Computational approaches for the analysis of chromosome conformation capture data and their application to study long-range gene regulation

A Dissertation Presented

By

BRYAN R. LAJOIE

Submitted to the Faculty of the University Of Massachusetts Graduate School Of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 10th, 2016

Computational approaches for the analysis of chromosome conformation capture data and their application to study long-range gene regulation

A Dissertation Presented
By
BRYAN R. LAJOIE

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

Job Dekker, Ph.D., Thesis Advisor

Jeffrey Bailey, M.D., Ph.D., Member of Committee

Konstantin Zeldovich, Ph.D., Member of Committee

Jeanne Lawrence, Ph.D., Member of Committee

Nils Gehlenborg, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Zhiping Weng, Ph.D., Member of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D., Dean of the Graduate School of Biomedical Sciences

Program in Systems Biology, Interdisciplinary Graduate Program

February 10th 2016

DEDICATION

This thesis is dedicated to my lovely wife Katherine, who has supported me tremendously throughout all of the years.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance from my advisor and the support from my family, friends, colleagues and collaborators. Throughout my time in the Dekker Lab I've had to privilege to work with a variety of extremely talented scientists. I list each of them in no particular order. Adriana Gelhart PhD, Nele Gheldof PhD, Ye Zhan, Nynke Van Berkum PhD, Natalia Naumova PhD, Amartya Sanyal PhD, Jon-Matthew Belton PhD, Emily Smith PhD, Rachel Patton-McCord PhD, Gaurav Jain, Johan Gibcus PhD, Noam Kaplan PhD, Anne-Laure Valton PhD, Yu Liu (Sunny) PhD, Houda Belagzhal, Marlies Oomen, Liyan Yang, Rasim Barutcu PhD, Mihir Ashok, Ozgun Uyan and of course, Job Dekker PhD. This thesis is also a result of my interaction with members of both the Program in Gene Function and Expression and the Program in Systems Biology. I thank Marian Walhout PhD for my time spent in her lab during my graduate rotation and all of her insightful feedback and advice over the years. I also thank all of my collaborators, both past and present, especially Edith Heard PhD and Luca Giorgetti PhD. And finally, I thank Job Dekker PhD. Together, we have published 31 manuscripts (most notably 6 Nature, 4 Cell, 3 Science and 3 Nature Methods) with many more manuscripts in revision/draft stages. I could not have dreamed of a more successful "tenure" and I will forever remember my time spent in the Dekker Lab. Thank you all.

ABSTRACT

Over the last decade, development and application of a set of molecular genomic approaches based on the chromosome conformation capture method (3C), combined with increasingly powerful imaging approaches have enabled high resolution and genome-wide analysis of the spatial organization of chromosomes. The aim of this thesis is two-fold; 1), to provide guidelines for analyzing and interpreting data obtained from genome-wide 3C methods such as Hi-C and 3C-seq and 2), to leverage the 3C technology to solve genome function, structure, assembly, development and dosage problems across a broad range of organisms and disease models.

First, through the introduction of cWorld, a toolkit for manipulating genome structure data, I accelerate the pace at which *C experiments can be performed, analyzed and biological insights inferred. Next I discuss a set of practical guidelines one should consider while planning an experiment to study the structure of the genome, a simple workflow for data processing unique to *C data and a set of considerations one should be aware of while attempting to gain insights from the data.

Next, I apply these guidelines and leverage the cWorld toolkit in the context of two dosage compensation systems. The first is a worm condensin mutant which shows a reduction in dosage compensation in the hermaphrodite X chromosomes. The second is an allele-specific study consisting of genome wide

Hi-C, RNA-Seq and ATAC-Seq which can measure the state of the active (Xa) and inactive (Xi) X chromosome. Finally I turn to studying specific gene – enhancer looping interactions across a panel of ENCODE cell-lines.

These studies, when taken together, further our understanding of how genome structure relates to genome function.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

THIRD PARTY COPYRIGHTED MATTER

The following figures were reproduced from journals: No permission required

Figure Number	Publisher
Figure 2.1	Nature
Figure 2.2	Nature
Figure 2.3	Nature
Figure 2.4	Nature
Figure 2.5	Nature
Figure 2.6	Nature
Figure 2.7	Nature
Figure 2.8	Nature
Figure 2.9	Nature
Figure 2.10	Nature
Figure 2.11	Nature
Figure 2.12	Nature
Figure 2.13	Nature
Figure 2.14	Nature
Figure 2.15	Nature

The following figures were reproduced from journals:

Figure Number	Publisher	License Number
Figure 4.1	Methods	3806581014447
Figure 4.2	Methods	3806581014447
Figure 4.3	Methods	3806581014447
Figure 4.4	Methods	3806581014447
Figure 4.5	Methods	3806581014447
Figure 4.6	Methods	3806581014447
Figure 4.7	Methods	3806581014447
Figure 4.8	Methods	3806581014447
Figure 4.9	Methods	3806581014447

The following figures were reproduced from journals: No permission required

Figure Number	Publisher
Figure 5.1	Nature
Figure 5.2	Nature
Figure 5.3	Nature
Figure 5.4	Nature
Figure 5.5	Nature
Figure 5.6	Nature
Figure 5.7	Nature
Figure 5.8	Nature
Figure 5.9	Nature

Chapter I adapted from manuscripts published in a journal.
Nature. 2015 Jul 9;523(7559):240-4. doi: 10.1038/nature14450. Epub 2015 Jun 1.
Methods. 2015 Jan 15;72:65-75. doi: 10.1016/j.ymeth.2014.10.031. Epub 2014 Nov 6.
Nature. 2012 Sep 6;489(7414):109-13. doi: 10.1038/nature11279.

Chapter II is adapted from a manuscript published in a journal.
Nature. 2015 Jul 9;523(7559):240-4. doi: 10.1038/nature14450. Epub 2015 Jun 1.

Chapter III contains unpublished text/figures. This paper is current under revision at Nature.

Chapter IV is adapted from a manuscript published in a journal.
Methods. 2015 Jan 15;72:65-75. doi: 10.1016/j.ymeth.2014.10.031. Epub 2014 Nov 6.

Chapter V is adapted from a manuscript published in a journal.
Nature. 2012 Sep 6;489(7414):109-13. doi: 10.1038/nature11279.

Chapter VI contains unpublished text/figures. This manuscript has not yet been submitted for publication.

Chapter VII adapted from manuscripts published in a journal.
Nature. 2015 Jul 9;523(7559):240-4. doi: 10.1038/nature14450. Epub 2015 Jun 1.
Methods. 2015 Jan 15;72:65-75. doi: 10.1016/j.ymeth.2014.10.031. Epub 2014 Nov 6.
Nature. 2012 Sep 6;489(7414):109-13. doi: 10.1038/nature11279.

No permission was required to use the material from the publication:
Nature. 2015 Jul 9;523(7559):240-4. doi: 10.1038/nature14450. Epub 2015 Jun 1.

A license was issued to use the material from the publication:
Methods. 2015 Jan 15;72:65-75. doi: 10.1016/j.ymeth.2014.10.031. Epub 2014 Nov 6.
Supplier: Elsevier Limited ,The Boulevard,Langford Lane, Kidlington,Oxford,OX5 1GB,UK
Registered Company Number: 1982084
Customer Name: Bryan R Lajoie
License Number: 3806581014447
License content publication: Methods
Licensed content title: The Hitchhiker's guide to Hi-C analysis: Practical guidelines
Licensed content author: Bryan R. Lajoie,Job Dekker,Noam Kaplan

No permission was required to use the material from the publication:
Nature. 2012 Sep 6;489(7414):109-13. doi: 10.1038/nature11279.

CHAPTER I: Introduction

Preface

This introduction is adapted from the manuscripts and discussions contained in chapters II, III, IV, V and VI of this thesis.

Introduction

The human genome consists of over 3 billion nucleotides and is contained within 23 pairs of chromosomes. If the chromosomes were aligned end to end and the DNA stretched, the genome would measure roughly 2 meters long. Yet the genome is not only contained within, but functions within a sphere smaller than one tenth the thickness of a human hair (10 micron). This suggests that the genome cannot exist as a simple one-dimensional polymer; instead the genome must fold into a complex compact three-dimensional structure while maintaining the ability to function in all capacities.

It is increasingly appreciated that a full understanding of how chromosomes perform their many functions (e.g. express genes, replicate and faithfully segregate during mitosis) requires a detailed knowledge of their spatial organization. It has been described [1] that chromosomes are packaged into two distinct chromatin compartments, namely the A and B compartments.

Genomic compartments have been found to be correlated with chromatin state, including DNA accessibility, gene density, replication timing, GC content and histone marks [1]. Thus, DNA belonging to the A-type compartment are interpreted as active, euchromatic regions while DNA belonging to the B-type compartment are interpreted as inactive, heterochromatic regions. Genomic compartments have been found to have high-plasticity, such that they change in different cell-types and biological condition, matching large scale changes in gene activity. Further, recent evidence suggests chromosomes appear to be folded as a hierarchy of nested chromosomal domains [1]–[6], and these are also thought to be involved in regulating genes, e.g. by limiting enhancer-promoter interactions to only those that can occur within a single chromosomal domain [7]–[11]. Topologically Associating Domains (TADs) represent evolutionarily conserved sub-megabase self-interacting domains. TADs have also been proposed to consist of multi-kilobase looping associations between regulatory and structural elements [7], [12]. It has been hypothesized that TAD organization can provide a micro-regulatory environment for each gene, thus enriching interactions with specific elements and depleting interactions with other (further away) genomic elements, which can significantly reduce the complexity of the regulatory network. Also, genes can be controlled by a network of regulatory elements, such as enhancers, which can be located hundreds of kilobases away from their promoter. In fact, a given enhancer's target gene is not always the closest possible gene, the target of the enhancer can be up to a megabase away, effectively ignoring the large set of proximal genes [13]. It is now understood that such regulation often involves physical chromatin looping between the enhancer and the promoter [13]–[19]. Together, these recent discoveries of chromosome folding have provided important insights into the nature of long-range gene regulation and the mechanisms underlying gene expression dynamics and genome function as a whole.

Methods to study the 3D genome

Indiscriminate methods such as microscopy or FISH can study the physical structure of genome, but have difficulty measuring multiple discrete contacts simultaneously. The Chromosome Conformation Capture (3C) method was the first method to capture and measure the structure of the entire genome in an unbiased manner [20] (Figure 1.1). 3C has since been further developed into various derivatives including 4C, 5C and Hi-C [21], [22], [23]. These methods use 3C as the core methodology by which they capture genomic interactions. The methods differ only in the actual technique by which the captured interactions are detected and measured, e.g. by PCR in 3C and by unbiased deep sequencing in Hi-C and 3C-seq. Though the 3C method does capture genome-wide data, it was not until the era of deep sequencing came about that one was able to survey all genome wide interactions in a single experiment, as is the case in Hi-C and 3C-seq.

In 3C, cells are cross-linked using formaldehyde, lysed and the chromatin is then digested with a restriction enzyme of choice (typically any 4-cutter HindIII or EcoRI). The chromatin is then extracted and the restriction fragments are ligated. The crosslinks are then reversed, proteins are degraded and DNA is purified. The newly generated chimeric DNA ligation products represent pairwise interactions and can then be analyzed by a variety of down-stream methods.

The Hi-C method includes one additional step that introduces a biotinylated nucleotide at the ligation junction that enables specific enrichment of the ligated DNA [1]. This has the important advantage in that it prevents sequencing DNA molecules that do not contain such junctions and are thus not informative. In 3C-seq one employs the classical 3C protocol and often a more frequently cutting enzyme (e.g. DpnII) followed by intra-molecular ligation without biotin incorporation [4]. The ligated DNA is then directly sequenced to identify pairwise chromatin interactions genome-wide. The 3C-seq methodology sequences all molecules including un-ligated molecules which can complicate the processing / filtering steps and can reduce the percentage of informative reads. However experimental strategies exist to minimize uninformative (un-ligated, self-ligated etc.), such as optimized crosslinked, digestion and ligation coupled with a larger size selection (less shearing) which results in larger molecule size (~1000 bp) and a larger percentage of informative reads.

Applications

3C methods have been applied to many different biological questions and contexts. This thesis will focus on the application of genome structure methods in the context of furthering the understanding of gene regulation. Chapter II and III will focus on dosage compensation, both in *C. elegans* (worm) and *Mus musculus* (mouse). Chapter V will focus on specific long-range looping interactions between genes and enhancers across a panel of ENCODE cell lines. Chapter IV will discuss a set a core guidelines one should follow to successfully complete a genome structure (3C, 5C, Hi-C) study, from initial conception through data analysis and data interpretation. It will also outline considerations one should be aware of when interpreting data produced by the 3C methods and discuss high-level analysis methods used to extract biological meaning from genome structure datasets. Finally, Chapter VI aims to expand the available tools and methodologies one can employ to process, filter, analyze and visualize genome wide structural data through the introduction and discussion of the cWorld toolbox.

Long range gene regulation

The vast non-coding portion of the human genome is awash in functional elements and disease-causing regulatory variants. The relationships between the genomic positions and order of regulatory elements and their impact on distal target genes remain unknown. Genes and distal elements can come together through looping to form higher order chromatin structures involved in gene regulation [24]. Mapping of these structures allows placing loci in three-dimensional context to reveal long-range and possibly functional relationships. Chapter V applies chromosome conformation capture carbon copy, 5C [23], to comprehensively interrogate interactions between transcription start sites (TSSs) and distal elements in 1% of the human genome representing the ENCODE Pilot regions [25]. 5C maps were generated for GM12878, K562, HeLa-S3 and H1-hES cells and results were integrated with other data from the ENCODE consortium (NCP0004) [26]. We discovered >1,000 long-range interactions in each cell line. In differentiated cells, interactions occurred preferentially between active promoters and distal elements that are enriched for chromatin features that are hallmarks of regulatory elements. In contrast, in H1-hES cells looping was not correlated with gene expression and often involved elements resembling poised enhancers. Looping interactions are related to the relative genomic positions of the elements and display directionality. First, Transcription Start Sites (TSSs) interact more frequently with enhancer-like and CTCF-bound elements located upstream than downstream, with a pronounced preference for elements located 100-200 Kb upstream. Second, only ~8% of interactions are with the nearest gene, and some skip as many as 20 genes. Third, in contrast to current insulator models, CTCF-bound elements do not block long-range interactions, implying that many of these sites do not demarcate physically insulated gene domains. Finally, interactions form complex long-range interaction networks. These analyses provide new insights into the links between linear genome sequence, three-dimensional chromatin architecture and gene regulation.

Sex Determination and Dosage

The vast majority of multi-cellular species/organisms have two or more sexes (males, females, hermaphrodites etc.). A recurring / shared system is the XY sex-determination system which is found in humans, most mammals, some insects and some plants [27]. Humans and mouse define males as having one X and one Y chromosome (or simply having a Y) and females as having two X chromosomes (or simply lacking a Y). Worms define males as XO (one X chromosome) and hermaphrodites as having two X chromosomes. In human and mouse, sexual reproduction occurs between a male and female, each contributing a single copy of each of their chromosomes (23 in human, 20 in mouse) to produce offspring. In worms, hermaphrodites can choose to either self-fertilize or to mate with a male worm to produce offspring. In the case of self-fertilization, the progeny is genetically identical to the parent, whereas when a hermaphrodite and male mate, the progeny contains a single copy of each chromosome from each of the two parents.

Dosage compensation

Chromosomes contain genes, which when expressed produce RNA, some of which can be translated into proteins, the workforce of the cell. In the case of female human/mouse, having two copies of the X chromosome (XX) would produce two doses of all X-chromosome genes. In the case of human/mouse males, there would only be a single dose of all X-chromosome genes, since they only have a single copy of the X chromosome (XY). This dosage imbalance between the sexes must be corrected to ensure that each of the sexes receive exactly the same gene dose across the sex chromosomes (XX and XY). To ensure the dose of all X chromosome genes is balanced between the sexes, one of the two female X chromosomes is randomly inactivated during development and thus will produce mostly no RNA and protein. The result is a single dose of most X chromosome genes in both females and males [27].

In worms, the hermaphrodite animal has two X chromosomes and thus a theoretical double dose of all X-linked genes. The male has the single X and thus a theoretical single dose of X linked genes. In contrast to the random inactivation solution for human/mouse, the worm instead, down-regulates the expression for both of the two female X chromosomes by one half, yielding a total dosage of 1 for hermaphrodites (X*0.5 + X*0.5 = 1X) and males (X*1 = 1X) [27]. Fly (drosophila) employs yet another creative solution to the dosage problem. In fly, the single male X chromosome is up-regulated by 2 fold to produce two doses of the single male X chromosome (2*X = 2X) which is equal to the two doses of the female's two X chromosomes (1*X + 1*X = 2X) [27]–[29]

C. Elegans dosage compensation

The three-dimensional organization of a genome plays a critical role in regulating gene expression, yet little is known about the machinery and mechanisms that determine higher-order chromosome structure. The dosage compensation complex (DCC), a condensin complex, binds to both hermaphrodite X chromosomes via sequence-specific recruitment elements on X (rex sites) to reduce chromosome-wide gene expression by half [30]–[34]. Most DCC condensin subunits also act in other condensin complexes to control the compaction and resolution of all mitotic and meiotic chromosomes [32], [33]. Thus an obvious hypothesis would be that the DCC complex alters the structure of the worm X chromosomes which in turn can modulate gene expressed down by one half. This thesis will demonstrate a DCC-dependent structural difference for the two X chromosomes in hermaphrodite worms which coincides with a marked difference in gene expression.

To compare the molecular topology of X chromosomes and autosomes in C. *elegans*, we generated genome-wide chromatin interaction maps from mixed-stage embryos using a modified chromosome conformation capture (Hi-C) protocol combining conventional chromosome conformation capture (3C) with paired-end sequencing [4], [20], [35]. To assess whether the DCC controls the spatial organization of hermaphrodite X chromosomes, we generated chromatin interaction maps for a dosage-compensation-defective mutant in which the XX-specific DCC recruitment factor SDC-2 was depleted, severely reducing DCC binding to X [30], [31], [36] (**Figure 2.8 a**) and elevating X chromosome gene expression (see below).

Since the hermaphrodite worm has two genetically identical X chromosomes, it is difficult to assign sequencing reads to one of the two X homologous chromosomes. Therefore we chose to pool the reads from the two X homologous chromosomes into a single X chromosome consensus structure. This strategy is employed for almost all genome-wide assays using NGS data (Chip-Seq, RNA-Seq, ATAC-Seq etc.) and the assumption that the two homologous chromosomes have a mostly similar genomic landscape and structure is correct as previously described in many allele-specific studies.

The mouse X chromosome

Important new insights into the 3D organization of mammalian chromosomes have come from recently developed and applied chromosome conformation capture approaches. These studies have revealed a hierarchy of structural organization spanning several genomic length scales, from multi-megabase ‘A/B’ compartments defined by blocks of chromatin that correlate with chromatin activity states, to topologically associating domains (TADs) which represent evolutionarily conserved sub-megabase self-interacting domains to multi-kilobase looping associations between regulatory and structural elements [7], [12]. This recent understanding of chromosome folding has provided important insights into the nature of long-range gene regulation and the mechanisms underlying gene expression dynamics.

However, less is known about the structure and organization of heterochromatin. To what extent does chromosome folding, TAD organization and long range looping differ in the context of a heterochromatic state? A classic example of facultative heterochromatin is the inactive X chromosome (Xi) in female mammals, which is condensed and organized into a distinct silent nuclear compartment. During early female development, X-chromosome inactivation (XCI) is triggered by up-regulation of the long non-coding Xist RNA from one of the two X chromosomes. Xist RNA coats the chromosome in cis and, via its A-repeat region [37], [38], induces transcriptional silencing of almost all of the ~1073 genes on the X. Interestingly, some genes (constitutive escapees) avoid this silencing in most cell types while others (facultative escapees) become reactivated from the Xi only in specific contexts [39]. The underlying mechanism(s) for both facultative and constitutive escape are not known. A role for Xist RNA in reshaping the organization of the entire Xi has been proposed [40], [41], with escape genes being excluded from the Xist-coated domain. However, the exact architecture of the Xi, for both its silent and expressed regions, is still unclear. Based on DNA FISH, the human Xi is a rather homogeneous structure with an overall compaction that is about 1.2-fold higher than that of the active X chromosome (Xa) [42]–[44]. Recent chromosome conformation capture approaches have pointed to some intriguing features of the

3D folding of the Xi, including formation of large mega-domains along the human Xi [45], and long-range associations between loci that escape inactivation and become expressed on the mouse Xi [41]. However detailed insights into the global molecular architecture of the Xi remain far from complete, due in part to the lack of chromosome-wide, high resolution, allele specific information. To this end, we have investigated the structure, chromatin accessibility and expression status of the Xi using allele-specific Hi-C, ATAC-Seq and RNA-Seq methods in embryonic stem cells (ESCs) and clonal neural progenitor cells (NPCs) both derived from a highly polymorphic (Cast x 129) F1 mouse. This F1 mouse cross contains 19,722,473 SNPs, averaging 1 SNP every ~140 bases which enables higher resolution analysis of allele-specific chromatin states and three-dimensional conformation than that previously performed in human cells (~10-fold higher SNP density) [45].

This thesis will demonstrate that the Xi lacks typical autosomal features such as active/inactive compartments and topologically associating domains (TADs), except around a small number of genes that escape XCI and remain expressed. Escaping genes form TADs and retain DNA accessibility at promoter-proximal and CTCF binding sites, indicating that these loci can avoid Xist-mediated erasure of chromosomal structure. We further show that gene-silencing competent Xist RNA is sufficient to induce segregation of the Xi into two ‘mega-domains’ separated by a boundary that includes the DXZ4 macrosatellite. Deletion of this boundary prior to XCI results in fusion of the mega-domains and altered patterns of escape that correlate with changes in TAD structure following differentiation and XCI. Our results suggest a critical role for the boundary locus and Xist RNA in shaping the structure of the Xi and modulating escape from XCI. Our findings also point to roles of transcription and CTCF binding in TAD formation in the context of facultative heterochromatin.

Practical guidelines for genome structure studies

Over the last decade, development and application of a set of molecular genomic approaches based on the chromosome conformation capture method (3C), combined with increasingly powerful imaging approaches have enabled high resolution and genome-wide analysis of the spatial organization of chromosomes. The aim of this thesis is to provide guidelines for analyzing and interpreting data obtained with genome-wide 3C methods such as Hi-C and 3C-seq that rely on deep sequencing to detect and quantify pairwise chromatin interactions genome-wide.

The chromosome conformation capture methodology (3C) is now widely used to map chromatin interaction within regions of interest and across the genome. Chromatin interaction data can then be interpreted to gain insights into the spatial organization of chromatin, e.g. the presence of chromatin loops and chromosomal domains. This thesis will discuss methods and considerations that are important for using deep sequencing data to build bias-free genome-wide chromatin interaction maps. Then I will touch on several approaches to analyze such maps, including identification of patterns in the data that reflect different types of chromosome structural features and their biological interpretations.

Development and application of computation tools: cWorld

This thesis will also introduce and discuss a set of tools for processing, manipulating, analyzing, visualizing and integrating genome structure datasets with other widely used genome-wide functional assays. The thesis will discuss guidelines for analyzing genome-wide chromatin interaction maps generated by Hi-C, but many of these considerations also apply to 3C-seq data or other genome structural datasets. First I will discuss the steps required to obtain high-quality unbiased interaction maps. Then, I will discuss analysis and interpretation of the interaction maps. Finally I will introduce and discuss a set of publically available perl, python and R scripts for manipulating genome structure data and all of the necessary assumptions and normalizations that must be considered to properly interpret this new data-type.

Concluding Remarks

Chapters II and III aim to expand the understanding of dosage compensation through the use of Hi-C, RNA-Seq, ATAC-Seq and Chip-Seq data. These chapters will demonstrate a unique structural organization that both the dosage compensated worm and mouse X chromosomes employ to facilitate tighter gene regulation of the X-linked genes. Chapter V will demonstrate the effectiveness of the 5C methodology in the context of mapping gene – enhancer looping interactions across a panel of ENCODE cells lines. From the data generated in 1% of the genome, the long-range interaction landscape of gene promoters can be inferred. This chapter will also provide insights into the complex network of gene regulation and begin to characterize and map specific functional long range looping interactions that control gene expression. Chapter IV will discuss guidelines to use when designing, executing and interpreting genome structure data. And finally chapter VI will introduce cWorld, a set of computation tools for manipulating genome structure data.

Figures

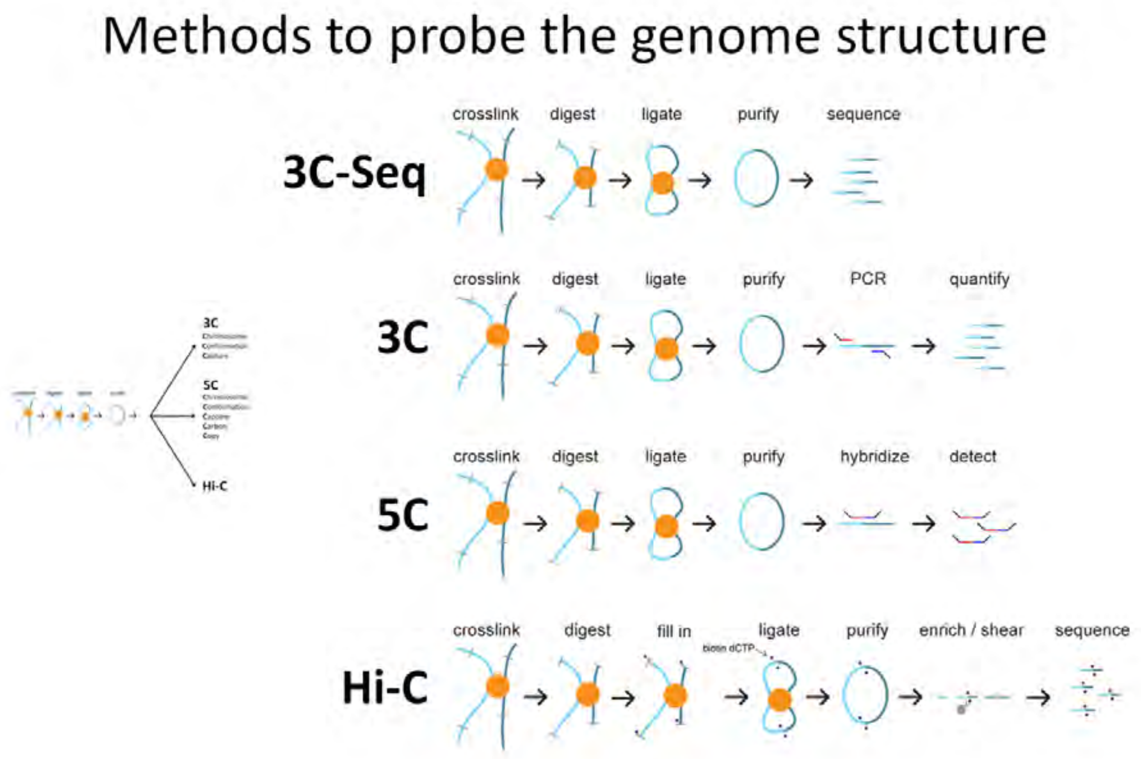


Figure 1.1 | Depiction of the *C method
Schematic of the 3C method and it’s various derivatives.

