Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Michael Everett Gehring Sauria Date

Three-Dimensional Chromatin Structure and Its Role in Cellular Function

By

Michael E. G. Sauria
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

_____
James Taylor, Ph.D
Advisor

_____ _____
Victor G. Corces, Ph.D
Committee Member

Lance Waller, Ph.D
Committee Member

_____ _____
William G. Kelly, Ph.D
Committee Member

Jeremy M. Boss, Ph.D
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D
Dean of the James T. Laney School of Graduate Studies

_____
Date

Three-Dimensional Chromatin Structure and Its Role in Cellular Function

By

Michael E. G. Sauria
M.S., Michigan State University, 2004
B.S., B.S., Carnegie Mellon University, 2001

Advisor: James Taylor, Ph.D

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

2014

Abstract

Three-Dimensional Chromatin Structure and Its Role in Cellular Function

By Michael E. G. Sauria

Cellular function is controlled by a complex interplay between genomic sequence and its surrounding context. A large force in establishing genomic context is the physical partitioning of the genome into defined neighborhoods that allows coordination of transcriptional activity, DNA interactions with RNA and proteins, and chemical modifications. Spatial organization is also involved in X-inactivation, cell fate determination, and senescence.

Recent high-throughput resequencing technologies have allowed investigation of chromatin architecture on a scale and resolution overcoming the previous limits of microscopy and inference at individual loci from less direct assessments. It is now possible to create a genome-wide map of DNA fragment interactions or investigate a protein-specific DNA interaction network. These approaches have revealed a complex hierarchical organization ranging from whole chromosomes down to short-range associations between adjacent features. Because these technologies generate large amounts of data representing a complex system, pose significant computational and analytical challenges.

We developed HiFive, a framework for analyzing HiC and 5C data, to address these challenges. HiFive allows handling of large amounts of data in an efficient manner and easy access to subsets of data for downstream analysis and plotting. We have also included an approximation approach to normalization that allows processing of data for a fraction of the computational cost and time. Compared to other available methodologies, HiFive performs as well or better across a variety of measures. To further validate the approaches used in HiFive, we also present downstream analyses locating significant structural signatures and analyzing gene spatial arrangements. We are able to increase sensitivity to detection of subdomain structures and their associated features. We also present a new approach to three-dimensional modeling that reveals a spatial partitioning of genes organized around transcriptional activity. Our results are consistent with our current understanding of chromatin architecture and suggest exciting possible avenues for future exploration.

Three-Dimensional Chromatin Structure and Its Role in Cellular Function

By

Michael E. G. Sauria
M.S., Michigan State University, 2004
B.S., B.S., Carnegie Mellon University, 2001

Advisor: James Taylor, Ph.D

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

2014

Acknowledgements

I would like to acknowledge the generous support of Emory University and Dr. Michael Zwick for nominating me for the Woodruff Fellowship. In addition, I would like to recognize the help of the PBEE faculty, especially Dr. Nicole Gerardo, in navigating challenges as they arose.

I would particularly like to thank my mentor and advisor Dr. James Taylor for allowing me to serve as his first student and to learn this process together. He treated me like a colleague and helped me to grow as a scientist and collaborator. I would also like to extend my thanks to the rest of the Taylor lab, particularly Drs. Jeremy Goecks and Olgert Denas, for their support, enlightening discussions, and help in maintaining sanity during the whole process.

Finally, I would like to extend my sincerest gratitude to my family and friends for their support and understanding as I dealt with the ups and downs of graduate school. Most importantly, I would like to single out my wife, Amber, for her endless patience in dealing with distance, delays, and many nights and weekends of working.

Table of Contents

Table of Figures

Table of Tables

Chapter 1 Introduction

Organization of the animal nucleus

Function of the nucleus

For close to a century and a half, we have known that the nucleus of the animal cell undergoes a complex set of operations associated with cell division (Schneider 1873; reviewed in Cremer & Cremer 2006). This process was further studied by Flemming (coiner of the terms chromatin and mitosis), who detailed the process of chromatid separation by the spindle apparatus (Flemming 1882). This was followed shortly by observations of the chromatin threads and the precise splitting of chromatid pairs into daughter cells (Heuser 1884, Waldeyer 1888). This work was subsequently synthesized into a theory of precise information division and transmission by Wilhelm Roux (1883). He proposed that the complex mechanism by which chromatids were paired and carefully separated between the newly formed cell pair spoke of more than an equal division of mass, but rather represented a carefully orchestrated system for maintaining the underlying composition of the chromosomes. In an insightful response to skeptics among his peers, Roux likened the apparent homogeneity of the chromatin to the view of a factory from far above in a balloon. It is necessary to infer the complex machinery given our knowledge of the complexity of the molecular function of the cell. Further, the need for such a complex process to divide the nucleus compared to the cytoplasm suggests a far more complex and less repetitive constitution (Cremer & Cremer 2006a). If he had

worked a century later, Roux's observation may have applied equally well to the state of the nucleus during the remainder of the cell cycle (interphase) and the necessary complexity of the arrangement of the chromatin during non-mitotic functions. It is this orchestration of chromatin structure and state required for cellular function that serves as the focus for this body of work.

Chromosome territories

During this same time period, the concept of the chromosome territory was proposed by two different scientists, Carl Rabl and Theodor Boveri (Boveri 1909, Rabl 1885). Rabl's theory consisted of two parts: First, that chromosome numbers remained constant across mitotic events; Second, chromosomes were anchored in place at the end of mitosis by the nuclear scaffold and remained in that position, each chromosome in its own space, throughout interphase until the next cell division. Further, he observed that chromosomes took on a polarized orientation with the regions interacting with the spindle (the centromeres) clustering on one side of the nucleus and the remaining chromatin stretched away from this focus, an arrangement now known as the "Rabl configuration". Boveri furthered this concept, and in the process coined the term "chromosome territory", by proposing that each chromosome has its own unique identity and remains intact through the cell's life and divisions (Boveri 1909). Based on observations about the persistence of chromosome positions from their disappearance at the beginning of interphase until their reappearance at the end of interphase, Boveri also theorized that chromosomes occupy individual and stable regions between cell divisions. He also noted that chromosome positions are not consistent from cell to cell or within a single cell across mitosis.

Although the concept of chromosome territories fell out of favor as the true length of DNA (~2 meters) became understood and the associated challenges with packing this into a nucleus with a 5-10 μm radius (Cremer & Cremer 2006b), work in the late 1970s provided new evidence for their existence (Stack *et al* 1977). This work was based on direct visualization of interphase chromosomes after fixation. Additional evidence mounted from a series of experiments using UV irradiation to induce local damage and visualization of the subsequent spatial organization of various markers of this damage (Cremer & Gray 1982, Cremer *et al* 1982, Hens *et al* 1983, Zorn *et al* 1979, Zorn *et al* 1976). Cell nuclei irradiated with tight-beam UV radiation showed shattered chromosomes with damage limited to a single or few chromosomes, demonstrating a lack of intermingling of chromosomes during interphase.

With the advent of resequencing and proximity-mediated ligation assays (PMLAs), there is ample evidence that chromosomes tend to occupy discrete and mutually exclusive regions within the nucleus (Dixon *et al* 2012, Hou *et al* 2012, Lieberman-Aiden *et al* 2009, Nagano *et al* 2013, Sexton *et al* 2012). Across a population of cells, associations in space are far more common between sequences occurring on the same chromosome and particularly on the same arm of the chromosome. This intra-chromosomal association bias is also seen within individual cells (Nagano *et al* 2013). Aside from specific loci, the only general inter-chromosomal associations that occur are between chromosome telomeres, as observed in *Drosophila* (Sexton *et al* 2012) and in mouse and human, along with inter-chromosomal association between centromeres (Imakaev *et al* 2012).

The fractal globule

Prior to the high-resolution genome wide view of chromatin nuclear arrangement afforded by PMLAs, Grosberg et al (1988, 1993) proposed a model of folding that they termed the "fractal globule". Under this model, small regions would be compacted along the chromosome resulting in a structure like a string of beads. This process is repeated, using the previously compacted state as the new base polymer until the entire mass is a single compact ball. The fractal globule has the added feature that because it does not contain knots (separate regions crossing each other and requiring a free end to separate them), the polymer can easily be unfolded and refolded either locally or globally, thereby facilitating local conformational changes as well as global separation for mitosis. Experimental association signal strength shows a relationship with inter-sequence distance that is consistent with type of conformation but not other proposed models (Lieberman-Aiden *et al* 2009). Because of its order of folding, the fractal globule also leads to behaviors consistent with the existence of chromosome territories.

Topological domains

Consistent with the folding principles of the fractal globules, a ubiquitous feature of eukaryotic genomes is the topological domain (TD, Dixon *et al* 2012), or topologically associated domain (TAD, Nora *et al* 2012). Observed in multiple cell types and species, these domains appear to underlie the spatial organization of the genome as the base unit of nuclear partitioning and are characterized by regions of highly self-interacting DNA and limited inter-region associations. The localization and associations of these discrete units within the nucleus appear to drive many different features of cell function including

transcriptional regulation, cellular differentiation, and quiescence. The boundaries of TDs appear to be conserved, not only through differentiation (Dixon et al 2012, Nora et al 2012), but also across species (Dixon et al 2012).

The boundaries of these domains show strong association with a variety of genomic features including transcription start sites (TSSs), histone marks, and insulator proteins (Dixon *et al* 2012, Hou *et al* 2012, Sexton *et al* 2012). There is a high density of TSSs occurring in close proximity to domain boundaries, particularly housekeeping genes, whereas tissue-specific genes do not show such clustering (Dixon *et al* 2012). Concurrent with these highly expressed genes, histone marks associated with active transcription are also associated with domain boundaries, including H3K4me3 and H3K36me3. In addition, the repressive histone mark H3K27me3 shows a striking drop-off when transitioning from regions associated with polycomb, a chromatin factor involved in maintaining inactive transcriptional states, to active regions (Sexton *et al* 2012). In addition, a class of proteins known as insulators shows a strong enrichment at domain boundaries. Across all eukaryotic species studied a large proportion of boundaries have the insulator CTCF binding in close proximity (Dixon *et al* 2012, Hou *et al* 2012, Sexton *et al* 2012). In Drosophila, which contains several additional known insulators, binding close to domain boundaries is also enriched for BEAF32, Chromator, and the accessory proteins CP190 and Mod (Mdg 4) (Hou *et al* 2012, Sexton *et al* 2012). Further, co-occurrence of BEAF32, CTCF, SuHW, and CP190 appears near boundaries far more than would be expected by chance (Hou *et al* 2012).

Organization of the nucleus

The eukaryotic nucleus exists as a collection of bodies contained within a double lipid bilayer, the nuclear envelope (NE), which is reinforced by a filamentous protein matrix (reviewed in Schooley *et al* 2012, Tapley & Starr 2013). Within the nucleoplasm, at least ten different types of bodies have been characterized, with functions ranging from transcriptional regulation to post-translational splicing and modifications (reviewed in Spector 2006). Much of the volume of the nucleus is occupied by chromatin in various states of condensation and this is reflected in the number of chromatin-associated structures observed in the nucleoplasm. A small number of nuclear structures do not appear to interact directly with chromatin, such as Cajal bodies (reviewed in Hebert 2013), PML bodies (reviewed in Rivera-Molina *et al* 2013), and paraspeckles (reviewed in Naganuma & Hirose 2013) and are primarily involved in RNA processing (both mRNA and noncoding RNA). The remaining structures within the nucleus are, at least in part, associated directly with the chromatin.

While a number of different attempts have been made to partition the genome based on genome annotation data, especially histone modifications (Filion *et al* 2010, Hoffman *et al* 2012, Wang *et al* 2012), partitioning can also be done on a structural basis. Using TDs as a base unit, the genome appears to be divided into at least 4 types of domains: 1) nucleolus associated domains; 2) polycomb bodies; 3) lamina associated domains; 4) RNA polymerase II (PolII) dependent transcriptionally active regions. While more types of domains are likely to exist, these are the clearly definable units supported by current evidence.

The nucleolus

The nucleolus represents the largest sub-nuclear body and consists of a heterogeneous collection of genes. The most distinctive set of genomic sequences are called the nucleolus organizing regions (NORs). These regions contain large numbers of repeated ribosomal RNA (rRNA) coding sequence in tandem and palindromic arrangements (Caburet *et al* 2005). In addition to NORs, nucleoli have been show to contain a variety of elements from across nearly all chromosomes (Nemeth *et al* 2010, van Koningsbruggen *et al* 2010). These additional regions, called nucleoli associated domains (NADs), contain a high proportion of satellite repeats, members of the zinc-finger olfactory and immunoglobulin gene families, 5S rRNA genes, and transfer RNA (tRNA) genes. Although most of the DNA contained in NADs is inactive, the NORs represent a mix of active and inactive genes (Caburet *et al* 2005). Transcriptional status appears to be regulated by DNA methylation and methylation of the ninth, twentieth, and twenty-seventh lysine residues of the third histone subunit (H3K9, H3K20, and H3K27, respectively). The other key features of the nucleolus are its association with RNA polymerase I, and involvement in ribosome biogenesis. Upon cell division NADs rearrange. Most of them return to nucleoli (not necessarily to the same groupings), though some shift to associating with the nuclear periphery (van Koningsbruggen *et al* 2010), suggesting that nucleoli association is stochastic in nature. CTCF, an important protein in chromatin architecture, has been shown to accumulate at the nucleolar periphery in close spatial association with the the protein nucleophosmin, a nucleic acid-associating protein involved in ribosome biogenesis in the nucleolus. Further, it appears

that an artificial CTCF binding site is sufficient to localize a GFP-expressing transgene to the nucleolar periphery the majority of the time (Yusufzai *et al* 2004).

Polycomb bodies

Polycomb bodies (PcBs) are foci found in cells and represent collections of sequences known as polycomb group response elements (PREs) that are capable of binding polycomb group (PcG) proteins. The purpose of PcBs appears to be the silencing of Hox and other developmentally important genes through a combination of aggregation and histone modification, especially H3K27me3 (Schuettengruber *et al* 2009, Schwartz *et al* 2006). PcB foci form from sequences separated by megabase (Mb) stretches of sequence (Bantignies *et al* 2011, Tolhuis *et al* 2011) and while some inter-chromosome associations occur, most are formed through intra-chromosomal interactions (Tolhuis *et al* 2011). Evidence suggests that this limitation is topological, such that rearrangement via inversion from one chromosome arm to the other causes a switch in foci association (Tolhuis *et al* 2011). Further, the size of PcBs is limited though a feedback process of SUMOylation and SUMO-deconjugation (covalent addition and removal of the protein SUMO) via the protein Velo (Gonzalez *et al* 2014). When SUMO is removed, massive PcBs form, whereas in the absence of velo, SUMO accumulates on the PcG protein Pc2, resulting in the failure of foci to form. Interestingly, Pc2 also appears to play a role in relocating genes from PcBs to a more transcriptionally permissive region of the nucleus (Yang *et al* 2011). This occurs by preferential binding to one of two non-coding RNAs (ncRNAs) depending on Pc2's methylation status. Although there is no direct evidence yet, indirect evidence suggests that PcBs may form stochastically, like NADs, associating with nearby PREs solely due to proximity.

Insulators, such as CTCF have a more complicated role in PcB formation and maintenance. There is limited evidence that PcB formation is dependent solely on the presence of insulators and that formation occurs independent of the PRE sequences (Li *et al* 2011). Deletion of the sequence associated with insulating function is sufficient to eliminate long-range association between two PcG targets, whereas deletion of PCEs has no effect on their association. Other work suggests that the process of foci formation is dependent on the PcG protein Eed (Denholtz *et al* 2013). Null mutants for Eed show a decrease in PcB association that is not attributable to disruption of the cell cycle. What is clear is that dCTCF (the *Drosophila* homolog of CTCF) and CP190 are necessary for proper maintenance of H3K27me3 levels and boundaries (Bartkuhn *et al* 2009, Van Bortle *et al* 2012). Disruption of either results in fluctuating levels of H3K27me3 across the boundary insulator sites (more often down than up) and a general decrease in H3K27 methylation across the entire polycomb region. This is likely due to a diffusion of the chromatin away from the methyl-transferase recruitment sites, preventing further condensation and histone mark maintenance.

Lamina associated domains

The nuclear lamina (NL) is a protein meshwork within the nuclear envelope that is used to bind regions of chromatin known as lamina-associated domains (LADs) and appears to function in general structural organization and possibly terminal differentiation. Association with the NL is mediated by the proteins lamin A, B1, B2, and C (Shimi *et al* 2008). Lamin A and lamin B form separate but interacting meshes on the nuclear envelope, both of which interact with chromatin. Lamin A is also found in the nucleolar periphery in addition to being found in the NL (Kind & van Steensel 2014).

Chromatin associating with lamins is characterized by a scarcity of genes and shows increased repressive marks, particularly H3K9me2. LAD boundaries are typified by CTCF binding, CpG islands, and outwardly oriented promoters (Guelen *et al* 2008). LAD association with the NL is stochastic, with an independent set of LADs associating with the NL after each cell cycle (Kind *et al* 2013). This process appears to be mediated by competitive binding of LADs by lamin A, B, and BAF (Kind & van Steensel 2014). Upon association with lamin A and BAF, LADs can associate with either the NL or the nucleolar periphery. The effects of knocking down either protein level further support this tug of war interaction between lamin A, lamin B, and BAF. When lamin A, BAF, or both are depleted, LAD associations with lamin B increase (Kind & van Steensel 2014). Conversely, knocking down lamin B1 results in increased chromosomal territory volumes, relocation of H3K27me3 marks from the NL to the nucleoplasm, blebbing of the nuclear envelope, and an increase in euchromatin association with these lamin A-rich blebs (Camps *et al* 2014, Shimi *et al* 2008). Additionally, there is an increase in active histone marks in this euchromatin that is concurrent with a decrease in RNA production, suggesting that promoter proximal stalling may be occurring. LAD association with the NL through lamin B is also mediated by a positive feedback loop with H3K9me2 (Kind *et al* 2013). LADs associated with the NL have higher levels of H3K9me2 but depletion of G9a, an H3K9 methyl-transferase, results in dissociation from the NL. There is also evidence from the disease Hutchinson-Gilford progeria syndrome that proper control of LAD association with the NL is necessary to maintain global chromatin architecture (McCord *et al* 2013). A deletion in the lamin A gene results in a mutant form called progerin that accumulates in the NL and has an impaired ability to bind chromatin. This

appears to interfere with LAD association with both lamin A and B, resulting in nuclear blebbing (a deformation of the nuclear membrane surface forming bubbles), decreases in H3K27me3, an increase in gene activity, and after numerous cell cycles a general loss of ordered chromatin compartmentalization. Working in the other direction, there is a theory that increases in the stochastic associations with lamin B drive terminal differentiation (Aranda-Anzaldo *et al* 2014). In this scenario, nuclear-matrix (NM) proteins bind the edges of LADs and form bridges. During the cell cycle, phosphorylation causes these protein bridges to disassociate, although the NM proteins remain bound to the LADs. As more LADs associate during differentiation, LADs are spatially closer (because of a higher density of them), allowing bridges to be formed over one or two NM proteins. With enough of these short bridges, phosphorylation is insufficient to overcome their binding strength, preventing the chromatin structure from breaking down and thus preventing mitosis to continue.

RNA polymerase II-dependent transcriptional foci

Regions of transcriptionally active polymerase II (PolII) dependent genes make up a significant portion of the nucleoplasmic genome and are defined by a more complex substructure than other domains (Ghamari *et al* 2013, Sofueva *et al* 2013). While their boundaries are marked by active TSSs and CTCF (Dixon *et al* 2012, Hou *et al* 2012), active domains also contain a large number of overlapping CTCF and cohesin sites that subdivide the domain into smaller sub-regions that show a degree of insulation from each other (Sofueva *et al* 2013). Unlike TD boundaries, this substructure is dependent on cohesin and is disrupted when cohesin is knocked down (Seitan *et al* 2013, Sofueva *et al* 2013). Within the nucleus, active regions are organized into mobile foci of high

transcriptional activity called transcriptional factories (Ghamari *et al* 2013). These associations appear organized into overlapping domains of transcription initiation and elongation, though they are not dependent on PolII occupancy for maintained association. In embryonic stem cells (ES cells or ESCs), groups of genes that are multiply bound by the pluripotency factors Nanog, Sox2, and Oct4 show significant association, suggesting that certain developmentally important groups of genes are assembled into special transcriptional factories (de Wit *et al* 2013, Denholtz *et al* 2013). A similar phenomenon occurs in plasma cells, where three active immunoglobulin genes from separate chromosomes show co-localization (Park *et al* 2014).

Within actively transcribed domains, multiple forms of regulation occur to orchestrate not only silencing of certain genes but also coordinating enhancer usage. Of critical importance to this substructure and especially enhancer-promoter (EP) interactions are the protein complex cohesin and mediator. Both show strong overlap with pluripotency factors and are associated with EP interactions in the absence of CTCF (He *et al* 2014, Phillips-Cremins *et al* 2013). Knocking down RAD21, a subunit of cohesin, results in a deregulation of genes, both up and down, as well as a disruption to domain sub-structure, but not TD structure (Seitan *et al* 2013, Sofueva *et al* 2013). In addition to cohesin, p300 and CBP also show occupation in loops with enhancer activity, binding with Nanog (Fang *et al* 2014). EP interactions appear to occur transiently, scanning within regions bounded by CTCF-cohesin pairs for interaction partners (Hughes *et al* 2014). The EP interactions that do occur happen regardless of PolII occupancy and are unchanged by external stimulus, appearing poised to activate the necessary genes (Berlivet *et al* 2013, Jin *et al* 2013, Palstra *et al* 2008). This suggests that other regulatory

mechanisms are fine-tuning the transcriptional activity of genes. One candidate is alternate forms of histone subunits (Chen *et al* 2013). H3.3 marks active enhancers, impairing higher-order chromatin folding and allowing increased transcription, whereas H2A.Z promotes chromatin compaction and repression of transcription. Another form of regulation that has been observed is transcription initiation RNAs (tiRNAs) (Taft *et al* 2011). These short 18 base pair (bp) sequences occur just downstream of the TSS and are highly associated with genes with proximal CTCF binding. Increases in a gene's tiRNA causes a reduction in CTCF binding which leads to a decrease in nucleosome organization, density, and H3K4 methylation. Increases in tiRNA have the opposite effect, decreasing CTCF binding. The mechanism by which this occurs is still unclear.

Assaying chromatin structure

Microscopic detection of chromatin structures

Identification of the structure of chromatin was greatly advanced by microscopic approaches. The classic "beads on a string" nucleosome structure (the 10 nm DNA strand) was first identified by a combination of X-ray diffraction and light microscopy (Kornberg 1974, Olins & Olins 1974). Thirty years later electron microscopy is being used to lend support to the still questionable notion of an in situ 30 nm DNA strand (Robinson *et al* 2006). While transmission electron, scanning electron, and atomic force microscopy have been invaluable in exploring the structure of condensed chromatin and topologies of individual molecules, interphase chromosomes and their complex relationships have still posed a major hurdle (reviewed in Daban 2011, Wanner & Schroeder-Reiter 2008). One of the challenges is the difficulty in determining three-dimensional structure beyond a single molecule. Additionally, the fixation process necessitates viewing the chromatin in a very different context than its native environment. Finally, the chromatin viewed is anonymous, meaning specific loci cannot be identified.

To address the limitations of direct visualization, a probe-based approach was developed called fluorescent in situ hybridization (FISH) (reviewed in Trask 1991). Sequence-specific probes are labeled with different fluorophores, allowing observation of relative physical positions of specific loci and localization within the nucleus. When coupled with confocal scanning laser microscopy, FISH allows good resolution between two probes (~100 nm) for positioning of multiple loci in three-dimensional space (reviewed in Solovei *et al* 2002, Tsuchiya 2011, Walter *et al* 2006). Probes for FISH are

long, on the order of hundreds of kilobases and multiplexing of distinguishable fluorophores has not yet broken 100. Although 3D-FISH serves as the gold standard in validating chromatin structural data, it is still insufficient for localization of targets less than tens to hundreds of base pairs long and can only examine a few unique targets at a time.

Chromatin conformation capture

To overcome the trade-offs between resolution, expense, and ability to target specific loci, Dekker et al (Dekker *et al* 2002) devised a PCR-based approach to interrogating pairs of loci to determine their relative association frequency within the nucleus. This method, which he and his colleagues called chromosome conformation capture (3C), has formed the foundation for a variety of approaches that have greatly advanced our understanding of the architecture of the nucleus and allowed us to pose completely new types of questions. Underlying this technique is the simple principle underpinning all PMLAs, the fact that two strands of DNA whose ends are close together in solution have a higher probability of undergoing ligation in the presence of a ligase than strands whose ends are further apart.

To understand 3C data, it is crucial to begin with a firm grasp of the underlying steps involved (Figure 1.1). The first step is to cross-link the genome in the condition of interest, covalently bonding together proteins and DNA using a fixative such as formaldehyde, ensuring that molecules in close proximity will remain so during subsequent steps. Cells are lysed, freeing the cross-linked DNA, which is then fragmented using a restriction enzyme (RE). The choice of RE determines not only the specific cut points within the genome but also the maximum resolution of the assay based

on the frequency of RE recognition sites. At this stage some fragments of DNA bound by RE sites are bound to other fragments via covalently linked protein bridges. The solution is highly diluted, decreasing the chances of random interactions between unbound fragments. A ligase is added, resulting in fragment ends being joined. Because of the dilution, pairs of cross-linked fragments have a chance of being ligated together, whereas free fragments are highly biased towards self-ligating, creating circles of DNA. After reversing the cross-linking, a pair of sequence-specific primers designed to anneal to two different RE fragments just inside of the RE sites are added and, if present, the hybrid sequence from the joining of the two targeted fragments is amplified via PCR. In order to normalize the resulting quantities of PCR products across different primer pairs, values are divided by the quantity of the corresponding PCR product that is produced from an equal-part mixture of all possible ligation products. The intensity of 3C signal is a function of two components: the rate of fragment cross-linking and the frequency of inter-fragment ligation. Dekker *et al* (2002) demonstrated that the lack of either step results in undetectable PCR product. Further, the distance between fragments within the genomic sequence is inversely correlated with the 3C signal in biological samples but not in equimolar mixtures of fragments, demonstrating that spatial proximity (due to tethering by intervening sequence) leads to an increase in fragment cross-linking and subsequent fragment ligation.

