, CCL locus polymer chains contains 340 monomers, spanning 575 HindIII fragments.

5.2.1.2 Converting Hi-C interaction frequencies into probability constraints

Following previous studies (Giorgetti et al., 2014; Kalhor et al., 2012), we assumed a direct relationship between Hi-C interaction frequencies and probability of interactions between monomers. For an interaction frequencies $f_{m,n}$ between HindIII fragments m and n, the interaction probability $p_{m,n}$ is

$$p_{m,n} = \alpha * f_{m,n}$$

,

$$p_{m,n} = 1 \text{ if } f_{m,n} = f^{max}$$

where α is the normalization constant. Beginning of each chain generation process, we decide if the interaction between fragments m and n will happen according to their probability. After deciding the occurrence of interaction, we randomly select monomers i and j in fragments m and n, respectively and enforce an interaction between these monomers (spatial distance between

them is less then 850 nm) during chain growth process. We repeat this procedure for all 30,000 chains of ensemble (Figure 22C).

5.2.1.3 Enforcing repulsive constraints for mutations

When we do a virtual mutation between sites, we make sure that every chain of the ensemble do not have the mutated interaction. During the chain growth process, we enforce that the monomers that are mutated have a spatial distance $\geq$ 850 nm between them.

5.3 Results

5.3.1 Structural modeling of Hi-C data

We used our nC-SAC model that enables to construct realistic ensembles of fiber conformations, which reproduce the interaction frequencies experimentally observed in chromosome conformation capture data sets. The same computational scheme was used to model 5C data (chapter 4) and can be used to model 3C or 4C data; here, we describe its application to Hi-C. A statistical interpretation of data is adopted, where Hi-C interaction frequencies are considered to be proportional to the probability of two genomic elements physically contacting each other within a cell population (See Materials and Methods).

We use the C-SAC polymer model (Gürsoy et al., 2014a) to model the CCL chromatin chain. In this model, we represent the chromatin as a collection of beads. We divide the 1 Mb locus into 340 beads, each corresponds to 3 kb DNA (Figure 21A). The original Hi-C data, based on pairs of interacting fragments that are 5–10 kb long, is thereby converted into a list of interacting pairs of beads (Figure 21A).

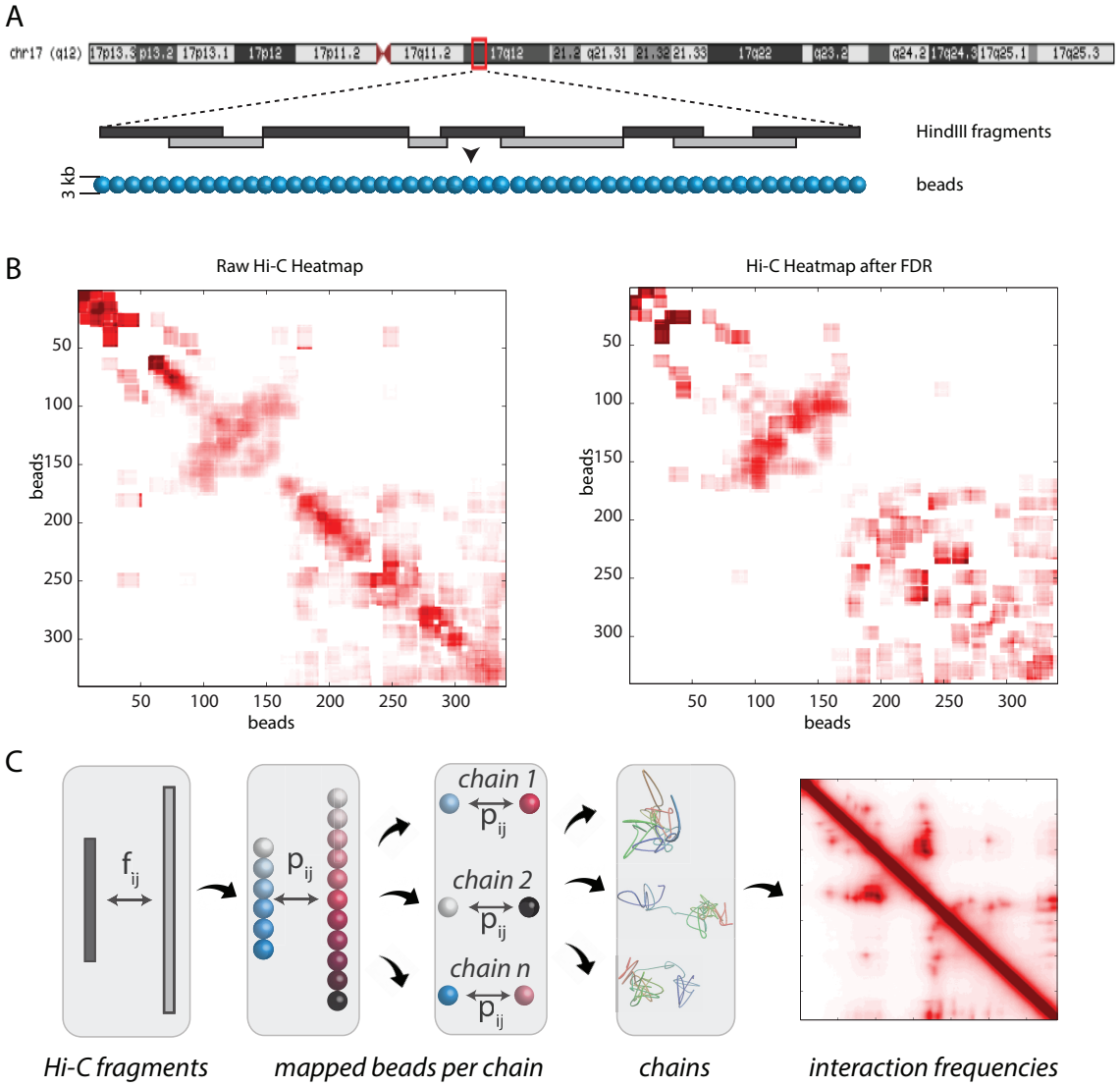

Figure 21. **Mapping Hi-C interactions onto polymer model and identifying non-specific interactions in CCL locus.** (**A**) Mapping the 1 Mb CCL locus and Hi-C interactions onto the polymer model of chromatin chain. Linear diagram of Chr 17 is shown and the genomic location of CCL locus (hg18:Chr17:29,500,000-30,500,000) is denoted in red box. HindIII restriction fragments within the locus are mapped onto sequences of adjacent beads in polymer model. (**B**) All reported Hi-C interactions between genomic elements in the CCL locus and significant Hi-C interactions remaining after interactions due to the polymer effect are excluded. (**C**) Details of mapping and chain generation process. After mapping the HindIII fragments onto beads, we select a bead from each fragment and assign an intreaction probability proportional to Hi-C interaction frequency. This procedure is repeated for every chain generated.

Before modeling 3D structures from Hi-C data, we excluded the interactions arising from the generic effects of constrained polymer chains as they will be intrinsically modeled in our polymer chains without further addition of Hi-C constraints (Belmont, 2014; Kang et al., 2015; Gürsoy et al., 2014a). To locate such interactions, we generated an ensemble of 100,000 random C-SAC chains using a chain growth strategy (Gürsoy et al., 2014a) (Figure 21B). Without *a prior* information, we model the confinement as a sphere of a diameter of 0.55 μm. This diameter is calculated from the volume that is expected to be occupied by a 1 Mb long DNA, assuming 6 billion bp of DNA is confined in an available space of diameter of 10 μm in crowded cell nucleus, which is in range of the size of an average human nucleus with a diameter of 6–20μm (Alberts, 2002). We bootstrap the chains in the random ensemble to generate 100,000 random ensembles of 100,000 C-SAC chains and calculate the probability of observing Hi-C interaction frequencies in these random ensembles and used these probabilities as p-values for the correction of multiple hypothesis testing at the False Discovery Rate (FDR) of $\alpha < 5\%$. A total of 540 interactions are assessed for their statistical significance and 194 Hi-C interactions (36%) are found to be enriched in constrained polymer chains and are therefore not used as constraints.

To build structural models of the CCL locus, we used nC-SAC algorithm that is discussed in detail in Chapter 4. Our goal is to construct chromatin chains from a distribution of samples that satisfy the interaction probabilities derived from Hi-C interaction frequencies. Following previous studies (Giorgetti et al., 2014; Kalhor et al., 2012), we assume a direct relationship between Hi-C interaction frequencies and the interaction probabilities, and map frequencies of significant Hi-C interactions to interaction probabilities between monomers (Figure 21C).

Every time a chain is generated, a subset of significant interactions are selected according to their probability and these interactions are regarded as physical constraints that the 3D chromatin chain needs to satisfy. An ensemble of 30,000 chromatin chains of CCL locus are then generated. The performance of model is determined by comparing the interaction frequencies of constructed ensemble with the Hi-C interaction frequencies and found a Pearson Correlation of 0.80 at 10 kb resolution (Figure 22A). The high resolution (3 kb) heatmap of contacts reveals several looping interactions associated with CTCF/cohesin binding as well as several histone modifications (Figure 22B) in perfect agreement with the original Hi-C study (Jin et al., 2013).