CHAPTER II: Condensin-driven remodeling of X chromosome topology during dosage compensation

Preface

This research chapter encompassed work published in Nature, by Emily Crane, Qian Bian, Rachel Patton MCcord, Bryan R Lajoie, Bayly Wheeler, Ed J. Ralston, Saturo Uzawa, Job Dekker, and Barbara J. Meyer. The publication is entitled "Condensin-driven remodelling of X chromosome topology during dosage compensation," *Nature*, vol. 523, no. 7559, pp. 240–244, Jun. 2015. [46]

Abstract

The three-dimensional organization of a genome plays a critical role in regulating gene expression, yet little is known about the machinery and mechanisms that determine higher-order chromosome structure [6], [8]. Here we perform genome-wide chromosome conformation capture analysis, fluorescent in situ hybridization (FISH), and RNA-Seq to obtain comprehensive three-dimensional (3D) maps of the Caenorhabditis *elegans* genome and to dissect X chromosome dosage compensation, which balances gene expression between XX hermaphrodites and XO males. The dosage compensation complex (DCC), a condensin complex, binds to both hermaphrodite X chromosomes via sequence-specific recruitment elements on X (rex sites) to reduce chromosome-wide gene expression by half [30]–[34]. Most DCC condensin subunits also act in other condensin complexes to control the compaction and resolution of all mitotic and meiotic chromosomes [32], [33]. By comparing chromosome structure in wild-type and DCC-defective embryos, we show that the DCC remodels hermaphrodite X chromosomes into a sex-specific spatial conformation distinct from autosomes. Dosage-compensated X chromosomes consist of self-interacting domains (1 Mb) resembling mammalian topologically associating domains (TADs) [2], [3]. TADs on X chromosomes have stronger boundaries and more regular spacing than on autosomes. Many TAD boundaries on X chromosomes coincide with the highest-affinity rex sites and become diminished or lost in DCC-defective mutants, thereby converting the topology of X to a conformation resembling autosomes. Rex sites engage in DCC-dependent long-range interactions, with the most frequent interactions occurring between rex sites at DCC-dependent TAD boundaries. These results imply that the DCC reshapes the topology of X chromosomes by forming new TAD boundaries and reinforcing weak boundaries through interactions between its highest-affinity binding sites. As this model predicts, deletion of an endogenous rex site at a DCC-dependent TAD boundary using CRISPR/Cas9 greatly diminished the boundary. Thus, the DCC imposes a distinct higher-order structure onto X chromosomes while regulating gene expression chromosome-wide.

Introduction

To compare the molecular topology of X chromosomes and autosomes in C. *elegans*, we generated genome-wide chromatin interaction maps from mixed-stage embryos using a modified chromosome conformation capture (Hi-C) protocol combining conventional chromosome conformation capture (3C) with paired-end sequencing [4], [20], [35] (**Figure 2.1, Figure 2.2 and Methods**). Interaction data, binned at both 10 kb and 50 kb intervals, revealed features observed in other organisms. Interactions occur most frequently in cis and decay with genomic distance (**Figure 2.2 and Methods**). Chromosome compartments comparable to active A and inactive B compartments [35], [47] are formed (**Figure 2.2, Figure 2.3, Figure 2.4, Figure 2.5**). Compartments at the left end of the X chromosome and both ends of autosomes align with binding domains for lamin [48], lamin-associated protein LEM-2 (**Figure 2.3, Figure 2.4, Figure 2.5**) [49], and the H3K9me3 inactive chromatin mark [50], suggesting their similarity to inactive B compartments of mammals.

Chromatin interaction maps also revealed self-interacting domains (~1 Mb), predominantly on X chromosomes. These domains are visible as diamonds along the interaction maps (**Figure 2.1 a, d**) and resemble TADs of mammalian and fly chromosomes [2]–[4]. To quantify TADs, we devised an approach of assigning an ‘insulation score’ to genomic intervals along the chromosome. The score reflects the aggregate of interactions occurring across each interval. Minima of the insulation profile denote areas of high insulation we classified as TAD boundaries (**Methods, Figure 2.1, Figure 2.6 a and Figure 2.7 a, b**).

The insulation profile of the X chromosome stands out compared to those of autosomes. The insulation signal amplitude is larger on the X chromosome (**Figure 2.1 a, d and Figure 2.7 d**), implying TAD boundaries are stronger. Also, TAD boundaries on the X chromosome are more abundant and regularly spaced (**Figure 2.7 d**).

Results

To assess whether the DCC controls the spatial organization of hermaphrodite X chromosomes, we generated chromatin interaction maps for a dosage-compensation-defective mutant (**DC mutant; Figure 2.1, Figure 2.2, Figure 2.6, Figure 2.6, Figure 2.3, Figure 2.4, Figure 2.5**) in which the XX-specific DCC recruitment factor SDC-2 was depleted, severely reducing DCC binding to X [30], [31], [36] (**Figure 2.8 a**) and elevating X chromosome gene expression (see below). The insulation profile of the X chromosome, but not autosomes, was greatly changed (**Figure 2.1 b, e, Figure 2.2, Figure 2.6, Figure 2.6, Figure 2.3, Figure 2.4, Figure 2.5**). Of a total of 17 TAD boundaries on the X chromosome, 5 were eliminated and 3 severely reduced in insulation. TAD boundary strength and spacing on the X chromosome in DC mutants resembled that of autosomes (**Figure 2.7 d**).

To characterize this transformation in conformation, we calculated the difference between chromatin interaction maps of wild-type and DC mutant embryos after converting the interaction data into genomic-distance-normalized

Z-scores. In DC mutants, interactions on X increased across TAD boundaries but decreased within TADs, revealing a DCC-dependent remodeling of X chromosome structure (**Figure 2.1 c, Figure 2.2, Figure 2.6, Figure 2.7, and Figure 2.4**). Weakening of TAD boundaries is expected to cause chromosome-wide changes in chromatin interactions. The largest changes in insulation on X occurred at TAD boundaries. Autosomes appeared unaffected (**Figure 2.1 c, f, Figure 2.8 a, Figure 2.2, Figure 2.6, Figure 2.7, Figure 2.8, Figure 2.5**).

TAD boundaries on the X chromosome are enriched for the highest DCC-occupied rex sites [30], [31], [51] (**Figure 2.8 a and Figure 2.9 d**). About 50% of all TAD boundaries and 90% of changed ones overlap the top 25 rex sites, a correlation higher than expected at random (**Figure 2.9 d**). In DC mutants, the largest insulation losses occurred in regions overlapping the strongest rex sites (**Figure 2.8 a**). These results imply the DCC plays a direct role in defining TADs by binding to rex sites to mediate formation of TAD boundaries. In contrast, genomic features such as highly occupied targets (HOT) sites [52] do not govern TADs (Supplementary Table 2).

Two TAD boundaries on X that overlap rex sites in the LEM-2 B-like compartment were not greatly reduced in DC mutants (**Figure 2.1 and Figure 2.8 a and Figure 2.4 e**). Although the DCC exerts a dominant influence on TAD formation, other forces act on the X chromosome to form TADs, as on autosomes.

To confirm the DCC-dependent topology of the X chromosome, we visualized TADs using quantitative 3D fluorescent in situ hybridization (FISH) in wild-type XX embryos and embryos lacking DCC binding on X: male XO and DC-mutant XX (**Figure 2.8 b–e**). We imaged fluorescent probes that tiled 500 kb regions within TADs or flanking TAD boundaries. Probe overlap was quantified by analyzing the distribution of Pearson's correlation coefficients between FISH signals from pairwise probe combinations8.

As expected for TADs in wild-type embryos, two adjacent probes within a TAD on either X chromosomes or autosomes overlapped to a greater extent than two adjacent probes on either side of a TAD boundary (**Figure 2.8 b–e and Figure 2.10 a–d**). For DCC-dependent TAD boundaries on X including rex-47, rex-32 and rex-8, adjacent probes flanking TAD boundaries overlapped and co-localized more in embryos lacking DCC binding than in wild-type XX embryos (**Figure 2.8 c, d and Figure 2.10 8b**). In contrast, the DCC-independent TAD boundaries on the X chromosome and autosomes did not change (**Figure 2.8 e and Figure 2.10 c, d**). FISH analysis also confirmed that some DCC-dependent TAD boundaries were eliminated (rex-47), and others reduced (rex-32) in DC mutants and XO males (**Figure 2.8 c, d**), showing that the DCC alters X chromosome structure by strengthening pre-existing TAD boundaries and creating new ones.

Robust correlation between rex sites, DCC-dependent TAD boundaries, and regions of greatest insulation loss in DC mutants (**Figure 2.8 a, Figure 2.9 d**

and Supplementary Table 2) led us to test whether rex sites interact in a DCC-dependent manner. We found rex–rex interactions to be among the most prominent interactions on the X chromosome by comparing the ranking (**Figure 2.9 a**) and cumulative distribution (**Figure 2.11 a, b**) of Z-scores for rex interactions with those for all other X chromosome interactions. In DC mutants, rex–rex interactions decreased more than any of the 1,000 random sets of X chromosome interactions (**Figure 2.11 a, c and Figure 2.9 b, c, e**). These observations support the hypothesis that DCC binding at rex sites facilitates rex–rex interactions.

The rex–rex interaction frequency was directly related to the level of DCC occupancy at rex sites, as shown by 3D profiles of Hi-C interaction frequencies made for pairwise combinations of 10 kb bins overlapping either the top 25 DCC-occupied rex sites or all 64 rex sites (**Figure 2.8 a and 2.11 d, Figure 2.9 f and Supplementary Table 2**). Interactions for the top 25 rex sites exceeded those for all rex sites.

The correlation between rex-interaction strength and DCC occupancy was reinforced by contrasting results with dependent on X (dox) sites. The DCC spreads to these lower affinity dox sites located in promoters of highly expressed genes once recruited to X by rex sites [30], [31]. Dox sites showed no substantial interactions in 3D plots (**Figure 2.9 g**).

The strongest rex–rex interactions occurred between rex sites at DCC-dependent TAD boundaries on the X chromosome (**Figure 2.11 e**). Weaker rex–

rex interactions also occur within TADs. In DC mutants, rex interactions within TADs and between TAD boundaries diminished to the level of non-rex interactions (**Figure 2.11 e**). For autosomes, in contrast, interactions between TAD boundaries were not greater than interactions within TADs, and neither set of interactions changed in DC mutants (**Figure 2.11 e and Figure 2.9 h**). These results suggest that DCC-dependent interactions between rex sites at TAD boundaries contribute more to boundary formation on X than rex interactions within TADs, although DCC-dependent rex interactions within TADs might contribute to TAD integrity.

Visualization of Hi-C interaction data via Circos plots shows that almost all rex sites engage in one or multiple strong DCC-dependent interactions with other rex sites, particularly at adjacent TAD boundaries (**Figure 2.11 f, g**). Together, our findings reinforce the model that rex sites contribute to TAD formation by recruiting the DCC and facilitating DCC-dependent looping interactions between rex sites at TAD boundaries. In contrast, TAD boundaries on autosomes do not appear to result from looping interactions between boundaries (**Figure 2.11 e, right panel and Figure 2.9 h**), suggesting that different strategies govern, in part, the formation of DCC-dependent and autosomal TADs.

The model that rex interactions play a critical role in establishing and reinforcing TAD boundaries makes specific predictions. First, rex interactions identified by Hi-C should be evident by FISH. Second, deletion of a strong rex site from a DCC-dependent TAD boundary should reduce or eliminate the boundary. Both predictions were verified by the data.

To confirm DCC-dependent rex–rex interactions and further assess X-chromosome topology, we devised a FISH assay using 3–6 kb probes to quantify the spatial separation between two sites (**Methods and Figure 2.12**). We compared distances between loci in XX embryos with (wild-type) and without (DC mutant) DCC binding on the X chromosome to quantify the level and DCC-dependence of interactions. We also compared distances in XO embryos with and without DCC binding on the X chromosome to quantify DCC-dependent interactions that occur between loci on the same chromosome (**Figure 2.12 legend**). Hi-C analysis did not distinguish between interactions within the same chromosome or across homologous chromosomes.

FISH analysis confirmed all categories of interactions shown by Hi-C: (1) strong DCC-dependent interactions between rex sites at DCC-dependent TAD boundaries (rex-32 to rex-23, rex-47 to rex-8, and rex-23 to rex-14); (2) strong DCC-dependent interactions between X loci lacking DCC binding (Xnb1 to Xnb2 and Xnb7 to Xnb8 (nb, not bound)); (3) strong DCC-independent interactions between loci on X (Xnb3 to Xnb4) or I (Inb1 to Inb2) that lacked DCC binding; and (4) weak DCC-independent interactions between distant loci on X (Xnb5 to Xnb6) or I (Inb3 to Inb4) that lack DCC binding (**Figure 2.12 b–g and Figure 2.13 a–f, i–k**). FISH and Hi-C results agreed, for both the strength and DCC-dependence of interactions (**Figure 2.13 g, h**).

The only discrepancy occurred for distantly spaced rex loci (rex-1 to rex-8 (6.7 Mb); rex-32 to rex-8 (8.1 Mb)), which showed greater DCC-dependent spatial proximity by FISH analysis than predicted by Hi-C (**Figure 2.13 l, m**). Loss of sensitivity in our Hi-C data for sites separated by .5 Mb may account for the difference.

Both FISH and Hi-C experiments showed that the DCC-dependent topology of the X chromosome brings many distant, non-rex sites into close proximity. If the DCC compacted the X chromosome uniformly, pairs of non-rex loci separated by similar distances should exhibit comparable levels of DCC-dependent interactions. However, they did not. For example, two pairs of non-rex loci (Xnb1 and Xnb2 (1 Mb); Xnb7 and Xnb8 (1.4 Mb)) showed strong DCC-dependent interactions (**Figure 2.12 e and Figure 2.13 g, h, k**), but the non-rex loci Xnb3 and Xnb4 (1.6 Mb) showed strong DCC-independent interactions (**Figure 2.12 f**). Thus, the DCC affects the overall topology of the X chromosome but does not cause uniform compaction across the X chromosome.

To test whether DCC-dependent interactions between rex sites create TAD boundaries, we deleted the endogenous rex-47 site from a DCC-dependent TAD boundary using genome editing with CRISPR/Cas9 (**Figure 2.10 e, f**) and assayed TAD structure with FISH (**Figure 2.11 h**). Chromatin immunoprecipitation followed by quantitative polymerase chain reaction (ChIP–qPCR) showed the deleted rex locus (rex-47 D) lacked DCC binding (**Figure 2.10 g**). The TAD boundary was greatly diminished, as predicted (**Figure 2.11 h**).

For FISH probes flanking the rex-47 TAD boundary, overlap was increased in rex-47 D and DC mutant embryos over that in wild-type embryos. In contrast, overlap was not statistically different between rex-47 D and DC mutant embryos. Thus, the DCC plays a key role in inducing and reinforcing TAD boundaries on X by mediating long-range interactions between its highest-affinity rex sites.

We explored the relationship between TAD structure and gene expression. Our prior work showed the DCC acts at a distance to repress gene expression [30], [31], [53], suggesting that a unique, DCC-dependent X-chromosome structure might mediate chromosome-wide gene repression, as supported by our Hi-C and FISH data. We assessed whether the structure of individual TADs affects gene expression locally or whether the chromosome-wide topology created from TADs regulates gene expression globally. Both RNA-Seq data derived from embryo preparations used for Hi-C analysis and GRO-Seq data from independent embryo preparations support the latter hypothesis for the following reasons.

First, in wild-type embryos, genes at TAD boundaries were not expressed at significantly different levels from genes within TADs, for either chromosome X (**Figure 2.14 b and Figure 2.15 a, d**) or chromosome I (**Figure 2.14 f**). Second, although the X chromosome is organized into DCC-dependent TADs in wild-type animals, no similarly coordinated block of genes exhibited elevated expression in DC mutants (**Figure 2.14 a**). That is, the changes in expression were not significantly different for X-linked genes within TADs, at all TAD boundaries, at changed TAD boundaries, or within regions of changed insulation (**Figure 2.14 c, d and Figure 2.15 b, c, e–i**). Similarly, DC mutations did not alter gene expression on chromosome I in any discernible pattern (**Figure 2.14 e, g, h and Figure 2.15 g–i**).

Conclusions

Our results support the model that TAD structure on the X chromosome mediated by DCC binding to rex sites creates a 3D topology that acts chromosome-wide to repress gene expression. Given that changes in TAD boundaries occur locally, while changes in gene expression occur chromosome-wide, a parsimonious model posits that DCC-dependent changes in X chromosome structure imposed by rex–rex interactions drive the chromosome-wide reduction in gene expression. Potential DCC-dependent nuclear positioning of the X chromosome might also affect gene expression, as speculated by others [54].

In summary, DCC-induced formation of TAD structure on the X chromosome demonstrates a striking remodeling of chromosome topology that reveals a central role for condensin in shaping the 3D landscape of interphase chromosomes. Not only does condensin compact and resolve mitotic and meiotic chromosomes, it acts as a key structural element to regulate gene expression. No other molecular complex or set of DNA binding sites is yet known to cause comparably strong effects on megabase-scale TAD structure in higher eukaryotes [55]–[57]. Our new understanding of the topology of dosage-compensated chromosomes provides fertile ground to decipher the detailed mechanistic relationship between higher-order chromosome structure and chromosome-wide regulation of gene expression.

Supplementary Information is available in the online version of the paper.

Figures

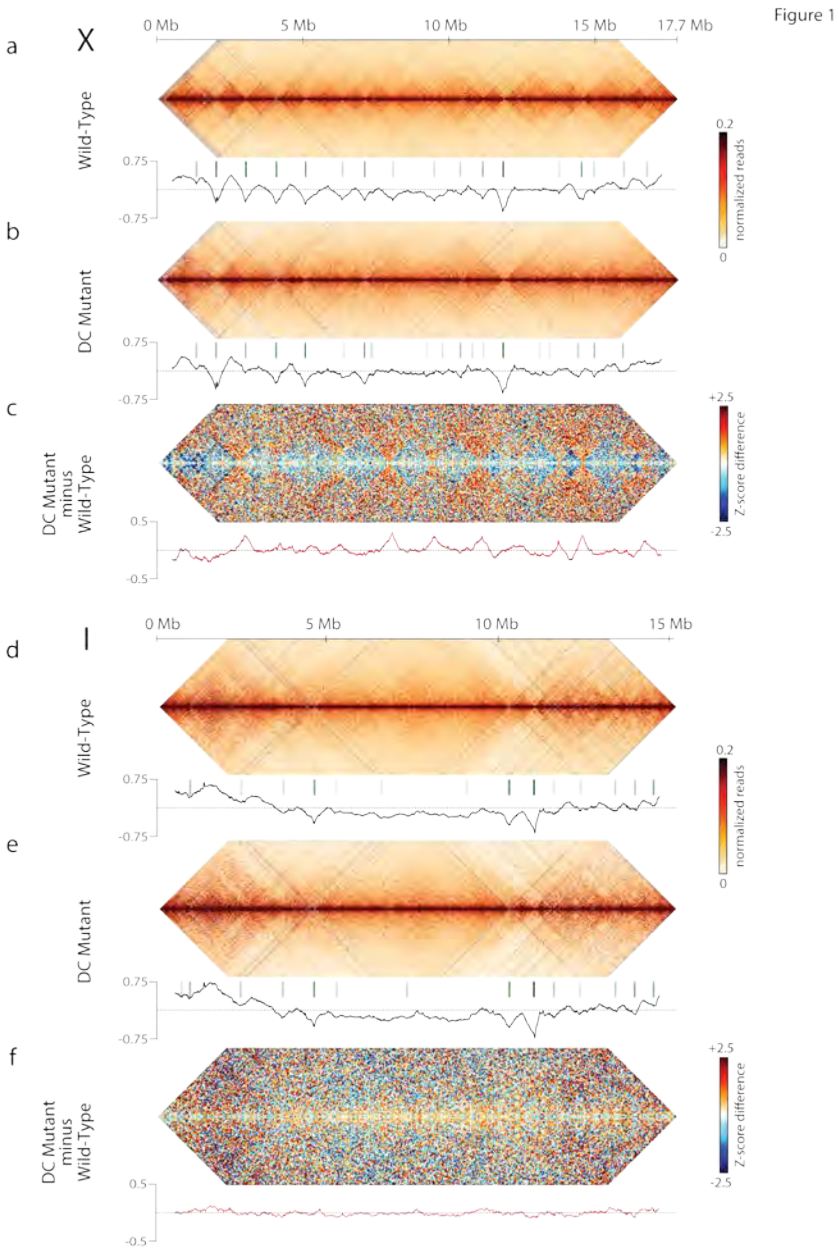


Figure 2.1 | DCC modulates spatial organization of X chromosomes.
a, b, d, e, Chromatin interaction maps binned at 10 kb resolution show interactions 0–4 Mb apart on chromosomes X and I in wild-type and DC mutant

embryos. Plots (black) show insulation profiles. Minima (green lines) reflect TAD boundaries. Darker green indicates stronger boundary. c, f, Blue–red Z-score difference maps binned at 50 kb resolution for X and I show increased (orange–red) and decreased (blue) chromatin interactions between mutant and wild-type embryos. Differential insulation plots (red) show insulation changes between mutant and wild-type embryos.

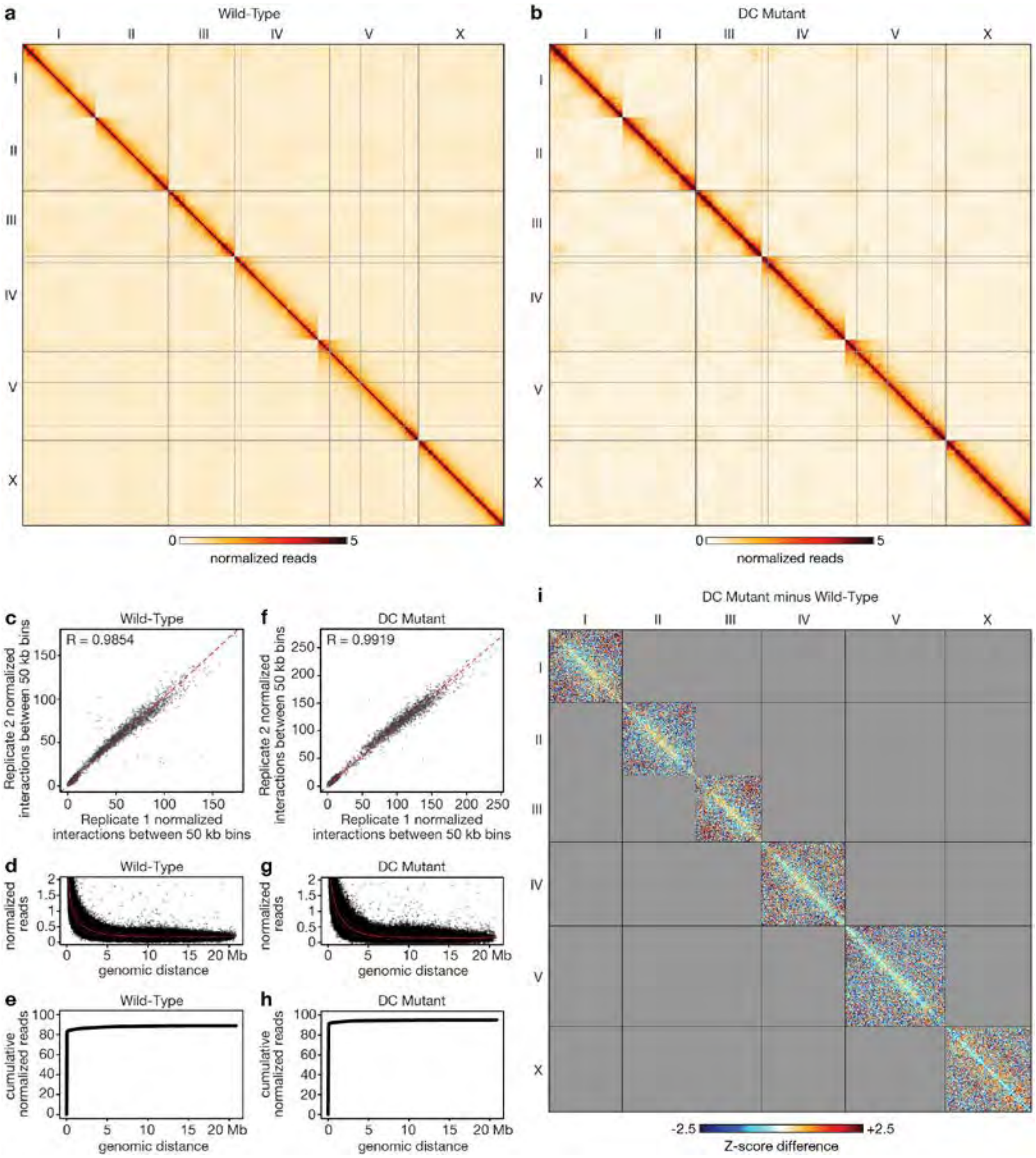


Figure 2.2 | Genome-wide chromatin interaction maps for wild-type or DC mutant embryos and genome-wide difference chromatin interaction map. a, b, Genome-wide chromatin interaction maps for wild-type embryos (a) and DC mutant embryos (b) from Hi-C data of two biological replicates pooled and binned at 50 kb and corrected with ICE. c, f, Scatter plots comparing normalized interactions between pairs of 50 kb bins in the two biological replicates from wild-type embryos (c) or DC mutant embryos (f) (both excluding x 5 y diagonal). A strong correlation between biological replicates is shown for wild-type embryos (Pearson's correlation coefficient 5 0.9854) and for DC mutant embryos (Pearson's correlation coefficient 5 0.9919). d, g, Overall interaction frequency

decays with increasing genomic distance in wild-type embryos (d) and in DC mutant embryos (g). e, h, Cumulative reads versus linear genomic distance in wild-type embryos (e) and in DC mutant embryos (h). i, Genome-wide difference chromatin interaction map. Shown is the 50 kb binned heatmap depicting the Z-score difference between wild-type and DC mutant embryos (see Methods for Z-score difference calculation). The most apparent differences are on the X chromosome: blue signal within TADs (loss of intra-TAD interactions) and red signal between TADs (gain of inter-TAD interactions).