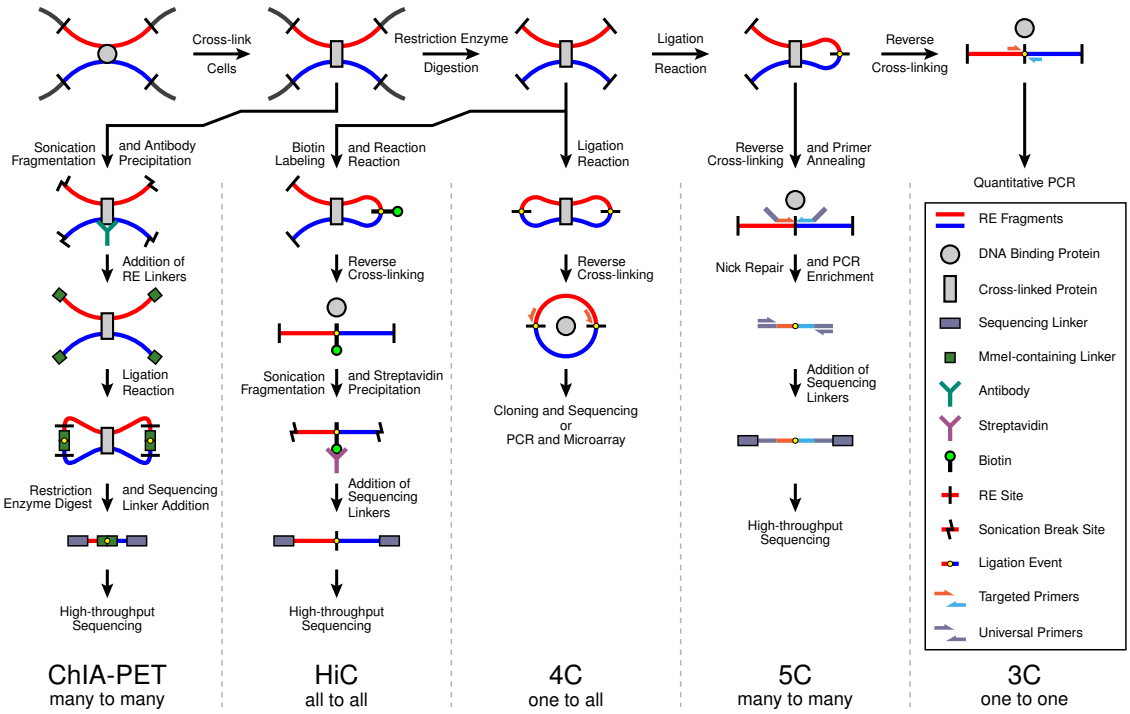

Figure 1.1 3C-based DNA association assays.
3C is a proximity-mediated ligation assay that underlies many sequence-based chromatin conformation assays. Spatially adjacent sequences are covalently linked together via intermediate proteins. DNA is fragmented either by restriction enzyme digest or fragmentation by sonication. Sequences are ligated together and then enriched by targeted primers and PCR amplification, pull-down by antibody or binding of biotin, or both. Finally hybrid sequences are analyzed by quantitative PCR, microarray hybridization, or sequencing.

While 3C has opened new avenues of investigation and is fairly accessible to any laboratory, it is majorly limited in terms of the scale and time involved. Because 3C is a relative measure of association frequency, a variety of loci must be tested in order to make any conclusions. In addition, for every test reaction a corresponding control reaction is needed. The advent of quantitative PCR has greatly helped with the precision of measurements, but scale remains the main challenge of this approach.

3C-derived approaches

In order to take advantage of the custom printed microarray and next generation sequencing (NGS) technologies while simultaneously expanding the scale of 3C, several variations of PMLAs have been developed. Although all of them still rely on some variation of producing 3C-type hybrid RE fragments, each exploits some aspect of primer design or chemical modification in order to allow multiplexing of target quantification. These extensions include circular chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), chromatin interaction analysis paired-end tag sequencing (ChIA-PET), and HiC (Dostie *et al* 2006, Fullwood *et al* 2009, Lieberman-Aiden *et al* 2009, Simonis *et al* 2006, Zhao *et al* 2006).

In order to avoid designing primers for both fragments involved in an association, 4C targets fragment pairs in which both loose ends have ligated together (Simonis *et al* 2006, Zhao *et al* 2006). This allows primers targeting the sequences just inside of the RE sites for a single "anchor" fragment to be used to amplify sequence from any other RE fragment that is ligated between the anchor fragment ends. These sequences can then be quantified either by hybridization to custom printed microarrays or high-throughput sequencing. While this approach allows interrogation of many possible interactions simultaneously there is no suitable control, meaning that the technique is limited to comparison studies between conditions. 4C has been used to shed light on a diverse range of topics, including X chromosome inactivation, maternally imprinted genes and polycomb body formation (Splinter *et al* 2011, Tolhuis *et al* 2011, Zhao *et al* 2006).

A different approach to up-scaling 3C known as 5C involves designing a set of primers much like those used in 3C but that are capable of multiplexing and have

universal primer sequences extending from them (Dostie *et al* 2006). After ligation of cross-linked RE fragments, the primers are annealed to the hybrid fragments such that when two targeted sequences have been ligated the primers abut each other with no gap. A second round of ligation is performed, resulting in a joining of these primers. PCR amplification using the universal primers corresponding to those attached to the designed primers allows selective amplification of fragments pairs for which there is a forward and reverse primer. The advantage of this approach is that an unbiased set of associations can be queried with a specific region or regions resulting in a survey of all forward by reverse primer target combinations. This allows high sequencing depth of a controllable number of interactions. This approach is limited by the ability to query fragment pairs only with complementary primer orientations. Interpretation of 5C results requires either a 3C-like control mixture of possible ligation products or a more sophisticated computational correction approach (Dostie *et al* 2006, Nora *et al* 2012, Phillips-Cremins *et al* 2013).

A genome-wide variant of 3C known as ChIA-PET uses antibody and biotin-streptavidin precipitations to specifically enrich cross-linked RE fragments associated with a specific protein target (Fullwood *et al* 2009). Initially, a protein of interest is chosen and after cross-linking and fragmentation of the genome by sonication, DNA cross-linked to the target protein is precipitated using an antibody targeting that protein. Fragment ends are ligated to linkers containing an MmeI RE site and a biotin label and then subsequently ligated again, joining nearby linkers. The cross-linking is reversed, releasing hybrid fragments from bound proteins and an MmeI digestion is performed. This RE has the characteristic of cutting 20 bp away from its recognition site. The result is a fragment containing a short stretch of DNA from the first fragment, a linker, and a

short stretch of DNA from the second fragment. These fragments are precipitated using streptavidin beads to bind the biotin label and the isolated fragments are sequenced and mapped back to the genome. This approach expands the scope of association mapping to a whole genome perspective but is limited in focus only to queried proteins and their associated DNA sequences.

For an unbiased genome-wide perspective, Lieberman-Aiden *et al* (2009) developed HiC, a non-PCR based expansion of 3C. Unlike previously described methods, HiC enriches hybrid RE fragments by incorporating biotin at the time of ligation and subsequently only pulling down fragments that have undergone ligation. In order to restrict the fragments that are sequenced and ensure that sequencing covers the ligation junction, the sample is fragmented via sonication prior to precipitation with streptavidin and isolated fragments are then size selected, usually for ~500 bp. This approach allows the detection of all combinations of both RE fragment ends across the whole genome, assuming that the sequence close to the RE site is uniquely mappable. This allows production of a fairly complete map of any given region and can still maintain fairly good resolution if a high enough sequencing depth is achieved. However, there is no particularly good control that can be run short of direct condition comparison. Instead, analysis of HiC data remains an almost entirely computational challenge.

PMLA data analysis

As described above, each variation of the 3C assay poses a different set of interpretation challenges. In small-scale approaches such as 3C and 4C, attempts can be made to run corresponding negative controls. However given the fact that cellular contaminants, differing reagent or template concentrations, and handling conditions can

have large influences on the efficiency of a PCR reaction, external controls may prove difficult to directly compare to samples (Farrell 1997, Mallet 2000). Further, as the complexity of the experiment increases from the number of possible interactions being queried, the chances of a significant looking sample-control comparison arising from stochasticity of experimental process increases dramatically. Thus external controls seem all but useless for large-scale analyses such as 5C or HiC. Instead, we propose that given the amount of information produced for each sample, a computational approach that relies solely on within-sample relationships is the most promising approach to data interpretation.

Chapter 2 Determining chromatin conformation from interaction frequency data using a probabilistic modeling approach*

Introduction

Although the vast majority of the human genome was sequenced more than a decade ago, it is clear that sequence alone is insufficient to explain the complex gene and RNA regulatory patterns seen across time and cell type in eukaryotes. The context surrounding sequences—whether from combinations of DNA binding transcription factors (TFs) (Arnone & Davidson 1997, He *et al* 2011, Zinzen *et al* 2009), methylation of the DNA itself (Cantone & Fisher 2013, Varriale 2014), or local histone modifications (Cantone & Fisher 2013, Kimura 2013)—is integral to how the cell utilizes each sequence element. Although we have known about the potential roles that sequentially distant but spatially proximal sequences and their binding and epigenetic contexts play in regulating expression and function, it has only been over the past decade that new sequencing-based techniques have enabled high-throughput analysis of higher-order structures of chromatin and investigation into how these structures interact amongst themselves and with other genomic elements, to influence cellular function.

Several different methods for assessing chromatin interactions have been devised, all based on the sequencing of hybrid fragments of DNA created preferentially between

* Parts of this chapter have been adapted from an article under review at *Genome Research*.

spatially close sequences. These approaches include ChIA-Pet (Fullwood *et al* 2010), tethered chromosome capture (Kalhor *et al* 2012), and the chromatin conformation capture technologies of 3C, 4C, 5C, and HiC (Dekker *et al* 2002, Dostie *et al* 2006, Lieberman-Aiden *et al* 2009, Zhao *et al* 2006). While these assays have allowed a rapid expansion of our understanding of the nature of genome structure, they also have presented some formidable challenges, including handling experimental biases and computational resources.

In both HiC and 5C, systematic biases resulting from the nature of the assays have been observed (van Berkum & Dekker 2009, Yaffe & Tanay 2011), resulting in differential representation of sequences in the resulting datasets leading to enrichment or depletion of sequence associations unrelated to biological causes. While analyses at a larger scale are not dramatically affected by these biases due to the large number of data points being averaged over, higher-resolution approaches must first address these challenges. Several analysis methods have been described in the literature and applied to correcting biases in HiC (Hu *et al* 2013, Hu *et al* 2012, Imakaev *et al* 2012, Jin *et al* 2013, Yaffe & Tanay 2011) and 5C data (Naumova *et al* 2013, Nora *et al* 2012, Phillips-Cremins *et al* 2013, Rousseau *et al* 2011, Sanyal *et al* 2012). There is still, however, room for improving our ability to remove this systematic noise from the data and resolve finer-scale features.

A second challenge posed by data from these types of assays is one of resources. Unlike other next-generation sequencing assays where even single-base resolution is limited to a few billion data points, these assays assess pairwise combinations, potentially increasing the size of the dataset by several orders of magnitude. For a three billion base-

pair genome cut with a six-base RE, the number of potential interaction pairs is more than half a trillion. Even allowing that the vast majority of those interactions will be absent from the sequencing data, the amount of information that needs to be handled and the complexity of normalizing these data still pose a major computational hurdle, especially for investigators without access to substantial computational resources.

Here we present HiFive, an analysis method developed for handling both HiC and 5C data using a combination of empirically determined and probabilistic signal modeling. We demonstrate that HiFive allows memory- and computationally-efficient HiC and 5C data handling and normalization while retaining high-resolution data for downstream analyses of interaction signals, making fine-scale chromatin structural analysis accessible to a wider range of investigators.

Material and Methods

Mapping of interaction data

5C datasets were downloaded from the Gene Expression Omnibus (GEO) archive [1] (Barrett *et al* 2009) and split into paired-end fastq files using Fastq-Dump version 2.1.18 from the SRA toolkit [2] (Wheeler *et al* 2008). Read ends were mapped independently to probe sequences, also obtained from GEO, using the alignment program Bowtie version 0.12.7 [3] and the mapping settings "--phred33-quals --tryhard -m1 -5 3 -3 2 -v 2" (Langmead *et al* 2009). The number of occurrences of each possible probe pair was tallied from all reads for which both ends were successfully mapped. Pairs for which the probes targeted the same strand were discarded as artifacts.

HiC data were obtained from GEO and split using Fastq-Dump. Read ends were mapped independently to the mouse genome build 9 using Bowtie2 version 2.1.0 [4] (Langmead & Salzberg 2012) coupled with HiCLib's iterative mapping function [5] (Imakaev *et al* 2012). Briefly, reads were truncated from the 3-prime (3′) end to the first 20 or 22 bp for reads of total 36 or 50 bp, respectively. Reads were mapped using Bowtie2 and uniquely mapping reads were kept. All multiply mapped reads were extended by 4 bp and mapped again. This process was repeated until all reads were uniquely mapped or the total read length had been attempted. All mapping was done using the Bowtie2 flag "--very-sensitive".

Only reads for which both ends corresponded to fragments bounded by RE sites were used. Reads mapping outside of the first and last fragment were excluded as the

fragment size associated with these boundary sequences is impossible to determine given the difficulty in assembling complete genomic sequence at chromosome ends.

All data described in this study were obtained from publicly available sources as detailed in Table 2.1.

Table 2.1 List of public datasets used.

Sample	Replicate	Cell Type	Data Type	Reference	GEO ID
Male ES E14	1	Male mES	5C	Nora et al 2012	GSM873934
Male ES E14	2	Male mES	5C	Nora et al 2012	GSM873935
mESC HindIII	1	mES cell line (J1)	HiC	Dixon et al 2012	GSM862720
mESC HindIII	2	mES cell line (J1)	HiC	Dixon et al 2012	GSM862721

Software Implementation

HiFive is based on a hierarchy of data modules stored in HDF5-formatted files (a management structure for handling complex and large sets of data) using the package h5py for easy, compact, and fast access that may be shared across experiments and analyses as shown in Figure 2.1 [6]. All aspects of the software are written in Python2 [7] and make use of the Cython [8] and NumPy [9] packages to accelerate computationally intensive operations. This gives HiFive speed similar to C-based code with the human readability and ease of use of Python. In addition, certain scripts and functions support parallelization using the package mpi4py [10] to utilize the Message Passing Interface (MPI), greatly increasing the scalability of analysis. This allows computationally

intensive processes to be automatically split across multiple processors on a cluster or multicore computer. The overall goal of this implementation is to reduce storage, memory, and processing power requirements without sacrificing analytical power.

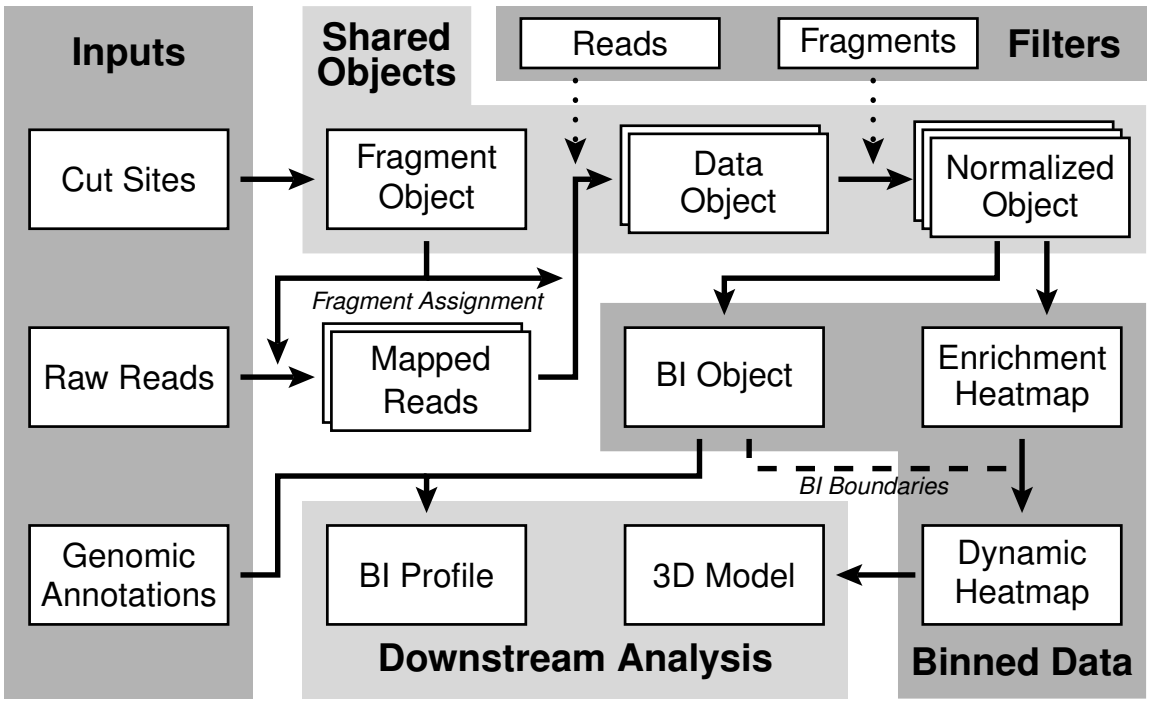

Figure 2.1 The software architecture of HiFive.
Self-contained units of information are denoted by boxes, with solid arrows denoting dependencies for object creation. The split line marked 'fragment assignment' depends on the type of data being handled and acts upstream or downstream of the mapped reads objects for 5C and HiC, respectively. Dotted lines denote filters limiting information passed from one object to the next. The dashed line denotes an optional input.

HiFive creates a base object for each data type describing the partitioning of the genome by a specified RE. For 5C data, the fragment object contains only information for the regions queried, importing information from a BED-format file containing RE fragment boundaries and which primers target them. HiC fragment-end (fend) objects are created from a HiCPipe-compatible fend text file (Yaffe & Tanay 2011). A fend is

defined as the sequence bounded by an RE site and the midpoint of the RE fragment. Fragment and fend objects also index genomic segments by region and chromosome, respectively, for fast access to subsets of data.

All fragment- or fend-pair data are stored in HDF5-formatted files with a fragment or fend lookup table for efficient data retrieval. Data for 5C experiments are loadable either from text files containing fragment pairs and counts or directly from BAM-formatted alignment files. Counts are separated into intra-chromosomal (*cis*) and inter-chromosomal (*trans*) interactions and are indexed by interaction fragments. HiC data are handled similarly, although reads can be loaded from MAT-formatted text files normally used with HiCPipe [11] (Yaffe & Tanay 2011), as well as BAM files and text files containing paired-end coordinate data. Because ligation fragment-pairs are randomly sheared, specific end coordinates are discarded after paired-reads are assigned to fragments, although strand information is retained and used to determine to which fend reads are assigned.

Data Filtering

In both experimental approaches, data are filtered in a two-stage process. After mapping is completed and read pairs have been assigned to fragments or fends for 5C and HiC respectively, a set of filters is applied to the data during data object creation (Figure 2.2 andFigure 2.3). Once a HiC or 5C object is created and linked to a data object, a second filter is applied as described below.

Because of the experimental design, 5C produces very few read pairs that, once successfully mapped, are not able to be included in the data object. This is a feature of the primer design such that, in theory, only opposite strand primers that are joined

subsequent to annealing to specific fragment targets should be amplified and sequenced. Next, reads are only mapped to probe sequences thus eliminating reads corresponding to non-queried sequences. Due to the alternating primer design, sites that the RE fails to cut are also excluded, as only one read of the pair will map to a probe sequence. Finally, read pairs mapping to same-orientation probes are removed as they represent ligation or sequencing artifacts and make up a very small portion of the mapped read pool.

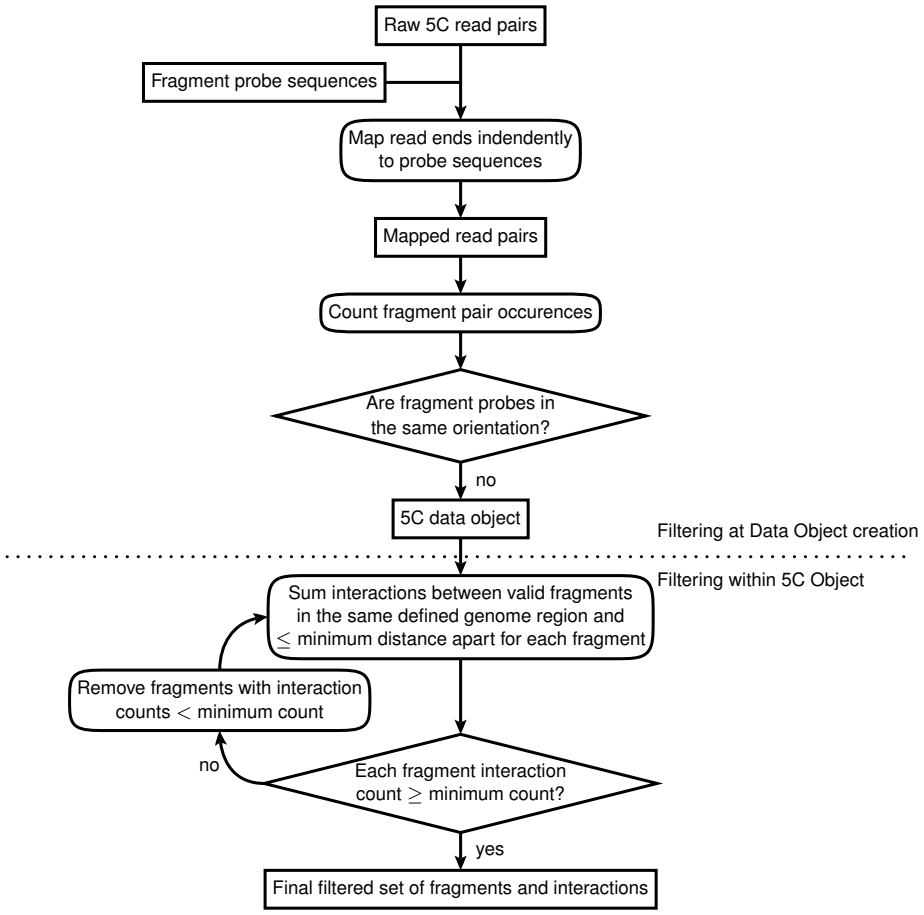

Figure 2.2 5C read filtering scheme.

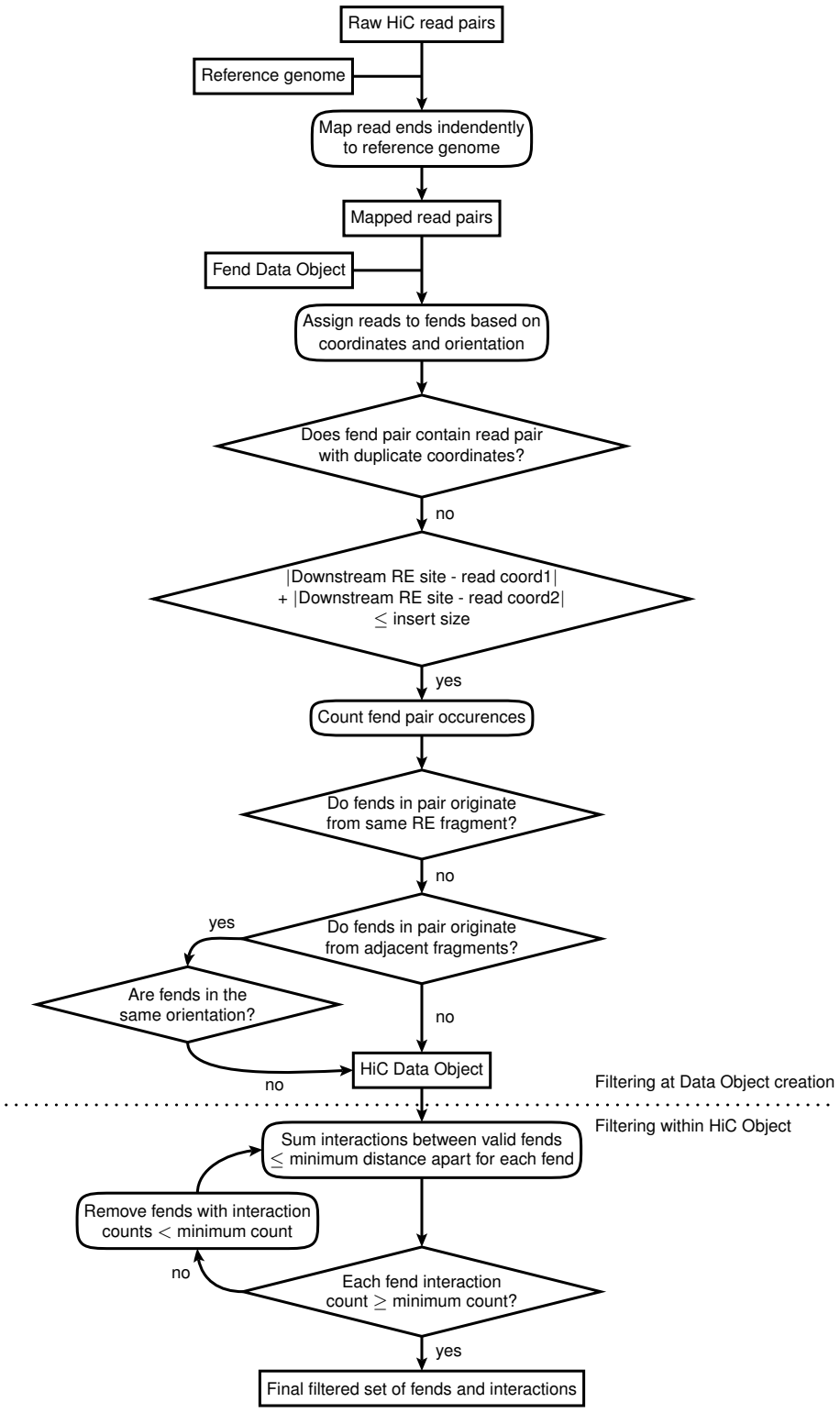

Figure 2.3 HiC read filtering scheme.

HiC data present a larger challenge in terms of artifactual reads (reads derived from sequence pairings of questionable validity) and requires a more comprehensive filtering process to produce a valid data object. Because reads are mapped directly to the genome, the pools of reads that have valid mappings at both ends include a variety of possible pairs (Figure 2.4). Based on the mapping position of the two read ends, there are three categories of pairs to consider: read pairs originating from the same fragment, pairs originating from adjacent fragments, and pairs from fragments with at least one intervening fragment. In the case of pairs from the same fragment, it is possible that the reads originated from the same fragment on different homologous chromosomes, but this appears to occur only rarely (Selvaraj *et al* 2013). Read pairs originating from adjacent fragments can map either to the same or opposite strands. In the case of opposite strand mapping, it is possible that there was failure to cleave the restriction site between the fragments. As it is impossible to tell uncleaved fragment pairs from legitimate ligation events, these reads cannot be considered without over-representing such interactions. Read pairs from adjacent fragments but the same strand must have successful cleavage of the restriction sites downstream of both read ends in order to generate the read in question and may all be safely considered. Read pairs that come from non-adjacent fragments also must be products of two successful cleavage events followed by ligation in order to form and are thus valid.