5.3.1.1 Identification of enhancers of CCL genes

Looping interactions between cis-regulatory elements and gene promoters were determined to be important for regulation of transcription (Fraser and Bickmore, 2007; Lieberman-Aiden et al., 2009; Dostie et al., 2006; Helmink and Sleckman, 2012). The identification of chromatin interactions of CCL locus in 3 kb resolution allowed us to examine the distal enhancers of the promoters of CCL genes. For this purpose, we generated virtual 4C plots for each CCL gene from the ensemble of structures we constructed using the nC-SAC approach and analyzed the long-range interactions of anchored gene promoters (Figure 23). We first identified the long-range regulatory genomic elements for all gene promoters. These are (black stars in Figure 23) the interactors of promoters that are also identified by the Hi-C study and are associated with CTCF binding as well as histone marks related to enhancer activity.

We first found that the promoters that are in close genomic proximity are regulated by same enhancers. For example, CCL2 and CCL7 genes are regulated by same enhancer, which is

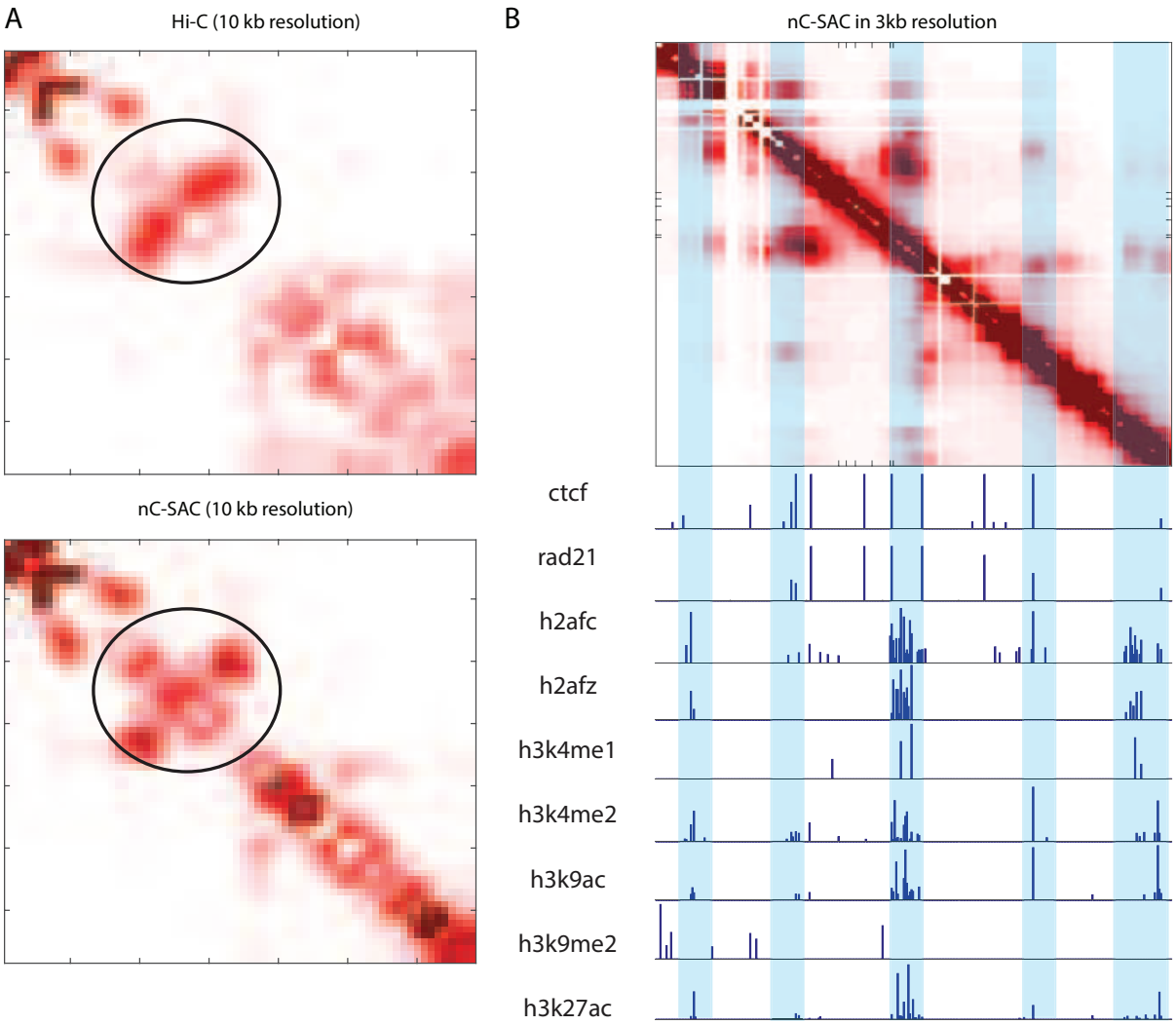

Figure 22. **Ensembles of predicted 3D chromatin chains of the CCL locus.** Interactions between genomic elements of the CCL locus from predicted structural ensembles of 30,000 chromatin chains and comparison against Hi-C and ChIP-Seq data. (**A**) Heatmaps of spatial interactions of CCL locus including Hi-C interaction frequencies and frequencies from the modeled structural ensembles. Red intensity indicates increased frequency in 10 kb resolution. (**B**) Heatmap of spatial interactions of CCL locus from the modeled structural ensemble in 3 kb resolution is depicted along with the available ChIP-Seq data on CTCF/cohesin binding as well as histone modifications.

around 75 kb away from these genes. We also found that genes CCL13 and CCL11 interact with multiple enhancers that are as far as ∼0.5 Mb apart from the promoters. These promoters and enhancers are also associated with CTCF binding sites. We also predicted potential long-range regulatory elements for CCL13 and CCL1 genes (red and green stars in Figure 23), located on the 3' and 5' of the locus. These putative enhancers were not captured by Hi-C due to the distribution of restriction enzyme sites, but are associated with CTCF/cohesin binding as well as histone marks related to enhancer activity. However, they rather make less frequent interactions with the genes compared to other identified enhancers. This analysis also showed that CCL13 and CCL1 genes might be regulated by multiple enhancers, some of which are almost 0.5 Mb away from these genes.

Shared Enhancers:

We first examined whether genes share the same enhancers simultaneously. For example, promoter of the CCL2 and CCL7 genes interact with an enhancer ∼150 kb upstream of them. We found that CCL2-enhancer and CCL7-enhancer interactions are observed simultaneously in 19.7% of the 3D models and this enhancer interacts with either of the CCL2 and CCL7 genes independently in more than 80% of the 3D models. Promoters of CCL11, CCL8, CCL13 and CCL1 genes interact with an enhancer as far as ∼180 kb downstream of CCL1 gene (black star in Figure 23). We found that this enhancer interacts with all four genes simultaneously only in 0.5% of the ensemble. It interacts with any three genes simultaneously in the 5.6% of the ensemble, with any two genes simultaneously in the 35.8% of the ensemble.

Multiple Enhancers:

We then examined the fraction of 3D models in the ensemble that multiple interactions between enhancers and a promoter of a gene are simultaneously observed. For example, promoter of CCL3 gene significantly interacts with three enhancers, two of which are speculated by our study (Figure 23). Among all 30,000 3D models, there is no single chain that all three interactions happen simultaneously. Only 5.5% of the ensemble have any two interactions simultaneously. Similarly, the promoter of CCL1 gene interacts with 2 other enhancers at the same time only in 5.3% of the ensemble and we observe each interaction independently in ∼90% of the ensemble for both CCL13 and CCL1 gene promoters. These results suggest a transcriptional activation mechanism that are independently backed-up by different enhancers.

These findings suggest that the conformation of the CCL locus is highly variable in the ensemble. A wide variety of locus configurations coexist within the ensemble, ranging from consisting of multiple promoter-enhancer interactions to single promoter-enhancer interaction. In the case of shared enhancers for CCL2 and CCL7 genes, additional single-cell data will shed light into the mechanism of whether these genes are active or not in the same cell simultaneously. For CCL13 and CCL1 genes that interact with multiple but same enhancers, additional experimental investigation will be fruitful to examine if a back-up transcription mechanism is at play for the competition between the promoters and the same enhancers.

5.3.1.2 CTCF on the folding of CCL locus chromatin

As the CTCF is an important element for the genome organization (Phillips and Corces, 2009) and is highly enriched on promoters and enhancers of CCL locus (Jin et al., 2013),