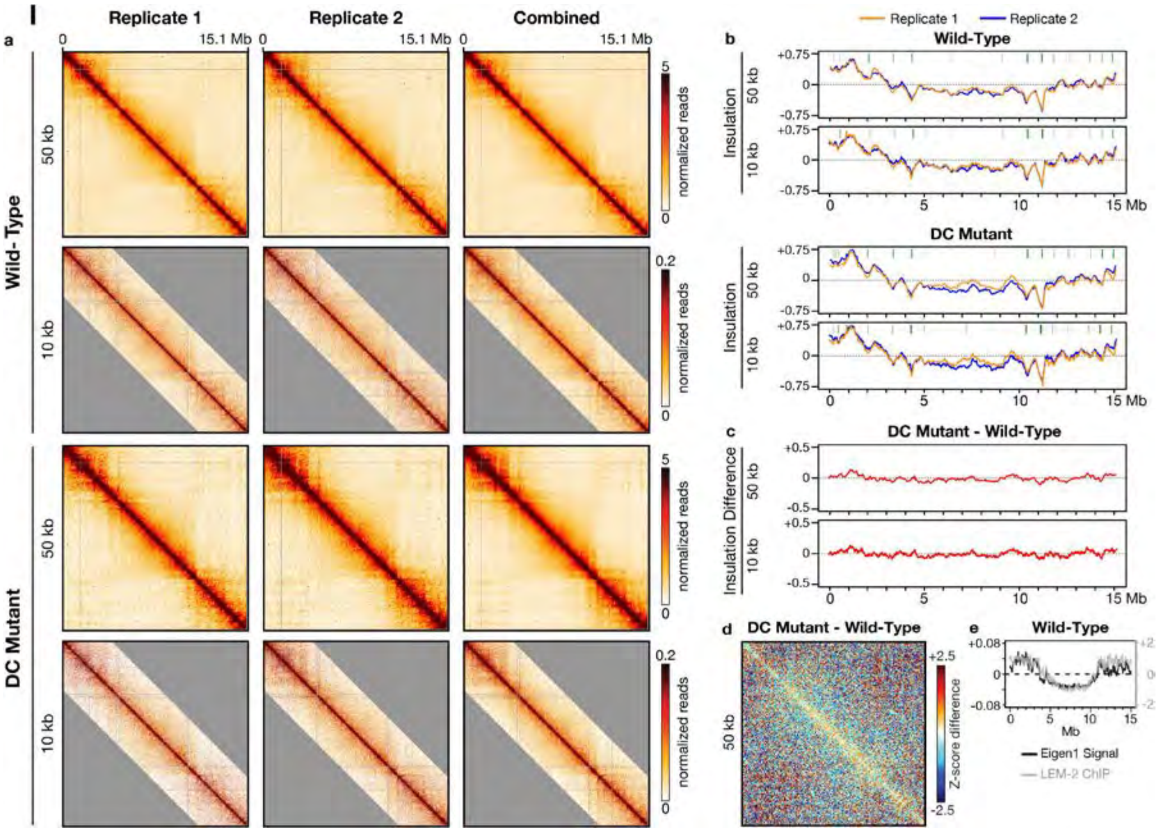


Figure 2.3 | Compartment and insulation analysis for chromosome I in wild-type embryos and DC mutant embryos.

a, ICE corrected chromatin interaction maps are shown for wild-type embryos and DC mutant embryos for both 10 kb binned and 50 kb binned data across replicate 1, replicate 2, and the combined replicates. b, Insulation profiles are shown for each biological replicate (replicate 1, orange line; replicate 2, blue line) for 50 kb and 10 kb binned data in wild-type embryos and DC mutant embryos. Insulation profiles are calculated using a 500 kb 3 500 kb insulation square (10 bins 3 10 bins for the 50 kb binned Hi-C data, and 50 bins 3 50 bins for the 10 kb binned Hi-C data). The insulation profiles are consistent across replicates. Green tick marks, TAD boundaries identified using combined replicate data. c, Differential insulation plots derived from the insulation profiles calculated above (50 kb binned and 10 kb binned Hi-C data). d, 50kb binned heatmap depicting the difference in chromatin interactions expressed as the difference in Z-scores between wild-type and DC mutant. e, Plot showing the compartment analysis calculated using the 50 kb binned wild-type Hi-C data. A/B compartment profile was determined by principle component analysis. First Eigen Vector value representing compartments (black) is plotted along the chromosome, revealing three zones for each autosome: two outer sections and the middle third of the chromosome. Positive Eigen1 signals represent the B (inactive compartment) and negative Eigen1 signals represent the A (active compartment). The compartments at chromosome ends display increased interactions with each other, both in cis and in trans (**see Figure 2.2 a**). Also shown is the average binding of the lamin-associated protein LEM-2 along the chromosomes (grey). Overall compartmentalization correlates with LEM-2 binding, showing that compartments at both ends of chromosome I are located near the nuclear periphery.

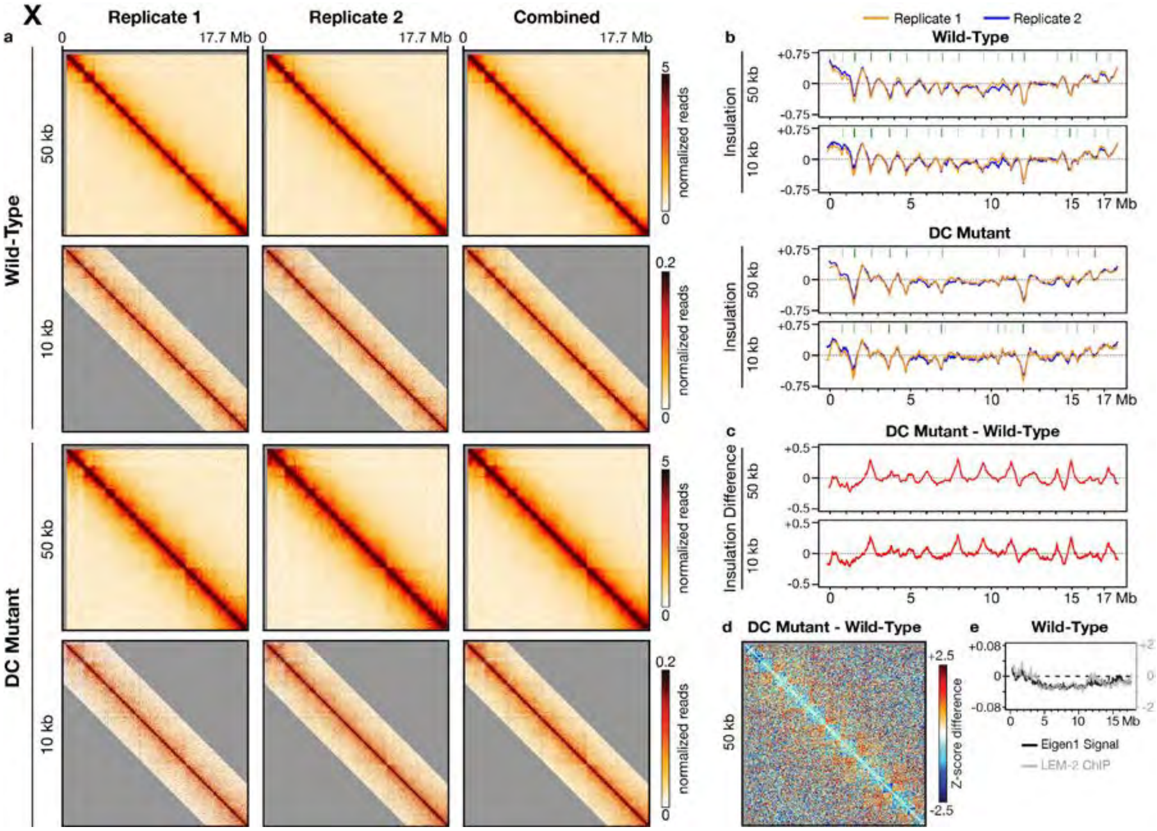


Figure 2.4 | Compartment and insulation analysis for chromosome X in wild-type embryos and DC mutant embryos.
a–e, See legend to Figure. 2.3. In e, only two compartments are observed for chromosome X, compared to three for chromosome I. Overall compartmentalization correlates with LEM-2 binding, showing that the compartment at the left end of chromosome X is located near the nuclear periphery.

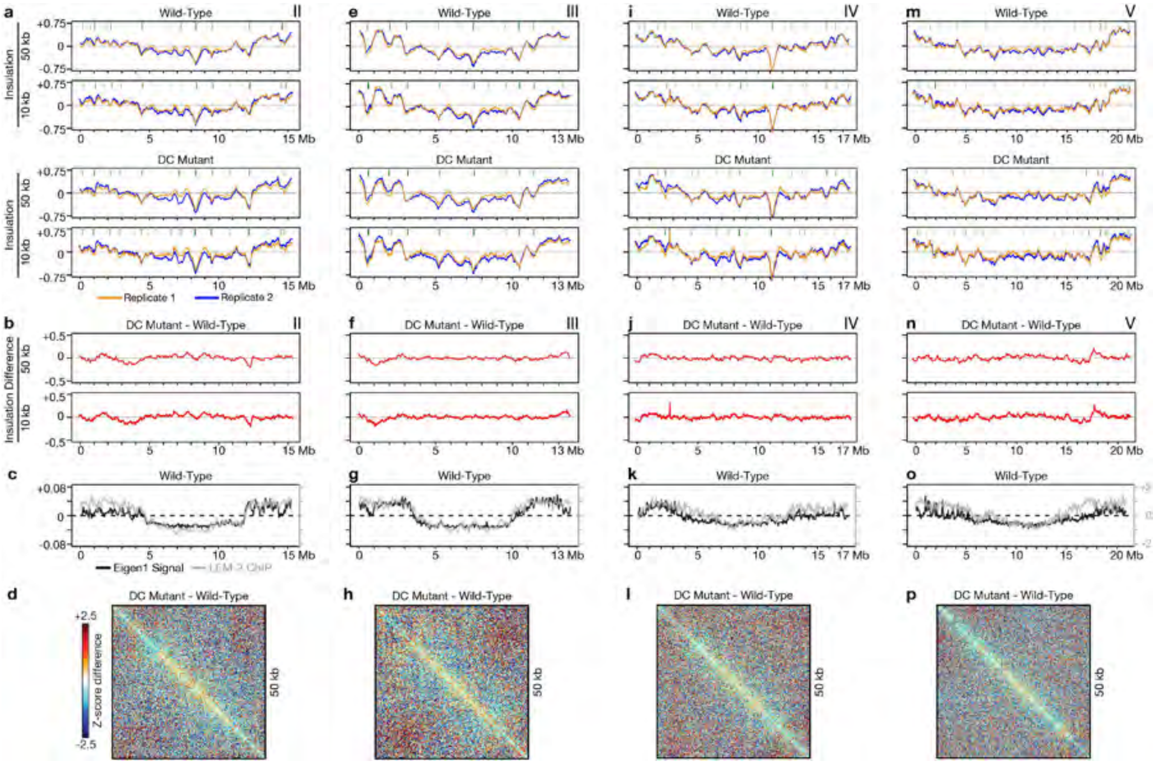


Figure 2.5 | Compartment and insulation analysis for chromosomes II, III, IV and V in wild-type embryos and DC mutant embryos.
a–d, Chromosome II. e–h, Chromosome III. i–l, Chromosome IV. m–p, Chromosome V. a, e, i, m, Insulation profiles for each biological replicate (replicate 1, orange line; replicate 2, blue line) for 50 kb or 10 kb binned Hi-C data in wild-type embryos and DC mutant embryos. Green lines, TAD boundaries identified from combined replicate data. b, f, j, n, Differential insulation plots made from insulation profiles (50 kb binned or 10 kb binned Hi-C data). c, g, k, o, Plots show chromosome compartment analysis calculated with 50 kb binned data. Average binding of the lamin-associated protein LEM-2 is shown along the chromosomes (grey). Compartmentalization correlates with LEM-2 binding; compartments at both ends of autosomes are near the nuclear periphery. d, h, l, p, Heatmaps (50 kb bins) show differences in chromatin interactions as the differences in Z-scores (DC mutant minus wild-type embryos).

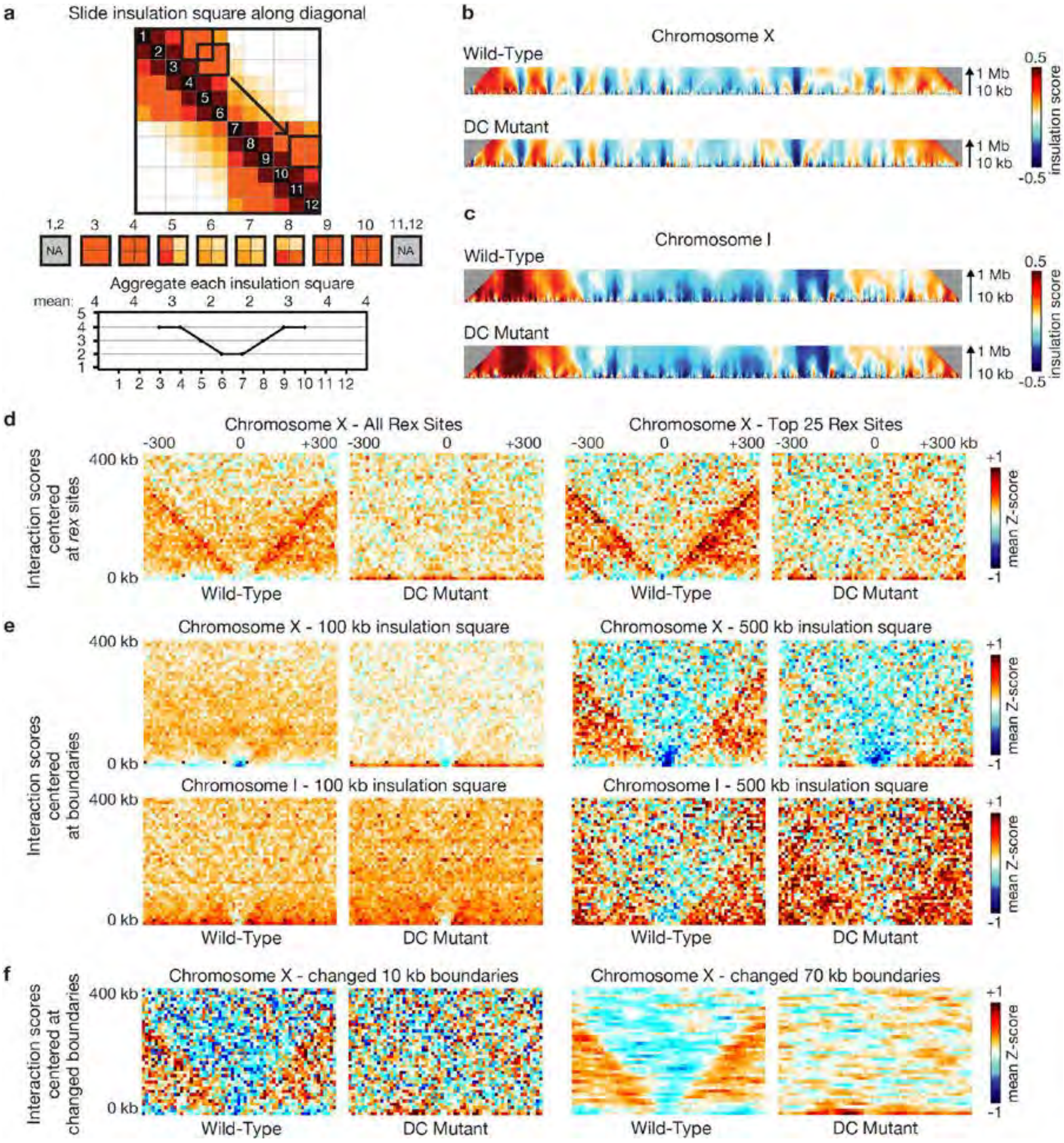


Figure 2.6 | Insulation profile calculation parameters and boundary calling. a, Cartoon shows approach for calculating the insulation profile. A square is slid along each diagonal bin of the interaction matrix to aggregate the amount of interactions that occur across each bin (up to a specified distance upstream and downstream of the bin). Bins with a high insulation effect (for example, at a TAD boundary) have a low insulation score (as measured by the insulation square). Bins with low insulation or boundary activity (for example, in the middle of a TAD) have a high insulation score. Minima along the insulation profile are potential

TAD boundaries. b, c, Heatmaps of chromosome X and chromosome I represent the insulation profiles calculated using insulation square sizes ranging from 10 kb to 1 Mb. At the 100 kb scale, weak boundaries are observed on the X chromosome and autosomes, but they are generally not changed in DC mutants. These boundaries cannot be detected at larger scales, meaning they do not insulate over distances beyond ~100 kb (see e). These smaller scale structures may represent sub-TAD domains not correlated with dosage compensation. Boundaries called using a 500 kb insulation square represent TAD boundaries that define domains observed in chromosome-wide interaction maps of the X chromosome at 10 kb resolution. These boundaries are used in this paper (**Figure 2.1**) and insulate over the larger distances defining the Mb-sized TADs. Boundaries on the X chromosome are the strongest and are DC dependent. d–f, Pile up plots depict aggregate (mean) Hi-C 10kb Z-score data centered on specified ‘anchors’ (for example, rex sites, boundaries, changed boundaries). d, Pile up plots centered on all rex sites or top 25 rex sites in wild-type and DC mutant. e, Pile up plots centered on all boundaries called using insulation squares of 100 kb (left) or 500 kb (right) for chromosome X and chromosome I in wild-type and DC mutant. f, Pile up plots using boundaries called with a 500 kb insulation square, centered (left) on the single 10 kb bin at the midpoint of all 8 changed boundaries or (right) on all seven 10 kb bins within changed boundaries.

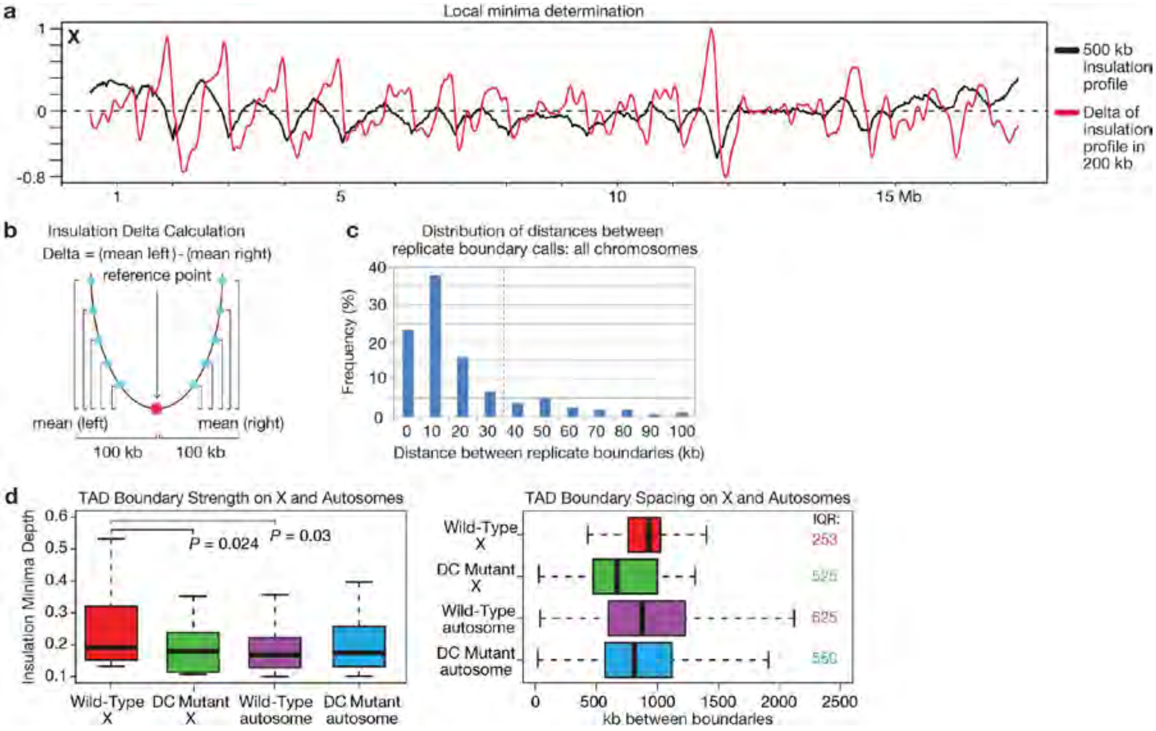


Figure 2.7 | TAD boundary analysis.

a, Insulation/delta plot of the 10 kb binned wild-type sample combined replicate chromosome X Hi-C data calculated using a 500-kb insulation square size. The insulation profile is depicted in black. In red, the ‘delta’ vector is depicted. It is derived from the insulation vector using a 200 kb delta window (see insulation methods). The ‘delta’ vector is used to facilitate the detection of the valleys/minima along the insulation profile. b, Cartoon example showing how the delta vector is calculated from the insulation data vector. For each bin (reference point) the average insulation differences are calculated between all points up to 100 kb left of the reference point relative to the reference point. The same is repeated for all points up to 100 kb right of the reference point. The delta value is then defined as the difference between the mean (left difference) and mean (right difference). c, Bar plot shows the distribution of distances between boundary calls obtained with biological replicate Hi-C data across all chromosomes. Dotted vertical line indicates that 630 kb was chosen for boundary definition, as it was the window in which the majority of replicate boundary calls (.80%) overlap. d, Boxplots compare boundary strength (left) and spacing (right) in wild-type versus DC mutant embryos. Wild-type boundary strength on chromosome X (defined as the distance from the insulation minimum to the largest neighboring maximum in the insulation profile) is higher than the DC mutant chromosome X boundary strength (P 5 0.024) and higher than the boundary strength on wild-type autosomes (P 5 0.03). TAD boundary strength on autosomes does not change in the DC mutant compared to the wild type (P 5 0.979). Boundaries on chromosome X have less variance in spacing (interquartile range (IQR) 5 253 kb) compared to the DC mutant (IQR 5 525 kb) embryos. DC mutant X chromosome boundary spacing is more similar to the boundary spacing on the autosomes in wild-type embryos (IQR 5 625 kb) and DC mutant embryos (IQR 5 550 kb).

Figure 2

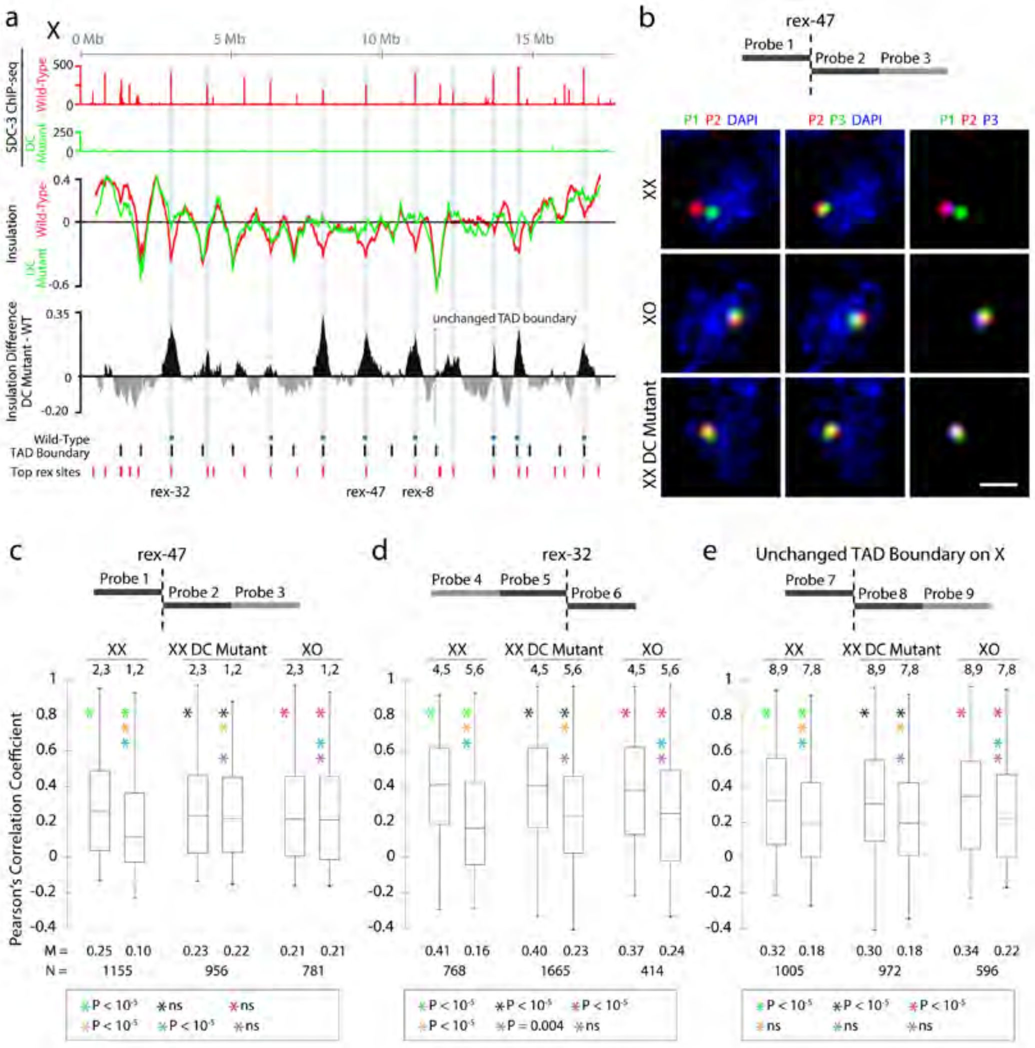


Figure 2.8 | FISH shows DCC-dependent TAD boundaries at high-affinity rex sites.

a, High DCC occupancy correlates with TAD boundaries lost or reduced upon DCC depletion. Top, ChIP-Seq profiles of DCC subunit SDC-3 in wild-type (red) and DC mutant (green) embryos. The y axis, reads per million (RPM) normalized to IgG control. Middle, insulation profiles of wild-type (red) and DC mutant (green) embryos. Bottom, insulation difference plot for wild-type insulation profile subtracted from DC mutant profile. Black lines, TAD boundary locations. Blue dots, boundaries with insulation changes .0.1 between wild-type and DC mutant embryos. Red lines, locations of 25 highest DCC-occupied rex sites. Cyan bars, sites with the largest insulation loss. b, Confocal images of embryonic nuclei of various genotypes stained with a DNA intercalating dye (blue) and 500 kb FISH probes around the rex-47 TAD boundary. c, d, e, Quantification of FISH probe co-localization confirms DCC-dependent and DCC-independent boundaries found by Hi-C. Box plots, distribution of Pearson's correlation coefficients between pairwise combinations of FISH probes within (blue) or across (orange) TADs. Boxes, middle 50% of coefficients. Center bars, median (M) coefficients. n, total number of nuclei. Asterisks of same color specify data compared using one-tailed Mann–Whitney U-test. NS, not significant.