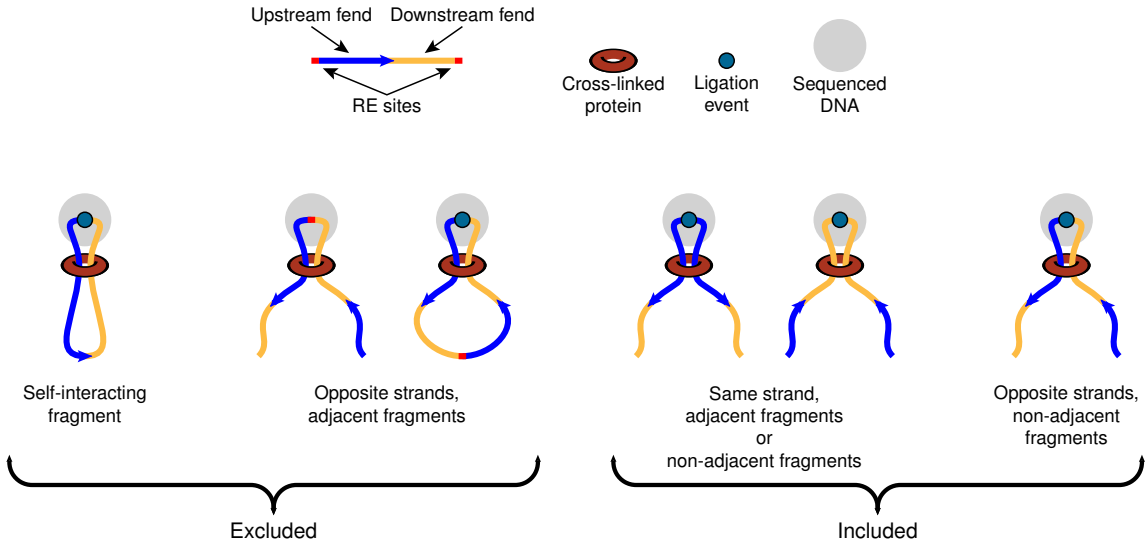

Figure 2.4 Possible HiC read pairs.
Schematic illustrating the arrangements leading to HiC read pairs and their inclusion or exclusion in HiFive analysis.

For HiC data, HiFive filters out any read pairs that could have been the result of a failed RE cut or from a single fragment's ends being ligated together. This includes identical fends, fends from the same fragment, and fends from adjacent fragments in opposite orientations. The HiC assay also includes a size-selection step which allow removal of read pairs that have an insert size that suggests a sequencing artifact, assuming that ligation between sequences occurred at the nearest downstream RE cut site for both reads. Finally, because reads are randomly sheared prior to sequencing, HiFive excludes likely PCR duplicates by removing pairs that map to identical coordinates of previously included pairs.

In the HiC analysis subsequent to filtering based on fend pairing, fend pairs were removed based on total sequenced fragment size. For each interacting pair, the distance from read end to nearest downstream RE site was determined and the sum for each read pair was calculated. If this value exceeded one kilobase, that read pair was removed from

the analysis. After all read filtering was performed, the number of occurrences for each fend pair was tallied and used for all subsequent analyses.

Once data objects are created and linked to a HiC or 5C object, a filter based on read coverage can be applied to either. Removal of fragments is accomplished using an iterative filtering process. For each fragment or fend, the number of non-filtered interacting sequences for which the pair had a non-zero read count is calculated within a specified maximum interaction distance range. Sequences with fewer than the specified minimum-count of valid interaction pairs are discarded and the process is repeated until all remaining sequences have at least the minimum number of valid interaction pairs.

In this study, all 5C, and HiC analyses employed this coverage-based filtering. For 5C analyses, no minimum distance was used and fragments were included if they had at least ten valid interactions. For the HiC analyses, two cutoffs were employed, depending on the normalization strategy being used. With the standard HiFive algorithm, a higher cutoff of 25 interactions within a range of five Mb was used. This is due to the relatively sparse coverage with HiC compared to 5C and the fact that only a subset of interactions is used in the normalization. When HiC data were normalized using the HiFive-Express approach, the cutoff was set to ten interactions within 5 Mb, as all intra-chromosomal interactions are included in the method.

Estimation of Distance-Dependent Signal

In both HiC and 5C, the largest influence driving interaction signal intensity is the genomic distance between interaction sequences (Figure 2.5). Given an experiment with dense-enough coverage at shorter ranges of inter-fragment distances, we observe a roughly power-law function such that the log of the interaction counts varies as a linear

function of the log of the inter-fragment distance. We find that this relationship holds best when the distance is measured from midpoint to midpoint of the fragment pair and when considering interactions covering ranges from 1 kilobase (Kb) to 1 Mb.

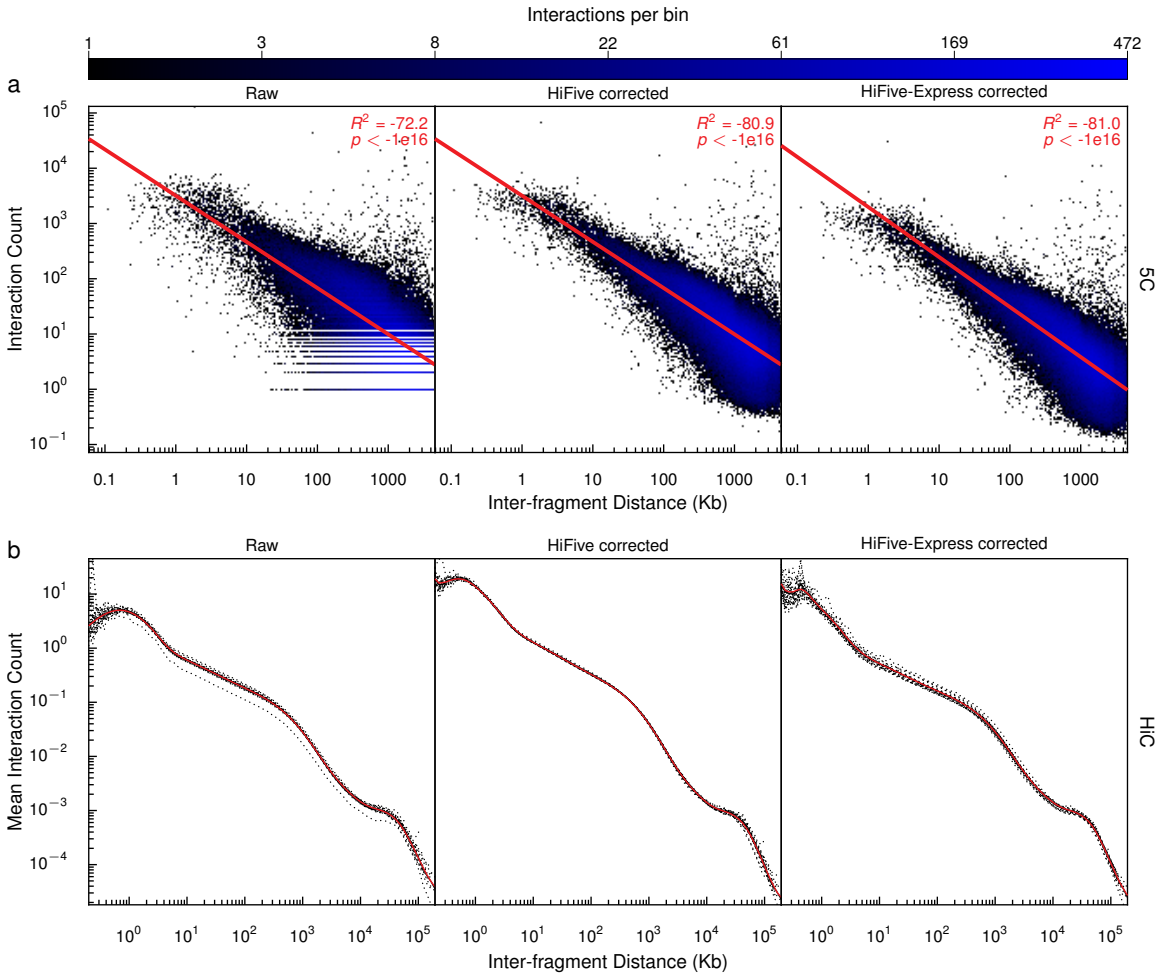

Figure 2.5 HiC and 5C signal-distance relationships and HiFive's approximation functions.
All non-zero interaction counts for mouse ESC cells, replicate one before and after fragment corrections are applied. Interactions were binned in a 200 by 200 grid for display. The red line shows the best-fit linear regression of interaction log-counts as a function inter-fragment log-distances. The corresponding regression r-squared and p-values are also shown in red. b) Mean interaction counts for hic data from the HiC mouse ESC HindIII replicate one dataset are shown split into 90 equal log-sized distance bins ranging from 200 bp to 194.2 Mb before and after fend correction. Each chromosome was binned separately (dotted lines) and all together (solid black line). The final distance-dependence function approximation after smoothing using a triangular binning approach of plus or minus two data points is shown in red.

RE sites are unevenly distributed throughout the genome leading to large variation between sets of interaction distances for each fragment or fend. Thus a fragment with many short neighboring-fragments is expected to have an elevated set of interactions as a direct consequence of this spacing rather than any biologically relevant feature. Thus, it is imperative that any normalization scheme must take fragment spacing into account in order to generate unbiased estimates of relative interaction frequency. To this end, HiFive finds a distance-dependent signal approximation function prior to normalization and incorporates this function into the calculation of expected counts.

In order to approximate the nearly linear relationship between the log of non-zero counts and the log of inter-fragment distances seen in 5C data (Figure 2.5a), HiFive uses a simple regression approach. Because the log of the counts is used, zeros are necessarily excluded from the calculations. We do not, however, believe that this has any negative impact on parameter estimation. To the contrary, we see increased variation at lower read counts suggesting that unobserved fragment pairs are the least accurate data in the experiment. For valid non-zero interaction count $s_{i,j}$ and distances $d_{i,j}$ between fragments i and j, the distance-dependence portion of the expected signal is estimated by function D with slope b and intercept a as:

$$D\left(d_{i,j}\right) = e^{\log\left(d_{i,j}\right)b+a}$$

Counts, s, are valid if not filtered in the creation of the data object and originate from non-filtered fragments. These counts are denoted by subset A. The slope is defined as:

$$b = \frac{\sum_{i,j}^{A}\left[\log\left(d_{i,j}\right) - \overline{\log\left(d_A\right)}\right]\left[\log\left(s_{i,j}\right) - \overline{\log\left(s_A\right)}\right]}{\sum_{i,j}^{A}\left[\log\left(d_{i,j}\right) - \overline{\log\left(d_A\right)}\right]^2}$$

and the intercept is defined as:

$$a = s_A - \log(d_A)\, b$$

Once calculated, the intercept is held constant. The slope can be updated throughout the normalization procedure.

HiC data have a similar general relationship between signal and inter-fend distance, though the relationship is not linear over the larger range of distances included (Figure 2.5b). In addition, the sparseness of non-zero interaction counts makes direct assessment of the distance-dependent relationship difficult. HiFive overcomes these challenges by using a piecewise approximation approach. Although other non-parametric options are available that may yield more precise estimates, the size of datasets involved in HiC analysis make them impractical from a computational and time standpoint. To find the piecewise approximation, HiFive begins by splitting interactions into *N* bins of equal log-distance sizes covering the complete range of distances for intra-chromosomal interactions. The mean of each bin is calculated from all valid counts, *s*, spanning the bin's range, where valid interactions exclude those filtered out when creating the HiC data object and those originating from removed fends and are denoted as subset *A*. Thus, for the bin *n* with an upper bound of *c*, the mean μ is calculated as follows:

$$\mu_n = \frac{\sum_{i,j}^{c_{n-1} < d_{i,j} \leq c_n} s_{i,j}}{\left|\{s_{i,j} : c_{n-1} < d_{i,j} \leq c_n\}\right|}$$

To account for noise in the data, HiFive can apply a triangular smoothing function to the bin means using smoothing parameter *k*, giving the smoothed mean value μ':

$$\mu'_n = e^{\frac{log(\mu_n)k + \sum_{i=1}^{k-1} \log(\mu_{n-i}\mu_{n+i})(k-i)}{k^2}}$$

The HiC distance-dependence function is defined as a piecewise linear approximation of these smoothed means, where a bin's mean corresponds to a specific distance *g* falling within its upper and lower boundaries denoted as c_u and c_l, respectively:

$$g = c_u e^{(c_l - c_u)(1-\sqrt{0.5})}$$

For an interaction between fends *i* and *j* with a distance falling between two defined points, such as between the distances denoted by g_n and g_{n-1} associated with bins *n-1* and *n*, respectively, the corresponding estimated distance-dependent signal is estimated:

$$D(d_{i,j}) = \frac{(g_n - d_{i,j})\mu'_{n-1} + (d_{i,j} - g_{n-1})\mu'_n}{g_n - g_{n-1}}$$

In this study, we divided the interaction ranges into 90 bins with a minimum interaction distance of 200 bp. Smoothing was done using a smoothing parameter of two in the triangular smoother.

HiFive Data Normalization

HiFive's normalization approach uses a combination of empirical distance-dependence estimation and probabilistic modeling to estimate expected signal and enrichment. HiFive works on two key assumptions. First, the majority of interaction reads are derived from a combination of signal that is dependent on inter-fragment distance and fragment-specific bias. Second, the effects of the fragment biases can be described as the product of the individual biases associated with the interaction fragment pair. Thus the expected signal *E* for the interaction between fragments or fends *i* and *j* is the product of the bias correction *f* for each end of the interaction and the distance-dependence function *D* as follows:

$$E_{i,j} = D(d_{i,j}) f_i f_j$$

This approach allows adjustment for factors known to contribute to differences in fragment observation rates, such as guanine and cytosine (GC) content, fragment length, and mappability (Hu *et al* 2012, Yaffe & Tanay 2011) without explicitly limiting the correction to a specific relationship of factors. Although the general framework is the same for both HiC and 5C, differences in the experimental procedures necessitate technique-specific variants of the distance-dependence function and underlying probability distribution.

The 5C observed interactions are modeled with a log-normal distribution around the predicted values with standard deviation σ such that:

$$\log(s_A) \sim \mathrm{N}(\log(E_A), \sigma^2)$$

The standard deviation is estimated at the same time as the distance-dependence parameters prior to learning fragment corrections such that:

$$\sigma = \sqrt{\frac{\sum_{i,j}^{A}\left[\log\left(s_{i,j}\right) - \log\left(E_{i,j}\right) - \overline{(\log(s_A) - \log(E_A))}\right]^2}{|A| - 1}}$$

Because estimated counts change as the corrections are learned, HiFive includes the ability to periodically update the parameters of the distance-dependence function. If specified by the user, the slope b in the distance-dependence function and σ are updated, although the term $s_{i,j}$ is replaced with the bias-corrected count $s'_{i,j}$ for calculating these parameters such that:

$$S'_{i,j} = \frac{s_{i,j}}{f_i f_j}$$

HiC interaction counts are modeled as a series of Poisson processes with λ being defined as the predicted value for a given interactions such that:

$$s_A \sim \mathrm{Pois}(E_A)$$

Unlike 5C, only a fraction of interactions are used to learn fend bias corrections. Specifically, counts (including zeros) are included in the model if they belong to the set of valid interactions, they are intra-chromosomal, and the distance between their fends is less than or equal to a user-specified maximum interaction distance range. This is done for two reasons; 1) the vast majority of observed interactions occur over short interaction distances and 2) including all possible interactions or simply all possible *cis* interactions is computationally unfeasible for this kind of model.

Like with 5C data, the HiC distance-dependence function can be updated during the learning of fend correction values. This is done by substituting the raw count term $s_{i,j}$ with the corrected interaction count term $s'_{i,j}$ as described above for finding the distance bin means.

HiFive learns bias parameters for both HiC and 5C data using a two-stage gradient descent approach that maximizes the likelihood of the observed data under the probability distributions described above. In the burn-in phase, fend or fragment bias parameters are updated using a constant learning rate. This is followed by an annealing phase in which the learning rate decreases in a linear fashion to zero.

For the normalization in this study, 5C data were normalized using all *cis* interactions over a 5,000 iteration learning phase and a 10,000 iteration annealing phase. An initial learning rate of 0.01 was used in both phases. The distance-dependence parameters were updated after every 100 iterations in both phases.

HiC normalization was done using cis interactions up to a maximum interaction range of 5 Mb. The learning and annealing phases were carried out for 5,000 and 10,000

iterations, respectively. The initial learning rate for both phases was 0.01 and the distance-dependence parameters were updated every 2,500 iterations in each phase.

HiFive-Express Iterative Approximation for Bias Correction

Due to the memory and computational requirements associated with rigorous HiC and 5C normalization, HiFive also includes a fast and computationally inexpensive iterative approximation normalization alternative for both data types referred to as HiFive-Express. HiFive-Express makes use of the same framework and underlying predicted value scheme as HiFive, changing only the approach to bias correction value calculations described above. While the results are not as robust as the more computationally expensive HiFive normalization, they are still sufficient for many applications, allowing resolution down to the individual fragment or fend level depending on assay type. In addition, HiFive-Express can be performed on a single processor in minutes with a comparatively small memory footprint and can easily make use of all data, up to and including *trans* interactions for HiC data.

Unlike HiFive's probability-based maximization, HiFive-Express attempts to minimize the distance between one and the fraction defined by interaction counts over predicted values. Predicted counts can either be found using bias values alone or with bias and distance-dependence function values. The advantage of this approach is two-fold: 1) the calculations are simpler (and therefore faster) than calculating the gradients and 2) because the cost is calculated in terms of a fraction with observed counts as the numerator, zero counts always have a fractional value of zero, reducing the needed calculations down to only observed interactions (a small fraction of the total possible interactions) and a single sum of the number of unobserved interactions, although this

second point only applies to HiC data. In addition, because HiFive-Express uses an iterative update, no learning rate parameter is necessary.

In 5C data, the normalization using HiFive-Express is still limited to non-zero interactions as the correction approximation is performed on the log of the observed counts. For each round, the fragment correction value f_i is updated as follows from the non-zero subset of observed interactions involving fragment i, A_i:

$$f_i' = \log\left(\frac{\sum_j^{A_j} \log(s_{i,j}) - \log\left(D(d_{i,j})\right) - \log(f_i f_j)}{2|A_i|}\right) f_i$$

If distance-dependence is not taken into account, $D(d_{i,j})$ (the distance-dependence estimation of the log count) is set to one.

HiFive-Express uses a similar approach for HiC data although because counts rather than log counts are used, all valid possible interactions are considered including unobserved interactions. For each round of HiC correction approximation, the fend correction factor f_i is updated as follows from the valid set of interactions involving fend i, A_i:

$$f_i' = f_i \sqrt{\frac{1}{|A_i|}\sum_j^{A_i} \frac{s_{i,j}}{D(d_{i,j}) f_i f_j}}$$

If distance-dependence is not taken into account, $D(d_{i,j})$ is set to one.

For this study, 5C data normalization using the HiFive-Express approach used all valid non-zero *cis* interactions. Predicted counts included a distance-dependent signal estimate. Learning was accomplished over 10,000 iterations with a recalculation of the distance-dependence parameters occurring every 100 iterations.

HiC data normalized using the HiFive-Express approach made use of all valid *cis* interactions and included the distance-dependence in the predicted count estimate. Learning occurred over 1,000 iterations and updates to the distance-dependence parameters were run every 200 iterations.

Neighboring Fend Correlations

In order to assess the relative similarity between neighboring fends originating on the same RE fragment versus those originating on adjacent fragments, we identified groups of three consecutive fends that passed all filtering steps. For each triplet set, we considered all non-zero interactions within 1 Mb of the center fend. For each adjacent fend pair within the triplet, the correlation of log counts or log corrected counts was calculated across all partners for which both fends had a non-zero interaction count.

Results

HiC Unit of Interaction

One of the ways that HiFive achieves high resolution of HiC data lies in its treatment of DNA fragments that result from RE digestion of the genome. Both HiC and 5C experiments rely on fractionation of the genome by REs. Unlike 5C, however, HiC data are composed of reads that, theoretically, can map anywhere along the restriction fragments. It has been shown that fragment length is inversely related to interaction signal intensity (Yaffe & Tanay 2011). In addition, the HiC assay maps reads with an orientation indicating which end of restriction fragments was ligated. With these two facts in mind, we assessed the similarity between interactions within 1 Mb of a set of fends compared to the adjacent fend on either side of them in the raw data and data corrected for distance-dependence, fend bias, and both (Figure 2.6). We found that fends originating from the same restriction fragment did not show any more similarity in their non-zero log-interactions with other fragments than adjacent fends originating from neighboring restriction fragments. We concluded that the nature of the assay coupled with the filtering of reads results in fends that originate from the same fragment acting as independent units of interaction. As such, for all normalization of downstream analysis, they were treated independently.

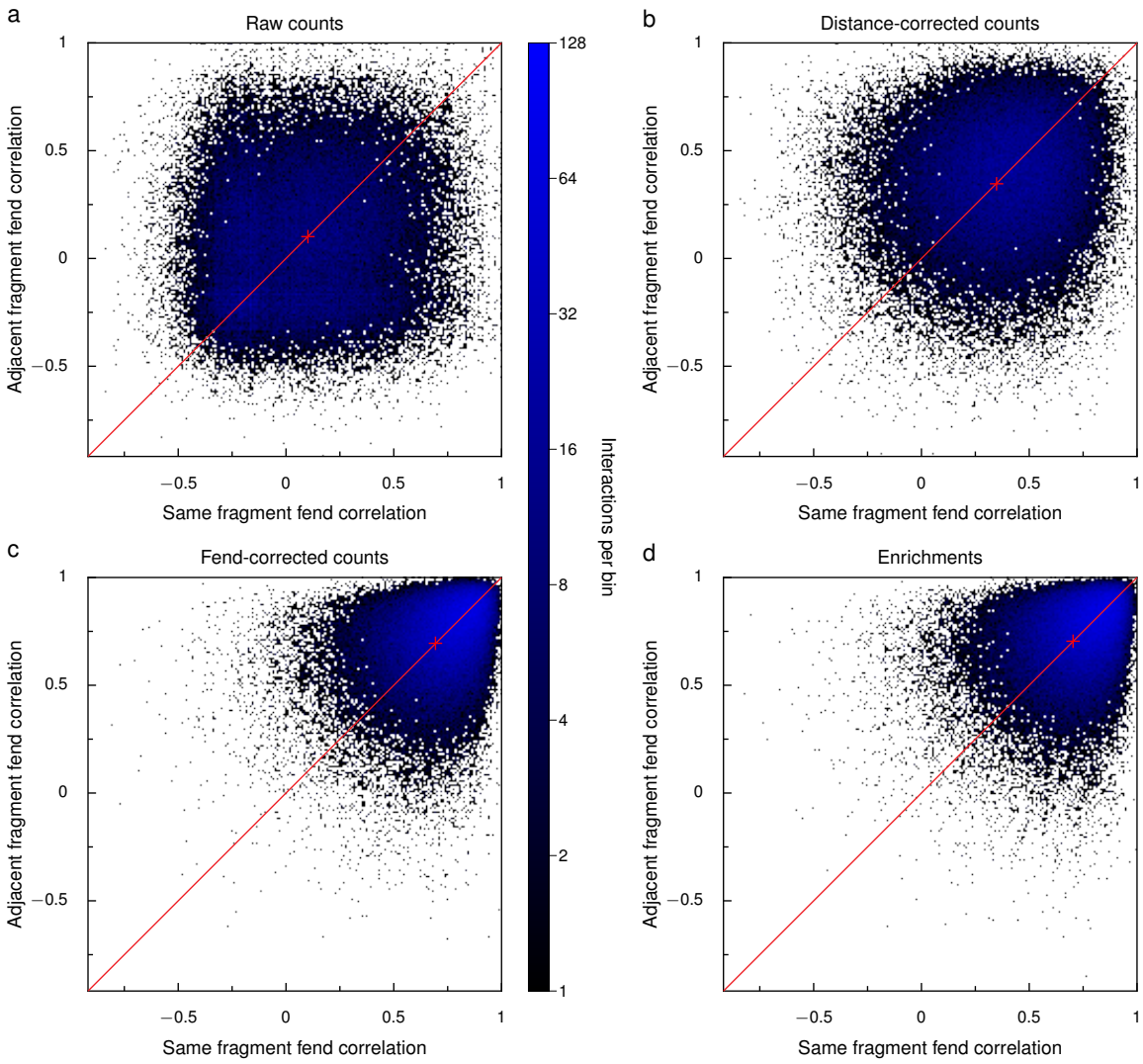

Figure 2.6 Correlation of neighboring fend interactions for adjacent vs. same fragment fends.
For all valid triplets of adjacent fends, correlations between each neighboring pair of fends and their non-zero log-interaction counts with all valid fends within one Mb were calculated. Interactions were binned in a 200 by 200 grid for display. A one to one relationship line is shown in red with the data center of mass marked by a red X. a) Correlations were calculated for each fend pair using raw reads. b) Correlations were calculated using reads after correction for fend biases using HiFive correction values.

The practical result of assessing fends independently is that the number of possible interacting sites is doubled (Figure 2.7). Further, this increases the number of

interactions and therefore the interaction map resolution by a factor of four compared to assessing whole restriction fragments.

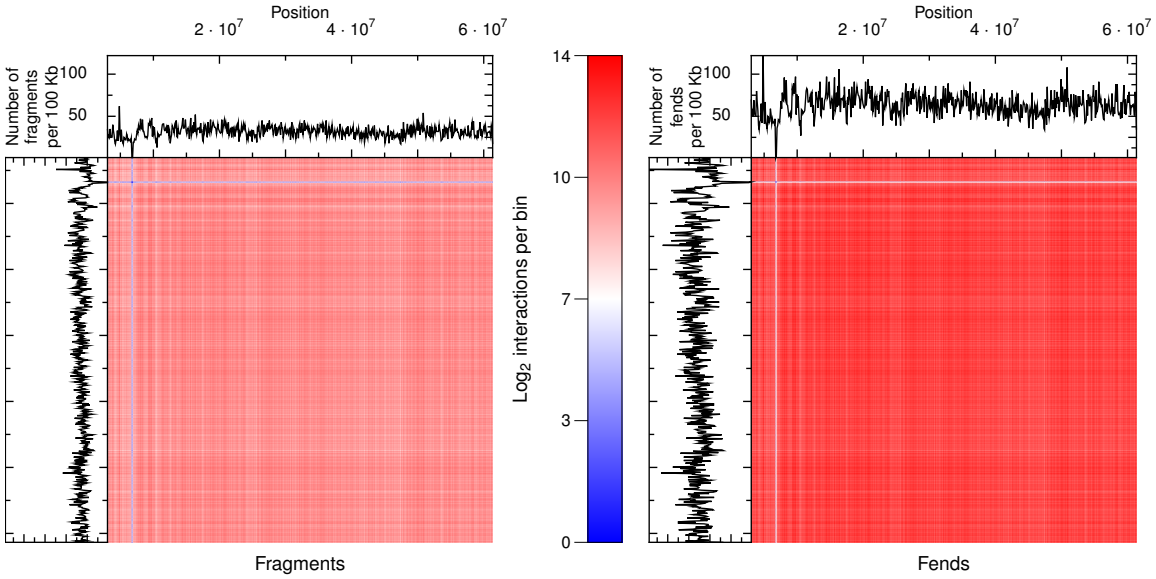

Figure 2.7 Fragment and fend density variation.
Mouse chromosome 19 is shown divided into 100 Kb bins. Fragments and fends produced by HindIII digestion are shown binned according to their midpoint coordinates. The numbers of fragments or fends per 100 Kb bin are shown along the border while the number of possible interactions are shown in the heatmap.