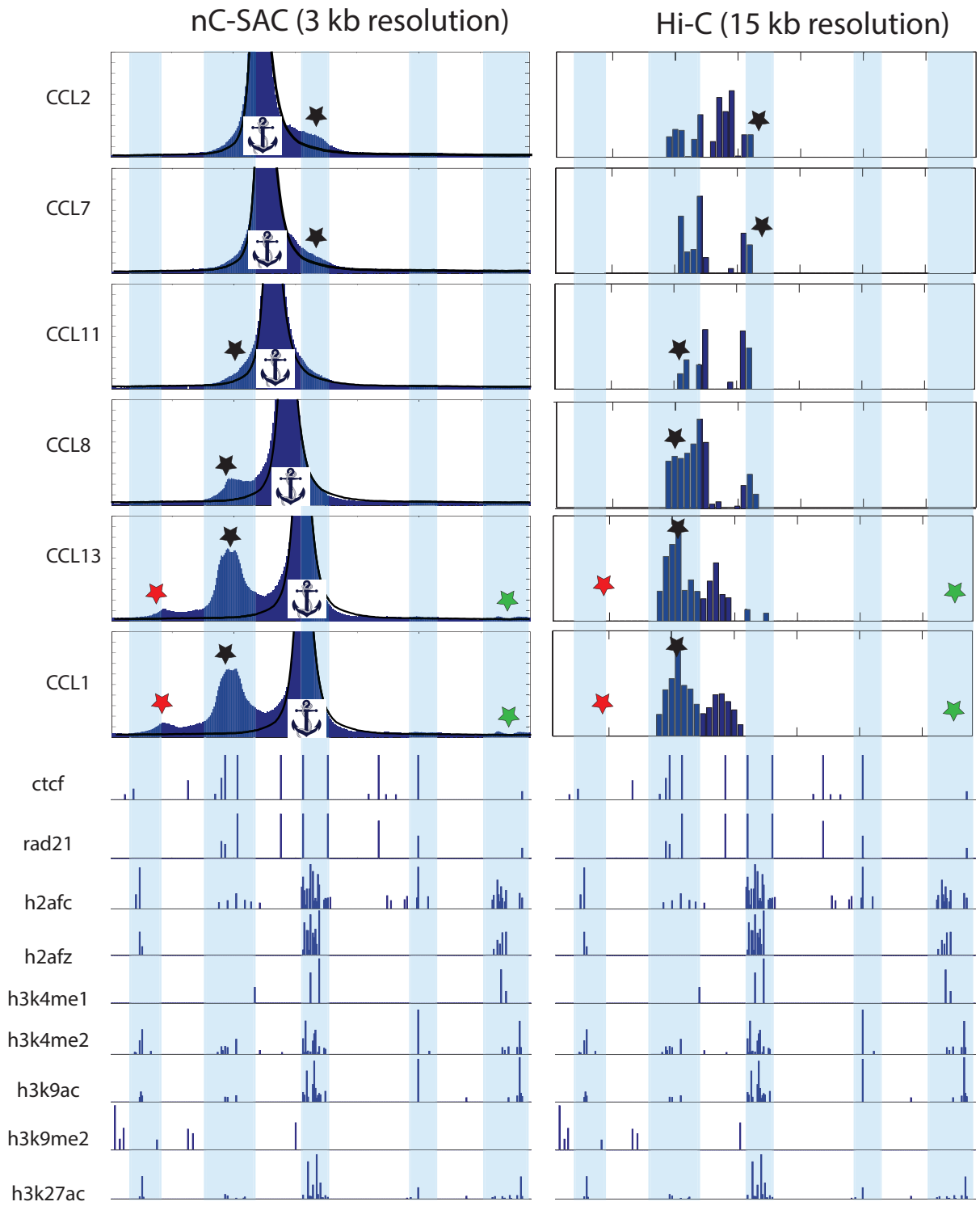

Figure 23. **Virtual 4C plots for the genes on CCL locus.** Interactions between promoters of the important genes and the rest of the locus is depicted along with the available ChIP-Seq data both in the nC-SAC ensemble and Hi-C data. Black stars denote the enhancers that are identified by nC-SAC model but absent in Hi-C data.

we examined the role of CTCF on determining the overall configuration of the locus. We first generated an ensemble of chromatin chains by allowing the interactions only between CTCF binding sites determined by genome-wide ChIP-Seq experiments and referred it CTCF ensemble (Figure 24A). We compared the interaction frequencies of CTCF ensemble with the interaction frequencies of the ensemble that we obtained by using Hi-C interactions which we will refer as the wild type (Figure 24A). We found that using just CTCF sites as interaction pairs is not adequate enough to capture the interaction patterns of the locus. Specifically, the interactions between the promoter of the genes and the enhancers cannot be captured by this CTCF ensemble (Figure 24B). To quantify the how different the wild type is from the CTCF ensemble, we counted the number of contacts that are lost at least in 15% of the ensemble and found that 380 interactions are observed more in wild type (Figure 24A). That is, the linear ChIP-Seq maps of CTCF binding sites do not provide information on which CTCF binding sites interact with each other. This result in a structural ensemble that equally satisfies all 256 possible CTCF binding site pairs and misses the other interactions measured by Hi-C experiment.

We then generated another ensemble by using all significant Hi-C interactions except the ones between CTCF sites. We enforced a repulsion between CTCF binding sites that are observed in Hi-C data and kept the rest of Hi-C constraints as they are during chain generation process. We generated an ensemble of chains and referred it noCTCF ensemble (Figure 24A). The comparison between wild type and noCTCF ensembles showed that a total of 350 interactions are observed in wild type significantly more than those in noCTCF ensemble. This shows

that mutations on the CTCF binding sites has less effects on the overall configurations of locus compared to effects of the ensemble generated using only CTCF binding site pairings (loss of 350 interactions *vs.* 380 interactions). These results suggest that CTCF interactions alone do not determine the 3D configuration of the locus.

We also compared the interactions of the promoters in the CTCF, noCTCF and wild type ensembles (Figure 24C). We already knew that promoter of the genes and enhancers are enriched with CTCF binding. CTCF alone is not adequate enough to drive the formation of the interactions between promoters and enhancers as those interactions are lost in the CTCF ensemble. However, CTCF still plays a key role in formation of these interactions as they are lost in noCTCF ensemble as well (Figure 24C). We concluded that even though CTCF is a major player of the promoter-enhancer interactions and overall configuration of the locus, remaining Hi-C interactions that are probably mediated by some other factors are necessary for bringing the promoters and enhancers spatially together.

5.3.1.3 Evolutionary conservation determines promoter-enhancer interactions and the internal structure of the CCL locus

Following Giorgetti et. al (Giorgetti et al., 2014), we asked the question whether we can identify important structural hot spots that are responsible of bringing the promoter-enhancer interactions together and determining the internal structure of CCL locus. After identification of interaction peaks that are also enriched with important histone modifications from the high-resolution (3 kb) heatmaps of ensemble of constructed CCL conformations (Figure 25a), we systematically constructed ensembles by adding repulsion between the beads of mutated

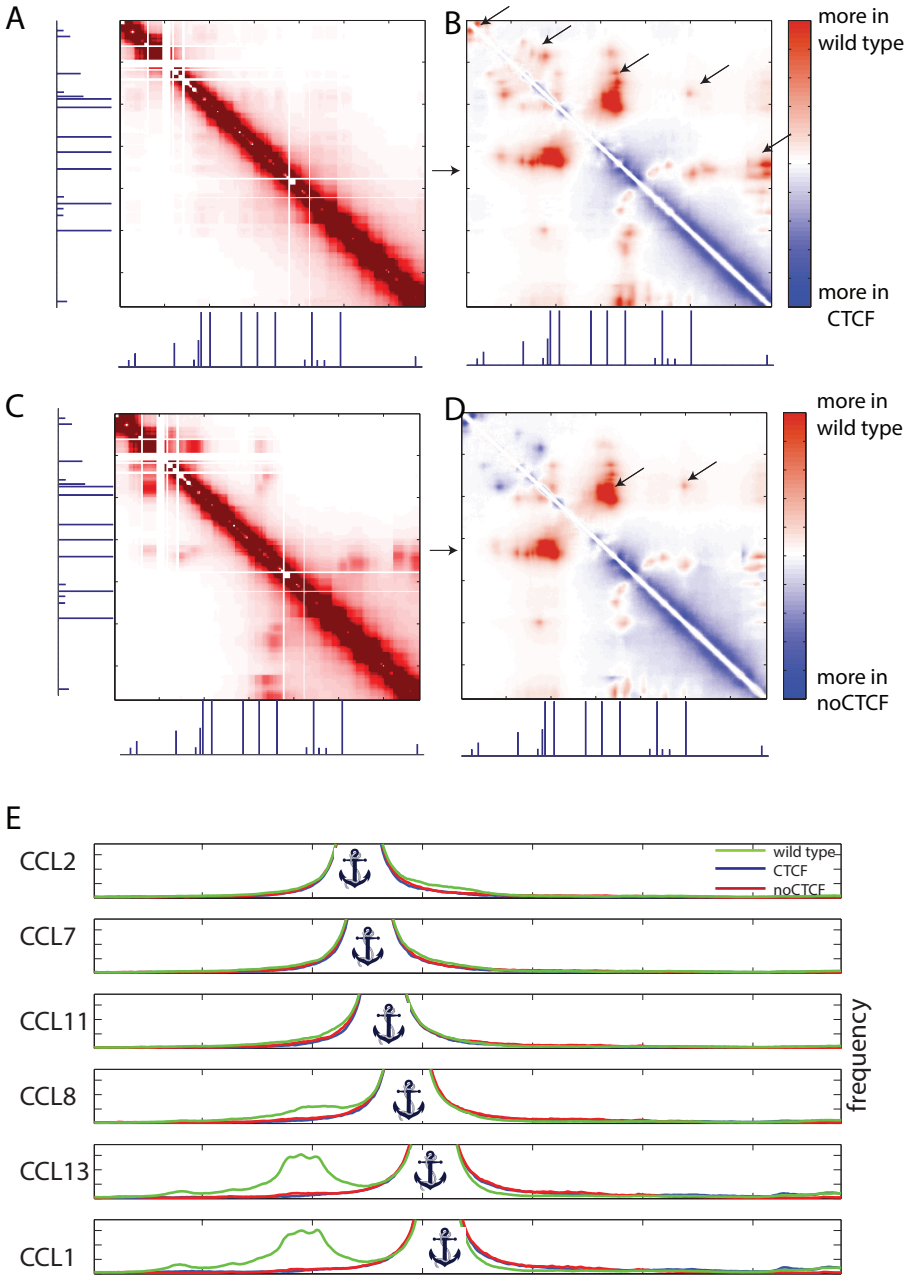

Figure 24. **The effect of CTCF on internal structure of CCL locus**
(**A**) Heatmap of interactions of the ensemble constructed using CTCF binding sites as interaction constraints. Bars show the CTCF enrichment in ChIP-Seq data. (**B**) Heatmap of difference of interactions between the CTCF ensemble and the wild type. Red intensity indicates that the interaction is enriched more in wild type and blue intensity indicates that the interaction in enriched more in CTCF ensemble. (**C**) Heatmap of interactions of the ensemble constructed using Hi-C interactions that are not between CTCF binding sites as constraints. (**D**) Heatmap of difference of interactions between the noCTCF ensemble and the wild type. Red intensity indicates that the interaction is enriched more in wild type and blue intensity indicates that the interaction in enriched more in noCTCF ensemble. (**E**) Interactions between promoters of the important genes and the rest of the locus is depicted for the wild type, CTCF and noCTCF ensembles for comparison.