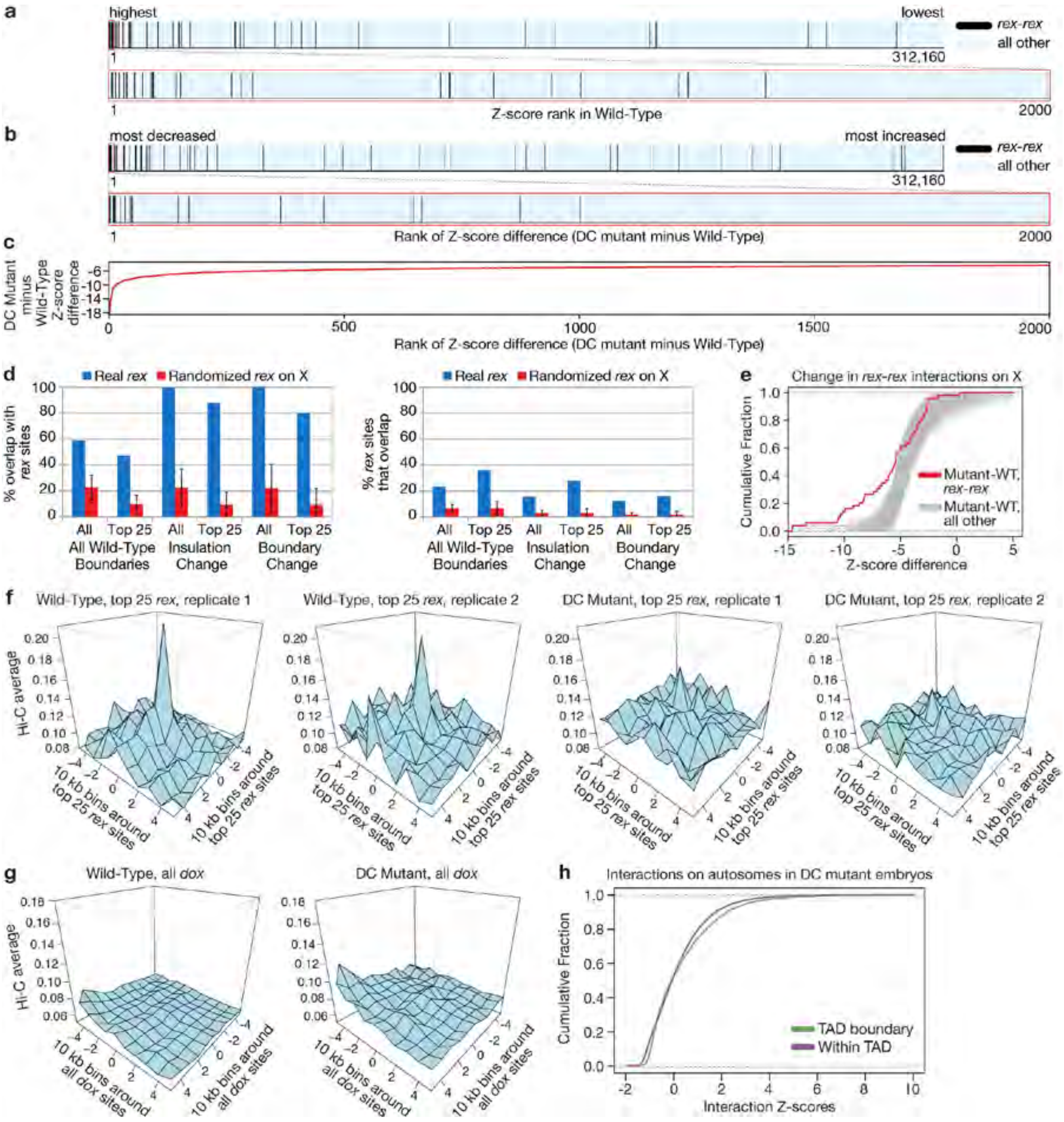


Figure 2.9 | rex sites are enriched at TAD boundaries and in top Hi-C interactions.

a, Tick plots rank the interaction Z-scores for the top 25 highest-affinity rex sites (black) among all other 10 kb bin Hi-C interactions on chromosome X (light blue). Bottom plot amplifies top 2,000 interactions. Density of black ticks (left) shows strong enrichment of rex–rex interactions among the most significant chromosome X interactions. b, Tick plots rank the Z-score differences (DC mutant minus wild-type embryos) for interactions between the top 25 rex sites among all other differences on chromosome X. Bottom plot amplifies top 2,000 changes. c, Quantification of Z-score differences for top 2,000 changes in (b). d, Bar graphs depict overlap between chromosome X TAD boundaries and rex sites. Three sets of TAD boundaries are shown: all 17 boundaries; 8 boundaries with an insulation change (DC mutant minus wild-type) .0.1; 5 boundaries present in wild-type embryos but absent in DC mutants. Overlap is calculated for the entire set of rex sites or just the top 25 rex sites. Percent of boundaries that overlap rex sites (left). Percent of rex sites that overlap each set of boundaries (right). Red bars, same sets of overlaps were calculated with 1,000 random sets of rex site positions along chromosome X. Average overlap and standard deviation are shown. No randomized set had as much overlap as the true rex set (P, 0.001). e, Cumulative comparison of Z-score differences for rex interactions and for 1,000 randomized sets of non-rex interactions (same number as in rex set). These rex or non-rex interactions had Z-scores .4in wild-type embryos. rex interactions are reduced more in DC mutants than other similarly strong chromosome X interactions (P 5 0.037; rex-interaction differences were significantly more reduced (KS test) than random interaction sets for 963 of 1,000 cases). f, 3D plots of Hi-C interaction profiles (normalized read counts) around top 25 rex sites for 2 Hi-C replicates of wild-type embryos or DC mutants. g, 3D plots of interactions between dox sites in wild-type embryos and DC mutants show no interaction peak. h, Cumulative plots show no difference in DC mutants for the distribution of autosomal Hi-C interaction Z-scores (10 kb bins) in TADs or at boundaries.

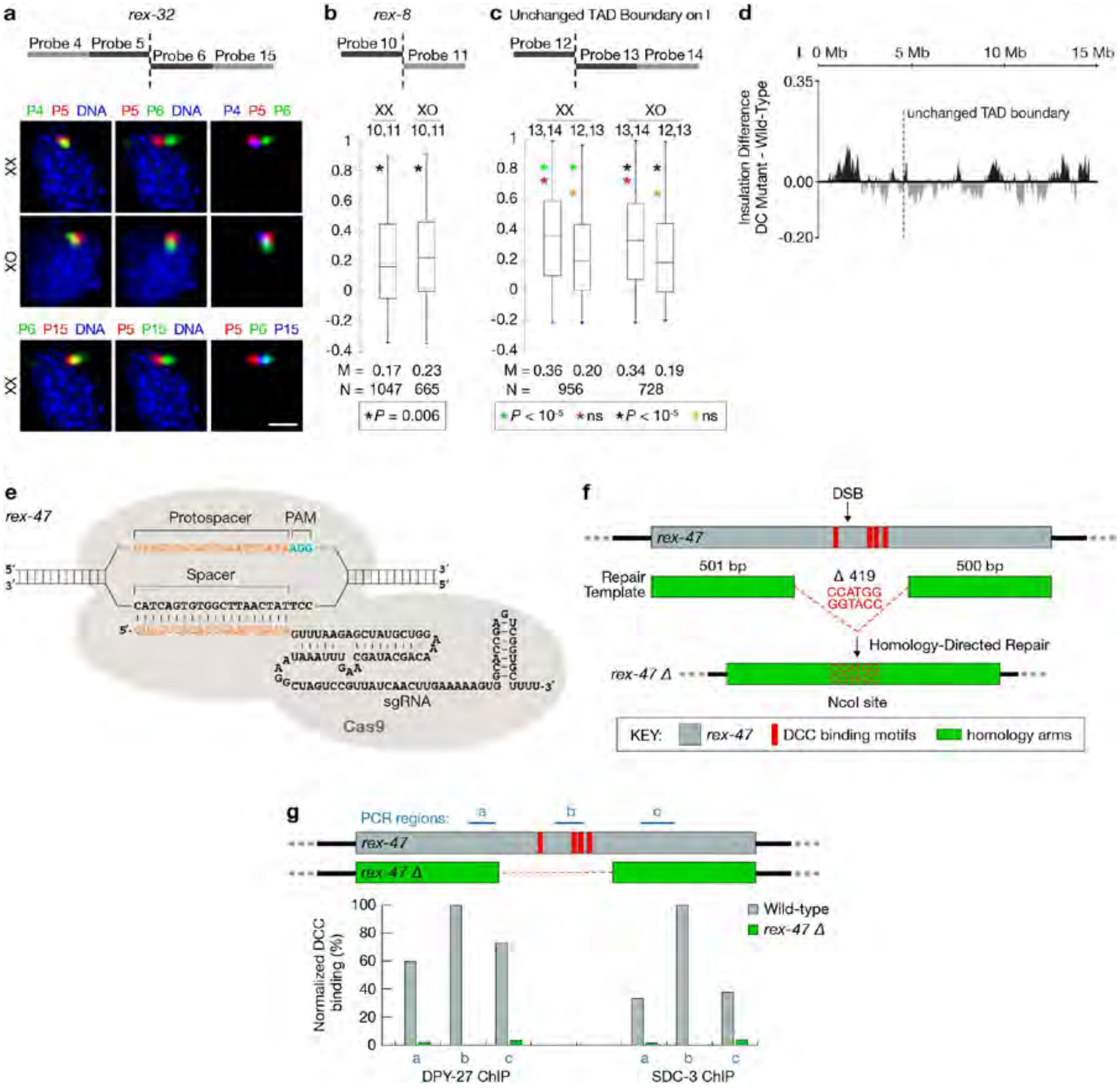


Figure 2.10 | Visualization and disruption of TAD boundaries.
a–d, Visualization of DCC-dependent TAD boundaries in single cells confirms Hi-C analysis. a, Representative confocal images of embryonic nuclei of different genotypes stained with a DNA intercalating dye (blue) and FISH probes surrounding rex-32. Scale bar, 1 mm. b, Quantification of co-localization between FISH probes flanking rex-8 (**see Figure 2.8 a**) in XX and XO embryos confirms the DCC-dependent boundary identified by Hi-C. Because TADs on either side of rex-8 are small, we could only use one 500 kb FISH probe for each TAD. c, Quantification of co-localization between FISH probes for a TAD boundary on chromosome I (dashed line in d) in XX and XO embryos confirms the DCC-independent boundary identified by Hi-C. b, c, Box plots show the distribution of

Pearson’s correlation coefficients between pairwise combinations of FISH probes. Boxes represent the middle 50% of coefficients, and the central bar within indicates the median coefficients (M). N, total number of nuclei. P values derived using the one-tailed Mann–Whitney U-test are shown below each graph. NS, not significant. d, Insulation difference plot of chromosome I for DC mutant insulation profile minus wild-type insulation profile. e–g, Deletion of endogenous rex-47 by Cas9 disrupts DCC binding and TAD boundary formation. e, Schematic illustration of the sgRNA–Cas9 complex interacting with the rex-47 target sequence. f, Cas9-mediated deletion of rex-47. Top, diagram showing the location of DCC binding motifs within rex-47 (red bars) and Cas9-induced double strand break (arrow). Middle, diagram of the double-stranded repair template containing two, 500 bp homology arms and an NcoI restriction site. Bottom, after precise homology-directed repair, a 419 bp region containing all DCC binding motifs was deleted and replaced with NcoI. g, Loss of DCC binding at endogenous locus carrying the rex-47 deletion. DCC binding at three, 100 bp regions located upstream (a), within (b) or downstream (c) of the 419 bp deletion was examined using ChIP–qPCR. Histogram shows the ChIP–qPCR signal for DCC components DPY-27 or SDC-3 at target regions relative to the level at region b in wild-type embryos.

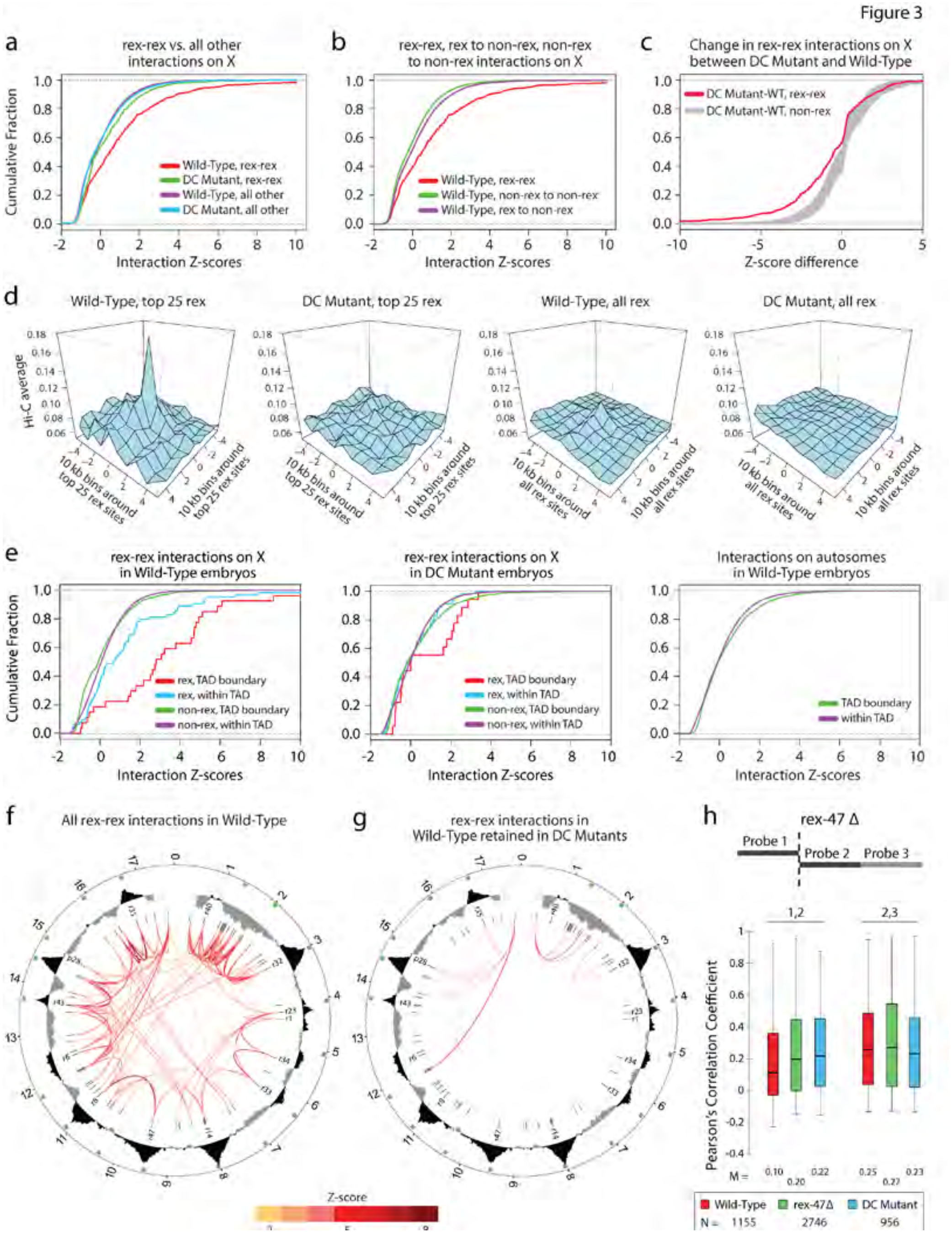


Figure 2.11 | Strong DCC-dependent interactions occur between high-affinity rex sites at TAD boundaries.

a, Cumulative distribution of Hi-C Z-scores for interactions between 10 kb bins with rex sites or with other X chromosome interactions in wild-type or DC mutant embryos. Interactions .4 Mb were excluded from panels a–e. P values are corrected for multiple testing. In wild-type embryos, rex–rex interactions are stronger than all other X chromosome interactions (P, 2 3 10216; two-sided KS test) and stronger than rex–rex interactions in DC mutants (P 5 1.5 3 1029; Wilcoxon signed rank test). b, Distributions of Hi-C Z-scores show that rex–rex interactions are stronger than non-rex interactions (P, 2 3 10216; two-sided KS test) or rex to non-rex interactions (P 5 1.7 3 10214; two-sided KS test). c, Distributions of Z-score differences (DC mutant minus wild-type) show that rex–rex interactions decrease more than any of 1,000 random sets of non-rex interactions of equal number (P, 0.001). d, Average Hi-C interaction profiles (normalized read counts) around pairs of top 25 rex sites or all known rex sites, in wild-type and DC mutant embryos. rex sites are centered at 0. e, Distributions of Hi-C Z-scores for interactions between bins with rex or non-rex sites at TAD boundaries or within TADs of wild-type (left) or DC mutant (middle) embryos. rex sites interact more at TAD boundaries than in TADs (P 5 0.0025). These sets of interactions are not different in DC mutants (P 5 0.348). Interactions at TAD boundaries or within TADs on autosomes (right). f, Circos plots depict all rex–rex interactions (Z-score .2, colored line) in 50 kb bins in wild-type embryos. Concentric circles show insulation difference plot (black and grey), wild-type TAD boundaries (green boxes), and rex sites (black lines, strongest sites named). g, rex–rex interactions in f that are retained in DC mutants. h, Deletion of rex-47 disrupts TAD boundary. Box plots of Pearson’s correlation coefficients for FISH probe combinations in wild-type, rex-47 D, and DC mutant. Probe overlap across TAD boundary increased in rex-47 D vs. wild-type (P, 0.01 ANOVA) but was not different in rex-47 D vs. DC mutants (P 5 NS, ANOVA). Probe overlap in TAD was not different in 3 strains (P 5 0.075, ANOVA).

Figure 4

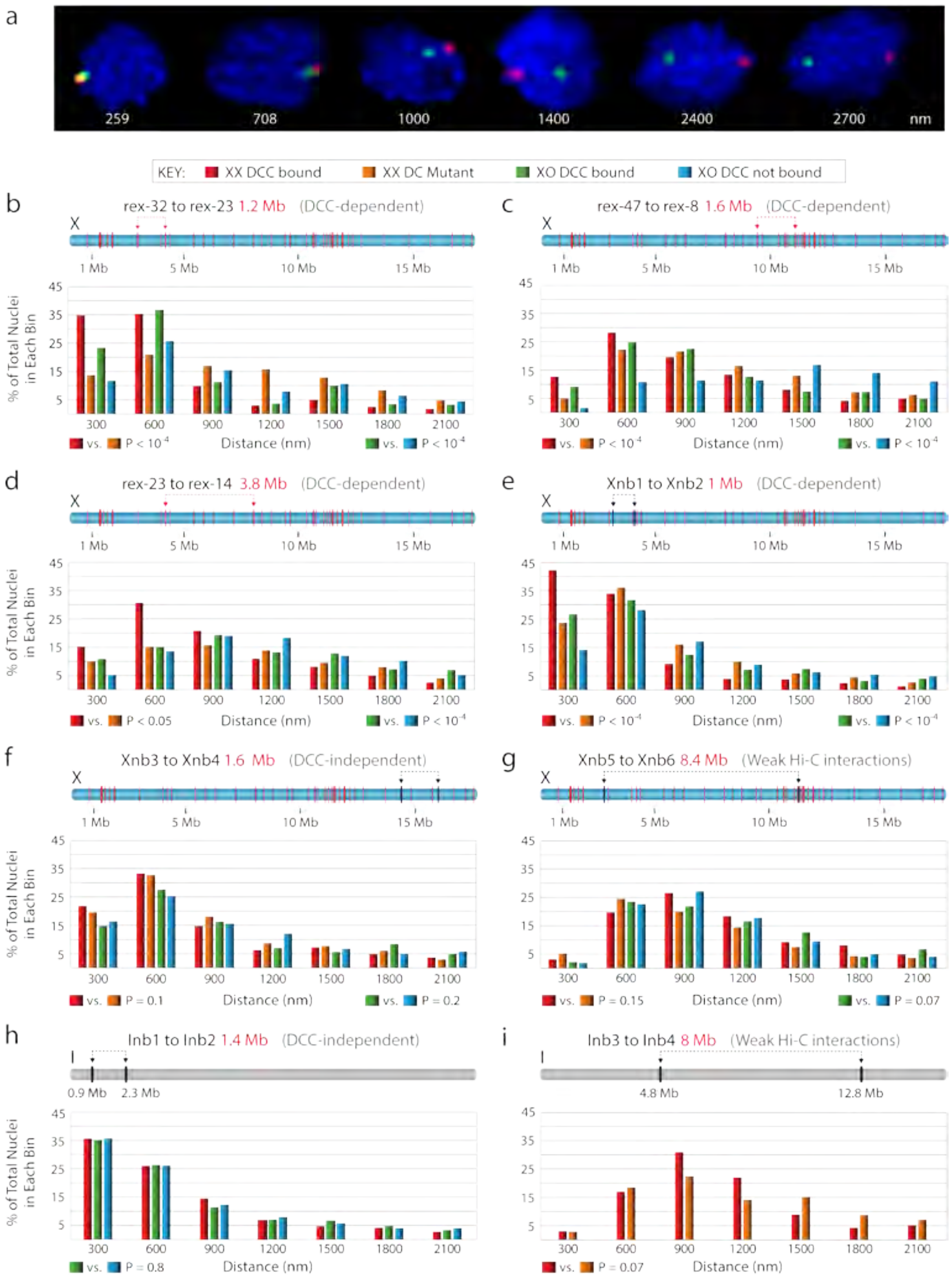


Figure 2.12 | Quantitative FISH shows DCC-dependent association of rex sites in single cells.

a, Representative embryonic nuclei show variability in spacing of FISH probes (red, green) targeting two rex sites. b–g, Quantification of the 3D distance between FISH probes in embryos of different genotypes. DCC binding to the single X chromosome of XO embryos was achieved using an XO lethal (xol-1) mutation, which activates sdc-2, the XX-specific trigger of DCC assembly20. Total number of nuclei is given in Figure 2.13 a–f. b–d, Pairs of rex sites at DCC-dependent TAD boundaries of varying genomic separation. e, A pair of sites on the X chromosome that lack DCC binding sites within 100 kb but have DCC-dependent Hi-C interactions. f, g, Loci on chromosome X and chromosome I that lack DCC binding sites within 80–90 kb and display DCC-independent Hi-C interactions. b–g, Distances between FISH spots were binned in 300 nm intervals and represented in relative frequency histograms. Schematic above each histogram depicts the locations of FISH probes (arrows), their genomic separation (red text), and the location of all rex sites (red bars) or sites lacking DCC binding (black). The DCC dependence or independence of the corresponding Hi-C interactions is indicated above the histogram (grey). P values comparing genotypes were calculated using the chi-square test to compare the 0–300 nm bin with 301–2,700 nm bins. The 0–300 nm bin contains FISH probes considered co-localized, because probes, 300 nm apart always overlap visually, while probes 700 nm apart appear only adjacent to each other.

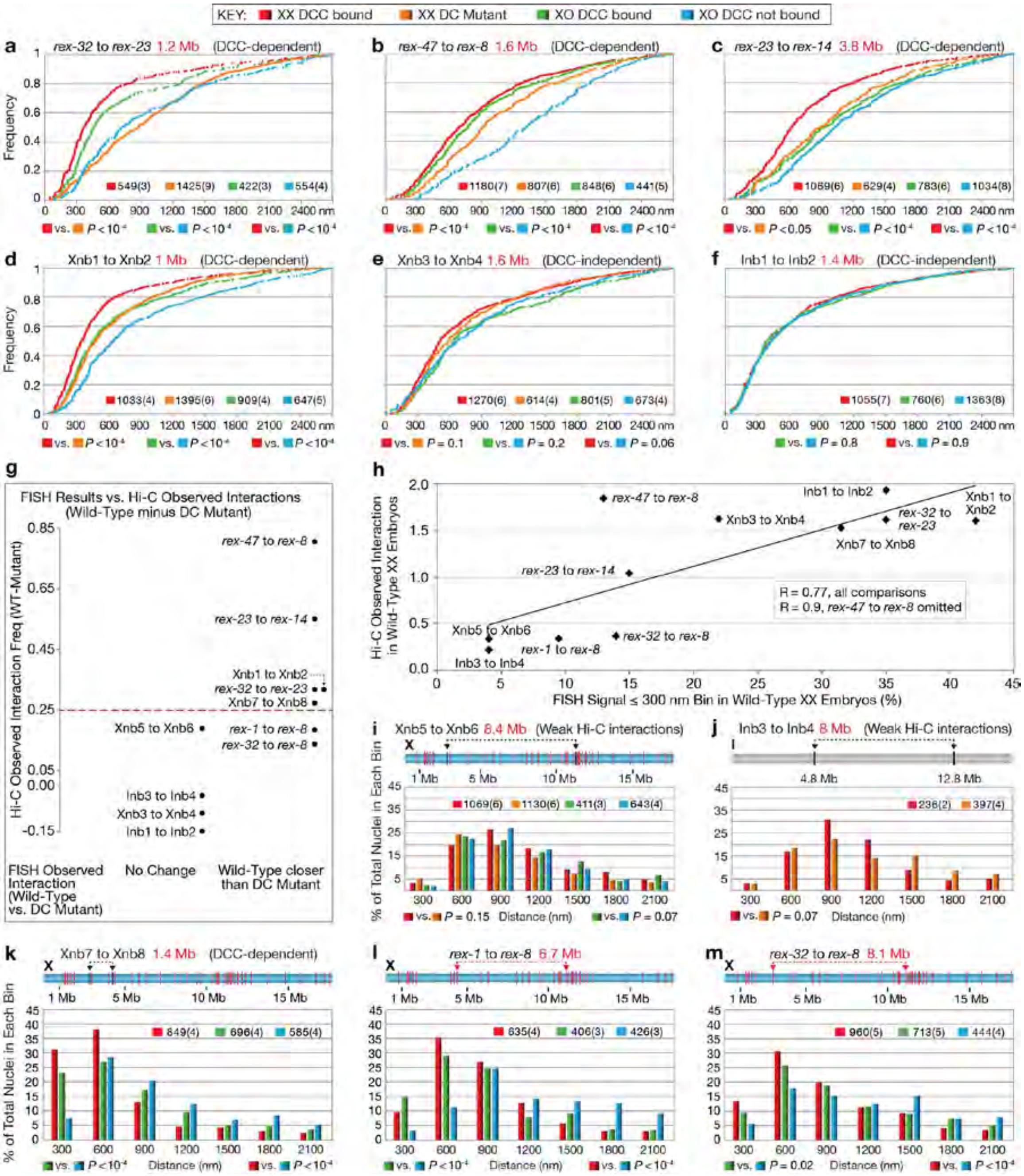


Figure 2.13 | Quantitative FISH shows that rex sites co-localize more frequently if the DCC is bound to chromosome X.

a–f, Data from histograms in Figure 2.12 b–g shown as cumulative plots. Number of nuclei and embryos (parentheses) assayed are shown (also for i–m). Distance between loci (red) and DCC dependence or independence of Hi-C interactions (black) are shown. P values (chi-squared test) compare values in the 0–300 nm bin to those in 301–2,700 nm bins. Same statistical analysis for (i–m). g, Correlation between DCC-dependent Hi-C interactions and DCC-dependent FISH co-localization. y axis, difference between wild-type and DC mutant Hi-C observed interaction frequency at 50 kb resolution. Higher number shows greater DCC-dependence. x axis shows two categories defined by FISH: sites with unchanged co-localization frequency in DC mutant (DCC-independent) (left); sites with less frequent co-localization in a DC mutant (DCC-dependent) (right). Red dotted line, cutoff for calling a Hi-C interaction ‘changed’ between the wild type and DC mutant. h, Scatter plot shows correlation between Hi-C and FISH data. y axis, Hi-C observed interaction frequency in 50 kb bins. x axis, percentage co-localization (that is, 300 nm bin) by FISH. R 5 0.77 for all comparisons; R 5 0.9 if the rex-47–rex-8 interaction is omitted. i–m, Histograms show quantification of 3D distances between two FISH probes. i, j, Distant loci on chromosome X or chromosome I with weak Hi-C interactions. k, DCC-dependent interaction between X sites lacking DCC binding. l–m, DCC-dependent interactions between distant rex sites.