HiFive's 5C Normalization Performance

In order to assess HiFive's performance in normalizing 5C data, we first examined the model fit across a variety of factors. These included assessing the effects of normalization on the distance-dependence relationship, the differences between fragments across their signal strength and variance, the relationship between genomic characteristics and signal strength, and the reproducibility of bias correction values across replicates and methods. Although it is expected that significant chromatin structural interactions are more likely to occur on longer sequences than shorter ones by chance, we

expect that these interactions should still make up a small fraction of the possible interactions being queried for any given fragment. Similarly, although some genomic features associated with chromatin structure have skewed GC content, we would still expect that this would have little impact on a fragment's overall interaction strength in the absence of technical biases. As such, we would expect any procedure that is adequately removing technical biases from the data to also eliminate any relationship between these factors and interaction signal strength when viewed across all interactions or a fragment's average interaction strength.

After normalization with HiFive, 5C interactions showed an improved fit to the distance-dependence line (Figure 2.5a). This improvement was true for both the standard HiFive and HiFive-Express algorithms. Reduction in variance amongst reads of similar inter-fragment distances was particularly strong at low counts. The algorithm used (HiFive vs. HiFive-Express) made little difference in the improvement of fit.

To determine whether systemic biases associated with fragment characteristics were removed, two previously cited sources of systematic noise, fragment length and GC content (van Berkum & Dekker 2009) were assessed across individual reads as well as for fragment averages (Figure 2.8). In both cases, the magnitude of the effects (slope of the regression line) was reduced, regardless of the correction algorithm. Additionally, there was a marked reduction in signal variance for fragment means for both approaches. While fragments showed a wide range of mean log interaction counts prior to bias correction, normalized mean log counts showed little difference and maintained a consistent but slightly lower variance between fragments for both correction algorithms (Figure 2.9).

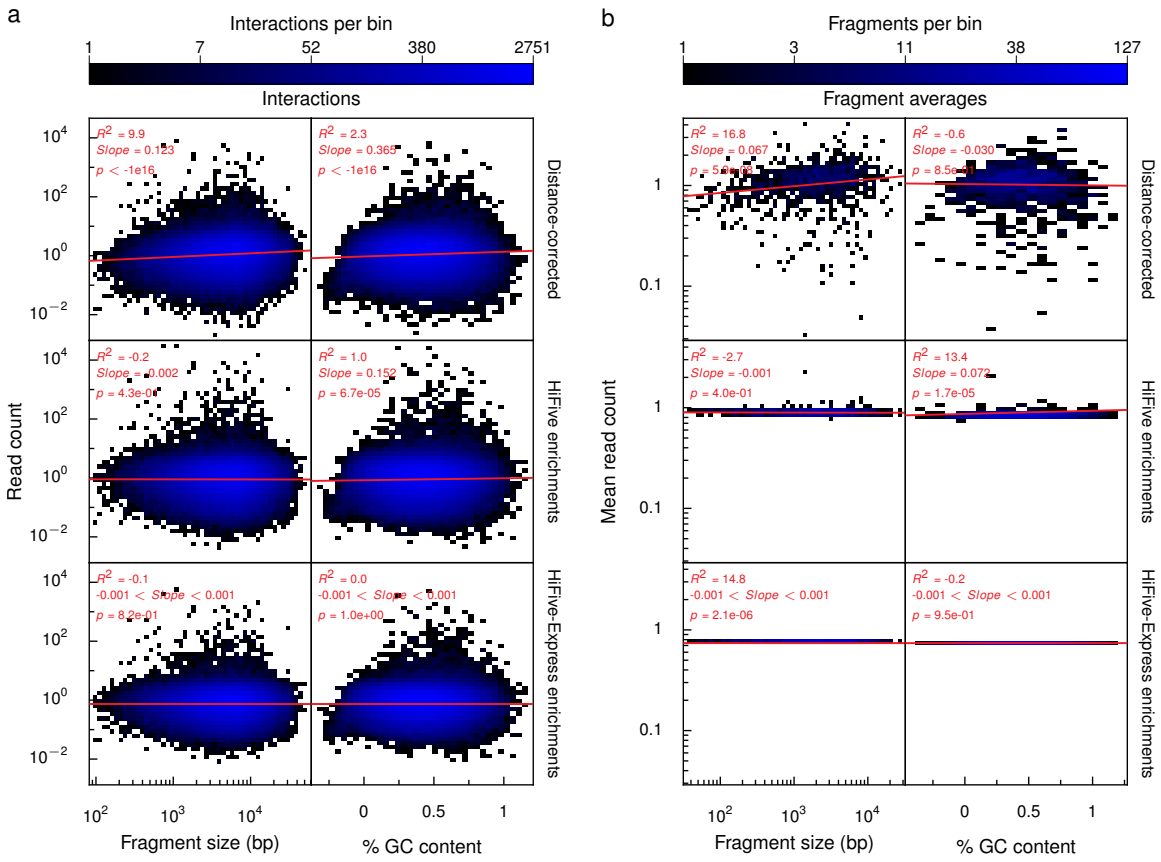

Figure 2.8 Fragment characteristics' effects on fragment bias in 5C data.
All non-zero interaction log counts for the mouse 5C ESC replicate one dataset were plotted as a function of the percent GC content of interacting fragments and the log of the sum of interacting fragment sizes. Interactions for each plot were binned in a 50 by 50 grid for display. A linear regression for each relationship was calculated and shown in red, along with corresponding r-squared, slope, and p-values. Interactions were corrected for distance-dependence (top) or distance-dependence and fragment bias (bottom). b) The means of all logged non-zero interactions for each fragment were plotted as a function of log fragment size or fragment GC content. For each relationship, a linear regression was performed as is plotted in red, along with the corresponding r-squared, slope, and p-values. Interactions were corrected for distance-dependence (top) or distance-dependence and fragment bias (bottom).

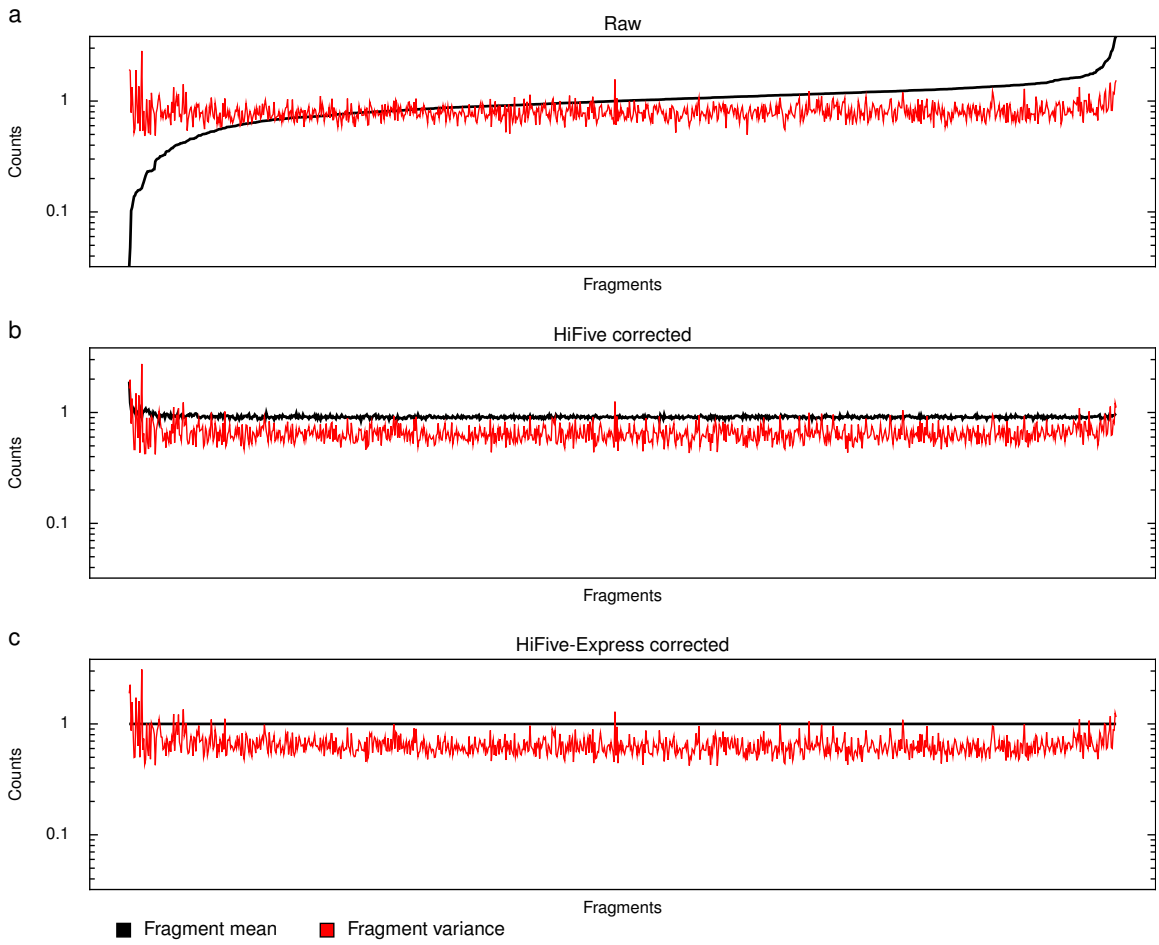

Figure 2.9 Fragment-associated bias in 5C data.
For each fragment included in the 5C analysis for the mouse ESC replicate one dataset, means and variances were calculated from the log of counts for all non-zero interactions in which that fragment participated. Fragments were ordered by their uncorrected means from lowest to highest. a) Means and variances are plotted for counts from which the distance-associated signal was removed to avoid skewing results given the limited number of interactions and relatively short distance range covered by the region. b) Means and variances are plotted for counts corrected for distance-associated signal and fragment bias.

Finally, we examined the consistency of learned correction values between replicates and algorithms (Figure 2.10). Across both replicates and correction algorithms, correction values were highly consistent. Unsurprisingly, variation between correction values was inversely related to the correction value (and therefore the mean signal). Comparison of correction values between algorithms for identical datasets showed a

nearly perfect correlation, suggesting that for 5C data, the choice of HiFive correction algorithm makes little difference.

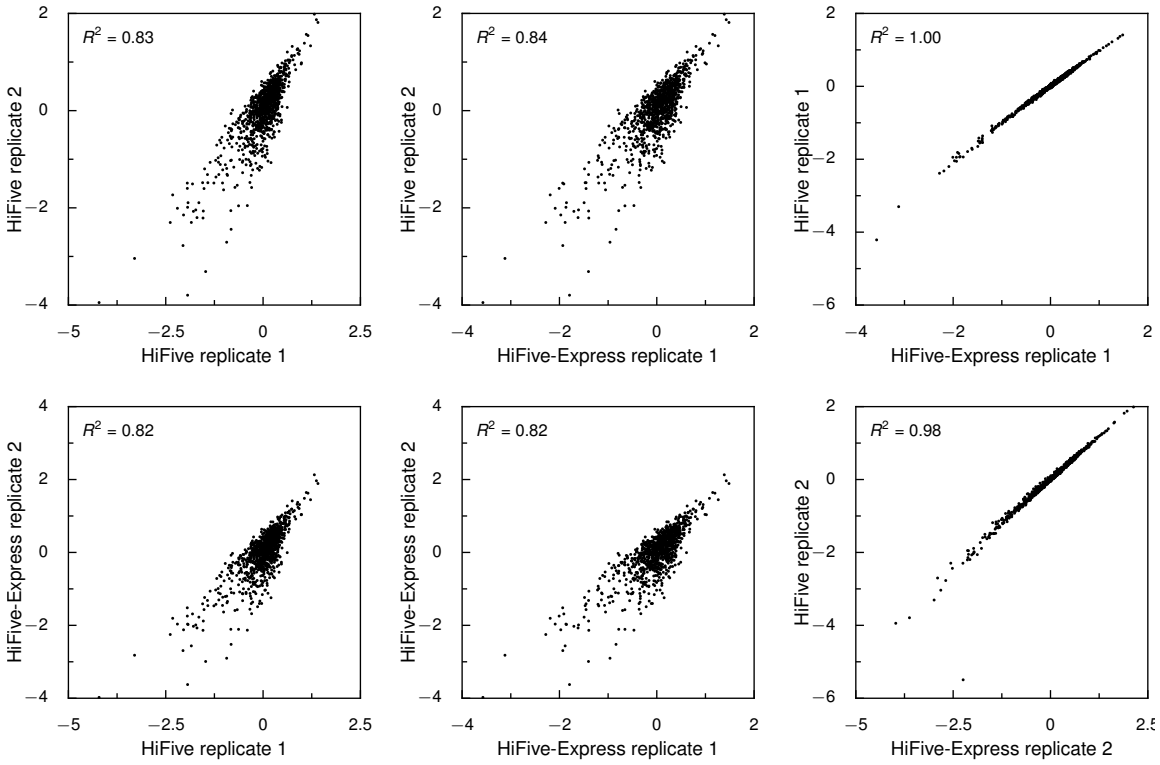

Figure 2.10 5C inter-replicate correlations of fragment bias correction values. Correlations between log fragment correction values for all pairwise combinations of correlations for normalizations performed with HiFive and HiFive-Express.

HiFive's HiC Normalization Performance

To assess HiFive's performance in normalizing HiC data, we examined the effects of correction across a number of factors. These included the following: distance-dependence relationship; correlations of GC content, sequence mappability, and fend length to signal; between-fend signal variation; and correction value differences between data replicates and algorithms. Similar to the effect described for 5C data, fend size and

mappability (the percent of 36 bp oligos within 500 bp of the RE site of a fend that are uniquely mappable) should show a positive correlation with signal strength due to likelihood of containing a structurally relevant feature and probability of observation, respectively. Unlike 5C, hybrid sequences produced by HiC vary greatly even between the same fends, due to the random fragmentation by sonication. This results in a variety of GC contents associated with any given fend. We include an analysis of the effects of GC content on interaction signal as measured across the 500 bp of the fend closest to the RE site.

The effects of HiFive normalization of HiC data on the distance-dependence were highly dependent on the algorithm used (Figure 2.5b). The standard HiFive normalization showed greatly reduced variation between distance bin means across different chromosomes and the genome-wise average. Conversely, inter-chromosome variation appears to have increased after HiFive-Express correction at shorter inter-fragment distance ranges. Interestingly, the reverse is true at very long-range interactions. HiFive-Express normalization appears to minimize inter-chromosome mean count differences for interactions at distances greater than 10 Mb, whereas the standard HiFive approach shows little improvement over raw signal at these ranges. Both approaches, however, show approximately the same transformation of the distance-dependence curve shape at short ranges.

Examining the effects of fend length, mappability and GC content, we find that only the former and latter show a strong influence on the mean signal of its associated fend (Figure 2.11), while average GC content does not. Correction using HiFive and HiFive-Express both greatly reduced effects of length and mappability, although only

HiFive reduced the variance among fend means for counts with and without the distance-dependent signal present (Figure 2.11a). When the distance-dependent signal was removed prior to finding fend means, the variance for interactions corrected with HiFive-Express was reduced to nearly zero (Figure 2.11b), reflecting fundamental differences in the underlying transformation of counts between these two approaches for HiC data.

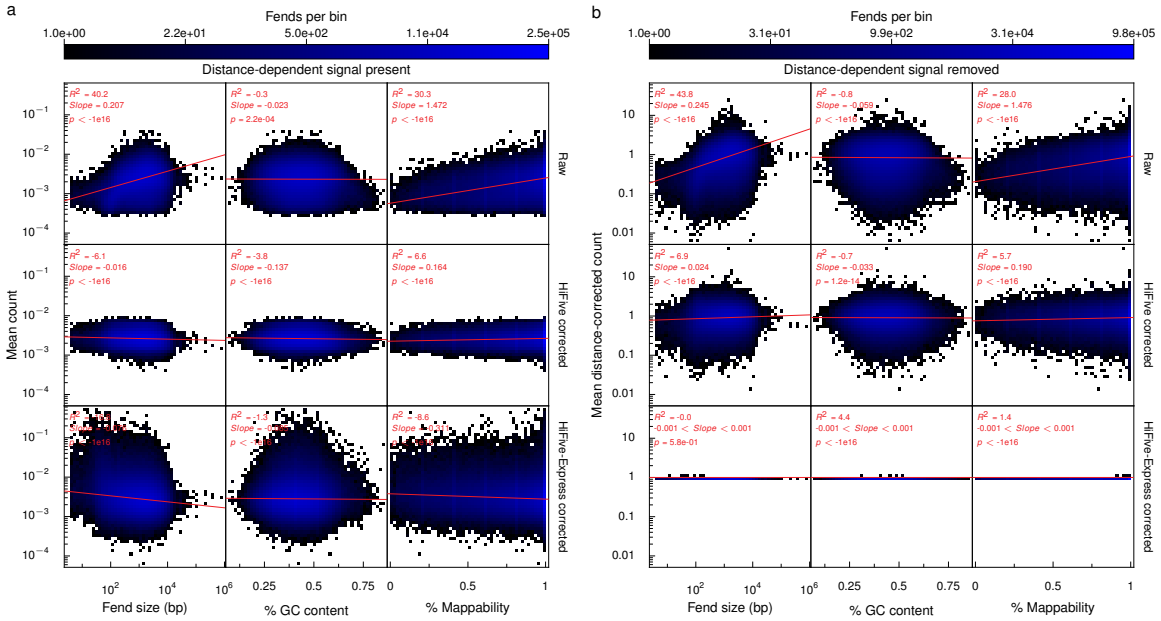

Figure 2.11 Fend characteristics' effects on fend bias in HiC data.
Using the HiC mouse ESC HindIII replicate one dataset, mean interaction counts for each fend were plotted as a function of fend size, mappability, and GC content before (top) and after (bottom) fend correction. Interaction means were binned into a 50 by 50 grid for display. For each relationship, a linear regression was performed with the logged interaction means and plotted in red along with the r-squared, slope, and p-values. a) Means for each fend were taken from raw counts or counts corrected only for fend bias. b) Means were calculated from counts after distance-dependent signal was divided out and, for correction conditions, fend biases were removed.

Next, we examined the range of fend means and variance across all valid fends before and after correction (Figure 2.12). Raw fend means showed a large range of means and variances for counts within one Mb interaction range for both *cis* and *trans*

interactions. After correction with either HiFive or HiFive-Express, fend means and variances were highly consistent, although variances were slightly higher when HiFive-Express was used in both *cis* and *trans* interactions.

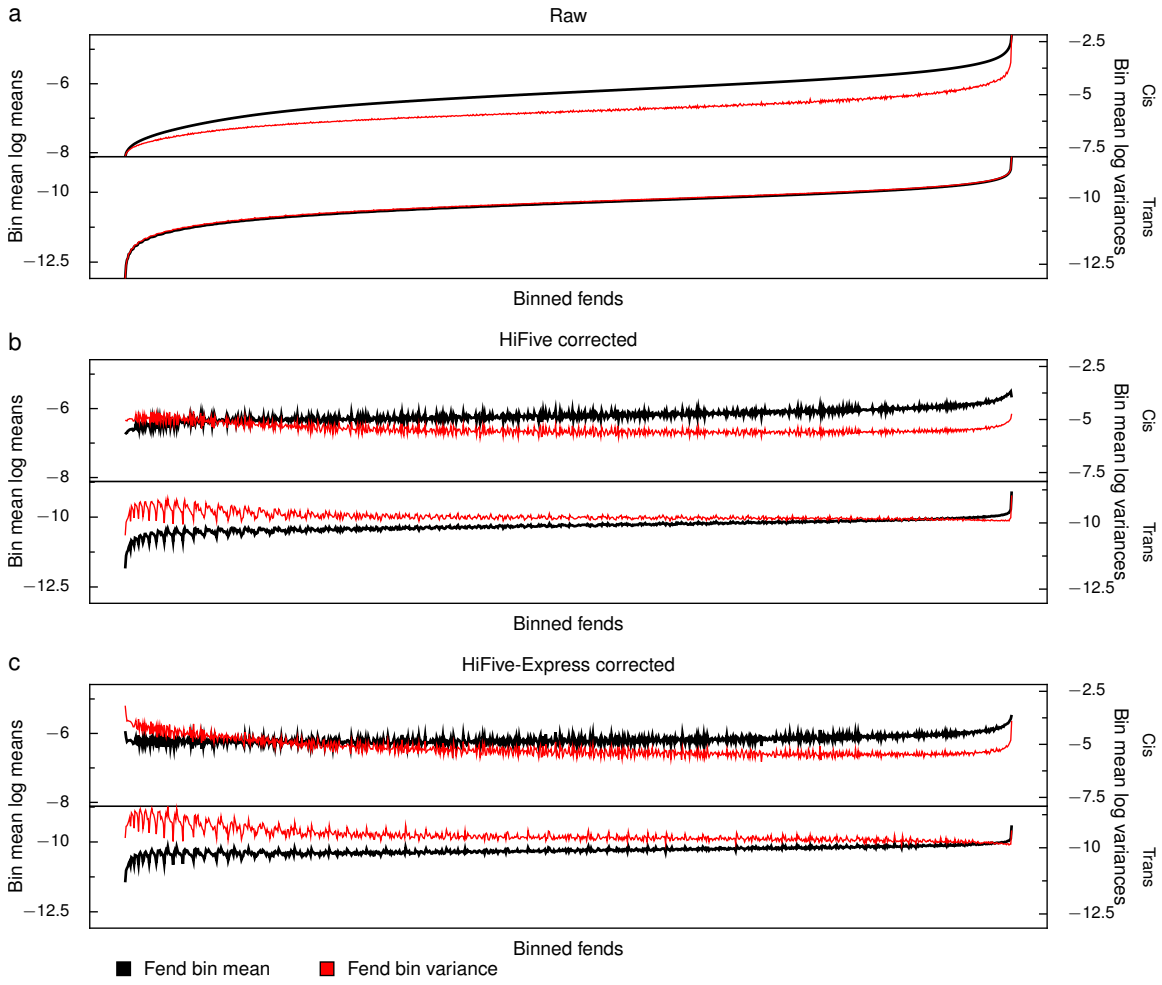

Figure 2.12 Fend-association bias in HiC data.
For each fend included in the analysis of the mouse ESC HindIII replicate one dataset, the count mean and variance were calculated for *cis* interactions and *trans* interactions involving that fend. Fends were ordered by their uncorrected means from lowest to highest and then binned into 4,000 equal-sized bins. Bin averages for mean and variance are shown. a) Counts are drawn directly from raw interaction counts. b) Counts were adjusted by fend correction values prior to calculation of mean and variance.

Finally, we examined inter-replicate and inter-algorithm correlations of learned corrections (Figure 2.13). HiFive and HiFive-Express both showed relatively high levels of consistency both between replicates as well as between each other. Corrections obtained using HiFive-Express showed a wider range of values, particularly at lower values that correspond to lower signal fends. Further lower correlations between replicates or across methods for replicate two reflect quality differences in datasets, and are consistent with sequencing depth differences between replicate one (107,811,834 valid reads) and replicate two (42,803,540 valid reads).

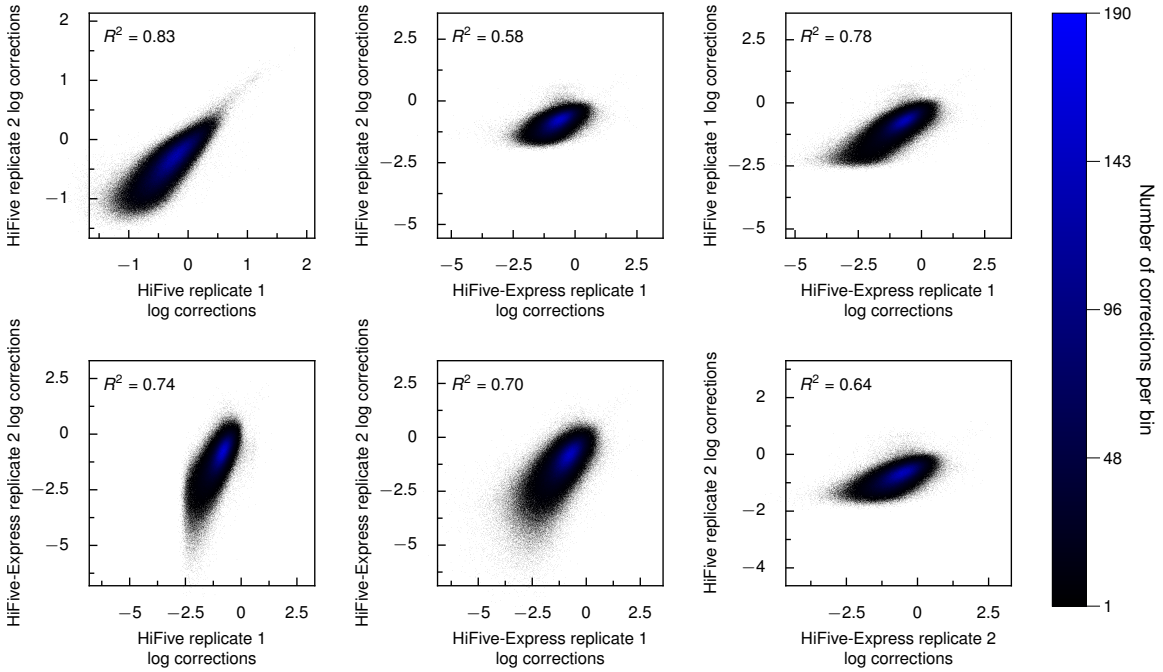

Figure 2.13 HiC inter-replicate correlations of fend bias correction values. Correlations between log fend correction values for all pairwise combinations of correlations for normalizations performed with HiFive and HiFive-Express. Corrections were binned in 1,000 by 1,000 grids for display.