regions while leaving other Hi-C interactions unchanged. This ensures that the resulting ensemble never contains the interactions between mutated sites. We first calculated the number of affected interactions by identifying the interactions that are lost in more than 10% of the ensemble after these virtual mutations. We found that there is no correlation between the Hi-C frequencies and the number of interactions that are lost after mutations. For example, when we removed the interactions between sites II and III (mutation d), which has the highest Hi-C interaction frequency among other interactions (Figure 25A), no loss of interactions between other sites (Figure 25B) were observed and the rest of the structure of the locus remained unchanged. We found that disrupting the interactions between sites II and V (mutation c) results in loss of 281 other interactions Figure 25B). These lost interactions encapsulate the ones between promoters and enhancers. The removal of interactions between sites I and III (mutation b) also causes disruption of internal structure of locus with loss of 228 other interactions (Figure 25B). Further analysis of these hotspots yield that sites II and V contain CTCF/cohesin enrichment peaks as well as high conservation scores. Similarly, site I and II have the highest conservation scores among all other sites.

This analysis suggests that a small number of hotspots control the overall organization of CCL locus and promote the interactions between important genes and their enhancers. Although, CTCF and cohesin are highly enriched architectural proteins in the locus, removal of interactions between the evolutionarily conserved CTCF binding sites result in bigger architectural changes in the structure of locus compared to the CTCF binding sites that are not conserved. We also showed that the interactions between conserved CTCF binding sites are

not necessarily the only key architectural elements of the locus, but other key loci that are highly conserved and might be mediated by other factors are important determinants the internal structure of the CCL chromatin, as well as promoting the contacts between enhancers and promoters of CCL genes.

5.4 Discussion

In this study we used our nC-SAC method (Chapter 4, Materials and Methods) to decipher the important structural components of CCL locus chromatin along with its sequence properties obtained from publicly available ChIP-Seq data (Jin et al., 2013). Our method generates large number of properly sampled self-avoiding chromatin chains that satisfy constraints imposed by Hi-C interactions as well as creates chromatin chains without selected interactions mimicking knock-out experiments. Consequently, this method enables us to examine structural properties of the CCL locus allowing exact comparison between knock-outs and wild type and to dissect structural hotspots associated wit h important epigenomic marks. Our results showed that interactions frequencies of our nC-SAC ensemble are correlated with Hi-C interaction frequencies with an R of 0.80 at 10 kb resolution. We further mapped the linear ChIP-Seq data on high-resolution heatmap of chromatin interactions and found an enrichment of CTCF/Cohesin binding as well as histone modifications on genomic regions with elevated interaction frequencies, in excellent agreement with original Hi-C and ChIP-Seq study (Jin et al., 2013).

We further predict enhancers for the important CCL genes using our detailed interaction frequencies and epigenomic data. We identified enhancers that are also identified by the original Hi-C study (Jin et al., 2013) an d other experimental studies (Bonello et al., 2011), as

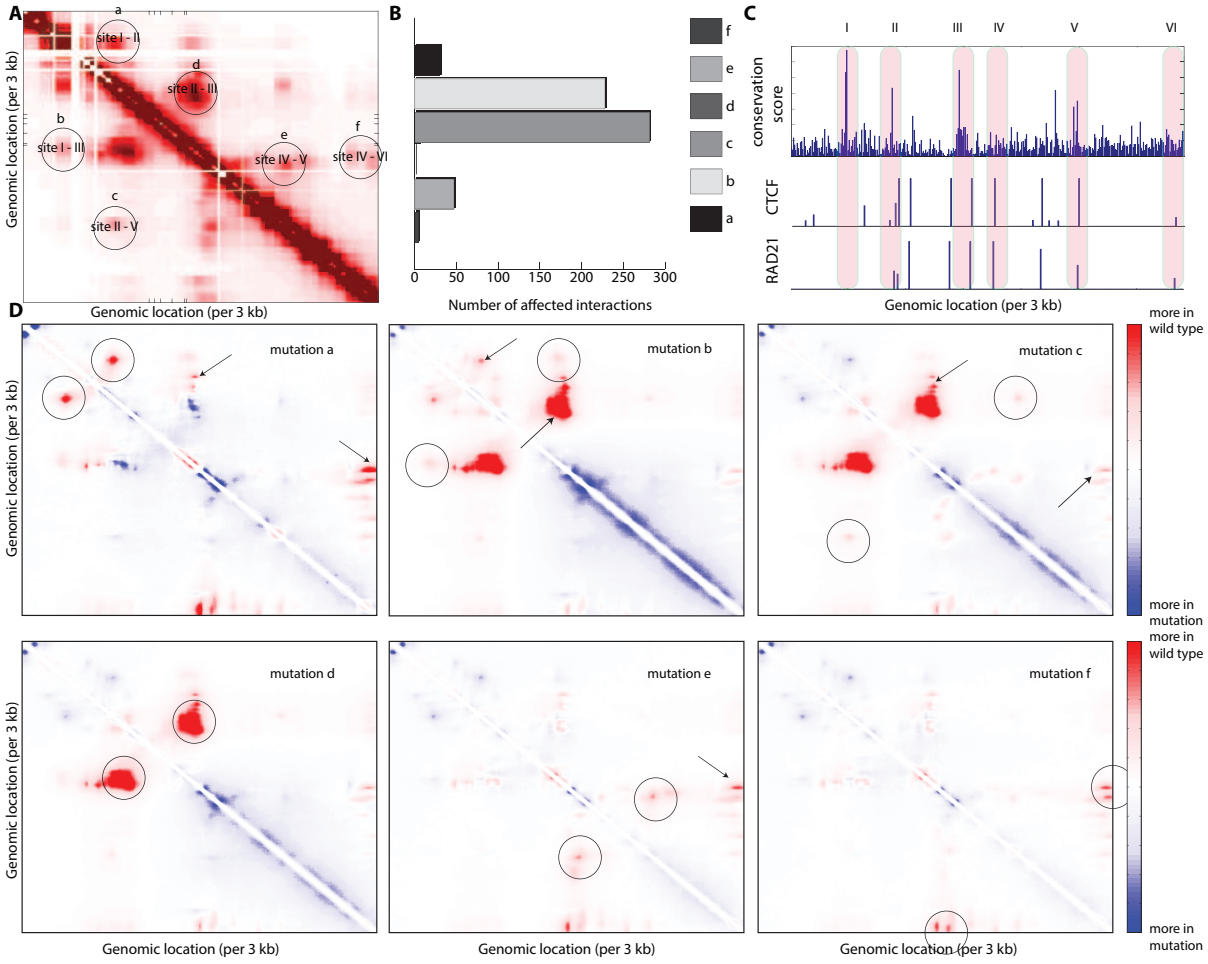

Figure 25. **Identification of hotspots of CCL locus**
(**A**) Selecting highly interacting pairs from the nC-SAC ensemble of CCL locus structures. Each candidate hotspot interaction is coded with letters, while sites are coded with numbers. (**B**) Bar plots of number of interactions that are lost due to the disruption of potential hotspot interactions. (**C**) Conservation, CTCF and RAD21 binding scores per 3 kb fragment of CCL locus. (**D**) Heatmaps of difference between the interaction frequencies of virtual mutations and wild type. Red color intensity indicates that the interactions are lost due to the mutations.

well as putative enhancers that are enriched with necessary histone marks. We speculated the abundance of two types of transcriptional units that are composed of (1) shared enhancers for multiple genes, and (2) multiple enhancers for a single gene. We showed that these transcriptional units are consequences of population-averaged nature of Hi-C data. Each unit takes several different configuration of chromatin fiber and only small amount of chromatin chains forms these multiple interactions simultaneously. Additional single-cell studies will be fruitful in understanding of expression of multiple genes are whether simultaneous or independent as well as if multiple enhancers are at play simultaneously in a single cell.

A major advantage of our sampling technique that we can generate of ensemble of structures with desired chromatin interactions. Here we exploited the role of CTCF in detail by generating ensemble of structures that contain only CTCF interactions and another ensemble of structures that are composed of Hi-C interactions that are not between CTCF enriched sites. We showed that linear maps of CTCF enrichment cannot provide enough information to generate ensembles of structures that capture the measured interactions. That is, it is extremely important to know which CTCF sites interact *a priori*. On the other hand, the ensemble with Hi-C interactions that are not between CTCF sites also do not capture the interactions between promoters and enhancers, while capturing the interaction patterns of rest of the locus. We concluded that as much as CTCF is a key player on shaping the 3D configuration of CCL locus, other interactions that are maybe mediated by other proteins or factors are as important for chromatin folding.

We also exploited other genomic elements that have elevated interactions with each other and overlap with epigenomic data. By creating virtual disruptions and generating ensembles

for each disruption, we measured the changes in the resulting chromatin structures compared against wild type ensemble. We unexpectedly found that interaction frequency of disrupted interactions has no correlation with the resulting changes in the ensemble. In contrast, we observed the largest changes in the configuration of the locus, when we disrupted low frequency interactions. This shows that the structural hotspots that shape the overall configuration of locus cannot directly be revealed from pairwise Hi-C data.