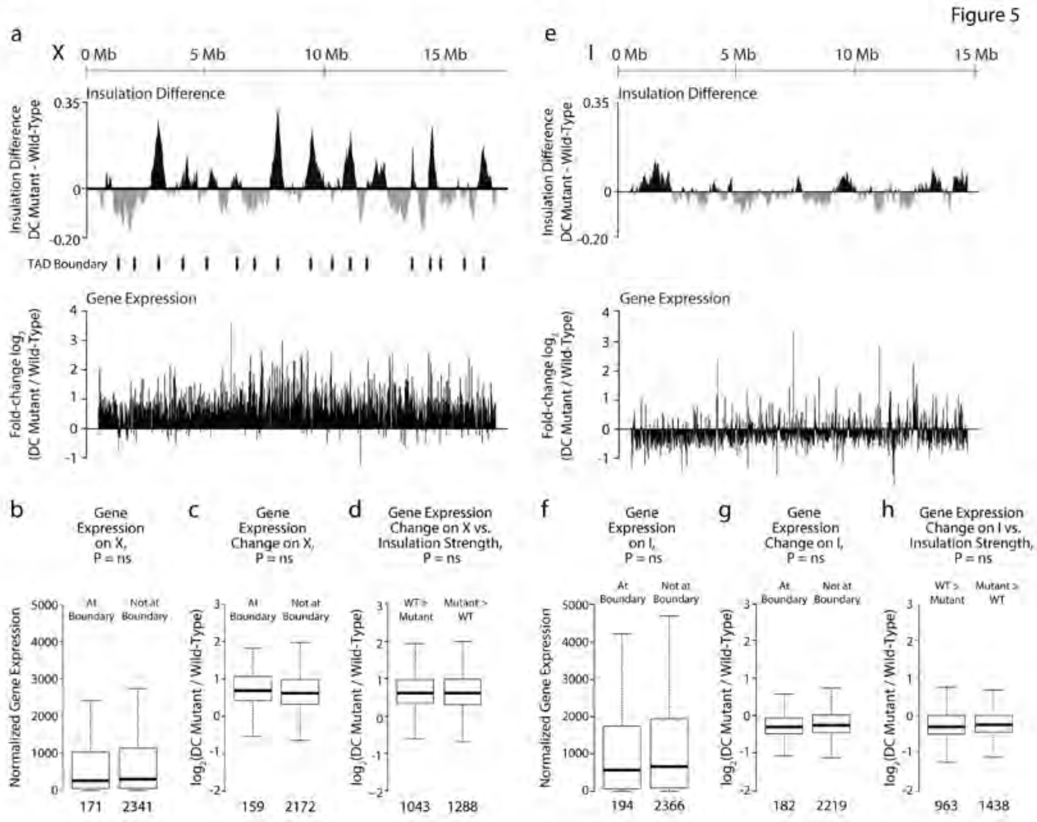
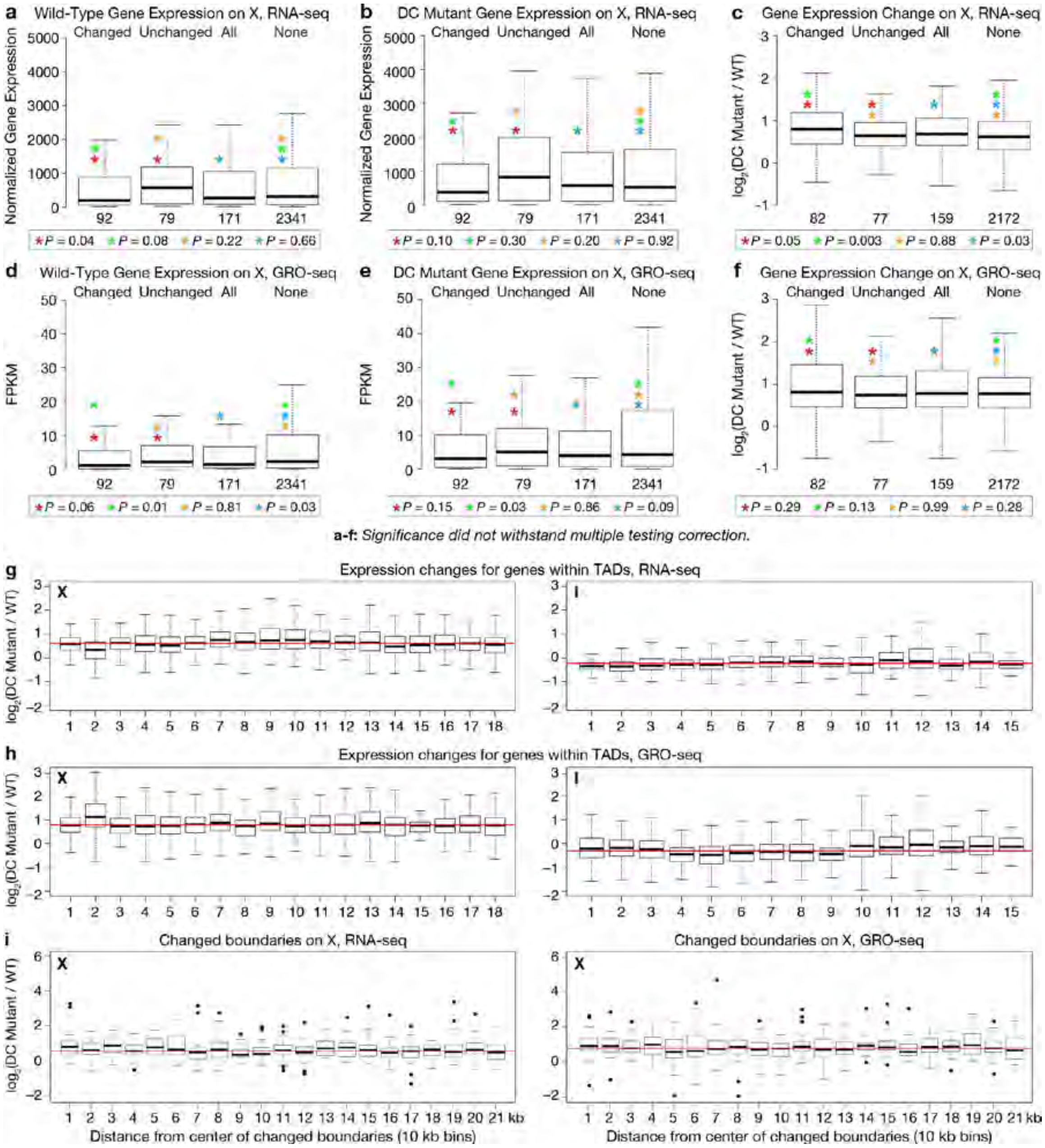


Figure 2.14 | DCC-dependent TADs influence global rather than local gene expression.

a, Insulation changes and TAD boundaries are compared to median fold-changes in expression (10 kb bins across chromosome X) between wild-type and DC mutant embryos. a–h, No discernible pattern was detected between mutant-induced changes in expression and gene locations relative to TADs or TAD boundaries. b, Box plots show comparison of expression levels for X chromosome genes within or outside TAD boundaries in wild-type embryos. Expression levels, normalized read number per kilobase of gene length. c, Box plots show expression changes for X chromosome genes within or outside TAD boundaries. d, Comparison of expression changes (DC mutant/wild-type) for X chromosome gene sets with greater insulation scores in wild-type embryos (grey domains in a) versus in DC mutants (black domains in a). e–h, Same as a–d but for genes on chromosome I. P values for b–d and f–h, Mann–Whitney U-test; no significant P values withstood multiple testing correction. NS, not significant. a, c, d, e, g, h, Lowest-expressed genes (bottom 10%) were removed from analyses.



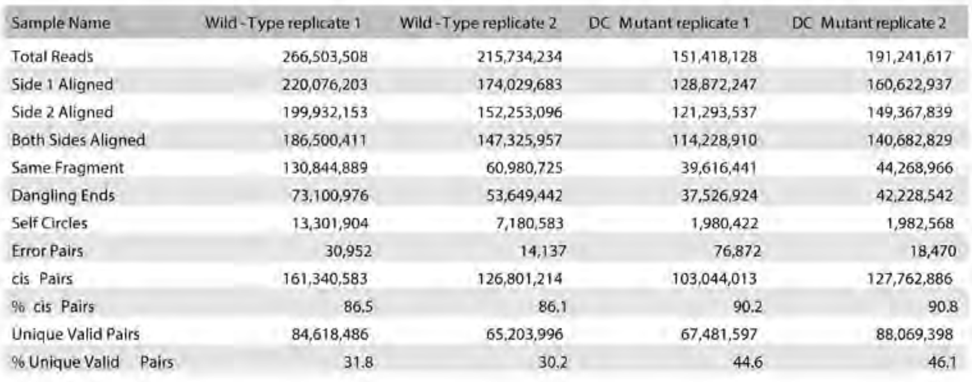
a-f: *Significance did not withstand multiple testing correction.*

Figure 2.15 | DCC-dependent TADs influence global rather than local gene expression. Gene expression analysis was assayed using RNA-Seq or GRO-Seq, as indicated.

a, b, Boxplots depict expression levels for wild-type or DC mutant embryos assayed by RNA-Seq for chromosome X genes at changed TAD boundaries, unchanged TAD boundaries, all TAD boundaries or genes not at TAD boundaries. Expression levels are given as normalized read number per kilobase of gene length. c, Boxplots depict the fold change in expression assayed by RNA-Seq between wild-type embryos and DC mutant embryos for genes at changed TAD boundaries, unchanged TAD boundaries, all TAD boundaries or genes not at boundaries. The lowest-expressing genes (bottom 10%) were removed from analysis. d–f, As in a–c, but assayed by GRO-Seq with gene expression levels given as fragments per kilobase of transcript per million mapped reads (FPKM). For a–f, P values were calculated using the Mann–Whitney U-test; significance did not withstand multiple testing correction. g, h, Boxplots depict the fold change in the gene expression between wild-type and DC mutant embryos based on RNA-Seq or GRO-Seq for chromosome X and chromosome I. Each box has genes from one TAD on chromosome X (left) or chromosome I (right). Lowest-expressing genes (bottom 10%) were removed from analysis. No discernible pattern was evident for expression changes versus gene location. i, Boxplots depict the fold change in chromosome X gene expression between wild-type embryos and DC mutant embryos relative to the distance from the TAD boundary. Each box contains genes in 10 kb bins radiating out from the center of each TAD boundary. The lowest-expressing genes (bottom 10%) were removed from analysis. No discernible pattern to the gene expression changes exists, as assayed by RNA-Seq (left) or GRO-Seq (right). Weak significance and lack of concordance between RNA-Seq and GRO-Seq data suggest no biologically relevant correlation between TAD boundaries and local regulation of gene expression.

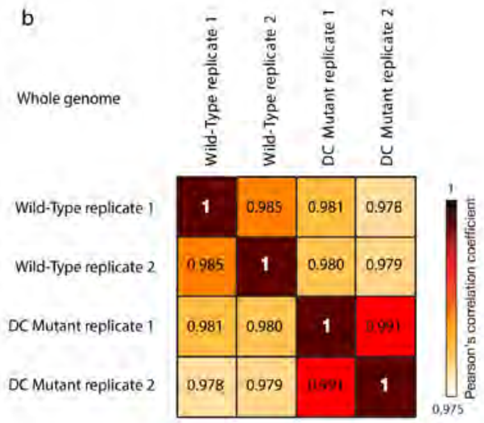
Tables

a



Sample Name	Wild-Type replicate 1	Wild-Type replicate 2	DC Mutant replicate 1	DC Mutant replicate 2
Total Reads	266,503,508	215,734,234	151,418,128	191,241,617
Side 1 Aligned	220,076,203	174,029,683	128,872,247	160,622,937
Side 2 Aligned	199,932,153	152,253,096	121,293,537	149,367,839
Both Sides Aligned	186,500,411	147,325,957	114,228,910	140,682,829
Same Fragment	130,844,889	60,980,725	39,616,441	44,268,966
Dangling Ends	73,100,976	53,649,442	37,526,924	42,228,542
Self Circles	13,301,904	7,180,583	1,980,422	1,982,568
Error Pairs	30,952	14,137	76,872	18,470
cis Pairs	161,340,583	126,801,214	103,044,013	127,762,886
% cis Pairs	86.5	86.1	90.2	90.8
Unique Valid Pairs	84,618,486	65,203,996	67,481,597	88,069,398
% Unique Valid Pairs	31.8	30.2	44.6	46.1

b



Whole genome	Wild-Type replicate 1	Wild-Type replicate 2	DC Mutant replicate 1	DC Mutant replicate 2
Wild-Type replicate 1	1	0.985	0.981	0.978
Wild-Type replicate 2	0.985	1	0.980	0.979
DC Mutant replicate 1	0.981	0.980	1	0.991
DC Mutant replicate 2	0.978	0.979	0.991	1

Table 2.1 | Hi-C Statistics

Name	Rank	Normalized SDC-3 embryo (RPM)	rex site midpoint on X	rex overlap with HOT or cold sites	rex overlap with boundary (B)
Prex -28	1	4256	14525823	HOT	not at boundary
rex-43	2	4025	13700797	HOT	B13 changed
rex-35	3	4792	16682015	HOT	B17 changed
rex-34	4	3934	5429511	HOT	not at boundary
rex-8	5	3494	11094129	cold	B11 changed
rex-40	6	4527	806703	HOT	not at boundary
rex-33	7	2788	6296501	HOT	B6 changed
rex-16	8	2820	11937457	HOT	not at boundary
rex-45	9	2579	1345039	HOT	B1
rex-23	10	1820	4209254	HOT	not at boundary
rex-32	11	2259	2997077	HOT	B3 changed
Prex -7	12	2081	1627111	HOT	not at boundary
rex-14	13	1645	8036363	HOT	B8 changed
Prex -1	14	1862	410189	HOT	not at boundary
rex-47	15	1581	9465796	HOT	B9 changed
rex-1	16	1649	4395609	cold	not at boundary
rex-24	17	1649	7181758	cold	not at boundary
rex-44	18	1910	1323202	cold	B1
Prex -30	19	1426	16056503	HOT	not at boundary
rex-39	20	1351	14813494	cold	not at boundary
rex-2	21	765	1909564	cold	not at boundary
rex-42	22	1329	17181330	cold	not at boundary
rex-46	23	1142	15736582	cold	not at boundary
rex-36	24	992	11899184	cold	not at boundary
rex-6	25	1059	12363153	HOT	not at boundary
rex-7	26	475	11924275	cold	not at boundary
Prex -31	27	793	16205944	cold	not at boundary
rex-28	28	612	10668006	cold	not at boundary
rex-18	29	431	1379678	cold	B1
Prex -22	30	562	13514564	cold	not at boundary
Prex -6	31	388	1389606	cold	not at boundary
rex-31	32	415	12730204	cold	not at boundary
rex-25	33	67	8404339	HOT	not at boundary
rex-41	34	329	17544537	cold	not at boundary
Prex -14	35	29	7334457	HOT	not at boundary
rex-37	36	31	8811062	cold	not at boundary
rex-21	37	262	1888595	cold	not at boundary
Prex -23	38	132	13696333	cold	B13 changed
rex-20	39	210	1682738	cold	not at boundary
Prex -11	40	21	2809799	HOT	not at boundary
Prex -3	41	59	1302480	cold	not at boundary
Prex -26	42	38	14027065	cold	not at boundary
rex-26	43	191	10353987	cold	B10
Prex -27	44	19	14480338	cold	B14 changed
Prex -16	45	17	8737190	cold	not at boundary
rex-29	46	161	10756297	cold	not at boundary
Prex -15	47	131	8019825	cold	B8 changed
Prex -20	48	29	12452335	cold	not at boundary
rex-4	49	21	11522205	cold	not at boundary
rex-17	50	98	8048784	cold	B8 changed
rex-38	51	70	5859903	cold	not at boundary
Prex -2	52	18	1223959	cold	not at boundary
rex-5	53	71	11472575	cold	not at boundary
Prex -25	54	8	13920715	cold	not at boundary
rex-19	55	66	1492424	cold	not at boundary
rex-27	56	12	10623247	cold	not at boundary
Prex -9	57	10	2583339	cold	not at boundary
rex-22	58	9	4010652	cold	not at boundary
rex-30	59	7	11223019	cold	not at boundary
Prex -33	60	19	16944007	cold	not at boundary
rex-3	61	4	11361260	cold	not at boundary
Prex -21	62	7	12890286	cold	not at boundary
Prex -8	63	1	2227265	cold	not at boundary
Prex -10	64	6	2731772	cold	not at boundary

		All HOT sites			non-rex HOT sites	
		Autosomes (all TADs)	X (all TADs)	X (changed TADs)	X (all TADs)	X (changed TADs)
% boundaries that overlap HOT sites	Observed	8.4	52.9	87.5	17.6	25
	Expected	21.2	25.5	26.4	19.4	20.1
	P-value	0.89	0.021	<0.001	0.667	0.212
% HOT sites that overlap boundaries	Observed	3.0	14.7	12	5.5	3.6
	Expected	6.7	6.7	3.3	6.7	3.3
	P-value	0.8314	0.019	<0.001	0.715	0.535

	No. of HOT sites	No. of non-rex HOT sites	No. of TAD boundaries	No. of changed TAD boundaries
Autosomes	301	301	83	0
X	75	55	17	8

Table 2.2 | REX / PREX / HOT Sites

Acknowledgements

We thank D. Mets for the computer script to analyse FISH data; A. Michel for the initial conformation experiments; D. Stalford for figures; Vincent J. Coates Genomics Sequencing Laboratory (NIH S10RR029668); and K. Brejc, T. Cline and A. Freund for manuscript comments. Research was supported in part by NIGMS grant R01 GM030702 to B.J.M. and NHGRI grant R01 HG003143 to J.D. B.J.M. is an investigator of the Howard Hughes Medical Institute.

Author Contributions

E.C. conducted Hi-C, ChIP-Seq and FISH experiments. Q.B. conducted FISH and rex-47 deletion experiments. R.P.M. conducted statistical and long-range interaction analyses. B.R.L. analysed Hi-C data and mapped TADs. B.S.W. conducted RNA-Seq studies. E.J.R., S.U. and all authors analysed data and edited the manuscript, J.D. guided and performed Hi-C analysis and wrote manuscript sections. B.J.M. guided the study and wrote the manuscript.

Author Information

Hi-C, ChIP-Seq, and RNA-Seq data are available at the NCBI Gene Expression Omnibus (GEO) repository, accession number GSE59716. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.J.M. (bjmeyer@berkeley.edu) or J.D. (Job.Dekker@umassmed.edu).

Methods

Nematode strains. The strains used in this study are as follows. Wild-type: TY125, N2 Bristol, XX. Dosage compensation mutants: sdc-2 (y93, RNAi) X (XX strain used in all experiments requiring a DC mutant strain, except those listed below using TY2222 or TY1996); TY1996, szT1/sdc-2(y74) unc-3(e151) X (XX DC mutant in Figures 2.8 b–e, 2.11 h and 2.12 b–f and Figure 2.13 a–e, i); TY2222, her-1(hvly101) V; xol-1(y9) sdc-2(y74) unc-9(e101) X (XX DC mutant used only in Figure 2.13 j); TY0810, sdc-2(y93) X (XX strain used to create sdc-2 (y93, RNAi) XX embryos); TY0525, him-8(e1489) IV; xol-1(9) X (used for XX and XO DCC bound). Strain to generate XO males lacking DCC binding: CB1489, him-8(e1489) IV (used for XO DCC not bound).

Sample size

No statistical methods were used to predetermine sample size. ChIP-Seq, RNA-Seq and chromosome conformation capture. To obtain wild-type control embryos, wild-type N2 worms were grown at 20 uC on NG agar plates with concentrated HB101 bacteria. For DC mutant embryos, 10 ml of packed synchronous sdc-2(y93) L1 worms were placed onto 10 cm RNAi plates (NG agar with 1 mM IPTG and 100 mg/ml Carbenicillin) seeded with 2–3 ml of concentrated HT115 (DE3) bacteria carrying the Ahringer feeding library plasmid

[58] expressing the coding region of sdc-2. The RNAi plates were incubated at 25 uC overnight before L1 larvae were added.

Immunofluorescence and FISH analysis

Animals were grown at 20 uConNG agar plates seeded with OP50 grown in Luria Broth (LB). The worms were grown at 20 uC until gravid adults, then dissected for their embryos and stained as described below.

Antibodies

Rat polyclonal SDC-3 (PEM4A) antibodies were made against amino acids 1067-1340 of SDC-3 fused to GST. Rabbit polyclonal antibodies against DPY-27 (rb699) and SDC-3 (rb1079) were as described previously [51], [59]. Mouse monoclonal Mab414 antibody (1 mg ml21) was obtained from Abcam (ab24609). Normal rabbit IgG (400 mgml21) was from Santa Cruz Biotechnology (sc-2027). Rabbit polyclonal LMN-1 antibody (500 mg ml21) was from SDIX (3853.00.02) ChIP-Seq library creation and analysis. Libraries were made and analysed from one batch of wild-type embryos (data consistent with all previously wild-type published ChIP-Seq data [53]) and two biological replicates of sdc-2(y93, RNAi) embryos as described previously [53].

Modified Hi-C embryo isolation and crosslinking

Worms of appropriate genotype, either wild-type worms (two biological replicates) or sdc-2(y93, RNAi) (two biological replicates), were grown until gravid adults. The worms were collected and bleached to release the embryos and remove the carcasses. Following bleaching, embryos were centrifuged for, 45 s at 1,500–1,800 rpm and washed 3 times in 13 M9 buffer to remove bleach solution. An equal volume of 13 M9 was added to the embryos and they were frozen in 1 ml aliquots and stored at 280 uC. The frozen embryos were thawed on ice and supplemented with 1 mM PMSF and 5 mM DTT. The embryos were then washed once in 50 ml formaldehyde solution (13 M9 solution with 2% (v/v) formaldehyde, Polysciences 18814-20). Embryos were cross-linked in 50 ml of formaldehyde solution for 30 min at room temperature while shaking. Following crosslinking, embryos were washed once with 50 ml of 100 mM Tris-HCl, pH 7.5, followed by two 50 ml washes of 13 M9. The embryos were then washed once in lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl and 0.2% (v/v) Igepal CA-630 (Sigma I8896)) supplemented with 5 mM DTT, 1 mM PMSF, 0.1% (v/v) protease inhibitors (EMD 539134) and 0.5 mM EGTA. To obtain extract, embryos were dounced 10 times using the large pestle (Kontes 2 ml glass dounce, Spectrum 985-44182; clearance 0.076–0.127 mm), and then 10 times using the small pestle (clearance 0.01–0.069 mm). All douncing steps were performed on ice. The dounced extract was spun for 5 min at 100g at 4 uC, and the supernatant was saved. The pellet was re-suspended in 750 ml of supplemented lysis buffer and dounced again. This procedure was repeated 7–10 times. After each spin, a 9 ml aliquot was taken from the supernatant, mixed with 1 mlof10ngml DAPI and visualized under a microscope. All supernatants containing only nuclei, and not broken carcasses, were combined. An aliquot of the combined supernatant was stained with DAPI and the nuclei were counted using a haemocytometer, and then spun down for 5 min at 2,000g at 4 uC. The nuclei were re-suspended in the appropriate volume of 1.253 DpnII buffer (NEB B0543S) to create a Hi-C library as described below.