Discussion

Here we presented HiFive, a computational framework for handling 5C and HiC DNA interaction data, removing systematic biases, and performing downstream analyses. We have demonstrated HiFive's ability to address and remove signal resulting from physical characteristics unrelated to chromatin structure and to produce estimates of enrichment and significance at the smallest possible level of resolution allowed by the assays. Further, we have designed an analysis alternative with fairly comparable performance but a small fraction of the computational requirements, opening high-resolution DNA-interaction analysis up to a new audience of scientists.

There have been three primary strategies in approaching normalization of these types of data. The first involves explicitly modeling known sources of systematic bias. There are two large drawbacks to this method. One drawback is in identifying all of the relevant sources of bias within the experiment. In their paper outlining HiCPipe, Yaffe and Tanay (2011) demonstrate that the three features detailed in this study account for a significant portion of systematic noise in HiC experiments and that a nonparametric model does perform well in terms of normalization. The fact remains that each fend was associated with a signal value for each term in the model. This gets to the second drawback, which is that because HiC involves random fragmentation, reads originate from a large stretch of possible sequence. To truly model the effects of GC content and mappability, each read would need to be considered separately in terms of the mappability and GC content associated with the particular bases actually sequenced. The second strategy employed is to use an external control or condition for direct assessment of signal differences. This approach assumes that the specific composition of the oligo (a

short stretch of nucleotides) mixture being sequenced does not impact the relative sequencing efficiencies of different oligos. Our own experience has not supported this idea, although a thorough study is needed to understand fully how mixture composition affects sequencing results in highly multiplexed reactions. Another challenge with the comparative approach is specific to HiC. Because of the number of possible combinations of interactions, a negative control would suffer from the same stochastic effects that are seen in experimental data and would be no guarantee of accurate correction values. The last strategy involves approximating correction values for each interaction unit (either each fend or interval if data have been binned) directly from the data. Because the results have been satisfactory and the computation demands are lower, approaches of this kind have used simple additive effects models for fend combinations. Improvements could likely be made by exploring more complex relationships between fend combinations and interaction signal. Despite this, we believe that this approach has the most promise for such data-rich experiments as 5C and HiC.

The differences in performance between HiFive and HiFive-Express reflect their relative strengths and weaknesses, particularly in HiC data. There is little practical difference in performance handling 5C data. We suspect this is due to the high coverage afforded by 5C experiments and the scarcity of unobserved data. The algorithmic differences become readily apparent when handling HiC data. HiFive-Express has a tendency to overcorrect due to its inability to differentiate between unobserved counts with high versus low expected signal (Figure 2.11a). However, when comparing average fend enrichment values versus different physical characteristics, features like GC content and mappability show no influence on signal and very little signal variation. Conversely,

HiFive is able to produce good expected read estimates for unobserved fend-pairings because of the way it models the probability of observing interaction reads. This results in lower variation with and without distance-dependent signal present across physical characteristic values and a dramatic lowering of the influence of these features on fend signal, though not to the same extend as HiFive-Express for enrichment averages (Figure 2.11b). This ability to find good estimates for unobserved counts is particularly important for finding binned enrichments, especially in sparsely observed regions such as happens with larger bins and at longer interaction ranges.

HiFive's major advantages lie in its flexibility in terms of computational requirements, ability to scale to any resolution without reanalysis, and organization of all data aspects associated with these DNA interaction assays. Because HiFive exists as a library rather than a stand-alone piece of software, it allows the user to easily construct custom pipelines. This is especially important as investigations into chromatin structure are becoming more sophisticated with the low-hanging fruit of macro-structural features giving way to attempts to understand the details of fine-scale folding principles and chromatin dynamics throughout the cell cycle. We believe HiFive will be a useful tool not only for focusing on higher resolution investigations, but also allowing for use of this additional layer of epigenetic information within a greater range of the scientific community.

Chapter 3 Validation of HiFive through method comparisons and biological findings*

Introduction

Our understanding of the physical configuration of chromatin has become increasingly sophisticated as a result of new assays that allow direct interrogation of spatially proximate DNA sequences. With the availability of NGS, assessment of chromatin associations and architecture has moved from solely the realm of microscopy into a large-scale modeling-based approach inferring relationships from millions of DNA-DNA associations through the use of PMLAs. This has resulted in an increase both in the amount of data being generated and the complexity needed for handling and understanding these data. In order to take advantage of the increased resolution and coverage of these newer experimental approaches, data correction methods are needed that can reliably separate technical and stochastic noise from structural signal.

A variety of computational methods have been developed to handle data produced from PMLAs using several different strategies. Methods such as HiCPipe and HiCNorm rely on physical characteristics of the queried sequences such as GC content and RE fragment length and an explicitly defined relationship with signal bias (Hu *et al* 2012, Yaffe & Tanay 2011). The alternative approach, taken by normalization methods including HiCLib, relies on learning fragment- or bin-associated correction values

* Portions of this chapter have been adapted from an article in review at *Genome Research*.

without defining specific driving factors behind the signal bias (Imakaev *et al* 2012). It is currently unclear how these methods stack up against one another and what the best methods are for assessing the relative merits of these computational approaches.

Testing the efficacy of these approaches is difficult because the experimental approaches that can achieve comparable levels of resolution are all based on the same underlying strategy. Although FISH has been used to verify structural predictions, the limitations of microscope resolution and processing time make this an impractical means of assessing performance. Further, because PMLAs assess entire populations of cells, which have been shown to vary dramatically in their specific chromatin structures (Nagano *et al* 2013), such single cell-based approaches are of limited value when looking at cell population-based assays. There are, however, several different ways to approach the challenge of validating a data normalization approach.

In this study we demonstrate the utility of HiFive, a probabilistic modeling approach for HiC and 5C data normalization and analysis through a combination of direct comparisons to other normalization strategies and detection of genomic features and associations described elsewhere in the literature. By assessing the reduction of variation, systematic biases, and inter-replicate differences, we show the efficacy of HiFive's normalization strategy. We also demonstrate HiFive's performance capabilities relative to other available methods by a comparison of inter-dataset similarity after normalization. Finally, we show how HiFive supports downstream analysis using several new tools developed for the HiFive library.

Materials and methods

Acquiring and Mapping Data

All 5C data were acquired from GEO [1] as SRA files and split using Fastq-Dump [2] (Table 3.1, Barrett *et al* 2009, Wheeler *et al* 2008). Using associated primer sequences also obtained from GEO, each read's paired ends were mapped independently using Bowtie [3] against primer sequences using the mapping options "--phred33-quals --tryhard -m1 -5 3 -3 2 -v 2" (Langmead *et al* 2009). Sequences with one end mapping to a forward primer and one end to a reverse primer were tallied and the resulting counts were used to generate filtered 5C datasets.

Table 3.1 List of public 5C and HiC datasets used.

Sample	Replicate	Cell Type	Data Type	Reference	GEO ID
Female MEF	1	Female E13.5 embryo-derived MEF	5C	Nora et al 2012	GSM873924
Female MEF	2	Female E13.5 embryo-derived MEF	5C	Nora et al 2012	GSM873925
Female ES day 2 PGK12.1	1	Female mES	5C	Nora et al 2012	GSM873926
Female ES PGK12.1	1	Female mES	5C	Nora et al 2012	GSM873927
Female ES PGK12.1	2	Female mES	5C	Nora et al 2012	GSM873928
Male MEF	1	Male E13.5 embryo-derived MEF	5C	Nora et al 2012	GSM873929
Male ES day 2 E14	1	Male mES	5C	Nora et al 2012	GSM873930
Male ES day 2 E14	2	Male mES	5C	Nora et al 2012	GSM873931
Female XO ES	1	XO Female mES	5C	Nora et al	GSM873932

DXTX				2012	
Female XO ES DXTX	2	XO Female mES	5C	Nora et al 2012	GSM873933
Male ES E14	1	Male mES	5C	Nora et al 2012	GSM873934
Male ES E14	2	Male mES	5C	Nora et al 2012	GSM873935
Male ES EED-	1	Male mES	5C	Nora et al 2012	GSM873936
Male ES EED-	2	Male mES	5C	Nora et al 2012	GSM873937
Male ES TT2 G9A-	1	Male mES	5C	Nora et al 2012	GSM873938
Male ES TT2 G9A-	2	Male mES	5C	Nora et al 2012	GSM873939
Male ES TT2	1	Male mES	5C	Nora et al 2012	GSM873940
Male ES TT2	2	Male mES	5C	Nora et al 2012	GSM873941
Male NPC E14	1	Male mES-derived NPC	5C	Nora et al 2012	GSM873942
Male NPC E14	2	Male mES-derived NPC	5C	Nora et al 2012	GSM873943
Primerpool_769_XIC3	N/A	N/A	5C	Nora et al 2012	GSE35721
mESC HindIII	1	mES cell line (J1)	HiC	Dixon et al 2012	GSM862720
mESC HindIII	2	mES cell line (J1)	HiC	Dixon et al 2012	GSM862721
mESC NcoI	1	mES cell line (J1)	HiC	Dixon et al 2012	GSM862722

HiC data were acquired from GEO [1] as SRA files and split using Fastq-Dump [2] (Table 3.1, Barrett *et al* 2009, Wheeler *et al* 2008). Read ends were mapped independently against the mouse genome build 9 using Bowtie [3] and an iterative mapping approach (Langmead *et al* 2009). Reads up to a maximum length of 50 bp were mapped allowing a maximum of mismatch. Uniquely mapping reads were kept and all unaligned reads were clipped by 4 bp from the 3′ end and mapped again. This process

was repeated until all reads were uniquely mapped or the total read length was less than 24 bases. All mapping was done using the Bowtie flags "--phred33-quals --tryhard -m1 -v 1". A read was included in the filtered dataset if both of its ends uniquely mapped to the genome and the total distance from both ends to their nearest downstream RE site was less than or equal to one Kb. In cases where two reads had the same mapping coordinates, only one instance was kept as the other was assumed to be a PCR duplicate.

All annotation data, with the exception of gene expression data, were downloaded from the GEO repository [1] and split using Fastq-Dump [2] (Table 3.2, Barrett *et al* 2009, Wheeler *et al* 2008). Reads were mapped using the same iterative process described above for HiC data. All annotation reads were single-ended. When multiple replicates were available, mapped reads for all replicates were pooled prior to processing.

Table 3.2 List of public annotation datasets.

Sample	Cell Type	Reference	GEO ID	Control	GEO ID
Smc1	mES Cell Line (V6.5)	Kagey et al 2010	GSM560341 GSM560342	Whole Cell Extract	GSM560357
CTCF	mES Cell Line (Bruce4)	Shen et al 2012	GSM723015	Input	GSM723020
PolII	mES Cell Line (Bruce4)	Shen et al 2012	GSM723019	Input	GSM723020
H3K4me3	mES Cell Line (Bruce4)	Shen et al 2012	GSM723017	Input	GSM723020
19 Tissue Gene Expression	mES Cell Line (Bruce4)	Shen et al 2012	GSM723776	N/A	N/A

5C Data Normalization with HiFive

Prior to normalization with HiFive or HiFive-Express, fragments were removed if they had fewer than ten interactions with valid fragments. This process was repeated until

a stable set of valid fragments was found. Normalization using the standard HiFive algorithm was carried out with a learning rate of 0.01 over 5,000 iterations for the burn-in phase and 10,000 iterations for the annealing phase. The distance-dependence function was updated every 100 iterations. For normalization by the HiFive-Express algorithm, 10,000 iterations were used with a distance-dependence function update occurring every 100 iterations.

5C Data Normalization with Alternate Methods

Two other methods of 5C normalization were used in this study. These included the method described by Nora *et al* (2012) and the use of the method HiCLib [5] as adapted for 5C as described by Naumova *et al* (2013).

For the Nora method, all 5C datasets (not just the mouse male ESC replicates) were used to calculate median primer counts and ranges. Any primer whose non-zero median exceeded 2.5 times the interquartile range of medians across all primers and samples was removed. Counts were included if both ends of the interaction were deemed valid. For all included counts, a loess non-parametric regression line was estimated for the relationship between counts and inter-fragment midpoints with an alpha of 0.01 and using R version 3.0.1 [12]. Normalized counts were calculated as the observed count divided by the expected value from the loess regression.

Normalization of 5C data using the modified HiCLib approach differed from that described by Naumova *et al* (2013) in order to account for differences in the nature of the data being processed. In their study, Naumova *et al* (2013) used primers widely spaced across chromosomes and signals were converted from counts to binary states based on whether any observations were made for a primer pair, reducing (but not eliminating) the

influence of the distance-dependence portion of the interaction signal. Because the 5C dataset we analyzed occurred over ranges short enough to have robust and meaningful differences between individual interactions, we maintained counts as numbers. To account for the influence of the distance-dependence at this shorter interaction range, we ran two versions of this normalization approach. The first used raw reads as input to the normalization algorithm. The second divided the counts by the distance-dependent signal estimate generated by the HiFive distance-dependence function on uncorrected reads. In both of our analysis versions the same set of read filters were applied as described under the HiFive methodology. Corrections were calculated over ten iterations. In each interaction, each fragment's mean adjusted count was calculated and divided from all interactions involving that primer. This was done in order, and each update was applied prior to the calculation of the next primer's mean signal.

HiC Data Normalization with HiFive

HiC normalization was performed using the standard HiFive algorithm and *cis* interactions up to a maximum interaction range of 5 Mb. The burn-in and annealing phases were carried out for 5,000 and 10,000 iterations, respectively. The initial learning rate for both phases was 0.01 and the distance-dependence parameters were updated every 2,500 iterations in both phases. Data from both HindIII replicates were pooled and the combined dataset was corrected using the above-described parameters.

HiC Data Normalization with Alternate Methods

In addition to using HiFive, HiC normalization was also performed using HiCPipe [11] (Yaffe & Tanay 2011), HiCNorm [13] (Hu *et al* 2012), and HiCLib [5] (Imakaev *et*

al 2012). For normalization using HiCPipe and HiCNorm, the reads used were identical to those produced using HiFive's read filtering. HiCLib requires mapped read ends as input, so while all methods used the same set of mapped reads, the resulting interaction pairs varied between HiCLib and other methods.

Normalization using HiCPipe was performed using a model optimized for GC content, mappability, and fragment length as described by Yaffe and Tanay (2011). GC content and fragment length ranges were broken into 20 bins and mappability was broken into 5 bins ranging from 0.5 to 1.0. Model parameters were optimized using all valid *cis* interactions.

HiCNorm normalization was performed as described by Hu *et al* (2012). We used a model with parameters for fragment end size (with insert length of one Kb), GC content, and mappability. HiCNorm only learns corrections associated with specific bin sizes so normalization was performed for all *cis* interactions binned at sizes 10 Kb, 25 Kb, 100 Kb, and 1 Mb. *Trans* interactions were also modeled for the 1 Mb binned data. Normally this method utilizes the general linear model method in R. For large genomes and small bin sizes, the number of variables in this approach becomes unmanageable, both in terms of time and computer memory usage. In order to perform HiCNorm across all desired bin size ranges, we modified the method to make use of the more efficient Python package statsmodels [14] instead of relying on R. We confirmed that parameter estimates from both approaches were identical.

Normalization using HiCLib was performed as described by Imakaev *et al* (2012). Reads were loaded with a maximum insert size of one Kb. PCR duplicates and reads with

only one valid mapped end were removed. Iterative correction was performed over 10 rounds and normalized to fragment length.

Annotation Data Processing

Annotation data were processed using Macs2 version 2.1.0 [15], a ChIP-seq peak calling program (Zhang *et al* 2008). When available, samples were processed with an appropriate control. All samples used the settings "-g mm –bw 200 –B". In addition, H3K4me3 was processed using the broad peak flag. Peak calls were used for all analyses with the exception of H3K4me3, which used broad peak calls. Processed gene expression data were downloaded from the Ren lab's Mouse Encode website [16] (Shen *et al* 2012).

Dynamic Binning

Post-normalization, two challenges remain in utilizing chromatin interaction data. Because observing a given interaction is a stochastic event, even normalized data contain noise. Further, in regions where observed interactions are expected with very low frequency, a lack of observed interactions is of limited use. The previous solution to both of these related issues has been to pool expected and observed counts into bins spanning an arbitrary set of size ranges and then calculate the ratio of the sums. While this reduces the signal to noise ratio, it may also break up relevant features between bins and create sets of bins with vastly different information content due to differing occurrence densities and characteristics of restriction fragments. By eliminating the need to select a bin size, resolution can be maintained in areas of high information density without suffering from empty or under-filled bins in data-sparse regions.

HiFive employs a dynamic binning approach that takes an initial partitioning of the data space and adjusts each bin size to meet a user-defined minimum number of interactions (Figure 3.1a). The search space can be limited, possibly resulting in invalid bins for data-sparse regions, or allowed to run until the criterion has been met for each bin. Initial partitioning can be based on fend or fragment boundaries, uniform-sized bins, or arbitrary user-defined bins. Data used when expanding bins can also be at fend- or fragment-level resolution, uniformly sized, or user-defined. In the case of fend- and fragment-resolution data arrays, bins would expand by a single interaction at a time given the non-uniform spacing, whereas a uniformly spaced data array would expand bins in all directions for a bin-size increase of (n + 1) * 4 per round, where n is the current number of steps the new boundary is from the bin. This process results in a more informative overview of the data, both visually and with respect to removing the effects of stochasticity from data-sparse regions (Figure 3.1b).

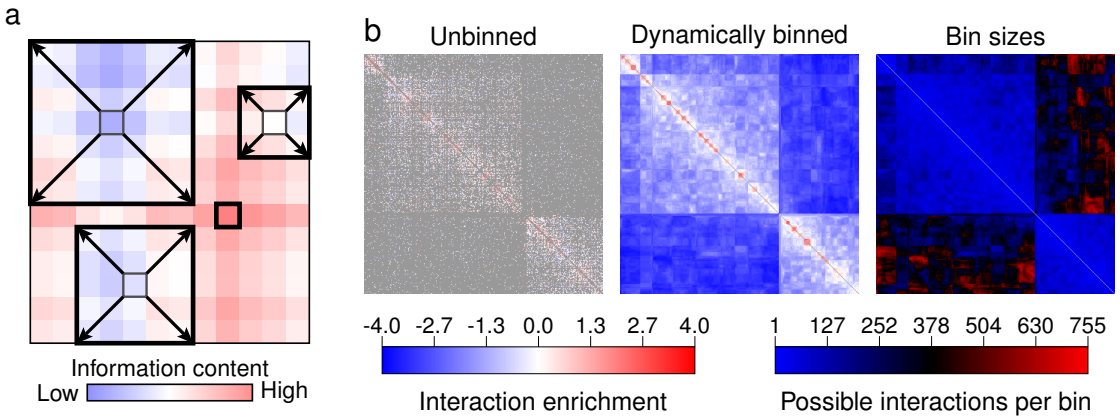

Figure 3.1 Expanding data interpretation using dynamic binning.
Dynamic binning considers each bin individually, expanding its borders in all directions until a user-defined minimum number of observations have been incorporated or a maximum bin size has been reached. This results in different sized bins containing roughly equivalent amounts of data. b) Use of dynamic binning results in a visually more- interpretable representation of the data with reduced stochastic noise while maintaining higher resolution in data-rich regions.

5C Data Correlations with HiC Data

For each method, the logs of all non-zero normalized reads were included in comparison to HiC data for calculation of the correlation. For HiFive, HiFive-Express, and raw reads, in the "distance-corrected" condition the distance-dependent portion of the signal as determined by the HiFive distance-dependence function was divided out of the reads prior to logging (although the distance-dependence was taken into account for all HiFive normalizations) as follows:

$$s'_{i,j} = \frac{s_{i,j}}{D(d_{i,j})}$$

For reads normalized by HiCLib in the "distance-corrected" condition, the distance-dependent signal was removed prior to normalization as described above. Comparisons were made to HiC data normalized with either HiFive or HiCPipe using the pooled HindIII dataset. Pearson correlations were calculated for two different binning strategies of HiC data. The first was a comparison to logged non-zero HiC data binned using coordinates corresponding to the interrogated fragments from the 5C experiment either corrected only for fend bias in the "not distance-corrected" condition or corrected for fend bias and distance-dependence in the "distance- corrected" condition. Because HiCPipe has no way of dealing with distance-dependent signal, estimations from HiFive data were divided out of HiCPipe corrected values prior to logging. Data were also compared to HiC data binned using 5C fragment boundaries and then dynamically binned using fend-level resolution for bin expansion and a minimum of 20 unique observed interaction combinations in each bin with no expansion limit.

HiC Inter-Dataset Correlations

After normalization using each HiC method (HiFive, HiFive-Express, HiCPipe, HiCNorm, and HiCLib), enrichment values were calculated for 10 Kb, 25 Kb, 100 Kb, and 1 Mb bin sizes. For all bin sizes intra-chromosomal interaction enrichments were calculated while inter-chromosomal enrichments were only found for the 1 Mb bin size. Binning was limited to chromosomes 1-19. For each bin size, Pearson correlations were calculated for log-enrichments of all non-zero bins between the HindIII pooled dataset and the NcoI dataset. Correlations were calculated for subsets of data including interaction distances of zero to a series maximum distances ranging from five times the bin size (e.g. 50 Kb for the 10 Kb binned data) up to the maximum interaction distance (~194 Mb) and broken into 11 equal-sized log steps, inclusive. Inter-chromosome log-enrichment correlations were calculated across all non-zero 1 Mb bins.

Calculating the boundary index

One approach that has been useful in marking structurally significant features in chromatin conformation data has been identifying shift-points where interactions move from one set of high-interacting partners to another. Dixon *et al* (2012) described a statistic called the directionality index (DI) that measured the difference in overall interaction strength for upstream versus downstream interactions with a set of fragments. This yields a positive or negative value, depending on the bias. Boundaries are then called using a hidden Markov model to determine transition points from upstream to down stream bias in the data. While this has proved useful for identifying what they labeled as topological domains, we find that this approach has limitations in identifying

smaller structures and boundary features nested within larger ones. In order to investigate these features, we have devised a variant of this statistic called the boundary index (BI).

The boundary index functions by capturing shifts in interaction partner preference without the assumption that such preferences are up or downstream. This is accomplished at each RE site by taking fends up and downstream of the site falling in equal sized intervals ("width"), calculating the log enrichment of interactions between fends in these intervals with groups of fends up and downstream of the site grouped into a specified size intervals ("height"), up to some maximum distance from the site ("window", Figure 3.2a). The BI for that site is the mean absolute difference between enrichments for interactions with fends in the upstream width versus the interactions with fends in the downstream width. Thus for boundary point *P* and width *W*, we define fends in sets *I* and *J*:

$$I = \{i: P - W \leq i < P\}$$

$$J = \{j: P \leq j < P + W\}$$

These fends interact with fends within a distance defined by the window upstream and downstream of *P* and divided into equal-sized intervals defined by the height, *H*, such that they make up *N* sets (the number of height-sized bins on one side of *P*) denoted by *K* and *M* for upstream and downstream sets, respectively:

$$K_n = \{k: P - Hn \leq k < P - H(n-1)\}$$

$$M_n = \{m: P + H(n-1) \leq m < P + Hn\}$$

Thus, the *BI* for point *P* is:

$$BI_p = \frac{1}{s|N|}\sum_{n}^{N}\left[\left|\log\left(\frac{\sum_k^{K_n}\sum_i^{I|i>k} s_{k,i}}{\sum_k^{K_n}\sum_i^{I|i>k} E_{k,i}}\right) - \log\left(\frac{\sum_k^{K_n}\sum_j^{J} s_{k,j}}{\sum_k^{K_n}\sum_j^{J} E_{k,j}}\right)\right|\right.$$

$$\left. + \left|\log\left(\frac{\sum_m^{M_n}\sum_i^{I} s_{i,m}}{\sum_m^{M_n}\sum_i^{I} E_{i,m}}\right) - \log\left(\frac{\sum_m^{M_n}\sum_j^{J|j<m} s_{j,m}}{\sum_m^{M_n}\sum_j^{J|j<m} E_{j,m}}\right)\right|\right]$$

A user-defined minimum number of bins must be included in the calculation or *BI* is not found for point *P*. A set of BI values can then be smoothed to reduce noise and enable better peak calling using a Gaussian smoother (Figure 3.2b) such that for all BI positions within a distance of 2.5 R of *P*, defined as set *T*, the smoothed value BI'_p is defined as:

$$BI'_p = \frac{\sum_t^T BI_t e^{\frac{-(p-t)^2}{2R^2}}}{\sum_t^T e^{\frac{-(p-t)^2}{2R^2}}}$$

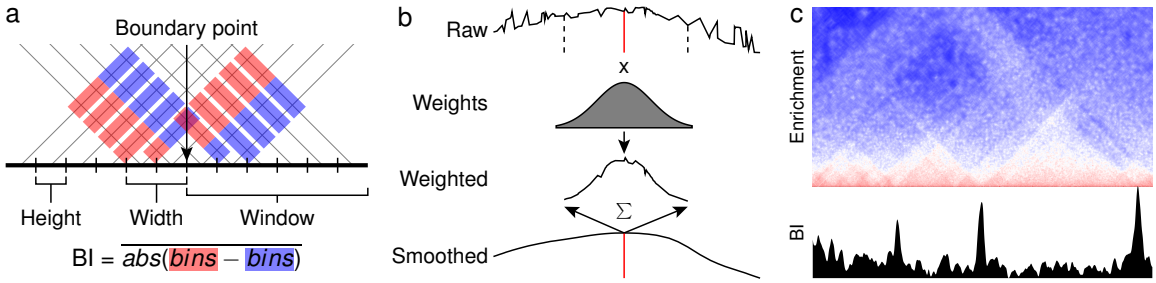

Figure 3.2 The boundary index statistic.
a) The BI is found by examining the difference between paired sets of reads from the red group versus the blue group. b) Raw scores are smoothed using a set of Gaussian weights. The raw score indicated by the red line is transformed by summing the product of the raw scores between dashed lines and the Gaussian weights, resulting in a smoothed set of scores. c) Smoothed BI scores shows peaks associated with transition points between domains and between subdomains.

Boundary index comparison to the directionality index

The BI scores were calculated for the HindIII pooled dataset and NcoI HiC dataset using a width of 100 Kb, a height of 100 Kb, and a window size of 500 Kb. BI's found with less than 10 valid paired bins were excluded. BI scores for the two datasets were combined and the joint set was smoothed using a Gaussian smoothing with a standard deviation of 10 Kb and a maximum range of 25 Kb. BI scores for each chromosome had the mean BI score across the chromosome subtracted and then all BI scores were divided by the genome-wide standard deviation. Peaks were called using the joint smoothed and adjusted set of BIs and defined as BI scores that had lower scores immediately adjacent to them and were at least 1.9 standard deviations greater than the adjacent trough on at least one side.