Another unexpected prediction was that the evolutionary conservation play important roles in the 3D structure of CCL locus. We found that mutations of only conserved CTCF binding sites or just conserved regions have more dramatic effects on the interaction patterns of CCL locus, specifically on the interactions between promoters and enhancers, than the mutations of other CTCF binding sites. This points to the importance of integration of epigenomics data with genomics variation and superimposing them on model 3D structures of chromatin.

By combining 3D structure modeling with epigenomics data, we have been able to dissect the structural hotspots of CCL locus and reveal information on configurational differences of single chromatin chains at the population level. Having defined key genomic elements establishing the interactions between promoters and enhancers, we can now suggest experimentally testable hypothesis on trasncriptional machinery of CCL locus. In summary, the nC-SAC approach along with the virtual mutations provide a powerful tool deciphering the structural units of chromatin present in cell population and exploring their impact on gene regulation.

CHAPTER 6

CONCLUSIONS

In this thesis, we have developed computational methods to construct model three-dimensional structures of chromatin for an understanding of gene regulation mechanisms involving the effect of nuclear space in maintaining the epigenetic state of the cell, long-distance DNA loopings that promote cell-specific gene expression, and the main physical factors and mechanisms that determine the genome organization.

6.1 Folding principles of human chromosomes

We have characterized the role of nuclear confinement on the overall organization of human chromosomes. We found that nuclear space that is available for genome is a major determinant of the statistical properties of genome that are observed experimentally. As nuclear size changes, there are significant differences in the chromosome architecture, which are reflected in variations in the scaling exponents. We showed that tentative formation of TADs mainly dictated by the polymer effects under the constraints of cell nucleus without the need of introducing additional binder molecules and fine tuning of their concentrations. Analysis of our predicted ensemble of three-dimensional structures showed that we can both capture the overall scaling properties of genome as well as the variation between different chromosomes. In addition, we have shown that randomly placed binders do not affect directly the scaling behavior. Biological binders such

as CTCF may play more specific roles of modifying or biasing chromosomes towards formation of specific domains required for cell function.

6.1.1 Future Work

As spatial confinement is a dominant factor in determining chromosome folding, the specific epigenetic state of genes and transcription activities in different cell types are likely influenced by the degree of nuclear confinement. Cell nucleus size at different developmental stages or physiological states may be altered to induce different chromosome folding landscape, enabling different genetic programming to be activated. How nuclear size and shape relate to cell size and shape, and how their relative ratio or pattern regulate the epigenetic programs of the cells at different developmental stages are important problems requiring further investigations.

In addition, current chromatin models are based on growing a single chromosome chain, and cannot be used to study inter-chromosomal interactions. Another question is how the 15 Mb sequence scale, and the parameter D are controlled in the cell. These issues will likely be resolved when available algorithm is further improved.

6.2 Folding principles of yeast genome

In this study, we explored computationally the structural properties of budding yeast genome under different combinations of landmark constraints and nuclear confinement. Our results showed that the overall patterns of chromatin interactions of budding yeast genome are well captured when only polymer effects under the spatial confinement of cell nucleus and landmark constraints are considered. We found that the size of the nuclear confinement is the key determinant of intra-chromosomal interactions, while centromere tethering is responsible

for much of the observed inter-chromosomal interactions and correlation of pairwise telomere distances to chromosomal arm lengths. Furthermore, novel chromatin interactions undetected in experimental studies can be uncovered from the ensemble of model genomes generated with nuclear confinement and landmark constraints, and are found to be stabilized by binding of a transcription factor and RNA polymerase. In addition, we found there are important specific genomic elements enriched with tRNA genes that were not captured by polymer properties under landmark constraints, but are detected in experimental studies. Overall, our findings define the specific roles of confinement and individual landmarks, and can uncover likely biologically relevant interactions from genome-wide 3C measurements that are beyond polymer effects.

6.2.1 Future Work

Although we showed that experimentally measured interactions can be recapulated by constraining random self-avoiding chromatin chains with nuclear confinement and landmarks, because of the coarse-grained nature of both current polymer models and genome-wide 3C techniques, our model does not contain detailed spatial information of yeast genome. Inferring structural units of gene regulatory machineries that span just a few kilo bases requires chromatin models of much finer resolution. As the advances in theory, model, and experimental measurements continues, it is envisioned that high resolution models of yeast genome can be computed in the future.

6.3 Identification of gene regulatory units of α-globin locus

We describe a method that can transform 2D maps of 5C frequencies of interactions into a population of 3D chromatin chains. Our method identifies the most significant spatial inter-

actions, overcomes the sampling problem, and generates a large number of properly sampled self-avoiding chromatin chains that satisfy constraints imposed by 5C interactions. The model described here allowed structural and distance measurements at the population level in a manner consistent with the basic requirement of the physical chromatin chains and the 5C interactions.

Our results showed that non-specific spatial interactions arising from nuclear confinement and crowding effect are pronounced in 5C measurements, as up to ∼50–70% of 5C interactions can be accounted for when self-avoiding chromatin chains are confined in the available space of the crowded nucleus. Our model further predicts global differences in the chromatin chains between cells at different expression levels. Our method can make detailed predictions of spatial interactions between distant genomic elements, which are validated by available ChIA-PET studies. While 5C measurement has identified many important chromatin interactions, details of our predicted chromatin chains suggest a complex many-body mechanism of gene regulation that is beyond a simple gene-enhancer model.

6.3.1 Future Work

The resolution of our model is limited to that of the HindIII fragments in this study (5–13 kb) and no direct information is provided on chromatin dynamics. These issues will likely be resolved with high resolution data as well as modeling. Live cell imagin techniques has been emerging lately, incorporating such data to the models will give further information on chromatin dynamics. Our method can be used for construction of configurations of other gene loci. However, successful predictions are limited by the availability, consistency, and resolution of experimental measurements. In addition, while our method can predict novel interactions,

such predictions can only be made in neighborhoods with rich contact information. As the density of experimentally captured interactions decreases, successful predictions become less likely. In regions devoid of primer coverage, spatial interactions will likely go undetected. Recent high-resolution ($\sim$ 1 kb) Hi-C studies provide great resources of information on 3D genome folding of different cells (Rao et al., 2014). With additional algorithm development, the nC-SAC method can be further improved so it can generate 3D ensembles of chromatins from high resolution Hi-C data.

6.4 Identification of chromatin hotspots of CCL locus

In this chapter, we used n Constrained-Self-Avoiding Chromatin (nC-SAC) computational method for constructing configurations of chromatin chains at the level of large ensembles based on the Hi-C interaction frequencies of the 1 Mb long CCL locus at 3 kb resolution. The interaction frequencies of our nC-SAC ensemble and the interaction frequencies of Hi-C measurements correlate with an R of 0.80 at 10 kb resolution. Our model identifies the interactions between promoters of CCL genes and distant genomic elements that are subject to histone modifications related to enhancer activity. Majority of identified enhancers are in excellent agreement with experimental studies. We further predicted putative enhancers that have elevated interactions with promoters of genes and are subject to necessary histone modifications, but have not been identified by Hi-C study. Our findings point to spatially clustered transcriptional units that are composed of many active genomic elements and further show highly variable conformations of these units in the cell population. We further integrated epigenomic profiles of the genomic elements to hypothesize putative structural hotspots that determine the internal structure of CCL

locus. Using the nC-SAC method, we created virtual mutations at hypothesized hotspots and measured the resulting changes in the chromatin structure. We proposed that a small number of genomic elements that are highly conserved and enriched with CTCF and cohesin determine the internal structure of the locus as well as are responsible for the interactions between the promoters and enhancers.

6.4.1 Future Work

By combining 3D structure modeling with epigenomics data, we have been able to dissect the structural hotspots of CCL locus and reveal information on configurational differences of single chromatin chains at the population level. Having defined key genomic elements establishing the interactions between promoters and enhancers, we can now suggest experimentally testable hypothesis. We will benefit from the experimental verification of these hypothesis, and will better understand the structural regulatory machineries in functional loci.

REFERENCES

ed. B. Alberts Molecular biology of the cell. New York, Garland Science, 4th ed edition, 2002.

Amendola, M. and van Steensel, B.: Mechanisms and dynamics of nuclear lamina-genome interactions. Current Opinion in Cell Biology, 28:61–68, June 2014.

Ay, F., Bunnik, E. M., Varoquaux, N., Bol, S. M., Prudhomme, J., Vert, J.-P., Noble, W. S., and Le Roch, K. G.: 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, June 2014.

Ay, F. and Noble, W. S.: Analysis methods for studying the 3d architecture of the genome. Genome Biology, 16:183, 2015.

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., and Nicodemi, M.: Complexity of chromatin folding is captured by the strings and binders switch model. Proceedings of the National Academy of Sciences of the United States of America, 109(40):1617316178, October 2012. PMID: 22988072.

Bau, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J., and Marti-Renom, M. A.: The three-dimensional folding of the -globin gene domain reveals formation of chromatin globules. Nature Structural & Molecular Biology, 18(1):107–114, January 2011.

Beliveau, B. J., Apostolopoulos, N., and Wu, C.-t.: Visualizing Genomes with Oligopaint FISH Probes: Visualizing Genomes with Oligopaint FISH Probes. In Current Protocols in Molecular Biology, eds. F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. Seidman, J. A. Smith, and K. Struhl, pages 14.23.1–14.23.20. Hoboken, NJ, USA, John Wiley & Sons, Inc., January 2014.