Modified Hi-C library preparation

The Hi-C libraries were made as described below. The protocol was based on a 3C library preparation followed by modifications [4], [60], [61]. Approximately 1.5 3 108 C. elegans nuclei were pipetted into 5–10 1.7 ml tubes and re-suspended in 300 ml of 1.253 DpnII buffer. 38 ml of 1% (w/v) SDS was added per tube and the tubes were incubated at 65 uC for 10 min. After the addition of 34 ml of 20% (v/v) Triton X-100, the tubes were incubated at 37 uC for 1 h, shaking at 1,000 rpm. 30 ml (1,500 U) of DpnII (NEB R0543M) were added to each tube, and they were incubated overnight at 37 uC while rocking. 26 ml of 20% (w/v) SDS was added to each tube and they were incubated at 65 uC for 20 min, shaking at 1,000 rpm. The reaction was then added to 7.6 ml of ligation master mix (745 ml of 10% Triton X-100, 745 ml of 10X T4 ligation buffer (500 mM Tris-HCl, pH 7.5, 100 mMMgCl2, 100 mMDTT), 80 mlof 10 mg ml21 BSA, 80 ml of 100 mM ATP, and 5.96 ml water). 100 ml (100 U) T4 DNA ligase (Invitrogen 15224-025) was added and the reactions were incubated for 4 h at 16 uC. After incubation 50 mlof10mgml21 proteinase K was added and the tubes were further incubated at 65 uC overnight. The next day, 50 mlof 10 mg ml21 proteinase K was added to the reactions, and they were incubated at 65 uC for an additional 2 h. 2 ml of RNaseA (1 mg ml21) was added to each sample and incubated for 30 min at 37 uC. The ligated DNA was then phenol-chloroform extracted and ethanol precipitated overnight. DNA was pelleted at 14,000g for 30 min at 4 uC, and then washed twice with 70% ethanol and air-dried. The DNA pellets from all Hi-C reactions were combined and dissolved in a total of 500 mlof 13 TE buffer, pH 8.0. Excess salt was removed from the samples via centrifugation using a filter unit (AMICON Ultra Centrifugal Filter Unit – 0.5 ml 30 kDa) following the manufacture's instruction. Briefly, the samples were spun at 18,000g for 10 min to reduce the volume to 40–50 ml. Flow through was discarded and 450 mlof13 TE, pH 8.0 buffer was added to each unit and spun as before. This wash step was repeated at least 5 times. The volume of the eluate was adjusted to 100 ml with water. The concentration of DNA was determined and 10 mg of the Hi-C library was resuspended in 100 ml of water. AMPure beads, supplied as a suspension of magnetic beads in a PEG solution (Beckman Coulter, A63880), were used to remove large DNA fragments (.10 kb), following the protocol provided by the manufacturer. Specifically, for the first DNA selection, 35 mlof AMPure beads were added to the 100 ml of DNA. The supernatant was kept and the beads, which bind only large DNA molecules under these PEG conditions, were discarded. To then remove smaller fragments, 65 ml of AMPure beads were added to the supernatant and the beads, which bind all DNA molecules greater than 100 bp due to the greater PEG concentration, were kept and washed with 70%ethanol. The DNA was eluted from the beads in 100 ml of 10 mM Tris-HCl, pH 8.5. The eluted DNA was then adjusted to 125 ml with 13 TE, pH 8.0 and sheared to 500–1,000 bp using a Covaris S2 (Covaris, 520045) in micro tubes with the following settings: duty cycle, 5%; intensity, 3; cycles/burst, 200; time, 65 s. The sheared DNA was then size selected for fragments larger than, 100 bp using AMPure beads and eluted in 34 ml of water. The DNA was quantified and 500 ng was used to make a paired-end Illumina sequencing library following the standard protocol (PE-930–1001), with the exception that we size selected 500–600 bp at the gel excision step before adding adapters for sequencing. The library was sequenced using 100 bp paired end reads with a HiSeq2500 s machine.

Read mapping/binning/ICE correction

Iterative mapping and error correction of the chromatin interaction data were performed as previously described [62]. Supplementary Table 1 summarizes the mapping results and lists the different categories of DNA molecules encountered in the libraries. We obtained around 70 million valid pairs that represent chromatin interactions per replicate. The frequency of redundant read pairs, due to PCR amplification were found to be below, 5% and were removed. The number of Hi-C interactions mapped to sequences belonging to homologous chromosomes (both intra-chromosomal (cis) and inter-homologue (trans) interactions) was much higher than the interactions mapped to non-homologous chromosomes (inter-chromosomal (trans) interactions). Assuming that inter-homologue interactions (trans) are as frequent as non-homologous inter-chromosomal interactions (trans), we estimate that 80–90% of interactions mapped to the same chromosomes are intra-chromosomal (cis) interactions, with DC mutants (90%) higher than wild type (.85%). Whether this difference reflects a biological phenomenon or is due to technical differences is currently not known. Conversion of interaction data into Z-scores eliminates this difference (see below).

The data were binned at both 10 kb and 50 kb non-overlapping genomic intervals. Binned data were normalized for intrinsic biases such as differences in number of restriction fragments within bins using the previously developed ICE method29. To normalize for differences in read depth of different data sets we summed the entire genome-wide binned ICE-corrected interaction matrix, excluding the diagonal (x 5 y) bins. We then transformed each interaction into a fraction of the matrix sum (minus diagonal x 5 y bins). Each fraction was then multiplied by 1,000,000. Biological replicates were highly correlated (Pearson's correlation coefficients .0.98 for 50 kb binned data excluding short-range interactions up to 50 kb). The correlations between biological replicates were higher than those between the wild type and DC mutant. Overall these numbers indicate that the modified Hi-C procedure was reproducible and performed as expected. For most analyses sequence reads obtained for biological replicates were pooled and ICE-corrected as described above to create a combined replicate data set.

At 10 kb resolution, very long-range interactions are not sampled deeply enough to provide robust and reliable data. Therefore, we truncated the 10 kb binned data to include only cis interaction pairs separated by 4 Mb or less in linear genomic distance. This distance cutoff was chosen based on the observation that beyond this point, both wild-type and DC mutant data sets have no observed reads in more than 50% of bin–bin interactions. In addition to limiting the dynamic range of interaction counts at these large distances, this high frequency of un-sampled interactions beyond 4 Mb causes a dramatic collapse in the standard deviation of the overall chromatin interaction decay over distance, making the LOWESS expected and Z-score calculations beyond 4 Mb unreliable. For 50 kb bins, all distances were included in analyses, because the coverage of cis interaction pairs never dropped below 50% for any distance at this resolution.

TAD calling (insulation square analysis). To calculate the ‘insulation’ score of each bin in the 10 kb binned Hi-C data, we calculated the average number of interactions that occurred across each bin. This can be visualized by sliding a 500 kb 3 500 kb (50 bins 3 50 bins) (Figures 2.6 and 2.7) square along the matrix diagonal, and aggregating all signal within the square. The mean signal within the square was then assigned to the 10 kb diagonal bin and this procedure was then repeated for all 10 kb diagonal bins. For any bins within 500 kb of the matrix start/end, an insulation score was not assigned, as the 500 kb 3 500 kb insulation square would extend beyond the matrix bounds. The insulation score was then normalized relative to all of the insulation scores across each chromosome by calculating the log2 ratio of each bin's insulation score and the mean of all insulation scores. Valleys/minima along the normalized insulation score vector represent loci of reduced Hi-C interactions that occur across the bin. These valleys/minima are interpreted as TAD boundaries or areas of high local insulation. The valleys/minima were detected as follows: first, a delta vector was calculated to approximate the slope of the normalized insulation vector. The delta vector is defined as the difference between the amount of insulation change 100 kb to the left of the central bin and 100 kb to the right of the central bin (relative to the central bin) (Figure 2.7 a, b). The delta vector crosses the horizontal 0 at all peaks and all valleys. All bins where the delta vector crosses 0 were extracted. Zero-crossings occurring at peaks were removed, and the remaining zero-crossings, all occurring at potential valleys were passed through a boundary strength filter. The boundary strength was defined as the difference in the delta vector between the local maximum to the left and local minimum to the right of the boundary bin. All boundaries with a boundary strength < 0.1 were removed. This method in practice is very similar to the widely used zero-derivative method for detecting peaks/valleys in various signal vectors.

The precision with which we could define a boundary was determined by comparing boundary calls across biological replicates (**Figure 2.7 c**). The final boundary zones were defined as 630 kb around the pooled replicate insulation minima bins (70 kb total) because most (.80%) replicate boundary calls overlapped within this window. Wild-type and DC mutant insulation profiles were compared by subtracting the wild-type insulation profile from the DC mutant insulation profile. We compared the insulation profiles and boundary calls resulting from a full range of alternative insulation square sizes (**Figure 2.6 b, c**). We find that a 500 kb square size captures best the major robust boundaries that change in the DC mutant. In contrast, boundaries detected by a 100 kb insulation square, for example, only affect interactions within a few bins of the boundary rather than insulating larger genomic regions from one another and do not change in the DC mutant (**Figure 2.6 e**).

Code availability

Code for Hi-C read mapping and processing is based on the published ICE method [62]. The code to calculate insulation profiles is publicly available at (https://github.com/blajoie/crane-nature-2015).

Z-score calculation

We modelled the overall chromatin interaction decay with distance using a modified LOWESS method (alpha 5 0.5%, ignore zeros, IQR filter), as described previously [63]. LOWESS calculates the weighted-average and weighted-standard deviation for every genomic distance by leveraging all data genome-wide. We transformed interaction data into a Z-score by calculating :((observed signal – LOWESS-average)/LOWESS-stdev). Observed signals with a count of 0 were excluded from the Z-score transformation. By expressing inter-action data as Z-scores, we corrected for minor differences in the overall decay with genomic distance that can vary slightly between samples.

To calculate the difference between the wild-type and DC mutant Hi-C data, we calculated the difference between the combined replicate DC mutant Z-score data and the combined replicate wild-type Z-score data (DC mutant Z-score minus wild-type Z-score). (**Figure 2.1 c, f, Figure 2.1, Figure 2.3, Figure 2.4 and Figure 2.5**).

Compartment analysis and comparison to LEM-2 associated domains. The presence and locations of A/B-compartments can be quantified using principle component analysis, where the largest eigenvector typically represents the compartment profile [35], [47], [62]. Applying this approach to 50 kb binned interaction data, we determined the positions of such preferentially associating compartments along each C. elegans chromosome (**Figures 2.3 e, 2.4 e and 2.5 c, g, k, o**). Compartment positions quantified in this manner closely align with the large sub-chromosomal domains that are visible in the chromatin interaction maps.

LEM-2 binding data15 (log2 ratio of ChIP signal over input) were lifted from the ce4 genome assembly to the ce10 assembly, and data were averaged in 50 kb bins. These bins correspond exactly to the coordinates of the binned chromatin inter-action data. Binned LEM-2 binding data were then plotted along each chromosome, and compared to the compartment profiles (**Figures 2.3 e; 2.4 e and 2.5 c, g, k, o**).

3D plots

To test for elevated levels of interaction between certain classes of sites in the genome, we constructed 3D plots. For each plot, a list was first made of all 10 kb bins meeting desired criteria: containing any rex or predicted rex (Prex) site (**Figure 2.11 d**), containing a rex or Prex site in the top 25 by ChIP-Seq signal (**Figure 2.11 d and Figure 2.9 f**), or containing any dox site (**Figure 2.9 g**). Prex sites are defined as those with very strong ChIP-Seq signal that was greatly diminished in sdc-2 mutants. Unlike rex sites, which also have these properties, Prex sites have not been tested for autonomous DCC recruitment in vivo through an array assay4. Next, sub-matrices of wild-type or DC mutant interactions were prepared for all possible pairs of bins in this list, extending 50 kb away from the central bin in all directions. Pairs of bins that were separated by less than 100 kb were excluded so that no sub-matrices would overlap the whole-chromosome interaction matrix diagonal (interactions within the same bin). All pairwise sub-matrices were then averaged together and the values plotted in 3D. If sub-matrices stretched past the end of the chromosome or overlapped bins with no data (un-mappable sequence, etc.), only the part of the sub-matrix containing data was included in the average. Cumulative plot randomization. To assess the significance of the decrease in Z-scores observed for the set of rex–rex interactions, we selected 1,000 random sets of 785 interactions (**Figure 2.11 c and Figure 2.9 e**). These random interaction sets were thus the same size as the rex–rex interaction set. The P value represents the fraction of the 1,000 randomized interaction sets that changed more from wild-type to DC mutant than the rex–rex set (according to the KS test statistic). Circos plots. Plots were generated using the Circos package to highlight the strength of various sets of rex–rex interactions in wild-type and DC mutant at 50 kb resolution. A Z-score threshold of 2 was selected and interactions were colored and given line thickness proportional to their Z-score. Z-scores greater than 8 were determined to correspond to ‘singleton’ outlier interactions and were excluded.

TAD FISH

FISH probes covering 400–500 kb genomic regions were prepared using pooled fosmids (BioScience LifeSciences), as described previously 8.1mg DNA was labelled with Alexa-488, Alexa-594, Alexa-555 or Alexa-647 using FISH Tag DNA Kit (Invitrogen). The genomic locations of tested regions are listed as follows: Probe1, chromosome X, 9.05–9.45 Mb; Probe2, chromosome X, 9.5–9.9 Mb; Probe3, chromosome X, 9.95–10.35 Mb; Probe4, chromosome X, 2.0–2.5 Mb; Probe5, chromosome X, 2.5–3.0 Mb; Probe6, chromosome X, 3.0–3.5 Mb; Probe7, chromosome X, 11.2–11.7 Mb; Probe8, chromosome X, 11.7–12.3 Mb; Probe9, chromosome X, 12.3–12.8 Mb; Probe10, chromosome X, 10.6–11.1 Mb; Probe11, chromosome X, 11.1–11.6 Mb; Probe12, chromosome I, 4.1–4.6 Mb;

Probe13, chromosome I, 4.6–5.1 Mb; Probe14, chromosome I, 5.1–5.6 Mb; and Probe15, chromosome X, 3.5–4.1 Mb FISH procedure. C. elegans embryos were obtained by dissecting gravid N2, him-8(e1489) or szT1/sdc-2(y74) unc-3(e151) adults in 13 ml of water on poly-lysine coated slides. A coverslip was added on top of the dissected worms, and the slides were then frozen in liquid nitrogen for at least 1 min. Coverslips were cracked off, and the samples were dehydrated in 95% ethanol for at least 10 min. 35 ml of fix (2% (v/v) paraformaldehyde in egg buffer (25 mM HEPES, pH 7.3, 118 mM NaCl, 48 mM KCl, 2 mM CaCl2, 2 mM MgCl2) was added and slides were incubated in a humid chamber for 5.5 min. Slides were washed 3 times for 10 min with 13 PBS-T (0.5% Triton X-100 in 13 PBS) at room temperature. Excess 13 PBS-T was then removed and 15 ml of hybridization solution (30% (v/v) formamide, 33 SSC, 10%dextran sulphate) containing approximately 50 ng of each FISH probe was added. Hybridization was performed in a temperature-controlled slide chamber (Bio-Rad ALD0211 Alpha Unit Block Assembly). The following FISH program was typically run overnight: 90 uC for 5 min, 0.5 uC per second to 50 uC, 50 uC for 1 min, 0.5 uC per second to 45 uC, 45 uC for 1 min, 0.5 uC per second to 40 uC, 40 uC for 1 min, 0.5 uC per second to 38 uC, 38 uC for 1 min, 0.5 uC per second to 37 uC, 37 uC overnight. Slides were then washed at 39 uC as follows: 3 times for 10 min with 30%(v/v) formamide in 23 SSC, 3 times for 10 min with 20% (v/v) formamide in 23 SSC, 3 times for 5 min with 10% (v/v) formamide in 23 SSC, 3 times for 5 min with 23 SSC, and 3 times for 1 min with 13 SSC. Slides were then washed 3 times for 10 min in 13 PBS-T. For N2 embryos, the slides were mounted in Prolong Gold antifade reagent (Invitrogen, P36934) containing DAPI (1 ng ml21). For him-8(e1489) and sdc-2(y74) embryos, immunostaining with SDC-3 antibody was performed following FISH to determine the sex and/or genotype of embryos as described below.

Immunofluorescence

Excess 13 PBS-T was removed and 35 ml of primary antibody (rat anti-SDC-3 antibody, 1:400) were added. Samples were incubated in a humid chamber for 6 h to overnight. Slides were washed 3 times for 10 min with 13 PBS-T at room temperature and then incubated in secondary antibody (Alexa-Fluor-647 goat anti-rat antibody (Invitrogen), 1:250) for 6 h to overnight. Slides were then washed 3 times for 10 min with 13 PBS-T at room temperature and then mounted.

Microscopy and co-localization analysis

Embryos were imaged on a Leica TCS SP8 microscope using 633, 1.4 NA objective lenses. The scanning settings for SP8 were: 1,024 3 1,024 pixels frame size, 51.5 nm pixel size, 3.5 zoom factor, 400 Hz scanning speed and 83.9 nm step size for z sections. Image deconvolution was performed using Huygens Professional Software.

After deconvolution, the homozygous sdc-2(y74) unc-3(e151) XX embryos were determined based on the lack of SDC-3 staining on the X chromosomes and their sex was further confirmed by examining the number of X-chromosome FISH signals. For all genotypes, embryos between 200-cell and 400-cell stages which match the developmental stage of Hi-C samples were selected for further analysis.

The deconvolved image stacks of embryos were manually segmented based on DAPI staining using Priism software [64]. FISH signals in individual embryos were thresholded to make the total signals from each probe occupy equal volume. The center-of-mass coordinates for the FISH signals from the probe in the middle of the probe set were determined using a built-in find points function in Priism. Regions of equal volume were then created around the FISH signals to encompass the entire sets of FISH signals on the same chromosomes using a Python script. Pearson's correlation coefficients between pairs of FISH probes were then calculated: the more the two probes overlap, the higher the correlation coefficient. 3D quantitative FISH for measuring the interaction frequency between genomic loci.

FISH experimental design

To examine the DCC dependence of interactions between genomic loci, and to distinguish between inter-homologue (trans) and intra-chromosomal (cis) interactions, we performed the 3D FISH analysis in both XX and the XO embryos in which the DCC was bound or not bound to X chromosomes. For these experiments, we acquired confocal images of embryos hybridized with FISH probes to two genomic loci and also stained with lamin (LMN-1) antibody and DAPI to help segment the nuclei. Newly developed soft-ware was used to measure the 3D distance between FISH probes automatically.

To assay XO embryos having DCC binding on the X chromosome, we per-formed the experiments using xol-1(y9); him-8(e1489) animals. These animals carried a deletion of the master switch gene (xol-1) that inhibits DCC binding to X chromosomes of XO embryos. DCC association with the X chromosome kills XO animals by the L1 larval stage. To enrich for XO male embryos in our experiments, we used mutation in him-8 (high incidence of males), which elevated the frequency of male progeny in a hermaphrodite brood from 0.02% to 37%. The XX embryos deficient in DCC binding were obtained from szT1/sdc-2(y74) unc-3(e151) animals, as described above.

To measure the distance between FISH foci in z stacks of confocal images, we developed software (Mets and Meyer, unpublished) that identified foci automatically, assigned foci to appropriate nuclei, and quantified the distance between foci in 3D space, thereby permitting the unbiased quantification of probe-interaction frequency. The quantification involved several steps. Each FISH spot was center fitted, and its location was recorded in x, y and z. For all nuclei, distances between all combinations of red and green FISH spots were calculated using a distance quantification algorithm that employs LMN-1 and DAPI co-staining to segment the nuclei. In XX embryos, four FISH spots (two red and two green) were generally apparent for X-linked probes in each nucleus, corresponding to the hybridization of both probes to their target sites on both homologous chromosomes. To eliminate the bias in our calculations for interactions caused by the inclusion of distances between probes on different chromosomes, we used only the shortest of the four possible distances between red and green probes in each nucleus for X-linked loci in XX embryos and for autosomal loci in all embryos.

We segmented the distances into 300 nm bins and plotted the relative contribution of each bin to the total number of measured distances. The limit of resolution of the confocal microscope is, 200 nm in x and y, making 300 nm a reasonable choice for the smallest bin. Furthermore, probes spaced ,260 nm apart appear overlapping by visual inspection, and probes spaced ,700 nm apart appear adjacent, indicating that the smallest bin size (300 nm) represents a degree of overlap that would be consider co-localized. Chi-square tests comparing the number of FISH pairs within 0–300 nm to those within 301–2,700 nm were used to assess the similarity of data sets from different classes of embryos. The unbinned data were also represented in cumulative plots (**Figure 2.13 a–f**). Preparation of FISH probes. Primers were created to amplify 3–6 kb sequences of DNA corresponding to each site. 1 mg of the probe DNA was labelled using the FISH tag DNA Red Kit (Molecular Probes, F32949) or the FISH tag DNA Green Kit (Molecular Probes, F32947) according to the manufacturer's protocol, with the following exceptions: the DNaseI was diluted 1:1,000, and the labelled probes were eluted in 10 ml but then diluted 1:10 for use in staining. Primers to make the probes are listed below: rex-23 F (gcccattcaacccattgtcc); rex-23 R (gcactcgcatattccaaaacg); rex-32 (cgcagctggccgttaaatg); rex-32 R (cattgcaggtgcgttcacaac); rex-47 F (ccgaaa cacaacaacaatgc); rex-47 R (agactggcgaagaggaacaa); rex-8 F (tgtgatgcaagccagagttgg); rex-8 R (cattgagccgaatttccaaagg); rex-14 F (ttgcagttgcgaaagaaatg); rex-14 R (tttttgaggagatcgggatg); rex-1 F (ctcaagagctgcgaagtgc); rex-1 R (aaagttcaacgaccagaatgc); Xnb1 F (tcgaatgacctcaagcactg); Xnb1 R (tcaccactgaaatcggcata); Xnb2 F (aaaacgcggtgaaacgatac); Xnb2 R (gttttcctctccccaacaca); Xnb3 F (gtatgcacacgcctcaaaaa); Xnb3 R (ttggaatctctcaccggagt); Xnb4 F (atggtaggacgttccgtttg); Xnb4 R (aatccagccctctggttttc); Xnb5 F (atttgcttgggcattaaacg); Xnb5 R (ttcaatgaagagacgcgatg); Xnb6 F (ccgtttttggcaatgaactt); Xnb6 R agaggatggtttggacgttg); Xnb7 F (gagcgacgattctgtcttcc); Xnb7 R (cgtcatgtccattttgcttg); Xnb8 F (atcgtgccaagacctattcg); Xnb8 R (ttttcgcatttcctgcttct); Inb1 F (aaaggaccctccccctaact); Inb1 R (tccatgcctacttgcctacc); Inb2 F (caggcgagcattctaccact); Inb2 R (ccggaaagagcattgattgt); Inb3 F (gcactgcaattgccaaccag); Inb3 R (ttcaaagacactcctcccatcc); Inb4 F (attgccgctaacccaagtgc); and Inb4 R (tccaacgccaacaaaactcc).

Combined FISH and immunofluorescence procedure

FISH followed by immunofluorescence was performed as described in the previous section. 5–10 ng (0.5–1 ml of 1:10 dilution) of each FISH probe was used for hybridization. For immuno-fluorescence, primary antibodies were applied at the following dilutions in 13 PBS-T: rat anti-SDC-3, 1:400; rabbit LMN-1, 1:400. Secondary Alexa-Fluor-555 donkey anti-rabbit and Alexa-Fluor-647 donkey anti-rat antibodies (Invitrogen) were used at a 1:200 dilution.

Microscopy and image analysis

Embryos were imaged on a Leica TCS SP2 AOBS confocal microscope or a Leica TCS SP8 microscope using 633, 1.4 NA objective lenses. The scanning settings for SP2 were: 1,024 3 1,024 pixels frame size, 46.5 nm pixel size, 5.0 zoom factor, 400 Hz scanning speed and 81 nm step size for z sections. The scanning settings for SP8 were as described in the previous section. The images were then deconvolved using Huygens Professional with the appropriate settings. The images were visualized and processed in Priism. The embryos were first cut out from the background using the edit polygon and cut mask function. Then the DAPI and LMN-1 channels were blurred using the 3D Filter Function to make the nuclear signal continuous and thus allow for the nuclei to be accurately segmented. This protocol permits each nucleus to be counted as one spot by the find points function. A new processed image was made by discarding the z sections in the top and bottom 10% of the image, and by substituting the new blurred channels for those in the original image. The find points function was then used to count and record the local center of mass (LCOM) of each nucleus and each FISH spot in x, y and z using user-defined threshold values. The data for the location of the nuclei and the FISH, along with the processed image are processed using the software described in FISH experimental design section above.

rex-47 deletion

Expression vectors for both codon-optimized Cas9 and sgRNA (Peft-3::cas9-SV40_NLS::tbb-2 39 UTR and PU6::unc-119_sgRNA [65] were obtained from Addgene. To enhance the expression and assembly of sgRNA, the sgRNA vector was modified by introducing an A-U flip in the sgRNA stem loop and extending the Cas9 binding hairpin [66]. To clone the protospacer sequence for the sgRNA targeting rex-47 (59-GTAGTCACACCGAATTGATA-39), the modified sgRNA vector was PCR amplified using primers GTAGTCACACCGAAT TGATAGTTTAAGAGCTATGCTGGAAACAGCATAG and AACAGCTATG ACCATGATTACGCCAAGCTTCACAGCCGACTATGTTTGGCGTCGAG or GACGTTGTAAAACGACGGCCAGTGAATTCCTCCAAGAACTCGTACAAA AATGCTCTGAAG and TATCAATTCGGTGTGACTACAAACATTTAGATT TGCAATTCAATTATATAG to generate two fragments with overlapping protospacer sequences. The two PCR products were then inserted into the sgRNA vector backbone generated by EcoRI/HindIII digestion using a previously described Gibson Assembly protocol [67]. To clone the repair template for making the 419 bp rex-47 deletion, two 500 bp homology arms flanking the target region were PCR amplified from C. elegans genomic DNA using primers ACGACG TTGTAAAACGACGGCCAGTGAATTCGACGTGTCGAAATTTTCAG and TTGAATTATTGACCATGGCAGACAGAGCGTAACGAGTAAT or ACGC TCTGTCTGCCATGGTCAATAATTCAATGCAATGAAG and CTATGACC ATGATTACGCCAAGCTTAATAATAAACTTCCATAAGA. The homology arms and the sgRNA vector backbone were assembled using Gibson Assembly. The resulting repair template contains an NcoI restriction site between the homology arms, which facilitates the identification of desired mutations.

Cas9-mediated mutagenesis and mutant screening

To generate Cas9-mediated heritable rex-47 deletion, DNA microinjection was performed according to standard protocols. The Cas9 expression vector, sgRNA expression vector, repair template and two co-injection markers: pCFJ90 (Pmyo-2::mCherry) and pCFJ104 (Pmyo-3::mCherry) were mixed and injected into the germline of 34 N2 young adults at the following concentrations: Cas9 (50 ng ml21), sgRNA (200 ng ml21), repair template (50 ng ml21), pCFJ90 (2.5 ng ml21) and pCFJ104 (5 ng ml21). Three days post-injection, 269 F1 s expressing both Pmyo-2::mCherry and Pmyo-3::mCherry markers were cloned into liquid culture in 96-well plates and propagated at 20 uC as described previously [68]. Worms from each well were lysed and PCR amplified using primers CCGAAACACAACAACAATGC and TGGTA GCCGTATGCACAGTT. We identified 8 deletion mutants from the 269 F1s (3%) based on the size of PCR products. These deletions were further verified by NcoI digestion of the PCR fragments. The progeny of the F1s carrying the rex-47 deletion alleles were then cloned into a new set of wells for the identification of homozygote mutants. PCR products from the homozygote mutants were sequenced to verify the precision of the deletions.