To determine how BI peak calls performed against DI boundary calls from Dixon *et al* (2012) [17], we began by assessing the similarity of calls. We defined peaks from the two datasets as overlapping with each other if they occurred within 40 Kb of each other, i.e. one DI bin. The numbers of overlapping boundaries differ between the two sets as both sets contain a small number of calls occurring at or less than 80 Kb apart such that two boundary calls in one set could straddle a peak from the other dataset.

We also found density profiles for all boundary calls, partitioned by whether they had a corresponding peak in the other dataset, across several features known to be associated with TDs. These features included CTCF, Smc1 (part of the cohesin complex), H3K4me3, TSSs, and Pol2. Feature densities were determined from annotation peak midpoints (or TSS coordinates) and binned into 10 Kb bins extending 500 Kb up and downstream of each boundary call.

Three-dimensional chromatin modeling

By using one of HiFive's various binning options and filling in bins with no observed reads (and increasing the reliability of bins with few observations) using dynamic binning, HiFive provides an easy, fast, and intuitive way to approximate the consensus 3-dimensional conformation of chromatin from the two-dimensional array of interaction data generated using either the HiC or 5C assay. By disabling search-space limitations, HiFive guarantees the production of a completely filled matrix of values. By taking these values as *n* examples with *n* features, we are able to make use of simple dimensionality-reduction approaches such as principal component analysis (PCA) to find sets of three coordinates to best describe the bin-similarities. To do this efficiently, we make use of the fast PCA function within the Python machine learning package mlpy [18] to compute only the first three components, enabling us to apply this to higher-resolution data for more detailed models without prohibitive computing time.

BI peaks for partitioning individual chromosomes were found using the BI scores from the combined BI dataset described above. BI scores were smoothed using a Gaussian smoothing with a standard deviation of 5 Kb up and downstream. Peaks were called with no minimum height cutoff. After data from the HindIII pooled HiC dataset for each chromosome were binned using the BI peaks to partition them, the data were dynamically binned using fend-level resolution for bin expansion up to a minimum of 15 unique interaction pairs and no maximum distance cutoff. Log enrichments were found for each bin. Prior to modeling, the bins on the matrix diagonal (fend self-interactions) were set equal to the highest enrichment value. Modeling was done using the PCA-fast algorithm from the Python package mlpy finding the first three component estimates.

Estimated model signal strengths were calculated as the negative log values of inter-coordinate distances. Signal strength based on the distance-dependence curve was also calculated using the peak partitioning of each chromosome to bin all chromosome estimated distance-dependent values. Pearson correlations were calculated for all off-diagonal bins between the dynamically binned signal and both the model signal estimates and the distance-dependent signal estimates.

BI peaks for partitioning of the whole genome were found using the BI scores from the combined HiC BI dataset as described above. Smoothing was done using a Gaussian smoother with a standard deviation of 10 Kb. Peaks were called using a peak-height to trough cutoff of 0.3. Data from the HindIII pooled HiC dataset were partitioned across all counts, *cis* and trans, for chromosomes 1-19 using the BI peaks. The binned data were then dynamically binned using the BI-partitioned binning for bin expansion until all bins had at least 15 non-zero interaction counts with no maximum distance. Log enrichments were calculated for each bin and the diagonal bins were set equal to the highest enrichment value. All *cis* interactions were downscaled by a factor ten. Modeling was done using the PCA-fast algorithm from the Python package mlpy finding the first three component estimates. Estimated model signal strengths were calculated as the negative log values of inter-coordinate distances. The Pearson correlation was calculated across all *trans* bins.

Modeling was also performed using data binned by fixed intervals. HindIII pooled HiC data were binned using 25 Kb bins for each individual chromosome and 120 Kb bins for whole genome binning. The resulting binned data were then dynamically binned and modeled using PCA as described above.

Correlations between models created by different binning methods were calculated as follows. A set of all bin midpoints was compiled for the pair of models. For each model, coordinates for midpoints originating from the opposing model were inferred based on the nearest up and downstream midpoints from the target model and a weighted coordinate mean was calculated based on the relative distance to each bounding midpoint. Once coordinates for all midpoints on both models were calculated, distance matrices for all pairwise combinations were calculated and the Pearson correlation was determined between the matrices.

Model rendering was done using Blender version 2.69 [19]. For each set of coordinates, a node was placed and connected to adjacent nodes via a straight-line segment. Each line segment midpoint was given a thickness equal to the square root of the ratio of sequence distance between bin midpoints divided by the coordinate distance between the nodes bounding the line segment. Transitions between thicknesses were linear and the line thicknesses from the first and last line midpoints to the first and last nodes were constant.

Genes were positioned along the spatial models according to their TSS coordinates. Genes with TSSs occurring before the first bin midpoint or after the last bin midpoint were excluded. TSS spatial coordinates *T* were calculated as follows for a TSS with genomic position *P* that falls between bin genomic midpoints *M* with coordinates *C* for bins *i* and *j*:

$$T = \frac{c_i(M_j - P) + C_j(P - M_i)}{M_j - M_i}$$

Calculating gene spatial arrangements

Genes were placed into one of 101 bins based on expression level. One bin contained all genes with no observable transcripts and the remaining 100 bins divided the range of observed fragments per Kb of exon per million fragments mapped (FPKM) values into equal-sized groups. In assessing intra-chromosomal gene spatial arrangement, for each pair of bins the log ratio of the sum of spatial distances for all *cis* gene-pair combinations between bins over the sum of sequence distances for all *cis* gene-pair combinations between bins was calculated. Inter-chromosomal spatial arrangements were calculated for each pair of bins as the mean log distance between all *trans* gene-pair combinations between bins. Because there was no reference to scale model coordinates to, all distance values represent relative distance differences.

Results

5C Method Comparison

We evaluated HiFive's 5C normalization performance by examining consistency between 5C data and HiC data for the same corresponding region and cell type. At the same time, we compared HiFive against two other methods, the approach described in Nora *et al* (2012), a variation of the approach using HiCLib as applied to 5C data described in Naumova *et al* (2013), and raw 5C data. The resulting normalized values for each method, along with raw 5C data, were compared to HiC data covering the same region normalized using either HiFive or HiCPipe, both using normal and dynamic binning to account for the lower coverage in the HiC dataset. For the comparisons to the Nora *et al* approach, the distance-dependent signal was removed from both HiFive and HiCPipe-normalized HiC data prior to comparison as the Nora et al approach necessarily removes this portion of the signal. HiFive and HiCLib results were tested both with and without the distance-dependent signal portion present.

Normalizations performed using HiFive showed improved correlation between HiC and 5C data compared to alternative methods and raw 5C signal across both replicates when data were dynamically binned, regardless of HiC normalization method (Figure 3.3a). HiFive-Express also performs well when comparing 5C data against dynamically binned HiC data. Further, in the absence of the distance-dependent portion of the signal, HiFive-Express still performs well, regardless of the HiC correction method. It is unclear, how valid comparisons to the normally binned HiC are, given the extreme differences in coverage between the two types of assays (Figure 3.3b). The dynamically binned HIC, though, clearly recapitulates all of the same structural features

seen in the 5C data, suggesting it is a better standard for validation against. Visually, the biases associated with individual fragments are clearly apparent as striations in the raw data. After correction with HiFive and HiFive-Express, these striations are almost entirely absent. Other methods still show marked striping indicative of incomplete fragment-bias removal.

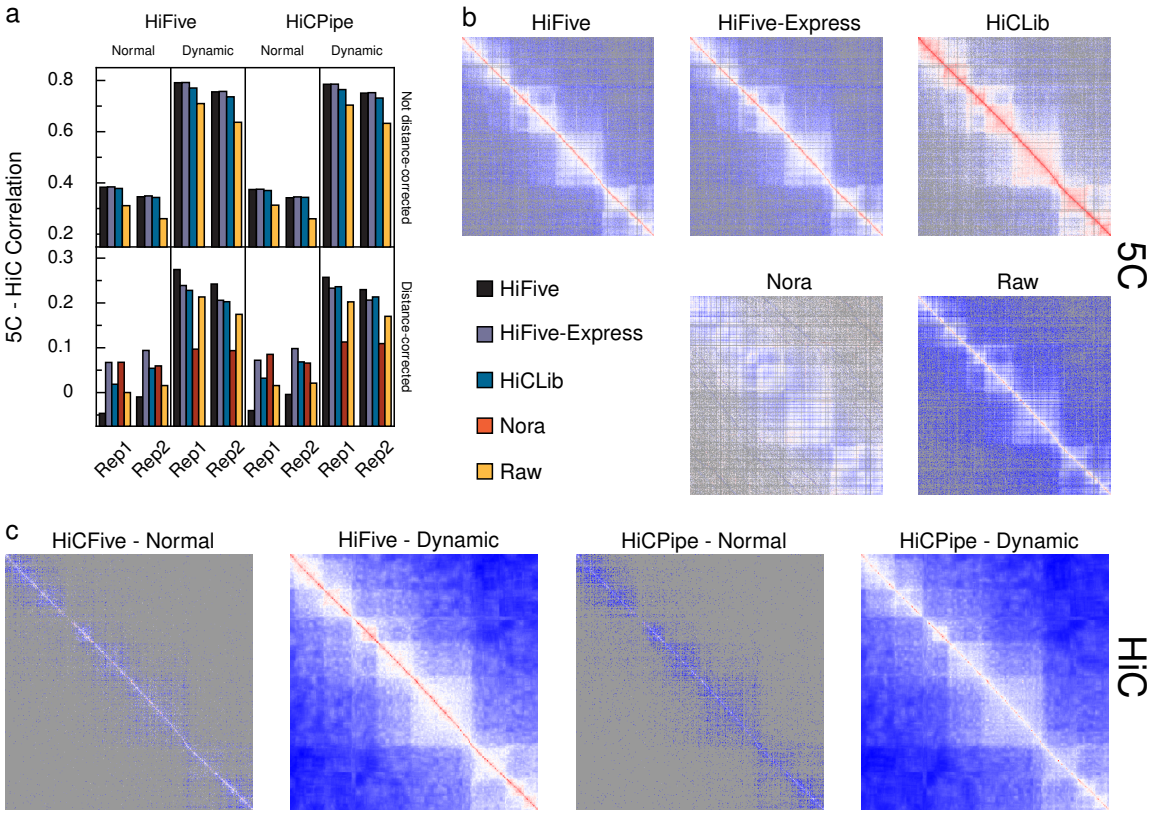

Figure 3.3 5C data normalization and correlation with corresponding HiC data.
5C data for two mouse ESC replicates normalized using several methods and correlated with correspondingly normalized and binned HiC data. a) Methods in the upper panel were compared without removal of the distance-dependent portion of the signal, whereas the lower panel shows the correlations in which the distance-dependent signal was removed. b) Visualization of normalized and logged 5C counts. Colors are scaled to maximize the dynamic range, with blue corresponding to the lowest counts, red to the highest, and white to the midpoint between the two. Gray denotes interactions where no reads were observed. Rows and columns correspond to forward and reverse probes, respectively. c) HiC data corrected using HiFive or HiCPipe and binned using the same

boundaries as the 5C data. Coloring is as described for b. Dynamically binned HiC data show bins sized to include 20 interactions per bin.

HiC Method Comparison

To assess the effectiveness of HiFive's HiC normalization compared to other methodologies, we examined correlations of corrected data across datasets produced using different REs. To do this, normalized data were generated for each dataset using HiFive, HiFive-Express, HiCPipe (Yaffe & Tanay 2011), HiCLib (Imakaev *et al* 2012), and HiCNorm (Hu *et al* 2012). Data were binned at four resolutions, 10 Kb, 25 Kb, 100 Kb and 1 Mb, and inter-dataset correlations were calculated across a series of maximum interaction distance ranges for *cis* interactions. *Trans* interaction correlations were also determined at the 1Mb resolution.

At the 1 Mb and 100 Kb resolutions, HiFive showed superior performance to all other methods across all interaction (Figure 3.4a). HiCPipe performed nearly as well at these lower resolutions, followed by HiCNorm. HiFive-Express performed equal to or better than HiCPipe using 100 Kb bins and just slightly worse than HiCPipe using 1 Mb bins across all interaction distance ranges. Neither HiCNorm nor HiCLib matched the performance of HiFive, HiFive-Express, or HiCPipe at the 1 Mb bin size. At the 100 Kb bin size, HiCLib was below the other methods and for all but the shortest interaction range cutoffs, HiCNorm performed nearly identically to HiCPipe. At 25 Kb, HiFive and HiFive-Express outperformed other methods, though HiFive-Express actually showed better inter-dataset correlations for show range interactions than the standard HiFive normalization. At the 10 Kb resolution, HiFive showed a slight performance dip relative to HiCPipe and HiFive-Express when comparing interactions under 200 Kb apart, though

HiFive-Express still outperformed HiCPipe across all ranges and both showed better performance than HiCNorm and HiCLib. Including interactions with ranges greater than 200 Kb, HiFive showed better agreement between datasets than HiCPipe.

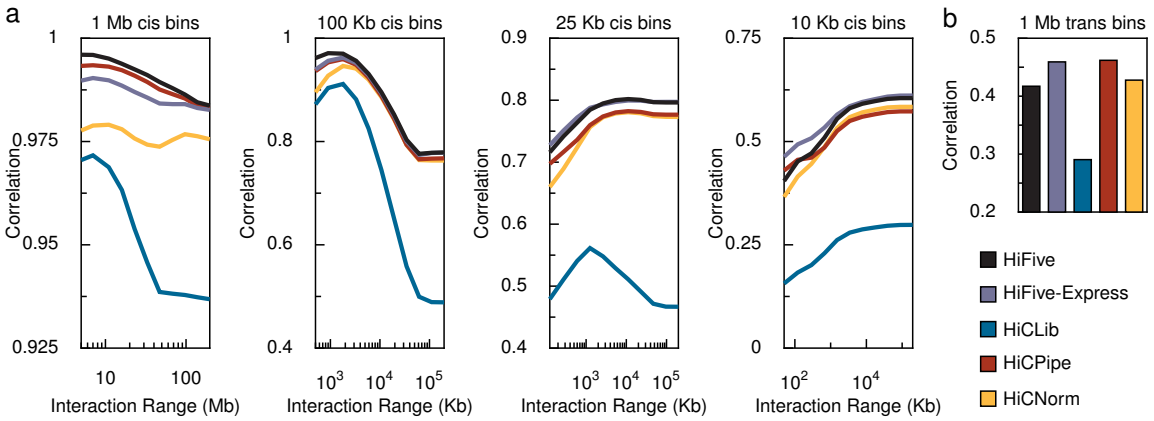

Figure 3.4 HiC normalization and inter-dataset correlation.
a) HiC data from mouse ESCs produced using HindIII and NcoI were normalized using a number of different analysis methods and intra-chromosomal interaction correlations were compared across a range of bin sizes and maximum interaction distances between the two datasets. b) Correlation between 1Mb-binned inter-chromosomal interactions across datasets after normalization using a variety of methods.

Although they were designed with short-range interactions and high-resolution analysis as their target purpose, HiFive and HiFive-Express still performed relatively well in normalizing *trans* interactions (Figure 3.4b). Interestingly, HiFive-Express performed better on this measure than HiFive, showing performance just slightly below that of HiCPipe.

The boundary index captures more significant features than the directionality index

To verify the performance of our boundary index statistic, we began by using a cutoff yielding 3,028 peaks and comparing these to the 3,051 TD boundaries generated

by Dixon *et al* (2012) across chromosomes 1-19. We found that there was a high amount of overlap between the two sets of boundary calls with more than 60% of the DI-based boundaries in regions with BI coverage falling within 40 Kb of a BI-based peak (Figure 3.5b). We partitioned boundaries into overlapping and unique and then found occupancy profiles for them based on CTCF, Smc1, H3K4me3, and PolII occupancy sites as well as TSS locations. Occupancy data were found up and downstream of each and binned at 10 Kb intervals. In sites that were considered equivalent between the two methods, we found nearly identical profiles. At the boundary sites common to both methods we saw an increase in signal for BI called sites, which we attribute to finer-scale positioning of the boundary. This increase in signal was particularly evident at transcriptionally associated features including TSSs, H3K4me3, and Pol2. Across sites unique to each method, the general profile shape was similar. Like the overlapping sites, though, we saw a stronger signal for boundary sites generated using the DI method, particularly at the boundary site itself. Interestingly, we saw an increased background signal in unique BI-called sites compared to unique DI boundaries. This may be indicative of BI-called sites having a higher sensitivity for finding structural features occurring with domains as a higher background suggests additional structure features are occurring in relatively close proximity to these sites.

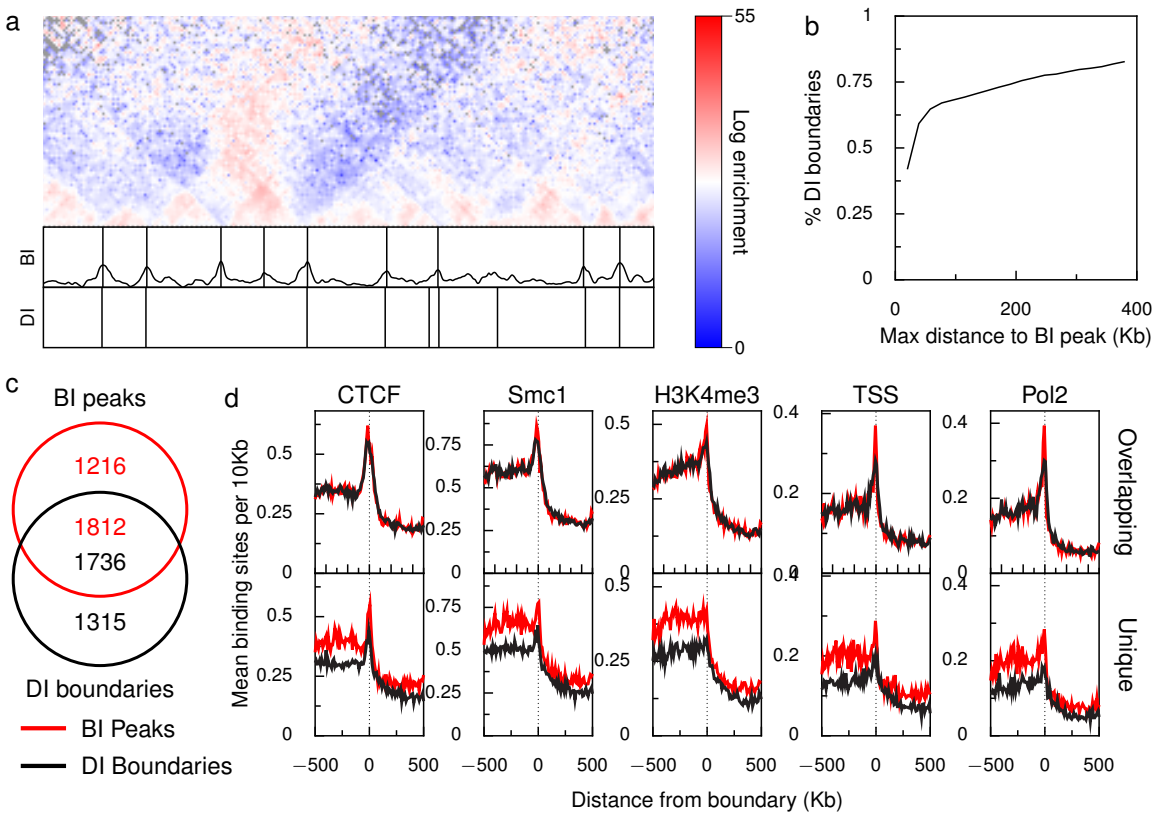

Figure 3.5 Boundaries identified using boundary index scoring and associated signals. Boundaries identified from peaks in BI signal across scores pooled from two sets of HiC data compared to topological domain boundaries found using DI. a) Interaction enrichment signal for a 5 Mb stretch of chromosome 19 and its associated BI signal, BI peaks, and DI domain boundaries. b) The percentage of DI boundaries that have a BI peaks within a given window size. c) Overlap of BI peaks and DI boundary sets using a 40 Kb cutoff for defining overlap. d) Frequency of annotation data peaks across a 1 Mb window centered on each boundary or peak and binned every 10 Kb.

Three dimensional chromatin models

The performance of HiFive's modeling strategy was assessed by determining the ability of its structural modeling approach to reproduce the fend-corrected interaction signal based on coordinate-distances alone for individual chromosomes as well as across the whole genome simultaneously. Two segmentations were used to create the distance matrices used for coordinate estimation. First, BI scores were calculated at fend-level

resolution and used to call peaks representing structural boundaries. These peaks were used to segment chromosomes into a set of bins containing fends of similar interaction patterns. Second, data were binned using a fixed bin width of 25 Kb and 120 Kb for individual chromosomes and the entire genome, respectively. These sizes were selected to yield a comparable number of bins to the BI peak-based binning approach. For both binning methods, binned fend-corrected interaction counts were calculated and dynamically binned with no bin expansion cutoff, producing complete interaction signal matrices. To account for the relative differences in signal reproducibility, and therefore intensity, between intra- and inter-chromosomal interactions in the whole genome modeling (Nagano *et al* 2013), *cis* interaction values were down-scaled by a factor of ten. For whole genome correlation calculations, only inter-chromosomal interactions were included. Because of poor coverage, poor inter-chromosomal signal, and structural divergence between homologous chromosomes, the X chromosome was excluded from modeling. Using the mlpy fast-PCA algorithm, coordinates were calculated and negative log-distances between coordinates were found for the model distances. Finally, correlations were calculated for the log signal of the dynamically binned data with both the model and binned distance-dependent signal estimates from the HiC distance model.

Partitioning chromosomes by 30 Kb fixed-width bins yielded 80,460 bins covering approximately 2.41 Gb of the genome. This was on par with partitioning using BI peaks, that resulted in 79,812 bins covering about 2.41 Gb of the genome with partition sizes that were approximately log-normally distributed around a mean of 30 Kb and with medians ranging from 22.7 Kb to 24.2 Kb (Figure 3.6a). The whole genome was partitioned into 20,126 bins using 120 Kb fixed width bins compared to 20,508 bins with

a mean partition size of 118 Kb and a median of 108 Kb using BI peak partitioning. The advantage of using the BI peak approach for partitioning instead of uniform-size binning is that it reduces the number of structural features that straddle bin boundaries, a problem that can either have the effect of washing out relevant signal if features are small or over-representing features across multiple bins, resulting in too-high a significance being placed on interactions associated with those features.

For individual chromosomes, interaction signals estimated solely from the distance-dependence estimates show striking divergence between fixed width and BI peak-defined binning with fixed width binning yielding correlations of 72.2-85.5% (78.7% mean) with dynamically binned data (Figure 3.6a). This was significantly better than the correlations of 47.6-59.8% (52.4% mean) from BI binned distance only estimates.

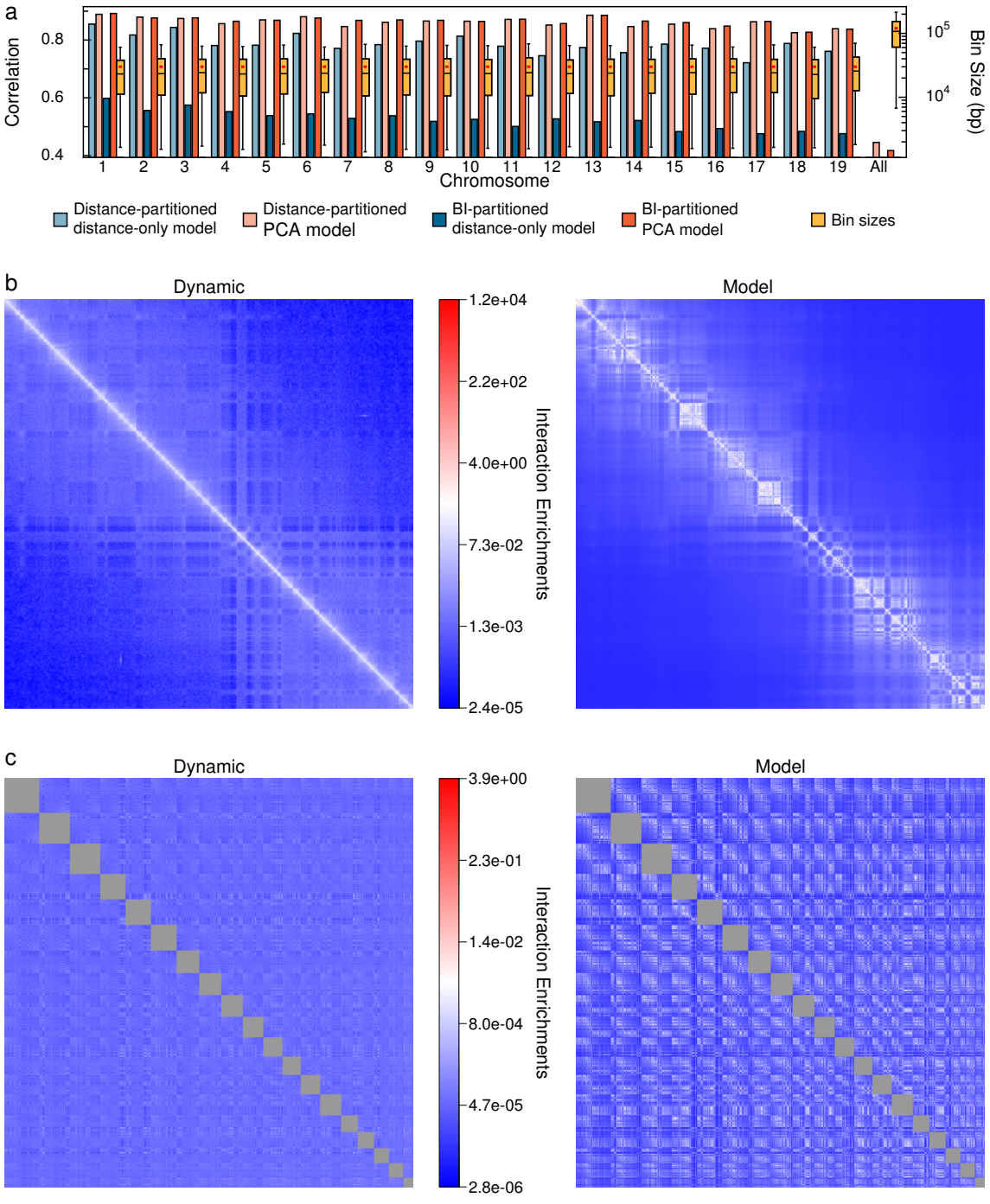

Figure 3.6 PCA-based chromosome and genome modeling.
Three-dimensional models were generated using fast-PCA for each chromosome from dynamically binned data partitioned using either fixed with bins or BI peak-defined bins. A whole genome model was also made using the same strategies including inter-chromosomal interactions and more stringent BI peaks. a) Correlations were calculated between the dynamically binned signal and the inverse log-distance between model bin locations as well as expected signal based only on sequence distance. Ranges of bin sizes resulting from BI partitioning are also shown. Red points indicate the fixed width bin

size. b) Dynamically binned signal for mouse chromosome 13 and the corresponding model predicted signal are shown. The model has been linearly rescaled to best match the observed data. c) Dynamically-binned inter-chromosomal signal for chromosomes 1-19 and the corresponding expected model signal. The model data are linearly rescaled.