Belmont, A. S.: Large-scale chromatin organization: the good, the surprising, and the still perplexing. Current Opinion in Cell Biology, 26:69–78, February 2014.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. 57(1):289–300, 1995.

Berger, A. B., Cabal, G. G., Fabre, E., Duong, T., Buc, H., Nehrbass, U., Olivo-Marin, J.-C., Gadal, O., and Zimmer, C.: High-resolution statistical mapping reveals gene territories in live yeast. Nature Methods, 5(12):1031–1037, December 2008.

Bohn, M. and Heermann, D. W.: Diffusion-driven looping provides a consistent framework for chromatin organization. PloS one, 5(8):e12218, 2010. PMID: 20811620.

Bohn, M., Heermann, D. W., and van Driel, R.: Random loop model for long polymers. Physical review. E, Statistical, nonlinear, and soft matter physics, 76(5 Pt 1):051805, November 2007. PMID: 18233679.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Mller, S., Eils, R., Cremer, C., Speicher, M. R., and Cremer, T.: Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLoS Biology, 3(5):e157, April 2005.

Bonello, G. B., Pham, M.-H., Begum, K., Sigala, J., Sataranatarajan, K., and Mummidi, S.: An evolutionarily conserved TNF-alpha-responsive enhancer in the far upstream region of human CCL2 locus influences its gene expression. Journal of Immunology (Baltimore, Md.: 1950), 186(12):7025–7038, June 2011.

Brackley, C. A., Johnson, J., Kelly, S., Cook, P. R., and Marenduzzo, D.: Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. Nucleic Acids Research, 44(8):3503–3512, May 2016.

Branco, M. R. and Pombo, A.: Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations. PLoS Biology, 4(5):e138, April 2006.

Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M.: Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proceedings of the National Academy of Sciences of the United States of America, 101(47):16495–16500, November 2004. PMID: 15545610.

Bystricky, K., Laroche, T., van Houwe, G., Blaszczyk, M., and Gasser, S. M.: Chromosome looping in yeast: telomere pairing and coordinated movement reflect anchoring efficiency and territorial organization. The Journal of Cell Biology, 168(3):375–387, January 2005.

Chen, M. and Gartenberg, M.: An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet., 43(7), 2011.

Cherry, J., Hong, E., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E., Christie, K., Costanzo, M., Dwight, S., Engel, S., Fisk, D., Hirschman, J., Hitz, B., Karra, K., Krieger, C., Miyasato, S., Nash, R., Park, J., Skrzypek, M., Simison, M., Weng, S., and Wong, E.: Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res., 40:D700–D7055, 2012.

Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M.: Polymer physics of chromosome large-scale 3d organisation. Scientific Reports, 6:29775, 2016.

Chuang, C.-H., Carpenter, A. E., Fuchsova, B., Johnson, T., de Lanerolle, P., and Belmont, A. S.: Long-Range Directional Movement of an Interphase Chromosome Site. Current Biology, 16(8):825–831, April 2006.

Cook, P. R.: The organization of replication and transcription. Science, 284(5421):1790–1795, June 1999.

Cremer, C., Mnkel, C., Granzow, M., Jauch, A., Dietzel, S., Eils, R., Guan, X.-Y., Meltzer, P., Trent, J., Langowski, J., and Cremer, T.: Nuclear architecture and the induction of chromosomal aberrations. Mutation Research/Reviews in Genetic Toxicology, 366(2):97–116, November 1996.

Cremer, T. and Cremer, C.: Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews. Genetics, 2(4):292301, April 2001. PMID: 11283701.

Cremer, T., Kreth, G., Koester, H., Fink, R. H., Heintzmann, R., Cremer, M., Solovei, I., Zink, D., and Cremer, C.: Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture. Critical Reviews in Eukaryotic Gene Expression, 10(2):179–212, 2000.

Dahl, K. N., Ribeiro, A. J. S., and Lammerding, J.: Nuclear shape, mechanics, and mechanotransduction. Circulation research, 102(11):1307–1318, June 2008. PMID: 18535268.

de Gennes, P.: Scaling Concepts in Polymer Physics. New York, Cornell University Press, 1979.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A.: Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews. Genetics, 14(6):390–403, June 2013.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N.: Capturing chromosome conformation. Science (New York, N.Y.), 295(5558):1306–1311, February 2002.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398):376–380, April 2012.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J.: Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research, 16(10):12991309, October 2006. PMID: 16954542.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S.: A three-dimensional model of the yeast genome. Nature, 465(7296):363–367, May 2010.

Dvorkin, N., Clark, M. W., and Hamkalo, B. A.: Ultrastructural localization of nucleic acid sequences in Saccharomyces cerevisiae nucleoli. Chromosoma, 100(8):519–523, September 1991.

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

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

Farkas, D. H. and American Association for Clinical Chemistry: DNA simplified: the hitchhiker's guide to DNA. Washington, D.C., AACC Press, 1996. OCLC: 36923450.

Farrell, C. M., West, A. G., and Felsenfeld, G.: Conserved CTCF Insulator Elements Flank the Mouse and Human -Globin Loci. Molecular and Cellular Biology, 22(11):3820–3831, June 2002.

Fawcett, D. W.: On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. American Journal of Anatomy, 119(1):129–145, July 1966.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., and Dostie, J.: Chromatin conformation signatures of cellular differentiation. Genome Biology, 10(4):R37, 2009.

Fraser, P. and Bickmore, W.: Nuclear organization of the genome and the potential for gene regulation. Nature, 447(7143):413417, May 2007. PMID: 17522674.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A.: Formation of Chromosomal Domains by Loop Extrusion. Cell Reports, 15(9):2038–2049, May 2016.

Fudenberg, G. and Mirny, L. A.: Higher-order chromatin structure: bridging physics and biology. Current Opinion in Genetics & Development, 22(2):115–124, April 2012.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E., and Ruan, Y.: An oestrogen-receptor-a-bound human chromatin interactome. Nature, 462(7269):58–64, November 2009.

Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., Ellis, J., and Bazett-Jones, D. P.: Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. EMBO reports, 13(11):992–996, November 2012.

Gavrilov, A., Eivazova, E., Priozhkova, I., Lipinski, M., Razin, S., and Vassetzky, Y.: Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification. Methods in molecular biology (Clifton, N.J.), 567:171188, 2009. PMID: 19588093.

Gerchman, S. E. and Ramakrishnan, V.: Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. Proceedings of the National Academy of Sciences of the United States of America, 84(22):7802–7806, November 1987. PMID: 3479765.

Ghavi-Helm, Y., Michaut, M., Acker, J., Aude, J.-C., Thuriaux, P., Werner, M., and Soutourina, J.: Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription. Genes & Development, 22(14):1934–1947, July 2008.

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E.: Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell, 157(4):950–963, May 2014.

Goetze, S., Mateos-Langerak, J., Gierman, H. J., de Leeuw, W., Giromus, O., Indemans, M. H. G., Koster, J., Ondrej, V., Versteeg, R., and van Driel, R.: The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. Molecular and cellular biology, 27(12):44754487, June 2007. PMID: 17420274.

Goloborodko, A., Marko, J. F., and Mirny, L. A.: Chromosome Compaction by Active Loop Extrusion. Biophysical Journal, 110(10):2162–2168, May 2016.

Gürsoy, G., Xu, Y., Kenter, A., and Liang, J.: Spatial confinement is a major determinant of the folding landscape of human chromosomes. Nucleic Acids Research, 42(13):8223–8230, July 2014.

Gürsoy, G., Xu, Y., and Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. Conference proceedings: IEEE Engineering in Medicine and Biology Society. Annual Conference, 2014:3945–3948, 2014.

Haeusler, R. A., Pratt-Hyatt, M., Good, P. D., Gipson, T. A., and Engelke, D. R.: Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. Genes & Development, 22(16):2204–2214, August 2008.

Hagge, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forn, T.: Quantitative analysis of chromosome conformation capture assays (3c-qPCR). Nature Protocols, 2(7):1722–1733, 2007.

Hahnfeldt, P., Hearst, J. E., Brenner, D. J., Sachs, R. K., and Hlatky, L. R.: Polymer models for interphase chromosomes. Proceedings of the National Academy of Sciences of the United States of America, 90(16):78547858, August 1993. PMID: 8356094.

Hartigan, J. and Wong, M.: Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society, 28(1):100–108, 1976.

Hartl, D. L., Freifelder, D., and Snyder, L. A.: Basic genetics. The Jones and Bartlett series in biology. Boston, Jones and Bartlett Publishers, 1988.

Hediger, F., Neumann, F. R., Van Houwe, G., Dubrana, K., and Gasser, S. M.: Live imaging of telomeres: yKu and Sir proteins define redundant telomere-anchoring pathways in yeast. Current biology: CB, 12(24):2076–2089, December 2002.

Heermann, D. W., Jerabek, H., Liu, L., and Li, Y.: A model for the 3D chromatin architecture of pro and eukaryotes. Methods, 58(3):307–314, November 2012.

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P.: Genome-wide map of regulatory interactions in the human genome. Genome Research, 24(12):1905–1917, December 2014.

Helmink, B. A. and Sleckman, B. P.: The response to and repair of RAG-mediated DNA double-strand breaks. Annual review of immunology, 30:175202, 2012. PMID: 22224778.

Iwasaki, O., Tanaka, A., Tanizawa, H., Grewal, S. I. S., and Noma, K.-I.: Centromeric localization of dispersed Pol III genes in fission yeast. Molecular Biology of the Cell, 21(2):254–265, January 2010.