ChIP–qPCR. Wild-type and rex-47 deletion embryos were obtained as described earlier. Input and ChIP samples using rabbit anti-DPY-27 or rabbit anti-SDC-3 antibody were prepared according to previously published protocols20. Three pairs of qPCR primers (ACTTTGCAAGAGTATGTAGTGAA/ACGAGTAATACTT TGAGCATACTT, TACGGCTACCAATCTTGTAA/TCTGTATCTCTAATCC CTAATAGT and TGTGACTACTTGCCCAATAAA/TATCTCTCCCTTCGCC TAAA) were used to amplify three, 100 bp regions located upstream, down-stream or within the rex-47 deletion region, respectively. qPCR was performed using iQ SYBR Green Supermix (Bio-Rad, 170-8880) on a CFX384 Touch Real-Time PCR Detection System (Bio-Rad).

FISH analysis of rex-47 deletion strain

The legend for Figure 2.11 h provides the quantification for three-way comparisons of FISH probe co-localization among wild-type, DC mutant, and rex-47 deletion strains. For two-way comparisons using the one-tailed Mann–

Whitney U-test, the rex-47 deletion strain differed significantly from the wild-type strain (P, 1025) for probes on each side of the TAD boundary, and the rex-47 deletion strain was not statistically different from the DC mutant strain (P 5 NS), as expected.

RNA-Seq library creation

Embryos of appropriate genotype, four total wild-type biological replicates (two from the Hi-C biological replicates) and three total sdc-2 (y93, RNAi) biological replicates (two from the Hi-C biological replicates), were isolated following the procedures above and frozen at 280 uCin13 M9 buffer. RNA was extracted using a protocol described previously [69], except that 10 mlofa 20 mg ml glycogen solution was used as a carrier. Libraries were prepared from 10 mg of total RNA. PolyA RNA was purified using the Dynabeads mRNA purification kit (Ambion) and fragmented using Fragmentation Reagent (Ambion). First strand cDNA was synthesized from polyA RNA using the SuperScript III Reverse Transcriptase Kit with random primers (Life Technologies). Second strand cDNA synthesis was performed using Second Strand Synthesis buffer, DNA Pol I, and RNase H (Life Technologies). cDNA libraries were prepared for sequencing using the mRNA TruSeq protocol (Illumina).

Gene expression analysis

Libraries were sequenced with Illumina's HiSeq2000 platform. Reads were required to have passed the CASAVA 1.8 quality filtering to be considered further. To remove and trim reads containing the sequencing barcodes, we used cutadapt version 0.9.5 (http://code.google.com/p/cutadapt/). Reads were aligned to the transcriptome using GSNAP [70] version 2012-01-11. Uniquely mapping reads were assigned to genes using HTSeq version 0.5.4p3 using the union mode. Gene expression levels and changes in gene expression were determined by analysis with DESeq [71]. Gene expression analysis were conducted both with these RNA-Seq data sets and published GRO-Seq data sets [53]20. For each chromosome, scatter plots analysed the log2 of the median fold-change in gene expression (DC-mutant expression/wild-type expression) calculated for each 10 kb bin along the chromosome versus the change in insulation score for that bin in wild-type versus DC mutant embryos. No significant correlation was found between the change in gene expression and the change in insulation score: for chromosomes I, II and X, R 5 0.04; for chromosome III and IV, R 5 0.00; for chromosome V, R 5 0.03.

CHAPTER III: Structural organization of the inactive X chromosome

Preface

This research chapter encompassed work performed by Luca Giorgetti, Bryan R. Lajoie, Ava C. Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, Noam Kaplan, Howard Y. Chang, Edith Heard, and Job Dekker. The manuscript is currently being revised at Nature (as of 02/2016).

Abstract

X-chromosome inactivation (XCI) entails a massive structural reorganization of the inactive X (Xi). However the molecular architecture of the Xi is unknown. Here we show that the Xi lacks typical autosomal features such as active/inactive compartments and topologically associating domains (TADs), except around a small number of genes that escape XCI and remain expressed. Escaping genes form TADs and retain DNA accessibility at promoter-proximal and CTCF binding sites, indicating that these loci can avoid Xist-mediated erasure of chromosomal structure. We further show that gene-silencing competent Xist RNA is sufficient to induce segregation of the Xi into two 'mega-domains' separated by a boundary that includes the DXZ4 macrosatellite. Deletion of this boundary prior to XCI results in fusion of the mega-domains and altered patterns of escape that correlate with changes in TAD structure following differentiation and XCI . Our results suggest a critical role for

the boundary locus and Xist RNA in shaping the structure of the Xi and modulating escape from XCI. Our findings also point to roles of transcription and CTCF binding in TAD formation in the context of facultative heterochromatin.

Introduction

Important new insights into the 3D organization of mammalian chromosomes have come from recent chromosome conformation capture approaches. These studies have revealed a hierarchy of structural organization spanning several genomic length scales, from multi-megabase ‘A/B’ compartments defined by blocks of chromatin that correlate with chromatin activity states, to topologically associating domains (TADs) which represent evolutionarily conserved sub-megabase self-interacting domains to multi-kilobase looping associations between regulatory and structural elements [7], [12]. This recent understanding of chromosome folding has provided important insights into the nature of long-range gene regulation and the mechanisms underlying gene expression dynamics.

However, less is known about the structure and organization of heterochromatin. To what extent does chromosome folding, TAD organization and long range looping differ in the context of a heterochromatic state? A classic example of facultative heterochromatin is the inactive X chromosome (Xi) in female mammals, which is condensed and organized into a distinct silent nuclear compartment. During early female development, X-chromosome inactivation

(XCI) is triggered by up-regulation of the long non-coding Xist RNA from one of the two X chromosomes. Xist RNA coats the chromosome in cis and, via its A-repeat region [37], [38], induces transcriptional silencing of almost all of the ~1073 genes on the X. Interestingly, some genes (constitutive escapees) avoid this silencing in most cell types while others (facultative escapees) become reactivated from the Xi only in specific contexts [39]. The underlying mechanism(s) for both facultative and constitutive escape are not known. A role for Xist RNA in reshaping the organization of the entire Xi has been proposed [40], [41], with escape genes being excluded from the Xist-coated domain. However, the exact architecture of the Xi, for both its silent and expressed regions, is still unclear. Based on DNA FISH, the human Xi is a rather homogeneous structure with an overall compaction that is about 1.2-fold higher than that of the active X chromosome (Xa) [42]–[44]. Recent chromosome conformation capture approaches have pointed to some intriguing features of the 3D folding of the Xi, including formation of large mega-domains along the human Xi [45], and long-range associations between loci that escape inactivation and become expressed on the mouse Xi [41]. However detailed insights into the global molecular architecture of the Xi remain far from complete, due in part to the lack of chromosome-wide, high resolution, allele specific information. To this end, we have investigated the structure, chromatin accessibility and expression status of the Xi using allele-specific Hi-C, ATAC-Seq and RNA-Seq methods in embryonic stem cells (ESCs) and clonal neural progenitor cells (NPCs) both derived from a highly polymorphic (Cast x 129) F1 mouse. This F1 mouse cross contains 19,722,473 SNPs, averaging 1 snp every ~140 bases which enables higher resolution analysis of allele-specific chromatin states and three-dimensional conformation than that previously performed in human cells (~10-fold higher SNP density) [45].

Results

Organization of the inactive X

In ESCs, prior to XCI, allele-specific Hi-C analysis revealed that autosomes as well as both active X chromosomes display prominent active/inactive (A/B) compartmentalization, quantified by eigenvector decomposition [1], [62], and TAD structure, quantified by insulation analysis [46] (**Figure 3.1 a-b**, **Figure 3.2**). In NPCs that were clonally derived from the same ESCs, compartments and TADs were similarly detected on autosomes and the active X chromosome. However striking differences were observed for the inactive X. First, TADs are largely absent on the Xi, as readily observed by visual inspection of the Hi-C interaction maps (**Figure 3.1 a**). To quantify the presence of TADs we calculated an insulation score (the number of interactions occurring across each bin) along the entire length of the chromosome [46]. TAD boundaries display low insulation scores (few interactions occurring across bin/boundary), while loci located within TADs display high insulation scores (many interactions occurring within TADs, across bin). The variance in insulation scores along the chromosome is a quantitative measure of the presence of TADs [72], with large variance indicating a strong TAD signature. In NPCs we detected an overall 2-3 fold decrease in the interquartile range (IQR) of the insulation score along the entire Xi as compared to the Xa (**Figure 3.1 a and Figure 3.3**), as well as marked differences in the pattern of local fluctuations in insulation scores, contrary to what is observed when comparing the two Xa's in ESCs or autosomes in ESCs and NPCs (**Figure 3.4)**. Secondly, the Xi does not display A/B compartments typically observed across the rest of the genome (autosomes and the Xa) (**Figure 3.1 b**). Rather, the Xi is partitioned into two massive domains of preferential interactions, spanning approximately 73 and 93 megabases, and separated by a region of approximately ~120-200kb that includes the DXZ4 macrosatellite locus [73], [74]. Thus, the mouse Xi has an unusual conformation characterized by the presence of two mega-domains, consistent with a previous study on the human Xi [45], and extensive loss of TAD structures and compartments as also shown by previous 5C work [3] as well as a recent Hi-C study [75]. Importantly, these previous studies did not investigate the precise nature of this unusual architecture nor its implications for X-chromosome inactivation and escape. Whether and how any of these Xi-specific structural properties are involved in regulating gene expression remains unknown.

To investigate the degree to which the Xi-specific mega-domains uncovered by Hi-C correspond to two spatially segregated chromosomal domains in single cells, we designed a DNA FISH assay using three probe sets (a, b, and c; labeled with Atto448/green or Atto550/red dyes), each spanning consecutive 18-Mb regions (**Figure 3.1 c**). We then performed FISH to assess the overlap between probe sets located in the same mega-domain (a-b) or located on either side of the mega-domain boundary (b-c). *Xist* RNA FISH was simultaneously performed to distinguish the inactive from the active X. Visual inspection and quantification of fluorescent images (**Figure 3.5 a**) revealed a significantly higher overlap between regions within the same mega-domain (probes a-b) on the Xi but not on the Xa (**Figure 3.1 c-d and Figure 3.5 b**), in agreement with the Hi-C data showing increased contact frequencies within mega-domains specifically (**Figure 3.1 c**). In contrast, regions on either side of the mega-domain boundary (probe sets b-c) showed a significantly lower overlap on the Xi, and no difference with the Xa (**Figure 3.1 c**). Similar results were obtained in astrocytes derived from the NPCs, and also using an independent NPC clone in which the Cast rather than the 129 X chromosome was inactivated, showing that either the 129 or cast X chromosome can display the presence of the boundary/mega-domains when inactivated (**Figure 3.5 c**). These data demonstrate that the Xi is spatially segregated into two mega-domains at the single-cell level and confirm the presence of a physically insulating boundary region on the Xi that is present both in stem cells (NPCs) and differentiated cells (astrocytes).

The above analyses also indicated significant chromatin compaction within each of the two mega-domains on the Xi compared to the Xa. To investigate this further, we quantified the volumes of DNA FISH signals from single probe sets a, b and c by two independent means. First, by using threshold-independent gyration tensor analysis (**Figure 3.5 d**) to estimate the gyration radius of FISH signals, we detected mildly but statistically significant larger gyration radii on the Xa than on the Xi (9% gyration volume difference on average, **Figure 3.5 e**). Second, we performed segmentation-based volume estimation analysis at multiple grayscale threshold intensities to measure the maximal 3D extension of FISH signals at each threshold (**Figure 3.5 f**). We found 25% larger maximal volumes on the Xa than on the Xi, with Xa signals being larger than those on Xi in the majority of cells (**Figure 3.5 f**) for a wide range of thresholds. Hence, contiguous regions within each mega-domain on the Xi appear to be slightly, but significantly more compact than homologous regions on the Xa, consistent with previous observations on the human Xi [42]. We also noted that individual 18-Mb regions within each mega-domain appeared more spherical than corresponding Xa regions (single probes a, b and c**, Figure 3.1 d and Figure 3.5 e**). However when considering composite signals from probes extending across the mega-domain boundary on the Xi, signals from the two mega-domains were clearly spatially segregated in the majority of cells (**Figure 3.1 c**, probes b-c), generating overall FISH signals that were more elongated than those from equally sized regions within the same mega-domain (**Figure 3.1 c**, probe sets a-b). The mouse Xi is thus very differently organized when compared to its active homolog, being partitioned into two large, spatially distinct regions of increased chromatin compaction.

Clustered genes that escape silencing are embedded in TADs

As a striking exception to the general absence of sub-megabase structure on the Xi within each mega-domain, Hi-C analysis revealed the presence of a small number of residual chromosome domains (see **Figure 3.1 a,** arrow in the bottom panel) that resemble TADs and stand out as regions of increased self-interactions. In order to investigate whether these TAD-like structures correspond to hotspots of biological activity on the otherwise inert Xi, we performed ATAC-Seq to identify all accessible, active elements on the Xi, and compared the data to Hi-C and previously published allele-specific RNA-Seq data produced on the same ESC and NPC clonal lines [76].

Upon differentiation (XCI) of ESCs into NPCs, RNA-Seq and ATAC-Seq profiles reflected a global loss of activity on the inactive X (**Figure 3.6 a**). Of the genes covered by ≥ 1 SNP, we could detect 87 expressed genes on the Xi, as opposed to 314 on the Xa by RNA-Seq (expressed defined as ≥ 3 RPKM). Concordantly we found 224 ATAC-Seq peaks on the Xi compared to 825 on the Xa (covered by ≥ 1 SNP), indicating a massive loss of active genomic elements on the Xi. As expected, the majority of the ATAC-Seq peaks on the Xi in NPCs fall near the X inactivation center (Xic) (from which Xist is specifically expressed only on the Xi), the pseudoautosomal region (PAR, expressed from both Xi and Xa), and at the promoters of genes that escape X inactivation, as identified by RNA-Seq (constitutive and facultative escapees, expressed from both Xi and Xa) **(Figure 3.6 a**). Strikingly, we noted that accessible sites on the Xi (based on ATAC-Seq) often lie within the Xi-specific TADs observed by Hi-C. Furthermore, the degree of local structure on the Xi was correlated with the number of transcribed loci in a particular chromosomal region, as shown in **Figure 3.6 b** by three examples: a dense cluster of 19 facultative escapees that include the *Mecp2* gene which shows a clear ~800kb TAD; part of the Xic region including *Xist*, with slightly increased interactions extending over a ~250 kb region 5' to the Xist promoter, in contrast to the homologous region on the Xa where *Xist*'s promoter lies in a well-defined ~500kb TAD; and a region of 5 escapees including the constitutive escapee *Kdm5c* (also known as *Jarid1c),* which lies adjacent to the facultative escapee *Huwe1*, which shows a prominent ~500 kb TAD.

We next analyzed the correlation between TAD structure, allele-specific expression [76] and allele-specific chromatin accessibility for each gene along the Xi, and found that genes located in regions with prominent TAD structure (high insulation scores) are correlated with elevated levels of chromatin accessibility and gene expression on the Xi (**Figure 3.6 c**). Moreover, genes that escape XCI (expressed on Xi) are located in regions with significantly higher insulation scores (indicative of prominent TAD structure) (**Figure 3.6 d**) as compared to silenced genes. We did not detect a significant difference in insulation scores for the corresponding sets of loci on the Xa (**Figure 3.6 e**). Similar results were obtained for ATAC-Seq signal at expressed versus silenced genes on the Xi (**Figure 3.7 a**). Importantly, the TAD-like structures detected on the Xi at expressed regions do not necessarily correspond to equivalent TADs on the Xa. This indicates that TAD structure is strictly related to gene expression status on the Xi, unlike TADs on the Xa which are present irrespective of expression status.

Interestingly, whereas on the Xa, only ~35% of ATAC-Seq peaks are promoter-proximal (< 5kb from promoters), on the Xi, more than half (51%) of accessible sites are promoter-proximal (p = 1.38 e-5; **Figure 3.6 f**). This indicates that escape from XCI is more often regulated at promoters or very proximal transcription factor binding sites. Almost all of the promoter-proximal and the largest class of promoter-distal (>5kb) ATAC-Seq peaks on Xi were found at CTCF binding sites (**Figure 3.7 b**), indicating that CTCF may play a role in escape from XCI. This may be related to the recent finding that loss of cohesin along the Xi, often co-located with CTCF, leads to loss of TADs [75], and indicates that escape from Xist-driven chromosome structure erasure may involve CTCF binding to facilitate TAD formation [56], [57], [77] and escape from silencing.

We also found that escapee loci on the Xi tend to interact with each other over long distances in cis, and even across the mega-domain boundary, consistent with a previous report based on 4C [41] (**Figure 3.6 g**). This implies that transcribed TAD-like regions on an otherwise heterochromatic X tend to associate together and could be related to the general phenomenon that active regions cluster together in (active) A-like compartments. In conclusion, our investigation of the molecular architecture, chromatin accessibility and transcriptional status across the heterochromatic Xi reveal a surprisingly complex organization with genes expressed from the Xi being embedded in TAD-like structures, tending to display chromatin accessibility that is often at or near their promoters and being engaged in long-range associations with one another.

Structural and transcriptional role of the mega-domain boundary locus on the Xi

We next explored the role of the DXZ4-containing boundary in the formation of the Xi mega-domains, and its implications for Xi structure and expression. For this we used a CRISPR/Cas9-based strategy to generate a 200-kb deletion of the boundary region (ΔFT), encompassing the DXZ4 macrosatellite repeat and unique flanking DNA from only the 129 allele in ESCs (**Figure 3.8 a, Figure 3.2 b and Figure 3.9 a**). When this ΔFT ESC line (D9) was differentiated into NPCs, Xist RNA coating of one of the two X chromosomes was induced as usual and we were able to derive multiple clonal NPCs in which the 129 (ΔFT) X chromosome was the inactive X. Hi-C was performed on one of these clones (D9B2) and visual inspection of the data revealed a massive reorganization of the ΔFT NPC Xi, compared to the WT NPC Xi, with the two mega-domains fusing into one single domain encompassing the entire chromosome (**Figure 3.8 b**). As expected, the Cast Xa chromosome in ΔFT NPCs showed no observable differences when compared to the Cast Xa in WT NPCs (**Figure 3.10 7a)**. To further validate our results, we again employed a FISH strategy (see above) to assess boundary formation and chromatin compaction. Loci on either side of the deleted boundary (probes b and c) overlap significantly more on the ΔFT Xi when compared to the normal Xi (**Figure 3.8 c-d**), consistent with the Hi-C data (**Figure 3.8 d**) and confirming that the two mega-domains had fused into a single chromosome-wide entity. We also assessed chromatin compaction and found that similar to the WT Xi, the ΔFT Xi showed increased chromatin compaction when compared to the Xa (significantly higher overlap of probes a and b, **Figure 3.8 c-d**). Thus the deletion of the mega-domain boundary results in increased intermingling between the two mega-domains of the WT Xi, but does not appear to change the overall chromatin compaction of the inactive X chromosome.

To assess the conformation of the ΔFT Xi in more detail and to determine whether the massive structural reorganization of the ΔFT Xi is accompanied by any functional changes, we also performed ATAC-Seq and RNA-Seq in the same mutant NPC clone (D9B2) for which Hi-C was performed. We could detect 29 expressed genes on the ΔFT Xi, as opposed to 313 expressed genes on the ΔFT Xa by RNA-Seq (expressed defined as ≥ 3 RPKM), whereas the WT NPC had 87 and 314 expressed genes on the Xi and Xa respectively. Surprisingly, we found that transcription and chromatin accessibility were lost at many locations along the ΔFT Xi (**Figure 3.11 a**) and notably at NPC-specific facultative escape genes such as the *Mecp2*-containing gene cluster (**Figure 3.11 b**, left). On the other hand, transcription and open chromatin were maintained on the ΔFT mutant for genes that are constitutively expressed from the Xi such as *Xist* (**Figure 3.11 b**, center) and *Jarid1c* (**Figure 3.11 b**, right). All (6 out of 6) constitutive escapee genes (5) found to be expressed from the wild-type Xi including *Jarid1c*, were still expressed from the ΔFT Xi, whereas only 21 out of the 87 facultative escapees (~24%) still showed some expression (≥ 3 RPKM) from the ΔFT Xi (**Table 3.1**). RNA FISH confirmed loss of transcription at facultative escapees and continued expression from the constitutive escape genes (**Figure 3.11 c and Figure 3.9 b-c**). These results were confirmed in two additional independent NPC clones (D9C7 and D9A3) derived from the ΔFT mutant ESC line although in one clone some facultative escape of some genes could still be detected (**Figure 3.9 d**). Strikingly, in the D9B2 NPC clone analyzed by Hi-C, TAD-like structures were also lost where transcription and chromatin accessibility were lost, but maintained at constitutive escape genes where expression is maintained (**Figure 3.11 b**). In particular, in the region including *Jarid1c*, 4 of the 5 facultative escape genes such as *Huwe1* and *Smc1a* became silenced on the ΔFT Xi, and this was paralleled by loss of TAD-like structure (**Figure 3.6 b**) despite the fact that the constitutive escapee *Jarid1c* remains expressed. X chromosome-wide comparisons in the ΔFT D9B2 NPCs showed strong correlations between the loss of escapee expression, loss of chromatin accessibility and reduction in TAD signal (**Figure 3.11 d**). Furthermore, although the massive restructuring of the Xi in the ΔFT mutant leads to greater overall intermingling of the previously segregated mega-domains (**Figure 3.8 b**), which might be expected to increase long-range interactions between regions that escape XCI, we find that most of the specific long range interactions between the 87 WT Xi escapee genes are lost on the ΔFT Xi in D9B2 NPC cells (**Figure 3.11 e**). Further no spatial clustering between the 29 ΔFT Xi expressed genes (escapees) is detected on the ΔFT Xi. This loss coincides with loss of expression and loss of TAD structures, indicating that very long range interactions between escapees on the WT Xi are indeed closely linked to expression status.

As an intriguing exception to the widespread loss of facultative expression on the ΔFT Xi, we noticed that seven genes that were silenced on the wild-type Xi were now found to be expressed on the ΔFT Xi in the D9B2 clone (**Table 3.1**). These *de novo* escapees do not occur in clusters and do not appear to be highly accessible by ATAC-Seq (**Figure 3.10 b)**, and thus are not expected to lie within strong TAD-like regions. Inspection of Hi-C maps showed a mild increase in local structure at some of these loci when compared to the wild-type Xi, more so for the most highly expressed of them such as *Maged1* and *Eda2r* (**Figure 3.10 b**), leading to an amount of local structure that is comparable to that seen for *Xist* and *Jarid1c* on the Xi (**Figure 3.8 b and Figure 3.11 b**). This result reveals that in clones where alternative sets of isolated genes escape silencing, corresponding mild amounts of local structure emerge. In addition, we found that the amount of local chromosome structure tends to increase at and around genes whose transcription level and accessibility increases on the ΔFT Xi as compared to the WT Xi, which include *de novo* escapees, but also a subset of the escapees whose expression level is increased (**Figure 3.11 d**). Thus, it appears that there is no fixed set of loci that consistently maintains TAD organization and expression in different clones, and that the presence and level of transcription is correlated with increased local structure. Further, our data suggest that the mega-domain boundary modulates the number and identity of facultative escapees.

When chromatin accessibility changes were assessed in more detail on the Xi in the D9B2 clones, of the 224 ATAC-Seq peaks that could be assigned allelically to the Xi, 139 were found to be lost in the ΔFT mutant (**Figure 3.11 f**, left panel). The set of accessible sites lost in ΔFT NPCs are enriched for promoter-proximal sites with 64% falling within 5kb of a TSS (**Figure 3.11 g**). 93% of these promoter-proximal sites contain a CTCF binding site, an enrichment compared to distal sites (>5kb from a TSS), comprised of 64% CTCF binding sites plus p300 binding sites, H3K27Ac sites, and other TF binding sites (**Figure 3.11 f** right panel). These CTCF sites are located closer to the TSSs of escape genes than sites that do not change in the ΔFT NPCs (**Figure 3.11 h**). These results again point to a role for CTCF in regulating escape from XCI.

Formation of mega-domains depends on gene-silencing competent Xist RNA

To investigate whether Xist RNA itself can induce the bipartite folding of the Xi, we induced *Xist* expression from one X chromosome in female ESCs carrying a tetracycline-inducible promoter at the endogenous *Xist* locus (TX1072) [78] (**Figure 3.12 a**), which has previously been shown to lead to gene silencing on the Xist-coated X chromosome. RNA/DNA FISH experiments as in **Figure 3.6 and Figure 3.11** revealed that the *Xist*-coated chromosome becomes partitioned in two domains separated at the DXZ4-containing region upon Xist RNA coating (**Figure 3.12 b-c**), although the strength of the boundary appears to be somewhat reduced as compared to NPCs. Thus, coating by the Xist mRNA is sufficient to induce formation of two mega-domains on the X chromosome. To determine whether this is due to gene silencing or to an independent architectural role of Xist, we performed the same experiment in cells carrying a wild-type or mutant form of the Xist RNA deleted for its A-repeat region, which is no longer able to induce gene silencing, but is still competent for Xist RNA coating and exclusion of RNA PolII [40]. For this we used previously characterized male ESC lines carrying tetracycline-inducible wild-type *Xist* [79], or mutant *Xist* (J1:XistΔA) [37] (**Figure 3.12 d and Figure 3.9 g**). Whereas wild-type *Xist* induction led to boundary formation at the DXZ4-containing region, the A-repeat mutant did not, indicating that gene silencing is in fact required for the establishment of the two mega-domains (**Figure 3.12 e-f**). Because the A-repeat is required for the interaction of a small subset of Xist-binding proteins [37], these results also suggest that one or more of these factors could initiate Xi-specific changes in higher-order chromosome folding upon Xist coating, followed by further events during differentiation. It has recently been proposed that the Xist RNA might bind to and thus somehow repel cohesin [75], which mostly binds DNA associated with CTCF. However cohesin and CTCF were not identified in other studies that identified Xist RNA protein partners, in particular those associated with the A-repeat motif [38]. Thus the precise mechanisms by which A-repeat containing Xist RNA induces global restructuring of the chromosome it coats remain open questions but our results suggest that Xist's gene silencing function may be tightly linked to its structural role.