Regardless of binning method, PCA-based models performed on par with each other and better than distance-only models. Fixed width binning produced models that correlated with dynamically binned data from 82.5-88.9% (86.1% mean) compared to BI-peak based binning whose correlations ranged from 82.7-89.2% (86.5% mean). Given that these HiC data represent a non-synchronous population of cells, this level of explanatory power at this resolution is excellent and enables the possibility of hypothesis testing on single distance values informed by a large set of interaction data instead of the small number of interaction counts around the interacting points of interest. Visually, it is clear that using this approach accounts for many of the larger long-range features that appear within the intra-chromosomal interactions (Figure 3.6b).

The whole genome model had inter-chromosomal signal correlations of 44.5% and 41.7% for fixed width and BI-peak binning, respectively. In light of the variable nature of nuclear chromosome structure (Nagano *et al* 2013), there are clearly still significant reproducible structural features and relationships between chromosomes. The signal produced by the whole genome structural model recapitulates the dominant signals for each interacting chromosome pair (Figure 3.6c) suggesting that the model is capturing spatial relationships shared by a significant portion of the cell population for some non-trivial period of their existence.

The PCA-based modeling approach, when coupled with dynamic binning for complete distance matrices shows a high degree of robustness. Comparisons between

distance matrices for the two alternate binning approaches showed extremely high correlation with individual chromosome model correlations ranging from 98.27-99.97% (99.67% mean) and 99.39% correlation between whole genome models. This consistence in resulting models is also apparent when directly comparing the shaped and structures produced in 3D renderings of the chromosomes (Figure 3.7a).

Examining the physical shape of the modeled chromosomes, two distinct features are evident both in the individual and whole genome models. The chromatin appears to exist predominantly as a combination of two different states: tight, dense coils occupying a small volume of space, and de-condensed stretches looped or folded back and forth creating accessible but still fairly low-volume compartments of chromatin (Figure 3.7b and d). In both individual and whole genome models, there appears to be a polarization of states with the tightly coiled condensed chromatin orienting in the same general direction as can be seen in the top portion of Figure 3.7b and d, suggesting either a common characteristic or external organizer restricting the spatial arrangement of these condensed domains. While we feel confident in the organization of small structures suggested by this modeling approach, we do acknowledge that it is based on data produced from a heterogeneous population and a mix of interactions from pairs of homologous chromosomes. That being said, the smaller structures are likely to be stable and reproducible, whereas the whole chromosome shape or relative positioning amongst chromosomes is an amalgam of configurations representing general organizational trends that likely do hold for large portions of the assayed cell population.

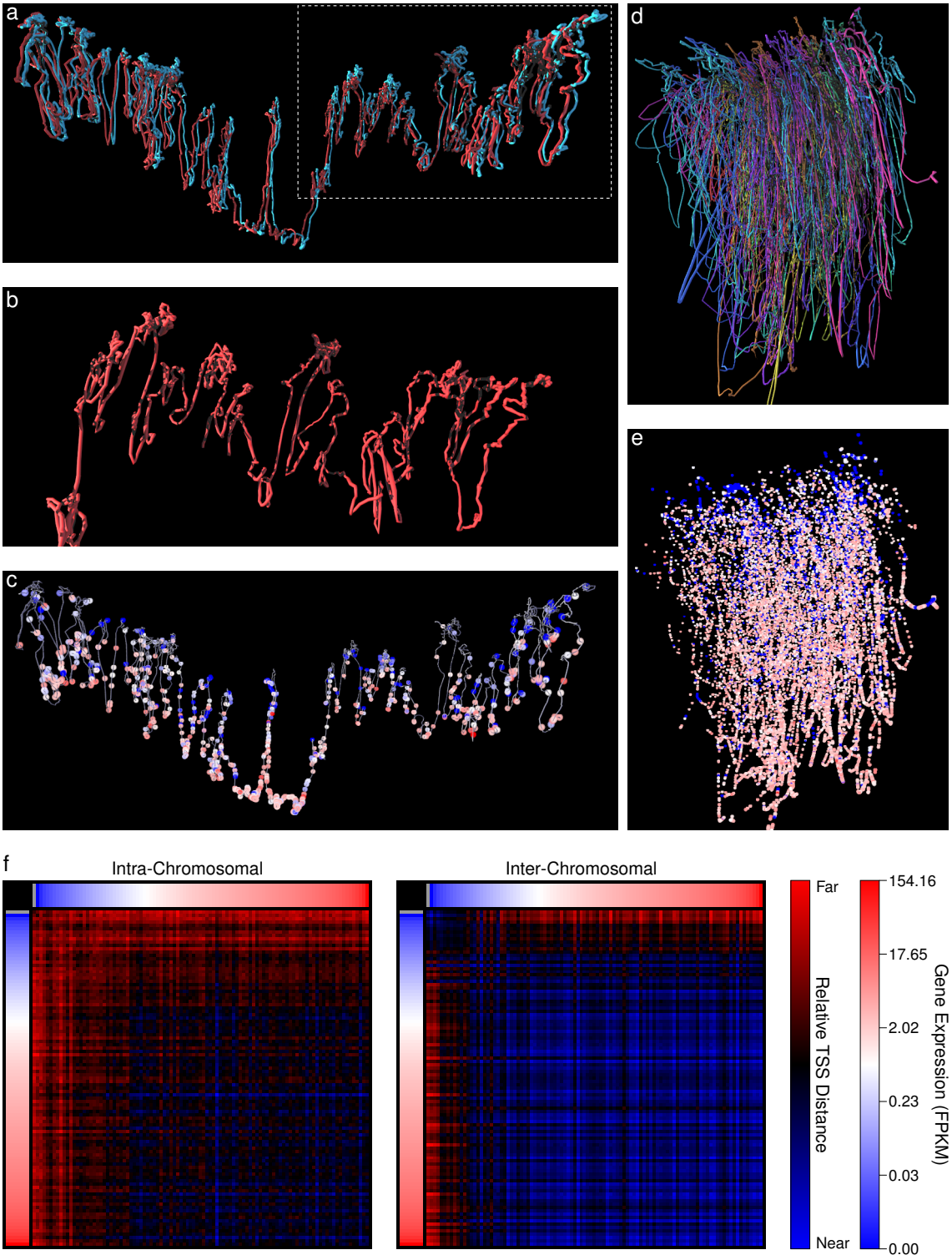

Figure 3.7 Spatial partitioning of genes by transcriptional activity.
a) A rendering of the chromosome 13 models showing the fixed width bin model in blue and the BI-peak bin model in red. The dotted white line indicates the section blown up in subfigure b. Strand thickness is proportional to the square root of the ratio of physical spacing of the points versus the sequential distance, such that a thicker strand when a

longer sequence length occurs between closer 3D endpoints. b) A close-up view of the BI peak bin model. c) A skeletal rendering of chromosome 13 with points indicating gene TSSs. Genes are color coded by expression level in FPKMs. d) A whole genome model rending chromosomes rendered in different colors and proportional strand thickness. e) All mouse genes in the same physical configuration as in panel b and colored according to expression levels. f) Mean relative distances between gene TSSs, binned by expression levels (top and left of plots). All non-observed genes are in a single bin and colored gray in the mean expression scale. Intra-chromosomal distances were determined by calculating the log-ratio of the mean physical distance over the mean sequential distance. Inter-chromosomal distances are simply the mean log-physical distances between TSSs.

Spatial partitioning of genes by transcriptional activity

In order to assess the interplay between physical conformation of the chromatin and transcription, we used expression data from Shen *et al* (2012), placing gene TSSs in their approximate locations in space according to our physical models. TSS locations were determined using sequence midpoints of the bin partitions such that a TSS was placed on the line between the coordinates of the closest up- and downstream partition midpoint with the ratio of physical distances to the two partition coordinates equal to the ratio of sequence distances. Genes whose TSSs fell outside the first and last partition sequence midpoints were excluded from these analyses.

The physical placement of gene TSSs within the modeled chromatin, both on an individual and whole genome basis, shows a polarization of gene clusters corresponding to expression levels. In Figure 3.7c, genes with few or no detectable transcripts are seen clustered together and oriented towards the top of the figure, whereas transcriptionally active genes are located on less condensed stretches of chromatin and tend towards the opposite region of chromosome-occupied space. This trend is also visible in the whole genome model, with the majority of low expression genes presenting in a sheet across the top of the model and active genes existing on parallel strands extending away from this

sheet and converging in space (Figure 3.7e). Interestingly, the highly compact coils appear mainly devoid of genes and are flanked by groups of inactive genes.

To quantify this spatial organization, we examined the physical spacing of TSSs with respect to each other partitioned by expression level. For intra-chromosomal organization, we used the individual chromosome models and partitioned genes into 100 equal-sized sets based on expression level (FPKM), with all non-observed genes in an additional bin. For each pair of bins, the log-ratio of the sum of all pairwise intra-chromosomal gene physical distances over the sum of all intra-chromosomal gene sequence distances was calculated. Physical distances were rescaled to have the same standard deviation as the sequential distances for each chromosome prior to summing. Because physical distances are not calibrated to actual units, the results show only relative distance relationships between expression bin pairs. A similar approach was used for inter-chromosomal gene spacing using the same binning of genes. However in this case only inter-chromosomal gene pairs for each pair of bins were considered and the average log physical distance was calculated using the whole genome model.

Using this approach we find that expression level is predictive of gene spacing such that low expression gene TSSs are spaced further apart from each other and from more highly expressed genes, given the intervening sequence between them, compared with pairs of more highly expressed genes (Figure 3.7f). This holds true for both intra- and inter-chromosomal gene spacing, suggesting that not only are transcriptionally active genes occupying a separate space from less or non-transcribed genes but also that active genes are being brought together in a way that silent genes are not. It is unclear whether this is an active shepherding process or a consequence of the genes' transcriptional

activity but these observations hold with other experimental data about the co-localization of co-expressed genes (Rieder *et al* 2014, Zhang *et al* 2013, Zhao *et al* 2014). In addition, although clustering of inactive genes is not seen in individual chromosomes, we do see shorter distances amongst inactive or very low expression genes.

Discussion

We have presented HiFive, a comprehensive framework for handling and analyzing structural data from 5C and HiC experiments. In addition to offering a robust and highly effective approach to signal normalization, HiFive enables users to achieve similar results in only a fraction of the time with an approximating algorithm that performs with comparable results.

Currently, limited attention has been paid to 5C data and particularly assessing the quality of processing the data prior to interpretation. The work presented in this study demonstrates not only the difficulty in assessing the quality of a 5C data analysis, but the specific shortcomings of comparison to binned HiC data without regard for coverage differences. By using an adaptive binning approach like dynamic binning, we demonstrate a way to maintain the fine scale resolution of the 5C assay while independently verifying the insights it provides. Further, we are able to show that HiC provides more information about structural features than is apparent from unbinned or naively binned HiC data.

In comparison to other HiC methods, HiFive demonstrates a superior approach, at least as measured by agreement between alternative restriction enzyme datasets. Like with the comparison to 5C data, this assessment is limited by the uncertainty of unobserved interaction pairs. We suspect that HiFive would show further improvements over other approaches, including HiFive-Express, if something like dynamic binning were applied allowing a greater influence of methods' abilities to predict low and unobserved counts on the performance assessment.

Discussion of both assay types highlight a feature unique to HiFive that we feel is of particular importance, its ability to dynamically bin data. In the context of 5C data, the ability to interpolate missing data is often crucial. This is the result of the assay's ability to only query fragment pairs that are associated with opposite-orientation primers. Because of this, the maximum possible coverage of possible fragment pairs is 50%. Dynamic binning offers the opportunity to fill in bins devoid of observed data making use of the relative amounts of information in the surrounding region. This feature is even greater relevance when addressing HiC data. The large number of possible interactions across the genome results in large numbers of missing data points. As demonstrated by the distance vs. signal relationship in Figure 2.5b, the mean counts for interaction pairs drop below one at ranges further apart than 10 Kb. In addition, because of the high variance in counts, the reliability of any given count is small, especially at low counts. This has two related consequences. First, there is a highly variable observed data density that drops off rapidly as interaction distances increase. Second, in regions of sparse interaction observations any counts appear as large enrichments when considered individually or at a high resolution of binning. Dynamic binning addresses these by allowing variable resolution to ensure that each bin contains enough observed reads to make a confident assessment of interaction strength while retaining fine-scale detail in areas of high observed read density.

Applying these data into a broader context requires the ability to connect structural features to other genomic annotations, something we've endeavored to do with the creation of the boundary index. Unlike other work, which gives binary states indicating the presence of absence of a boundary (Dixon *et al* 2012, Filippova *et al*

2014), we have presented a statistic that gives a continuous signal indicating the strength of structural similarity across sites. In addition to having tunable parameters allowing the adjustment of feature scale the statistic is sensitive to, the BI can be used in conjunction with a tunable cutoff allowing targeting of the types of transition features of interest. Further, one can change from a boundary-centric view and focus on BI strength associated with DNA binding sites or other genomic features to allow exploration of new structurally relevant factors.

In addition to creating a one-dimensional interpretation of the distance matrices produced by these assays, we have also presented an approach to move in the opposite direction and expand into a three-dimensional interpretation of the data. As demonstrated by the lack of sensitivity to binning approach, using PCA on a completed distance matrix provides a robust way of approximating feature structures and relationships. Clearly this is limited by the heterogeneous nature of the data, though using a hierarchical approach, modeling domains and proximal inter-domain relationships is an enticing approach for discovering underlying structural drivers of cellular function. Further, the gene spatial organization suggests that this modeling approach has potential to confirm, if not yield new insights into genome spatial organization.

Current understanding of gene spatial organization suggests that most Pol2-dependendent genes are found in clusters known as transcription factories. It is not surprising that our modeling finds actively transcribed genes in closer association with each other than low or inactively transcribed genes (Ghamari *et al* 2013). This extends across chromosomes, suggesting that the creation of such transcribed gene foci are not driven only by sequence proximity but by transcriptional state, a concept supported by

the literature, at least in stem cells (de Wit *et al* 2013). We also see that inactive genes cluster at the genome scale, consistent with observations of polycomb bodies.

Taken together, our work shows not only the efficacy of HiFive for data normalization but also the integration of downstream analysis tools that provide a way of connecting structural data with more traditional sequence annotation. The flexibility of this set of tools should open an additional dimension of analysis to a broader community of researchers for uncovering the interplay between cellular function and structure.

Chapter 4 Discussion

Explaining nuclear organization

The wealth of data that have been generated detailing various aspects of genome spatial arrangement has given us a broader understanding of the nature of nuclear organization and its role in shaping cell function and fate. While typically done independent of one another, experiments exploring the nature of different compartments of the nucleus such as LADs, transcription factories, or the nucleolus all point to a common set of organizing principles. Although many of the specific proteins involved are different between these compartments, some proteins, such as CTCF, show a common function across the nucleus.

We propose the "Velcro model" to explain how nuclear organization arises, a general set of behaviors that lead to the observed properties of chromatin conformation. Under this model, there is a hierarchy of interacting proteins. Interactions can be composed of one or more proteins that have an associated strength of interaction that denotes their rank in the hierarchy. For example, interactions between proteins associated with enhancers and those associated with promoters that directly interact would fall low in the hierarchy as the strength of their interaction appears to be weaker than other interaction pairings leading to more transient associations. Conversely, CTCF-CTCF interactions appear to be highly stable and would rank much higher in the hierarchy. The purpose of the hierarchy is to acknowledge that while some interaction proteins may be more common and thus are more likely to find partners be chance, when stronger interactions are formed that put strain on the chromatin, the lower in the hierarchy an

interaction falls, the more likely it is to release. This leads to organization proceeding generally in a top-down fashion.

Although interactions are specific between compatible proteins, under this model specific pairings amongst compatible proteins are stochastic in nature. This behavior by itself should lead to a fractal globule arrangement of the chromatin. Spatial proximity should lead to interactions within small neighborhoods of DNA, followed by associations of these neighborhoods with each other. This should lead to chromosomes that are fairly self-interacting but do have some inter-chromosome associations. Further, chromosome domains would be random from cell to cell.

We know that the fractal globule arrangement does not account for the domain structure observed in cells, which is where the hierarchy of interactions becomes important. Domains of a given type are bound by proteins that will interact with each other or a common substrate, such as lamin B creating an interaction bridge between the DNA and nuclear envelope. These interactions will lead to agglomerations of like domains such as transcription factories or NADs.

As the cell specializes, changes occur that are mediated by domain-specific recruitment of histone modifying proteins and DNA methylases. These allow access to new binding sites while restricting access to others. Production of tissue-specific transcription factors also alters the interaction potentials of any given region of DNA. The result is that the amount of accessible DNA is reduced as more regions become associated with LADs or PcBs, while at the same time new enhancers take over control of genes that have become or remain accessible, tailoring the cell's activities to its lineage. It is possible to imagine a region of DNA having the potential for recruitment by lamin

proteins but in a pluripotent state being bound by proteins with high interaction strength. As those sites are occluded by methylation, binding to the NE would become dominant and the entire region would become an inactive LAD.

We believe this model explains many of the observations made about chromatin organization in the interphase nuclease, although many details are still unknown. There are still many blanks to be filled in by experimental work, such as relative strength and stability of different interacting proteins, filling in the role of ncRNAs, and the limitations imposed by DNA flexibility on chromatin conformation. As it becomes more fleshed out, the Velcro model should also lend itself to making predictions that can be tested, allowing us to assess how well it describes the underlying mechanisms guiding chromatin topological arrangements.

Areas of future inquiry

Delineating the conservation of boundaries

One of the largest open questions in the field of chromatin architecture is how do different topological domains and subdomains change or persist across time, both from a cellular differentiation and evolutionary standpoint. To date only limited work has focused on comparisons between cell types or across species (Denholtz *et al* 2013, Dixon *et al* 2012, Kieffer-Kwon *et al* 2013, Nora & Heard 2010, Nora *et al* 2012, Phillips-Cremins *et al* 2013, Rousseau *et al* 2014a, Zhao *et al* 2006). Of these studies, the primary focus was almost always examining shifting interactions between genes and enhancers or other genes. Dixon *et al* (2012) performed a cursory analysis of domain similarity of homologous regions of mouse and human in comparable cell types, but little else has been explored in this realm.

One of the primary drawbacks has been a lack of suitable datasets for such explorations. When multiple cell types are tested, they are rarely at the same, or even sufficient, depth of coverage for proper analysis and comparison. This is likely a question of resources. Detection of precise domain boundary placement and subdomain structure is highly sensitive to depth of coverage. In order to make reasonable comparisons between cell types it is crucial not only that both samples have sufficient depth to allow fine-tuning of inferred boundary positions but also comparable sensitivity between samples. The precision of boundary calls between cell types is of paramount importance in differentiating between static boundaries and subtly shifting ones. Without an answer to this question, trying to determine the underlying mechanism of boundary formation and maintenance is far more difficult.

An additional consideration when looking at evolutionary conservation of domain structure is evolutionary distance between species. Currently the closest multicellular species pair that has genome-wide data available is mouse and human. This is far from sufficient for a meaningful depth of understanding about the persistence of domain structure or how domains may affect other aspects of genome evolution.

In order to address these challenges, we propose employing a learning approach that would use multiple cell types and addition genomic annotation data to create informed partitions of structural data. Relevant annotations and annotation combinations could be learned by looking at correlations between cell types of annotation features and structural signal. These features could then be used to inform the likelihood of alignment of domain edges during a partitioning of the genome into domains in parallel across all structural datasets. Examples already exist for strategies to partition the genome into domains that could be extended to handle multiple datasets simultaneously (Filippova *et al* 2014, Phillips-Cremins *et al* 2013). Further, we already have evidence for a set of good candidate annotations to inform our estimations of shared structure, such as gene expression, CTCF and cohesin binding, and histone modification transition points. What remains is assuring that input structural data are of comparable quality. In addition, care will have to be taken to ensure that inclusion of multiple similar or identical cell types does not bias boundary finding because of over-representation. This can be done by careful construction of weighted probabilities of shared states.

By creating a catalog of precise boundary points for a variety of cell types and species, our hope is that more comprehensive studies can be preformed regarding the conservation of structural features. There is no reason not to try and extend this same

basic approach across species. As long as the interaction between partitions is limited to influencing the probability of boundary location, then all that is required is a sequence alignment between all pairwise combinations of species. Of course, the questions still remain as to how boundaries are created and maintained and whether the underlying drivers are different between boundaries found in different contexts.

Defining boundary types and elements

Although there is conservation of many of the general domain structures across various cell types, it is not clear how these boundaries are formed or maintained (Lin *et al* 2012, Ryba *et al* 2010). Although experiments knocking out CTCF have shown an increase in inter-domain interactions not seen with other structural protein knockdowns, it is still unclear what other factors are involved or how CTCF is maintaining these domains (Seitan *et al* 2013, Sofueva *et al* 2013, Zuin *et al* 2014). CTCF has a high occupancy rate across the genome (55,000 – 65,000 sites) with 40-60% of sites being cell-type specific, but only around 15% of CTCF sites are associated with observed domain boundaries (Chen *et al* 2012, Cuddapah *et al* 2009, Dixon *et al* 2012). Given the large percentage of CTCF that occurs within domains and the number of possible interaction partners for CTCF-mediated structures, there is necessarily some other organizing principle involved in reforming domains after each round of mitosis.

There are two possibilities that seem most likely, though not mutually exclusive. The first is that CTCF sites that occur at boundaries have additional factors and modifications creating a specific boundary type that limits the range of possible interaction partners. For example CTCF binding sites flanking a LAD may have additional factors that interact with each other, strongly biasing these CTCF sites to

specifically bind to other LAD boundary sites. A second possibility is that other factors and modifications internal to the domain create an interaction preference for the domain, such as clusters of Nanog binding together (de Wit *et al* 2013). In this case CTCF would have a much higher probability of interacting by chance with a CTCF at a similar boundary type despite retaining the ability to interact with any other CTCF. As for the presence of intra-domain CTCF, around half of these occur within genes and are likely to spatially co-occur in transcriptionally active domains (Chen *et al* 2012, Chen *et al* 2008).

While there have been a great many pattern finders for DNA sequence, primarily as a result of looking for binding sites from ChIP-chip and ChIP-seq data (for reviews, see Das & Dai 2007, Robins *et al* 2008, Tran & Huang 2014), it is only recently with the availability of higher quality and more numerous annotation datasets that effective pattern finders have become available for combinatorial epigenetic patterns (Cha & Zhou 2014, Nair *et al* 2014, Teng *et al* 2014, Wong *et al* 2014). We believe that these approaches can be applied or modified to address questions of epigenetic signatures of structural features in the genome.

We propose three different approaches to this challenge, each focusing on the problem in a slightly different way. First, using domain boundary calls from above, annotation patterns can be partitioned into groups of similar sets of features in an undirected manner. This could be further focused by centering on observed CTCF sites, given the implication of CTCF in establishing such boundaries. Second, this type of pattern finding could be done in a supervised manner by partitioning domains based on their common features, much like whole genome partitions have been performed using histone modifications (Buske *et al* 2011, Filion *et al* 2010, Hoffman *et al* 2012). This

would allow boundary states labels to be applied prior to pattern finding. Finally, rather than identifying patterns from a subset of the genome, regions of sequence could instead be weighted by their likelihood in being involved in a boundary, for example their BI score. This would allow the elimination of patterns common to the genome outside of the target regions and thus not unique to structurally significant locations. This is the only one of the three approaches that does not require making definitive boundary calls, a distinct advantage when the precise definition as to what qualifies as a physical domain has yet to be decided.

Deciphering between targeted and stochastic association

Like the nature of boundary interactions, it is unclear how other structural interactions are formed. Because of their critical role in cellular function and the high amount of substructure observed surrounding them, regions containing transcriptionally active genes are of particular interest. Specifically, how do enhancers target particular genes and not others?

We know that combinations of cohesin and CTCF together play a crucial role in establishing the substructure of actively transcribed domains through specific and apparently stable interactions, but how cohesin and mediator, another key protein in EP interactions, act in the absence of CTCF binding sites is much less clear (He *et al* 2014, Phillips-Cremins *et al* 2013, Seitan *et al* 2013, Sofueva *et al* 2013). Evidence from Hughes *et al* (2014) suggests that promoters search along their genomic neighborhood, interacting with other promoters, enhancer elements, and CTCF sites more often but by no means exclusively. Genes within subdomain partitions also appear to require interaction between cohesin associated with their promoters and CTCF bound at the

partition boundaries (Majumder & Boss 2011). Further, these interactions are dependent on DNA to form. Finally, when cohesin is knocked down, expression becomes deregulated as intra-domain substructure is lost (Seitan *et al* 2013, Sofueva *et al* 2013). Taken together these observations suggest that while expression timing and location of binding sites for transcription factors are controlled, active enhancers produce effects that are stochastic in nature and the limit of their effects are determine by spatial proximity and partitioning by stable chromatin-organization structures such as CTCF-cohesin looping interactions. Answering this question crucial to our understanding of the basic function of transcriptional regulation. As such we propose two different approaches, one computational and one experimental, to try and address this question.

Multiple groups have produced cohesin knockdown conditions in a variety of cell lines, providing an ideal situation for assessing the impact of structure on the nature of EP interactions (Seitan *et al* 2013, Sofueva *et al* 2013). Because of the loss of intra-domain boundaries, these data provide an excellent opportunity to explore how the presence or absence of compartmentalization affects transcriptional activity. In order to accomplish this, there are three important components that must be modeled. First, variability in promoter strength must be assessed and accounted for in downstream calculations. Variations can occur as the result of sequence differences, transcription factor binding, histone modifications, and DNA methylation (Cheng & Gerstein 2012, Mijakovic *et al* 2005, Yada *et al* 2011). Similarly, enhancer effect must also be modeled. Like promoters, the effect that an enhancer can have on initiating transcription is dependent on its sequence, the binding of transcription factors, and histone modifications (Brown *et al* 2013, Wilczynski *et al* 2012, Zhu *et al* 2013). Luckily there appear to be few changes in

these factors for both promoters and enhancers upon disruption of cohesin, allowing us to predict change in expression based on cohesin status rather than predicting absolute expression. This should make assessing the accuracy of the final component of the model, the relationship between enhancers and promoters in a spatial context, simpler. In order to determine this relationship, we must take into account the promoter and enhancer distances from the subdomain boundary, their distance from each other, and the total size of the subdomain. For added complexity, enhancers in adjacent subdomains could also be incorporated into predictions for each promoter's activity. In may also be necessary to account for the number of promoters and their spacing that each enhancer can interact with, as these could affect the interaction frequency of the enhancer with any given promoter.