Iyer, B. and Arya, G.: Lattice animal model of chromosome organization. Physical Review E, 86(1), July 2012.

Jhunjhunwala, S., van Zelm, M. C., Peak, M. M., Cutchin, S., Riblet, R., van Dongen, J. J., Grosveld, F. G., Knoch, T. A., and Murre, C.: The 3d Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions. Cell, 133(2):265–279, April 2008.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A., and Ren, B.: A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature, 503(7475):290–294, November 2013.

Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C.: Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Research, 42(15):9553–9561, September 2014.

Junier, I., Dale, R. K., Hou, C., Kps, F., and Dean, A.: CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the -globin locus. Nucleic Acids Research, 40(16):7718–7727, September 2012.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L.: Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature biotechnology, 30(1):90–98, January 2012. PMID: 22198700.

Kang, H., Yoon, Y.-G., Thirumalai, D., and Hyeon, C.: Confinement-Induced Glassy Dynamics in a Model for Chromosome Organization. Physical Review Letters, 115(19):198102, November 2015.

Kreth, G., Finsterle, J., von Hase, J., Cremer, M., and Cremer, C.: Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. Biophysical journal, 86(5):2803–2812, May 2004. PMID: 15111398.

Lander, E. S.: Initial impact of the sequencing of the human genome. Nature, 470(7333):187–197, February 2011.

Langmead, B., T. C. P.-M. and Salzberg, S.: Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol., 10(3):R25, 2009.

Lehninger, A. L., Nelson, D. L., and Cox, M. M.: Lehninger principles of biochemistry. New York, W.H. Freeman, 4th ed edition, 2005.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M., and Ruan, Y.: Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell, 148(1-2):84–98, January 2012.

Liang, J., Zhang, J., and Chen, R.: Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. The Journal of Chemical Physics, 117(7):3511, 2002.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326(5950):289–293, October 2009.

Lin, M., Chen, R., and Liang, J.: Statistical geometry of lattice chain polymers with voids of defined shapes: sampling with strong constraints. The Journal of Chemical Physics, 128(8):084903, February 2008.

Lin, M., Lu, H.-M., Chen, R., and Liang, J.: Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. The Journal of Chemical Physics, 129(9):094101, September 2008.

Lin, M., Zhang, J., Lu, H.-M., Chen, R., and Liang, J.: Constrained proper sampling of conformations of transition state ensemble of protein folding. The Journal of Chemical Physics, 134(7):075103, February 2011.

Lisby, M., Mortensen, U. H., and Rothstein, R.: Colocalization of multiple DNA double-strand breaks at a single Rad52 repair centre. Nature Cell Biology, 5(6):572–577, June 2003.

Liu, J. S. and Chen, R.: Sequential Monte Carlo Methods for Dynamic Systems. Journal of the American Statistical Association, 93(443):1032, September 1998.

Lucas, J. S., Zhang, Y., Dudko, O. K., and Murre, C.: 3d trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. Cell, 158(2):339–352, July 2014.

Mateos-Langerak, J., Bohn, M., de Leeuw, W., Giromus, O., Manders, E. M. M., Verschure, P. J., Indemans, M. H. G., Gierman, H. J., Heermann, D. W., van Driel, R., and Goetze, S.: Spatially confined folding of chromatin in the interphase nucleus. Proceedings of the National Academy of Sciences of the United States of America, 106(10):38123817, March 2009. PMID: 19234129.

Mekhail, K. and Moazed, D.: The nuclear envelope in genome organization, expression and stability. Nature Reviews. Molecular Cell Biology, 11(5):317–328, May 2010.

Meluzzi, D. and Arya, G.: Recovering ensembles of chromatin conformations from contact probabilities. Nucleic Acids Research, 41(1):63–75, January 2013.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B.: Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Research, 23(2):270–280, February 2013.

Mirny, L. A.: The fractal globule as a model of chromatin architecture in the cell. Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 19(1):3751, January 2011. PMID: 21274616.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., and Grewal, S. I. S.: Cohesin-dependent globules and heterochromatin shape 3d genome architecture in S. pombe. Nature, 516(7531):432–435, December 2014.

Montefiori, L., Wuerffel, R., Roqueiro, D., Lajoie, B., Guo, C., Gerasimova, T., De, S., Wood, W., Becker, K. G., Dekker, J., Liang, J., Sen, R., and Kenter, A. L.: Extremely Long-Range Chromatin Loops Link Topological Domains to Facilitate a Diverse Antibody Repertoire. Cell Reports, 14(4):896–906, February 2016.

Naumova, N., Smith, E. M., Zhan, Y., and Dekker, J.: Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods (San Diego, Calif.), 58(3):192–203, November 2012.

Nmeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Pterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Lngst, G.: Initial Genomics of the Human Nucleolus. PLoS Genetics, 6(3):e1000889, March 2010.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blthgen, N., Dekker, J., and Heard, E.: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 485(7398):381–385, April 2012.

O'Toole, E. T., Winey, M., and McIntosh, J. R.: High-voltage electron tomography of spindle pole bodies and early mitotic spindles in the yeast Saccharomyces cerevisiae. Molecular Biology of the Cell, 10(6):2017–2031, June 1999.

Pajerowski, J. D., Dahl, K. N., Zhong, F. L., Sammak, P. J., and Discher, D. E.: Physical plasticity of the nucleus in stem cell differentiation. Proceedings of the National Academy of Sciences of the United States of America, 104(40):15619–15624, October 2007.

Phillips, J. E. and Corces, V. G.: CTCF: master weaver of the genome. Cell, 137(7):1194–1211, June 2009. PMID: 19563753.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J., and Corces, V. G.: Architectural protein subclasses shape 3d organization of genomes during lineage commitment. Cell, 153(6):1281–1295, June 2013.

Prokocimer, M., Davidovich, M., Nissim-Rafinia, M., Wiesel-Motiuk, N., Bar, D. Z., Barkan, R., Meshorer, E., and Gruenbaum, Y.: Nuclear lamins: key regulators of nuclear structure and activities. Journal of Cellular and Molecular Medicine, 13(6):1059–1085, June 2009.

Ptak, C., Aitchison, J. D., and Wozniak, R. W.: The multifunctional nuclear pore complex: a platform for controlling gene expression. Current Opinion in Cell Biology, 28:46–53, June 2014.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L.: A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7):1665–1680, December 2014.

Renda, M., B. I. B.-B. B. E. S. F. R. F. G. and Pedone, P.: Critical dna binding interactions of the insulator protein ctcf: A small number of zinc fingers mediate strong binding, and a single finger-dna interaction controls binding at imprinted loci. J Biol Chem, 282:33336–33345, 2007.

Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A., and Misteli, T.: Spatial proximity of translocation-prone gene loci in human lymphomas. Nature Genetics, 34(3):287–291, July 2003.

Rosa, A. and Everaers, R.: Structure and dynamics of interphase chromosomes. PLoS computational biology, 4(8):e1000153, 2008.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E.: Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature Biotechnology, 33(11):1165–1172, November 2015.

Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J., and Blanchette, M.: Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC bioinformatics, 12:414, 2011.

Sachs, R.K., v. d. E.-G. T. B. Y. H. H. J.: A random-Walk/Giant-Loop model for interphase chromosomes. Proceedings of the National Academy of Sciences, 92:2710–2714, 1995.

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L.: Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proceedings of the National Academy of Sciences of the United States of America, 112(47):E6456–6465, November 2015.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J.: The long-range interaction landscape of gene promoters. Nature, 489(7414):109–113, September 2012.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G.: Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell, 148(3):458–472, February 2012.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A.: Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature, 498(7453):236–240, June 2013.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P., and Regev, A.: Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature, 510(7505):363–369, June 2014.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W.: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation captureon-chip (4c). Nature Genetics, 38(11):1348–1354, November 2006.

Song, W., Dominska, M., Greenwell, P. W., and Petes, T. D.: Genome-wide high-resolution mapping of chromosome fragile sites in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America, 111(21):E2210–2218, May 2014.

Taddei, A. and Gasser, S. M.: Structure and function in the budding yeast nucleus. Genetics, 192(1):107–129, September 2012.

Taddei, A., Hediger, F., Neumann, F. R., and Gasser, S. M.: The function of nuclear architecture: a genetic approach. Annual Review of Genetics, 38:305–345, 2004.

Taddei, A., Schober, H., and Gasser, S. M.: The budding yeast nucleus. Cold Spring Harbor Perspectives in Biology, 2(8):a000612, August 2010.

Taddei, A., Van Houwe, G., Nagai, S., Erb, I., van Nimwegen, E., and Gasser, S. M.: The functional importance of telomere clustering: global changes in gene expression result from SIR factor dispersion. Genome Research, 19(4):611–625, April 2009.

Tark-Dame, M., van Driel, R., and Heermann, D. W.: Chromatin foldingfrom biology to polymer models and back. Journal of cell science, 124(Pt 6):839845, March 2011. PMID: 21378305.

Therizols, P., Duong, T., Dujon, B., Zimmer, C., and Fabre, E.: Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres. Proceedings of the National Academy of Sciences of the United States of America, 107(5):2025–2030, February 2010.

Tjong, H., Gong, K., Chen, L., and Alber, F.: Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome research, 22(7):1295–1305, July 2012. PMID: 22619363.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., Le Gros, M. A., Larabell, C. A., Chen, L., and Alber, F.: Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. Proceedings of the National Academy of Sciences of the United States of America, 113(12):E1663–1672, March 2016.