Based on predictions made from the wild-type cells, it is then possible to explore how changes in spatial partitioning affect EP interaction frequencies. Specifically, do interactions, as assessed by transcriptional activity, occur at a frequency explainable by random encounters or do some EP combinations occur with higher or lower than expected rates, suggesting compatible or incompatible functional pairing, respectively?

This question could also be approached without the surrogate of transcriptional activity and be explored strictly from the standpoint of encounter rates. Given the enhancer and promoter features described above and a model of the effects of spatial configuration, can we predict interaction rates? This could easily be tested using the cohesin knockdown condition paired with its corresponding wild-type condition. This would eliminate the need to determine the transcriptional consequences of different enhancers and promoters and allow an assessment strictly of contact probabilities. The

largest drawback is the lack of information about functional consequences based on the results.

A complementary approach would be to alter subdomain structure, either by the addition, subtraction, or shifting of enhancer elements, to determine the functional consequences. The simplest approach would be to create a series of cell lines with shifts in the relative positions between subdomain boundaries, an enhancer, and a promoter. Another option would be to create an inversion between an enhancer and a cohesin site, switching the enhancer from one partitioned group of genes to another. All of these could shed light on the interplay of EP interactions and the role of cohesin-dependent partitioning of topological domains.

Applications for chromatin structural understanding

As our understanding of chromatin conformation expands, so to does our ability to apply that knowledge to new areas and applications. Because of the intimate link between structure and gene regulation, changes in the architecture of an organism's DNA have massive potential to alter both function and timing. This can result in catastrophic disruptions, but also may be exploited for our own ends.

Associations between chromatin structure and disease

Despite extensive interest and investment in research into human diseases, it has only been in the past half-decade that investigation has extended into examining how spatial alterations may play a role in disease progression. Such applications have ranged from understanding the underlying mechanisms to exploiting structural alterations for diagnostic purposes.

Because of the many rearrangements, translocations, duplications, and deletions, cancer has been of particular interest for chromatin topology. Research into the underlying mechanisms of cancer has focused on two primary aspects, how spatial organization influences rearrangements and how changes in oncogene expression influences chromatin structure (Elemento *et al* 2012, Engreitz *et al* 2012, Fudenberg *et al* 2011, Rickman *et al* 2012). In most cancers, key preliminary steps involve either fusion of different coding genes, switching of control of a gene to a new promoter or enhancer, or both (for review, see Zheng 2013). These changes reflect the spatial organization of the genome, as genomic rearrangements occur between sites that show close spatial proximity in normal cells (Engreitz *et al* 2012, Fudenberg *et al* 2011). In an opposite

effect, overexpression of the oncogenic transcription factor ERG led to extensive rearrangements of chromatin structure and transcriptional activity alterations (Rickman *et al* 2012). Of course, it is unclear why such changes are occurring. More extensive study is needed to explore whether alteration of the epigenetic landscape is causing domain or subdomain boundary shifts, whether newly bound enhancers are creating previously unseen EP interactions, or some other cause exists.

In the case of cancer, structural variation is also being used as a diagnostic tool. Rousseau *et al* (2014b) were able to use 5C data from a section of the HoxA cluster to create a classifier that not only could distinguish between normal and leukemia cells, but was also able to separate samples from different subtypes as indicated by their MLL protein fusion partner. In a different application, 3C is being applied to multiple loci to create an "epigenetic barcode" enabling a rapid blood-based test for detection of melanoma (Bastonini *et al* 2014).

Chromatin conformation has also been implicated, both directly and indirectly, in numerous other conditions. Because of their role in gene silencing and chromatin organization, lamins play a crucial part in cellular function and have been implicated in a variety of diseases resulting from misregulation or mutation in one or more of the proteins of this family (for review, see Davidson & Lammerding 2014). One such disease, a premature aging disease called Hutchinson-Gilford progeria syndrome, has recently been examined in detail using HiC to map the changes in chromatin topology as the disease progresses (McCord *et al* 2013). As cells aged, this analysis showed a build up of a mutant form of lamin A, leading to a loss of lamina-chromatin interaction and loss of spatial partitioning into TDs.

Given the role of gene regulation in disease, previous studies suggest that diagnostics based on chromatin topology have the potential to be extended far beyond cancer. The ability to detect EP interactions as early indicators before clinical symptoms would be a boon to early intervention medicine for any number of maladies. This type of approach also holds promise for customized treatment options in cancer. Because of the fast accumulation and variability of genomic alterations, 3C-based technologies could provide a cost-effective and highly informative approach to fast response treatment. While the tools are catching up, the largest hurdle now is scaling up the sensitivity. Currently small sample or single cell interaction mapping provides very sparse results (Nagano *et al* 2013). Until this has been solved, such diagnostics will continue to rely on larger sample sizes.

Synthetic biology

Another area where understanding the effects of chromatin conformation on cellular function will have a potentially large impact is the field of synthetic biology. Currently work on artificial cells is restricted to prokaryotic cells, although synthetic gene networks have been created in eukaryotic cells (for reviews, see Blount *et al* 2012, Haynes & Silver 2009, Michalodimitrakis & Isalan 2009, Wieland & Fussenegger 2012). In the near future, it is possible to imagine creating cells that fulfill not one function, but are adaptable to a range of inputs. One way to do this would be to take a cue from nature and coordinate clusters of genes based on function and create response modules that can be turned on and off as a result of external signal. It is also possible to imagine using something akin to LADs to affect a "timer" on artificial life, slowly shutting the cell down over time. While this is currently beyond the scope of synthetic biology, work in

understanding the interplay between spatial organization and cellular function, differentiation, and signal response will lay the underpinnings for future work in synthetic biology applications.

Conclusions

Our understanding of the architecture of the nucleus has grown in leaps and bounds over the past decade. With NGS, spatial and temporal resolution of genomic positioning has become possible on a scale scarcely imagined a mere 15 years ago. Further, we have begun to branch into single cell dynamics, seeing past the veil of the cellular population with something other than a microscope. In parallel to this, computation approaches to manipulating, modeling, and analyzing these data have grown in complexity and power. With sequencing advances such as pore sequencing and optical tweezers manipulators for analyzing single molecules, new avenues of research should continue to open. The potential for continuing to expand our understanding of the physical nature of the cell and the interplay between form and function is vast but will also require our ability to make sense of increasingly complex data to keep pace.

Chapter 5 References

Aranda-Anzaldo A, Dent MA, Martinez-Gomez A. 2014. The higher-order structure in the cells nucleus as the structural basis of the post-mitotic state. *Prog Biophys Mol Biol* 114: 137-45.

Arnone MI, Davidson EH. 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124: 1851-64.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, et al. 2011. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* 144: 214-26.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 37: D885-90.

Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, et al. 2009. Active promoters and insulators are marked by the centrosomal protein 190. *EMBO J* 28: 877-88.

Bastonini E, Jeznach M, Field M, Juszczyk K, Corfield E, et al. 2014. Chromatin barcodes as biomarkers for melanoma. *Pigment Cell Melanoma Res* 27: 788-800.

Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. 2013. Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet* 9: e1004018.

Blount BA, Weenink T, Ellis T. 2012. Construction of synthetic regulatory networks in yeast. *FEBS Lett* 586: 2112-21.

Boveri T. 1909. Die Blastomerenkerne con Ascaris megalocephala und die Theorie der Chromosomenindividualität. *Arch Zellforsch* 3: 181-268.

Brown CD, Mangravite LM, Engelhardt BE. 2013. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet* 9: e1003649.

Buske OJ, Hoffman MM, Ponts N, Le Roch KG, Noble WS. 2011. Exploratory analysis of genomic segmentations with Segtools. *BMC Bioinformatics* 12: 415.

Caburet S, Conti C, Schurra C, Lebofsky R, Edelstein SJ, Bensimon A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res* 15: 1079-85.

Camps J, Wangsa D, Falke M, Brown M, Case CM, et al. 2014. Loss of lamin B1 results in prolongation of S phase and decondensation of chromosome territories. *FASEB J*

Cantone I, Fisher AG. 2013. Epigenetic programming and reprogramming during development. *Nat Struct Mol Biol* 20: 282-9.

Cha M, Zhou Q. 2014. Detecting clustering and ordering binding patterns among transcription factors via point process models. *Bioinformatics* 30: 2263-71.

Chen H, Tian Y, Shu W, Bo X, Wang S. 2012. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One* 7: e41374.

Chen P, Zhao J, Wang Y, Wang M, Long H, et al. 2013. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev* 27: 2109-24.

Chen X, Xu H, Yuan P, Fang F, Huss M, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133: 1106-17.

Cheng C, Gerstein M. 2012. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 40: 553-68.

Cremer C, Gray JW. 1982. DNA content of cells with generalized chromosome shattering induced by ultraviolet light plus caffeine. *Mutat Res* 94: 133-42.

Cremer T, Cremer C. 2006a. Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem* 50: 161-76.

Cremer T, Cremer C. 2006b. Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *Eur J Histochem* 50: 223-72.

Cremer T, Cremer C, Baumann H, Luedtke EK, Sperling K, et al. 1982. Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum Genet* 60: 46-56.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19: 24-32.

Daban JR. 2011. Electron microscopy and atomic force microscopy studies of chromatin and metaphase chromosome structure. *Micron* 42: 733-50.

Das MK, Dai HK. 2007. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 8 Suppl 7: S21.

Davidson PM, Lammerding J. 2014. Broken nuclei--lamins, nuclear mechanics, and disease. *Trends Cell Biol* 24: 247-56.

de Wit E, Bouwman BA, Zhu Y, Klous P, Splinter E, et al. 2013. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* 501: 227-31.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* 295: 1306-11.

Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, et al. 2013. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* 13: 602-16.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376-80.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16: 1299-309.

Elemento O, Rubin MA, Rickman DS. 2012. Oncogenic transcription factors as master regulators of chromatin topology: a new role for ERG in prostate cancer. *Cell Cycle* 11: 3380-3.

Engreitz JM, Agarwala V, Mirny LA. 2012. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One* 7: e44196.

Fang F, Xu Y, Chew KK, Chen X, Ng HH, Matsudaira P. 2014. Coactivators p300 and CBP maintain the identity of mouse embryonic stem cells by mediating long-range chromatin structure. *Stem Cells*

Farrell RE, Jr. 1997. DNA amplification. *Immunol Invest* 26: 3-7.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* 143: 212-24.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* 9: 14.

Flemming W. 1882. *Zellsubstanz, Kern und Zelltheilung*. Leipzig, Germany: F.C.W. Vogel.

Fudenberg G, Getz G, Meyerson M, Mirny LA. 2011. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol* 29: 1109-13.

Fullwood MJ, Han Y, Wei CL, Ruan X, Ruan Y. 2010. Chromatin interaction analysis using paired-end tag sequencing. *Curr Protoc Mol Biol* Chapter 21: Unit 21 15 1-25.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462: 58-64.

Ghamari A, van de Corput MP, Thongjuea S, van Cappellen WA, van Ijcken W, et al. 2013. In vivo live imaging of RNA polymerase II transcription factories in primary cells. *Genes Dev* 27: 767-77.

Gonzalez I, Mateos-Langerak J, Thomas A, Cheutin T, Cavalli G. 2014. Identification of Regulators of the Three-Dimensional Polycomb Organization by a Microscopy-Based Genome-wide RNAi Screen. *Mol Cell* 54: 485-99.

Grosberg A, Nechaev SK, Shakhnovich EI. 1988. [The role of topological limitations in the kinetics of homopolymer collapse and self-assembly of biopolymers]. *Biofizika* 33: 247-53.

Grosberg A, Rabin I, Khavlin S, Nir A. 1993. [Self-similarity in the structure of DNA: why are introns needed?]. *Biofizika* 38: 75-83.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453: 948-51.

Haynes KA, Silver PA. 2009. Eukaryotic systems broaden the scope of synthetic biology. *J Cell Biol* 187: 589-96.

He A, Kong SW, Ma Q, Pu WT. 2011. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci U S A* 108: 5632-7.

He B, Chen C, Teng L, Tan K. 2014. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* 111: E2191-9.

Hebert MD. 2013. Signals controlling Cajal body assembly and function. *Int J Biochem Cell Biol* 45: 1314-7.

Hens L, Baumann H, Cremer T, Sutter A, Cornelis JJ, Cremer C. 1983. Immunocytochemical localization of chromatin regions UV-microirradiated in S phase or anaphase. Evidence for a territorial organization of chromosomes during cell cycle of cultured Chinese hamster cells. *Exp Cell Res* 149: 257-69.

Heuser E. 1884. Beobachtungen, über Zellkerntheilung. *Botanisches Centralblatt* 17: 27-32, 57-59, 85-95, 117-28, 54-57.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9: 473-6.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* 48: 471-84.

Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, et al. 2013. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol* 9: e1002893.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 28: 3131-3.

Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, et al. 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* 46: 205-12.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, et al. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 9: 999-1003.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, et al. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503: 290-4.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467: 430-5.

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 30: 90-8.

Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, et al. 2013. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* 155: 1507-20.

Kimura H. 2013. Histone modifications for human epigenome analysis. *J Hum Genet* 58: 439-45.

Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, et al. 2013. Single-cell dynamics of genome-nuclear lamina interactions. *Cell* 153: 178-92.

Kind J, van Steensel B. 2014. Stochastic genome-nuclear lamina interactions: Modulating roles of Lamin A and BAF. *Nucleus* 5: 124-30.

Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* 184: 868-71.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-9.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.

Li HB, Muller M, Bahechar IA, Kyrchanova O, Ohno K, et al. 2011. Insulators, not Polycomb response elements, are required for long-range interactions between Polycomb targets in Drosophila melanogaster. *Mol Cell Biol* 31: 616-25.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289-93.

Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, et al. 2012. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat Immunol* 13: 1196-204.

Majumder P, Boss JM. 2011. Cohesin regulates MHC class II genes through interactions with MHC class II insulators. *J Immunol* 187: 4236-44.

Mallet F. 2000. Comparison of competitive and positive control-based PCR quantitative procedures coupled with end point detection. *Mol Biotechnol* 14: 205-14.

McCord RP, Nazario-Toole A, Zhang H, Chines PS, Zhan Y, et al. 2013. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res* 23: 260-9.

Michalodimitrakis K, Isalan M. 2009. Engineering prokaryotic gene circuits. *FEMS Microbiol Rev* 33: 27-37.

Mijakovic I, Petranovic D, Jensen PR. 2005. Tunable promoters in systems biology. *Curr Opin Biotechnol* 16: 329-35.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, et al. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502: 59-64.

Naganuma T, Hirose T. 2013. Paraspeckle formation during the biogenesis of long non-coding RNAs. *RNA Biol* 10: 456-61.

Nair NU, Kumar S, Moret BM, Bucher P. 2014. Probabilistic partitioning methods to find significant patterns in ChIP-Seq data. *Bioinformatics* 30: 2406-13.

Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, et al. 2013. Organization of the mitotic chromosome. *Science* 342: 948-53.

Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, et al. 2010. Initial genomics of the human nucleolus. *PLoS Genet* 6: e1000889.

Nora EP, Heard E. 2010. Chromatin structure and nuclear organization dynamics during X-chromosome inactivation. *Cold Spring Harb Symp Quant Biol* 75: 333-44.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485: 381-5.

Olins AL, Olins DE. 1974. Spheroid chromatin units (v bodies). *Science* 183: 330-2.

Palstra RJ, Simonis M, Klous P, Brasset E, Eijkelkamp B, de Laat W. 2008. Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One* 3: e1661.

Park SK, Xiang Y, Feng X, Garrard WT. 2014. Pronounced cohabitation of active immunoglobulin genes from three different chromosomes in transcription factories during maximal antibody synthesis. *Genes Dev* 28: 1159-64.

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153: 1281-95.

Rabl C. 1885. Über Zelltheilung. *Morph Jb* 10: 214-330.

Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, et al. 2012. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A* 109: 9083-8.

Rieder D, Ploner C, Krogsdam AM, Stocker G, Fischer M, et al. 2014. Co-expressed genes prepositioned in spatial neighborhoods stochastically associate with SC35 speckles and RNA polymerase II factories. *Cell Mol Life Sci* 71: 1741-59.

Rivera-Molina YA, Martinez FP, Tang Q. 2013. Nuclear domain 10 of the viral aspect. *World J Virol* 2: 110-22.

Robins H, Krasnitz M, Levine AJ. 2008. The computational detection of functional nucleotide sequence motifs in the coding regions of organisms. *Exp Biol Med (Maywood)* 233: 665-73.

Robinson PJ, Fairall L, Huynh VA, Rhodes D. 2006. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A* 103: 6506-11.

Rousseau M, Crutchley JL, Miura H, Suderman M, Blanchette M, Dostie J. 2014a. Hox in motion: tracking HoxA cluster conformation during differentiation. *Nucleic Acids Res* 42: 1524-40.

Rousseau M, Ferraiuolo MA, Crutchley JL, Wang XQ, Miura H, et al. 2014b. Classifying leukemia types with chromatin conformation data. *Genome Biol* 15: R60.

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

Roux W. 1883. *Über die Bedeutung der Kerntheilungsfiguren. Eine hypothetische Erörterung*. Leipzig, Germany: Wilhelm Engelmann.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, et al. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* 20: 761-70.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* 489: 109-13.

Schneider A. 1873. Untersuchungen über Plathelminthen. *Jahresberichte der Oberhessischen Gesellschaft für Natur- und Heilkunde in Geißen* 14: 69-140.

Schooley A, Vollmer B, Antonin W. 2012. Building a nuclear envelope at the end of mitosis: coordinating membrane reorganization, nuclear pore complex assembly, and chromatin de-condensation. *Chromosoma* 121: 539-54.

Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, et al. 2009. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol* 7: e13.

Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. 2006. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat Genet* 38: 700-5.

Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, et al. 2013. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res* 23: 2066-77.

Selvaraj S, J RD, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31: 1111-8.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, et al. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148: 458-72.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* 488: 116-20.

Shimi T, Pfleghaar K, Kojima S, Pack CG, Solovei I, et al. 2008. The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes Dev* 22: 3409-21.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, et al. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38: 1348-54.

Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, et al. 2013. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* 32: 3119-29.

Solovei I, Cavallo A, Schermelleh L, Jaunin F, Scasselati C, et al. 2002. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Exp Cell Res* 276: 10-23.

Spector DL. 2006. SnapShot: Cellular bodies. *Cell* 127: 1071.

Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, et al. 2011. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* 25: 1371-83.

Stack SM, Brown DB, Dewey WC. 1977. Visualization of interphase chromosomes. *J Cell Sci* 26: 281-99.

Taft RJ, Hawkins PG, Mattick JS, Morris KV. 2011. The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization. *Epigenetics Chromatin* 4: 13.

Tapley EC, Starr DA. 2013. Connecting the nucleus to the cytoskeleton by SUN-KASH bridges across the nuclear envelope. *Curr Opin Cell Biol* 25: 57-62.

Teng L, He B, Gao P, Gao L, Tan K. 2014. Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets. *Nucleic Acids Res* 42: e24.

Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, et al. 2011. Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet* 7: e1001343.

Tran NT, Huang CH. 2014. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol Direct* 9: 4.

Trask BJ. 1991. DNA sequence localization in metaphase and interphase cells by fluorescence in situ hybridization. *Methods Cell Biol* 35: 3-35.

Tsuchiya KD. 2011. Fluorescence in situ hybridization. *Clin Lab Med* 31: 525-42, vii-viii.

van Berkum NL, Dekker J. 2009. Determining spatial chromatin organization of large genomic regions using 5C technology. *Methods Mol Biol* 567: 189-213.

Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. 2012. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res* 22: 2176-87.

van Koningsbruggen S, Gierlinski M, Schofield P, Martin D, Barton GJ, et al. 2010. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell* 21: 3735-48.

Varriale A. 2014. DNA Methylation, Epigenetics, and Evolution in Vertebrates: Facts and Challenges. *Int J Evol Biol* 2014: 475981.

Waldeyer W. 1888. Über Karyokinese und ihre Beziehung zu den Befruchtungsvorgängen *Archiv für mikroskopische Anatomie* 32: 1-122.

Walter J, Joffe B, Bolzer A, Albiez H, Benedetti PA, et al. 2006. Towards many colors in FISH on 3D-preserved interphase nuclei. *Cytogenet Genome Res* 114: 367-78.

Wang J, Lunyak VV, Jordan IK. 2012. Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res* 40: 511-29.

Wanner G, Schroeder-Reiter E. 2008. Scanning electron microscopy of chromosomes. *Methods Cell Biol* 88: 451-74.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36: D13-21.

Wieland M, Fussenegger M. 2012. Engineering molecular circuits using synthetic biology in mammalian cells. *Annu Rev Chem Biomol Eng* 3: 209-34.

Wilczynski B, Liu YH, Yeo ZX, Furlong EE. 2012. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput Biol* 8: e1002798.

Wong KC, Li Y, Peng C, Zhang Z. 2014. SignalSpider: Probabilistic Pattern Discovery on Multiple Normalized ChIP-Seq Signal Profiles. *Bioinformatics*

Yada T, Yoshida K, Morita M, Taniguchi T, Irie T, Suzuki Y. 2011. Linear regression models predicting strength of transcriptional activity of promoters. *Genome Inform* 25: 53-60.

Yaffe E, Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 43: 1059-65.

Yang L, Lin C, Liu W, Zhang J, Ohgi KA, et al. 2011. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* 147: 773-88.

Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. 2004. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell* 13: 291-8.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9: R137.

Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, et al. 2013. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 504: 306-10.

Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38: 1341-7.

Zhao ZW, Roy R, Gebhardt JC, Suter DM, Chapman AR, Xie XS. 2014. Spatial organization of RNA polymerase II inside a mammalian cell nucleus revealed by reflected light-sheet superresolution microscopy. *Proc Natl Acad Sci U S A* 111: 681-6.

Zheng J. 2013. Oncogenic chromosomal translocations and human cancer (review). *Oncol Rep* 30: 2011-9.

Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. 2013. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res* 41: 10032-43.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462: 65-70.

Zorn C, Cremer C, Cremer T, Zimmer J. 1979. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. *Exp Cell Res* 124: 111-9.

Zorn C, Cremer T, Cremer C, Zimmer J. 1976. Laser UV microirradiation of interphase nuclei and post-treatment with caffeine. A new approach to establish the arrangement of interphase chromosomes. *Hum Genet* 35: 83-9.

Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, et al. 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* 111: 996-1001.

Chapter 6 Non-printed sources

1. Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo

2. NCBI Sequence Read Archive: SRA Toolkit. Version 2.1.16. Retrieved from http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.1.16/sra_sdk-2.1.16.tar.gz

3. Bowtie: an ultra fast, memory-efficient short read aligner. Version 0.12.7. Retrieved from http://sourceforge.net/projects/bowtie-bio/files/bowtie/0.12.7/bowtie-0.12.7-src.zip

4. Bowtie 2: fast and sensitive read alignment. Version 2.1.0. Retrieved from http://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.1.0/bowtie2-2.1.0-source.zip

5. HiCLib. Version 0.9. Retrieved from https://bitbucket.org/mirnylab/hiclib/get/tip.tar.gz

6. HDF5 for Python. Version 2.1.1. Retrieved from https://h5py.googlecode.com/files/h5py-2.1.1.tar.gz

7. Python. Version 2.7.2. Retrieved from https://www.python.org/ftp/python/2.7.2/Python-2.7.2.tar.bz2

8. Cython – C Extensions for Python. Version 0.12.1. Retrieved from http://cython.org/release/Cython-0.12.1.tar.gz

9. NumPy. Version 1.6.1. Retrieved from http://sourceforge.net/projects/numpy/files/NumPy/1.6.1/numpy-1.6.1.tar.gz

10. MPI for Python – Python bindings for MPI. Version 1.2.2. Retrieved from https://mpi4py.googlecode.com/files/mpi4py-1.2.2.tar.gz

11. HiCPipe. Version 0.9. Retrieved from http://www.wisdom.weizmann.ac.il/~eitany/hicpipe/hicpipe_0.9.tar.gz

12. The R Project for Statistical Computing. Version 3.0.1. Retrieved from http://lib.stat.cmu.edu/R/CRAN/src/base/R-3/R-3.0.1.tar.gz

13. HiCNorm: removing biases in Hi-C data via Poisson regression. Version 0.0. Retrieved from http://www.people.fas.harvard.edu/~junliu/HiCNorm/cis.rar

14. Statsmodels. Version 0.5.0. Retrieved from https://pypi.python.org/packages/source/s/statsmodels/statsmodels-0.5.0.tar.gz

15. Macs2. Version 2.1.0. Retrieved from https://pypi.python.org/packages/source/M/MACS2/MACS2-2.1.0.20140616.tar.gz

16. Mouse Encode Project at Ren Lab: Gene expression of 19 mouse tissues. Retrieved from http://chromosome.sdsc.edu/mouse/download/19-tissues-expr.zip

17. Dixon et al (2012) Supplementary materials table 2: Topological domain boundaries. Retrieved from http://www.nature.com/nature/journal/v485/n7398/extref/nature11082-s2.xls

18. MLPY – Machine Learning Python. Version 3.5.0. Retrieved from http://sourceforge.net/projects/mlpy/files/mlpy 3.5.0/mlpy-3.5.0.tar.gz

19. Blender. Version 2.69. Retrieved from http://download.blender.org/release/Blender2.69/blender-2.69-linux-glibc211-x86_64.tar.bz2

Chapter 7 Abbreviations

3C	chromosome conformation capture
4C	circular chromosome conformation capture
5C	chromosome conformation capture carbon copy
BI	boundary index
bp	base pair
ChIA-PET	chromatin interaction analysis paired-end tag sequencing
cis	intra-chromosomal
DI	directionality index
EP	enhancer-promoter
ES	embryonic stem
ESC	embryonic stem cells
fend	restriction enzyme fragment end
FISH	fluorescent in situ hybridization
GC	guanine and cytosine
GEO	Gene Expression Omnibus
Kb	kilobase
LAD	lamina-associated domain
Mb	megabase
MEF	mouse embryonic fibroblast
mES	mouse embryonic stem cell
MPI	message passing interface
NAD	nucleoli associated domain
NE	nuclear envelope

NGS	next generation sequencing
NL	nuclear lamina
NM	nuclear matrix
NOR	nucleolus organizing region
PCA	principle component analysis
PcB	polycomb body
PcG	polycomb group
PMLA	proximity-mediated ligation assay
PolII	polymerase II
PRE	polycomb response element
RE	restriction enzyme
rRNA	ribosomal RNA
TAD	topologically associated domain
TD	topological domain
TF	transcription factor
tiRNA	transcription initiation RNA
trans	inter-chromosomal
tRNA	transfer RNA
TSS	transcription start site