Tokuda, N., Terada, T. P., and Sasai, M.: Dynamical modeling of three-dimensional genome organization in interphase budding yeast. Biophysical Journal, 102(2):296–304, January 2012.

Trieu, T. and Cheng, J.: Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. Nucleic Acids Research, 42(7):e52, April 2014.

Vernimmen, D., Marques-Kranc, F., Sharpe, J. A., Sloane-Stanley, J. A., Wood, W. G., Wallace, H. A. C., Smith, A. J. H., and Higgs, D. R.: Chromosome looping at the human alpha-

globin locus is mediated via the major upstream regulatory element (HS -40). Blood, 114(19):4253–4260, November 2009.

Wang, S., Xu, J., and Zeng, J.: Inferential modeling of 3d chromatin structure. Nucleic Acids Research, 43(8):e54, April 2015.

Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C.-t., and Zhuang, X.: Spatial organization of chromatin domains and compartments in single chromosomes. Science (New York, N.Y.), 353(6299):598–602, August 2016.

Wedemann, G. and Langowski, J.: Computer simulation of the 30-nanometer chromatin fiber. Biophysical journal, 82(6):2847–2859, June 2002. PMID: 12023209.

Wong, H., Marie-Nelly, H., Herbert, S., Carrivain, P., Blanc, H., Koszul, R., Fabre, E., and Zimmer, C.: A predictive computational model of the dynamic 3d interphase yeast nucleus. Current biology: CB, 22(20):1881–1890, October 2012.

Yang, C. H., Lambie, E. J., Hardin, J., Craft, J., and Snyder, M.: Higher order structure is present in the yeast nucleus: autoantibody probes demonstrate that the nucleolus lies opposite the spindle pole body. Chromosoma, 98(2):123–128, August 1989.

Youngson, R. M.: Collins dictionary of human biology. London, Collins, 2006. OCLC: ocm63185739.

Zhang, B. and Wolynes, P. G.: Topology, structures, and energy landscapes of human chromosomes. Proceedings of the National Academy of Sciences of the United States of America, 112(19):6062–6067, May 2015.

Zhang, J., Dundas, J., Lin, M., Chen, R., Wang, W., and Liang, J.: Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. RNA, 15(12):2248–2263, December 2009.

Zhang, J., Chen, R., Tang, C., and Liang, J.: Origin of scaling behavior of protein packing density: A sequential Monte Carlo study of compact long chain polymers. The Journal of Chemical Physics, 118(13):6102, 2003.

Zhang, J., Chen, Y., Chen, R., and Liang, J.: Importance of chirality and reduced flexibility of protein side chains: a study with square and tetrahedral lattice models. The Journal of chemical physics, 121(1):592603, July 2004. PMID: 15260581.

Zhao, Z., Tavoosidana, G., Sjlinder, M., Gndr, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R.: Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature genetics, 38(11):13411347, November 2006. PMID: 17033624.

Zhou, V. W., Goren, A., and Bernstein, B. E.: Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics, 12(1):7–18, January 2011.

Zuleger, N., Robson, M. I., and Schirmer, E. C.: The nuclear envelope as a chromatin organizer. Nucleus (Austin, Tex.), 2(5):339–349, October 2011.

Zwaka, T. P. and Thomson, J. A.: Homologous recombination in human embryonic stem cells. Nature Biotechnology, 21(3):319–321, February 2003.

APPENDICES

Appendix A

The license to publish text from the publication Gürsoy, G., Xun, Y., Kenter, A., Liang, J.: Spatial confinement is a major determinant of folding landscape of human chromosomes. In Nucleic Acids Research, 42(13):8223-30, 2014.

Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Gamze Gursoy 511 W 36th street CHICAGO, IL 60609 United States Attn: Gamze Gursoy
Total	0.00 USD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL
FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be

necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.
9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.
10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:
v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B

The license to publish figures and tables from the publication Gürsoy, G., Xun, Y., Kenter, A., Liang, J.: Spatial confinement is a major determinant of folding landscape of human chromosomes. In Nucleic Acids Research, 42(13):8223-30, 2014.

Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Gamze Gursoy 511 W 36th street CHICAGO, IL 60609 United States Attn: Gamze Gursoy
Total	0.00 USD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL
FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be

Appendix C

The permission grant for the publication Gürsoy, G., Xu, Y., Liang, J.: Computational predictions of structures of multichromosomes of budding yeast. In Conf Proc IEEE Eng Med Biol Soc. 3945-8, 2014.

VITA

Gamze Gürsoy

Education

University of Illinois at Chicago, Chicago, IL
Ph.D., Bioinformatics, (Aug'08 - Dec'16)
Dissertation: Three-Dimensional Structures of Chromosomes in Eukaryotes: Novel Computational Approaches

Bogazici University, Istanbul, Turkey
B.S., Chemical Engineering, (Sep'02 - Mar'08)

Teaching

University of Illinois at Chicago, Chicago, IL

Teaching Assistant
Bioe596, *Datamining* (Summer 2009)
Bioe483, *Molecular Modeling in Bioinformatics* (Spring 2009)
Bioe240, *Modeling Physiological Data & Systems* (Fall 2008)

Research

University of Illinois at Chicago, Chicago, IL

Research Assistant (Aug 2008 - Nov 2016)

Northeastern Uni, NASA Center for Microgravity Materials Processing Boston, MA

Visiting Undergraduate Researcher (Aug 2007 - Nov 2007)

Awards

UIC Chancellor's Graduate Research Fellowship, (2013 – 2015)

UIC Graduate Student Council Travel Awards, (2012/2013/2014/2015/2016)

UIC Presenters Awards, (2012/2014)

NSF Fellow at 11th International Summer School on Biocomplexity from Gene to System , (2012)

Mathematical Sciences Research Institute full-level academic support for the workshop Algebraic, Geometric, and Combinatorial Methods for Optimization , (2010)

Fulbright Pre-doctoral Travel and Settling Award , (2008)

Turkish Petroleum Foundation Travel and Settling Award , (2008)

Northeastern University Honorarium , (2007)

Turkish Education Foundation Scholarship Undergraduate Scholarship (Full tuition & stipend, (2002–2007)

Publications

Peer-reviewed Publications

1. **Gürsoy, G**, Girardi, A., Liang, J. *Computational predictions of chromatin hotspots using n-Constrained Self-Avoiding Chromatin model.* In preparation.
2. **Gürsoy, G**, Xu, Y., Liang, J. *Spatial organization of budding yeast genome in cell nucleus and identification of specific chromatin interactions from multi-chromosome constrained chromatin model.* Submitted.
3. Xu, Y.*, **Gürsoy, G***, Kenter, A., Liang, J. *Constructing 3D chromatin ensembles and predicting functional interactions of α-globin locus from 5C data.* In review.

   * Contributed equally
4. **Gürsoy, G**, Liang, J. *3D Chromosome Structures from Energy Landscape.* Proc Natl Acad Sci U S A, In press.
5. **Gürsoy, G**, Terebus, A., Cao, Y., Liang J. *Mechanisms of Stochastic Focusing and Defocusing in Biological Reaction Networks: Insight from Accurate Chemical Master Equation (ACME) Solutions.* Conf Proc IEEE Eng Med Biol Soc, Accepted, 2016.
6. Liang, J., Cao, Y., **Gürsoy, G.**, Naveed, H., Terebus, A., Zhao, J.J *Multiscale modeling of cellular epigenetic states and tissue patterning: stochasticity in molecular network chromatin folding in cell nucleus, and cell-cell interactions in tissue patterning.* Crit Rev Biomed Eng., 43(4):323-46, 2015.
7. Camp S.M., Ceco E., Evenoski C.L., Danilov S.M., Zhou T., Chiang E.T., Moreno-Vinasco L., Mapes B., Zhao J., **Gürsoy G.**, Brown M.E., Adyshev D.M., Siddiqui S.S., Quijada H., Sammani S., Letsiou E., Saadat L., Yousef M., Wang T., Liang J., Garcia J.G. *Unique Toll-Like Receptor 4 Activation by NAMPT/PBEF Induces NFB Signaling and Inflammatory Lung Injury.* Sci Rep., 14;5:13135, 2015.
8. **Gürsoy, G**, Xu, Y., Liang, J. *Computational predictions of structures of multichromosomes of budding yeast.* Conf Proc IEEE Eng Med Biol Soc, 2014:3945-8, 2104.
9. **Gürsoy, G.***, Xu, Y.*, Kenter, A., Liang, J. *Spatial confinement is a major determinant of the folding landscape of human chromosomes.* Nucleic Acids Res., 42(13):8223-30, 2104.

   * Contributed equally
10. Albert, R., DasGupta, B., Hegde, R., Sivanathan, G.S., Gitter, A., **Gürsoy G.**, Paul, P., Sontag, E. *Computationally efficient measure of topological redundancy of biological and social networks.* Phys Rev E Stat Nonlin Soft Matter Phys., 84(3 Pt 2):036117, 2011.
11. Genchev, G.Z., Källberg, M., **Gürsoy, G.**, Mittal, A., Dubey, L., Perisic, O., Feng, G., Langlois, R., Lu, H. *Mechanical signaling on the single protein level studied using steered molecular dynamics.* Cell Biochem Biophys., 855(3):141-52, 2009.

12. Ismail, M.N., Fraiman, N.G., Callahan Jr., D.M., **Gürsoy, G.**, Viveiros, E., Ozkanat, O., Ji, Z., Willey, R.J., Warzywoda, J., Sacco Jr., A. *First unseeded hydrothermal synthesis of microporous vanadosilicate AM-6* Microporous and Mesoporous Materials, 120: 454-459, 2009.