Conclusion

Our study reveals that the inactive X chromosome is a surprisingly elaborate entity, with a global partitioning into two mega-domains and loss of TAD organization, except at clusters of genes that are still expressed from the otherwise silent Xi. TADs were previously thought to be highly stable across cell generations and differentiation [2], [3], and their presence or maintenance not to require transcription in general. However our study demonstrates that 1) TADs can indeed be lost in some contexts (as also observed on mitotic chromosomes [72], although in the case of the Xi, TAD loss is not a transient state but is stably transmitted through cell division) and that 2) gene expression and/or binding of factors such as CTCF can enable their maintenance and/or *de novo* re-creation. Our findings show that gene silencing and loss of accessibility is accompanied by loss of structure, but that *de novo* gain of escape corresponds to re-creation of local structure, and further that transcription at clusters of genes coincides with TAD formation. Together these findings suggest that gene expression and DNA binding factors may be driving forces of TAD organization in the context of the inactive X, which is otherwise devoid of TADs. The Xi may therefore represent a sequence-independent chromosome state at the structural level, from which sequence specific TADs can arise.

The reduced level of facultative escape in cells where the mega-domain has been deleted is intriguing. Although escape can be quite variable even in normal cells, three NPC clones derived from the D9 ΔFT mutant ESC line showed reduced escape by RNA FISH (**Figure 3.9 d**). These results suggest that during XCI the mega-domain boundary and the bipartite folding of the Xi that it induces, may modulate or affect the process leading to facultative escape. Constitutive escapees are much less affected by the boundary deletion and presumably have an intrinsic capacity to override the XCI process [80]. Facultative escapees on the other hand are first silenced during XCI and then re-expressed ([81], [82] and unpublished data). Although the mega-domain boundary region does not appear to interact with escapee regions in NPCs and is transcriptionally silent in NPCs, this region is transcribed and possibly euchromatic at the onset of XCI (MA and EH, unpublished observations). Transient interaction of this region with facultative escape loci during differentiation may thus occur and may be sufficient to regulate the local amount of escape and/or re-establish TADs at escape loci due to its unusual chromatin status (**Figure 3.12 g**) and atypical enrichment in CTCF binding [83]. An additional, but not mutually exclusive model is that the boundary region helps position the Xi in a particular sub-nuclear location during or after XCI, that facilitates the establishment of a given escape pattern. These results establish the Xi as a powerful model system for studying the mechanistic interrelationships between chromosome conformation and gene regulation, and point to a key role for gene activity in the establishment of chromosome structure at the level of TADs in the context of facultative heterochromatin.

Figures

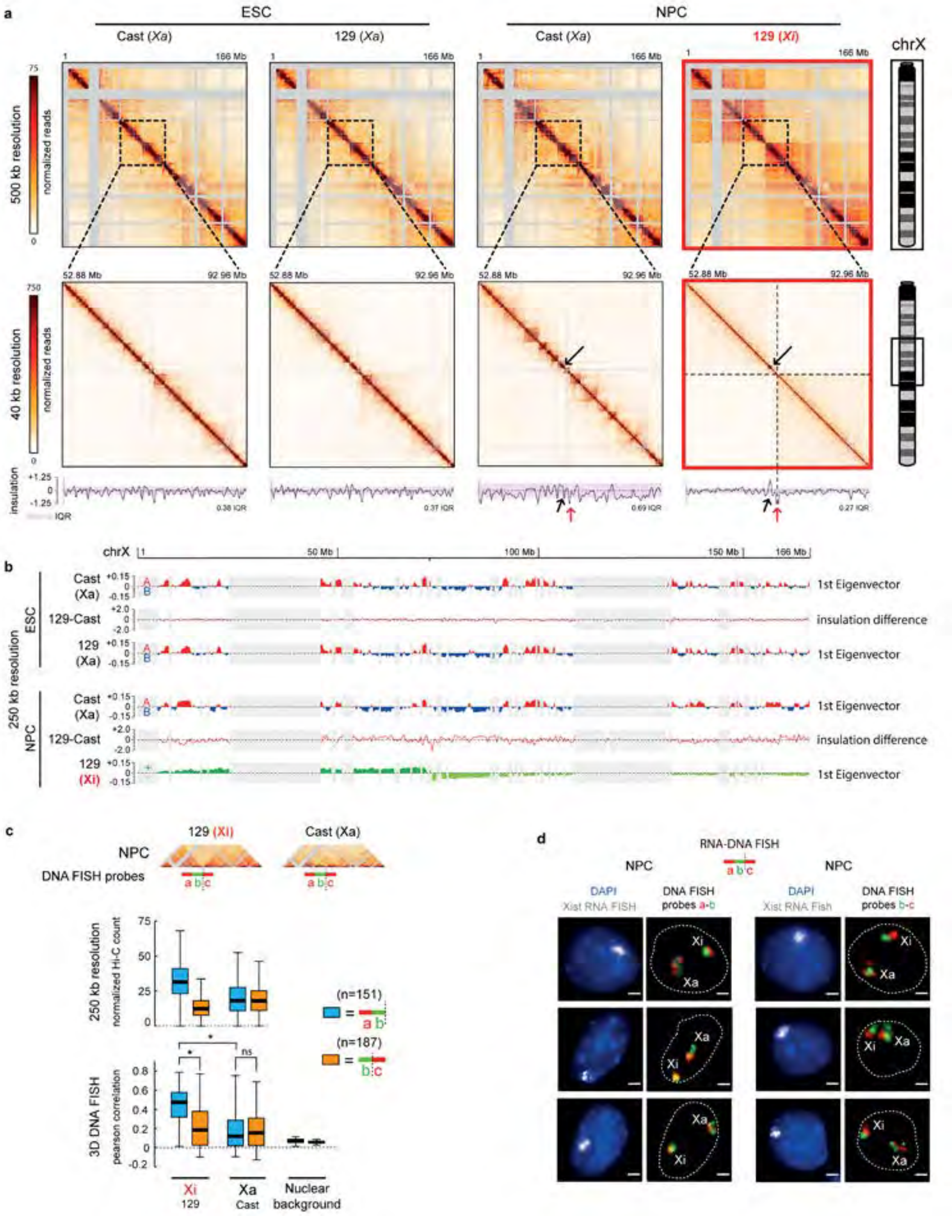


Figure 3.1 | The distinct conformation of the Xi, Xa and autosomes.
a, Allele-specific Hi-C contact maps for chromosome X in ESCs and NPCs at 500-kb resolution (top), and for a ~40-Mb region centered around the DXZ4-containing locus at 40-kb resolution (bottom). The insulation score is plotted at the bottom of each 40-Mb heatmap. Purple shaded areas indicate the IQR range

of insulation scores along the chromosome to illustrate the reduced insulation scores along Xi, indicative of a loss of TAD structure. Black arrow: position of the residual TAD in the 40-Mb region. Red arrow: position of DXZ4. b, Compartment profiles of chromosome X in ESCs and NPCs. The first eigenvector (PC1) of each allele-specific Hi-C contact map, obtained with Principal Component Analysis, is shown, together with the difference in chromosome-wide insulation score between the 129 and Cast allele. A/B-compartments are evident in ESCs and NPCs along both Xa (Red and Blue signal), whereas first eigenvector corresponds to the two mega-domains for the Xi in NPCs. In ESCs both Xa display comparable insulation profiles (difference is close to zero along the chromosome), whereas in NPCs large differences are observed (difference in insulation fluctuates along the chromosome). Grey areas indicate regions with low SNP density that were excluded from analysis. c, Top: Scheme of the DNA FISH probe sets (*a-b*: inside the same mega-domain, *b-c*: across the boundary). Bottom: Loci detected by probe set a-b are more interacting than b-c both in Hi-C (left) and in 3D-DNA FISH (right). * denotes $p<8e-17$ in a Wilcoxon's rank sum test corrected with Bonferroni for multiple hypothesis testing. d, DNA FISH signals from probe set a-b are more overlapping and spherical on the Xi than on the Xa, whereas signals from b-c show a clear partitioning on the Xi into two separate domains.

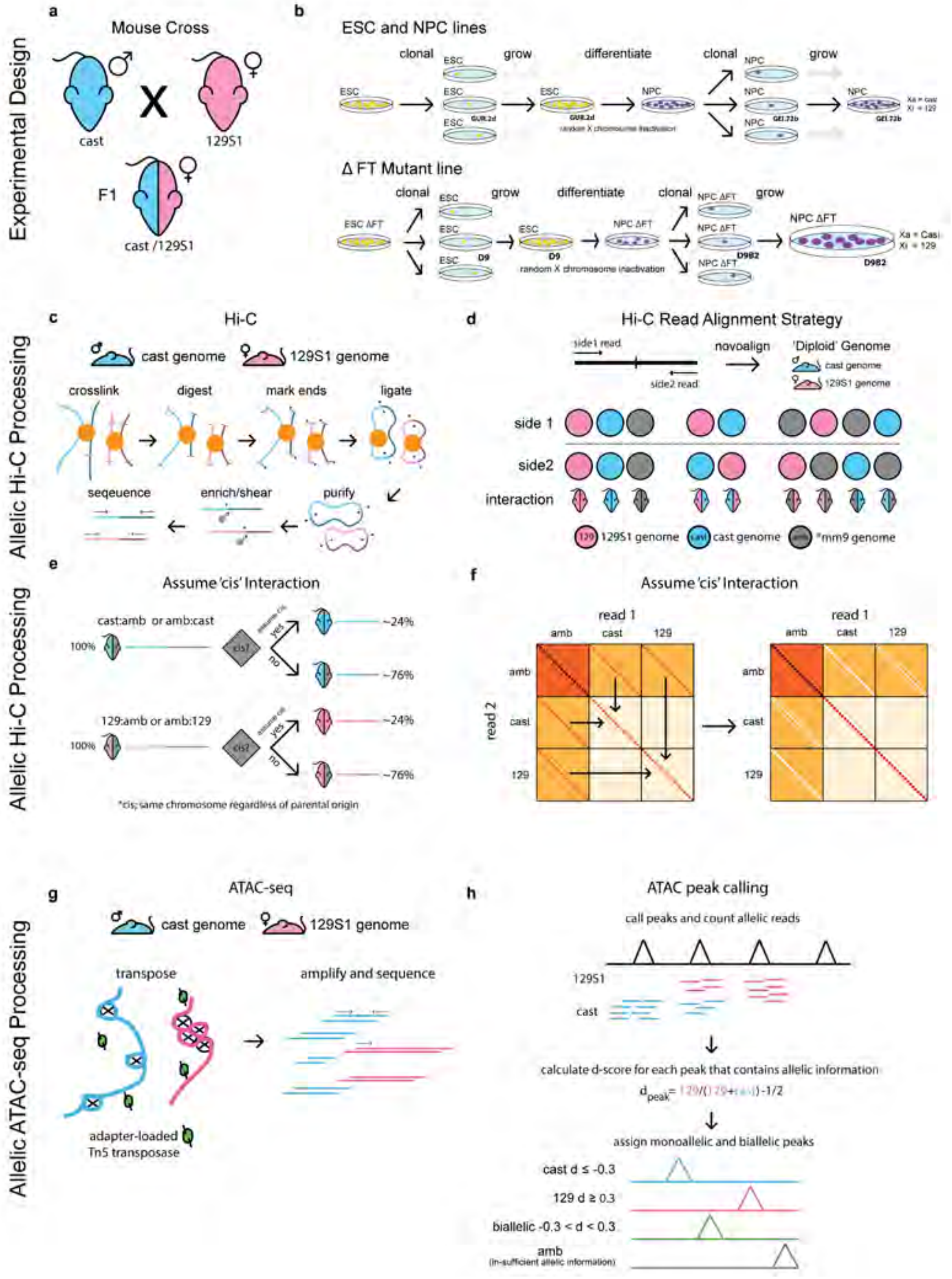


Figure 3.2 | Experimental Design
a, Schematic of hybrid mouse strains used for all experiments. b, Scheme outlining differentiation of ESCs to NPCs and picking of clones. Scheme outlining CRISPR deletion of the mega-domain boundary in ESC, differentiation to NPC and the picking of clones. c, Schematic of Hi-C library generation. d, Schematic of the Hi-C Alignment Strategy. PE reads are aligned to a 'diploid'

genome consisting of 22 chromosomes from Cast, and 22 chromosomes from 129 (1-19 X,Y,M). The interaction row shows all possible PE read combinations between the 129, Cast and Ambiguous genomes. e, Schematic showing the re-assignment of certain ‘cis’ interactions. PE reads where one side uniquely aligned to an allele and the other side aligned equally to both alleles (AMB), were re-classified as an allelic reads, only if both reads aligned to the same chromosome (cis). f, Cartoon explaining the re-assignment of 129:amb or cast:amb cis interactions. g, Scheme for ATAC-Seq library preparation. h, Scheme for allele-specific ATAC-Seq data analysis

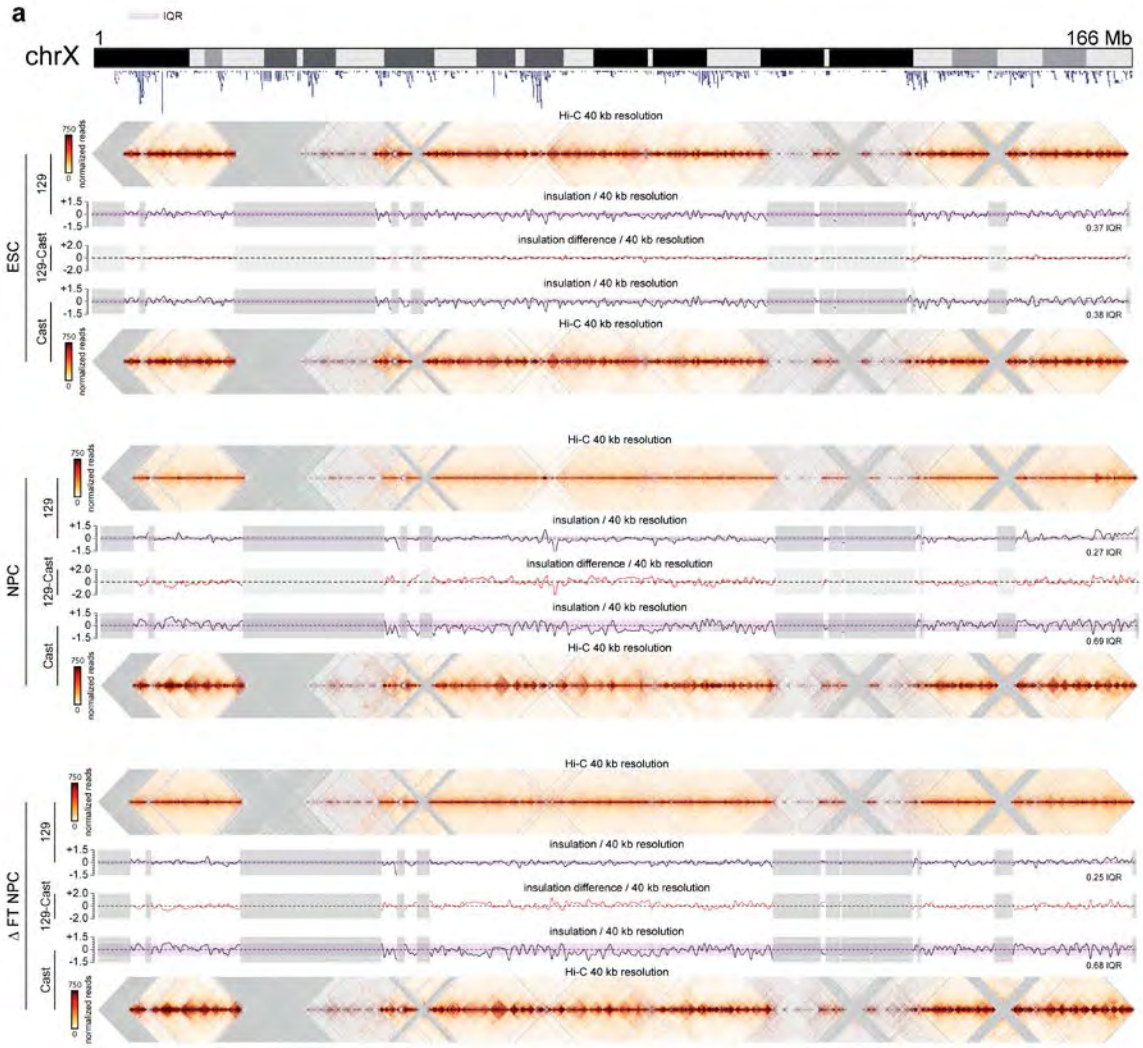


Figure 3.3 | chrX

a, Hi-C data, insulation scores, and the difference in insulation scores (129-cast) are shown for ESC (GUR.2d), NPC (GEI.72b) and mutant NPC (D9B2/B129T3) for both alleles (Cast and 129) for chrX. Large dips in the insulation vector are found at TAD boundaries. Peaks in the insulation vector are found towards the center of each TAD. The insulation difference plot highlights areas of differential TAD structure between the alleles (many differences as compared to the allelic differences along autosomes).

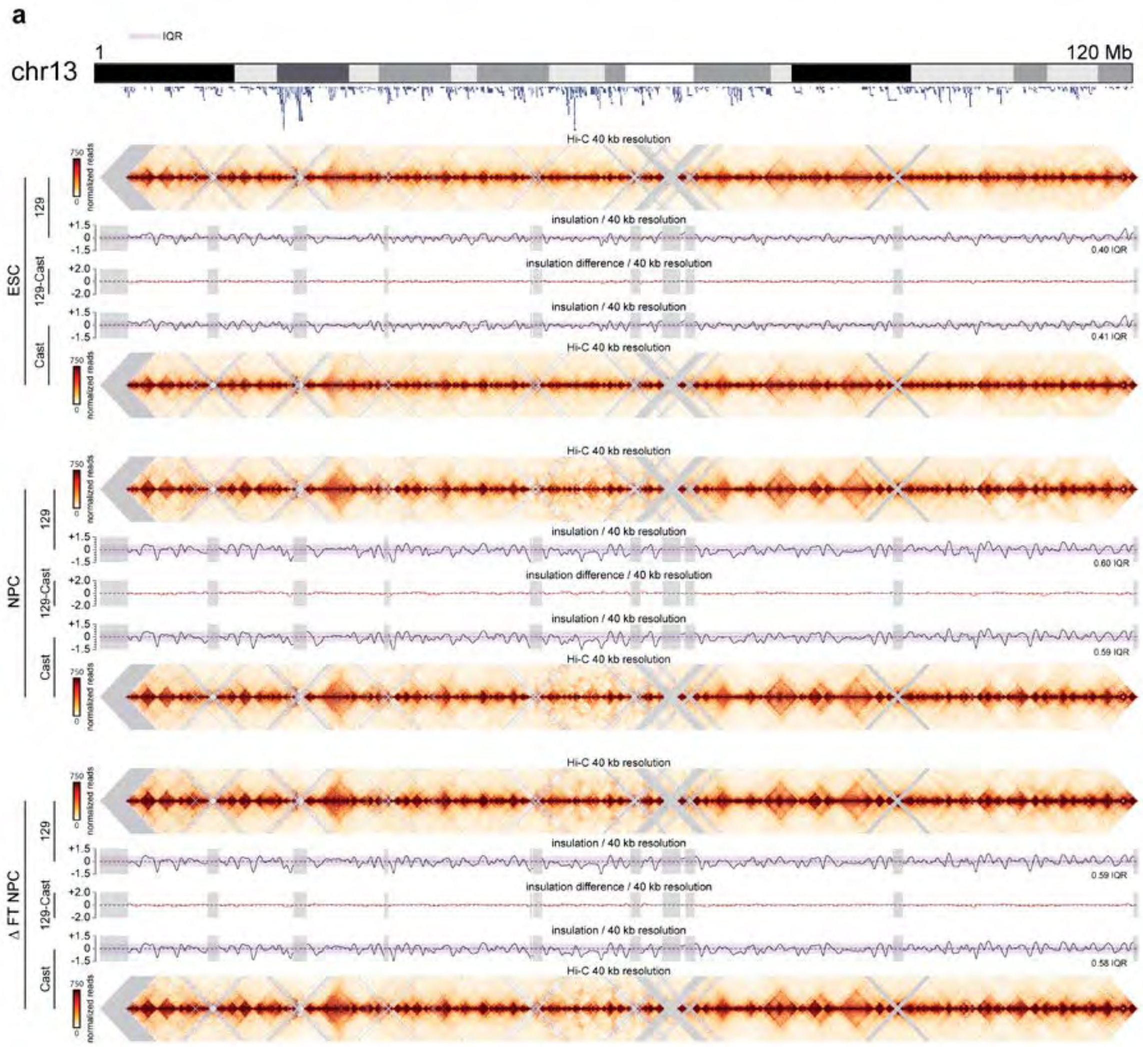


Figure 3.4 | chr13

a, Hi-C data, insulation scores, and the difference in insulation scores (129-cast) are shown for ESC (GUR.2d), NPC (GEI.72b) and mutant NPC (D9B2/B129T3) for both alleles (Cast and 129) for chr13. Large dips in the insulation vector are found at TAD boundaries. Peaks in the insulation vector are found towards the center of each TAD. The insulation difference plot highlights areas of differential TAD structure between the alleles (rare).

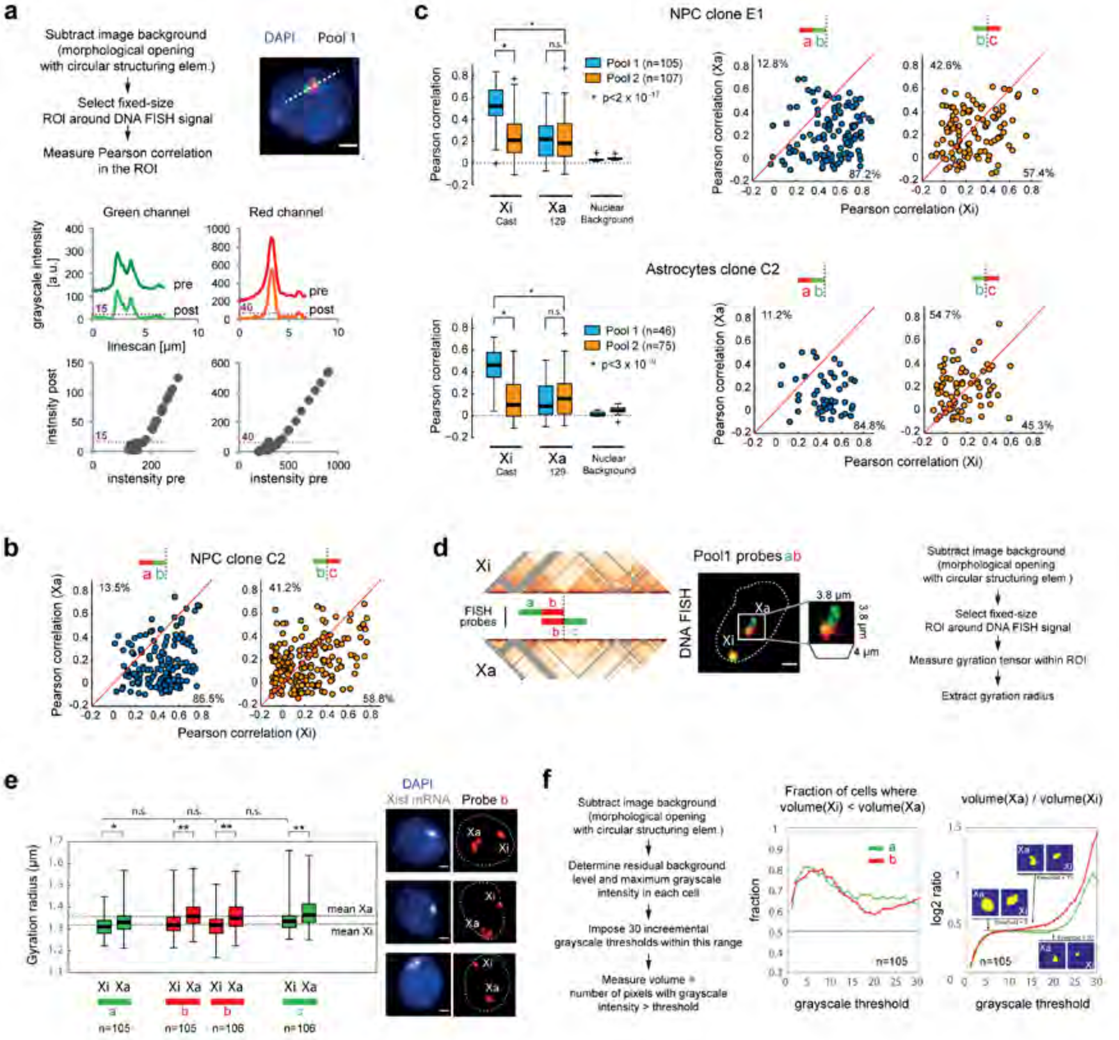


Figure 3.5 | DNA / RNA FISH

a, Top left panel: Scheme of the procedure used to quantify Pearson correlation. A background is generated for each xy plane in a three-dimensional z-stack by morphological opening the image with a circle of 5 pixels in radius, and subtracted from it. Pearson correlation between red and green pixel intensities is measured inside a fixed-size region of 40x40x20 pixels (5.16 x 5.16 x 4 µm) centered on each FISH signal. To demonstrate that background subtraction does not impact on the measured correlations, we show here a line-scan of 10 µm across a typical DNA FISH signal (top right panel). The shape of the signals along the line scan, as well as their relative intensities, is not affected by background subtraction (bottom). b, In more than 80% of nuclei in NPCs, Pearson correlations is higher on the Xi than on the Xa. Shown is NPC clone C2 (the same where Hi-C was performed). c, Same quantification as in Figure 1c (and panel b) for an independent NPC clone (E1) where the active X is on the 129 allele and the inactive X on the Cast, and in astrocytes derived from NPC clone C2. d, Scheme of the gyration tensor based analysis of FISH volumes (see methods). e, Left panel: Gyration radii of DNA FISH signals from probes a, b and c. Probe b was used in combination with both probes a and c separately in two independent experiments. Statistical significance was assessed by Wilcoxon's rank sum test (*=$p<0.05$, **=$p<1e-5$). The mean gyration radii for Xa and Xi signals are indicated by dotted lines as a guide for the eye. Right panel: representative images of probe a, showing smaller size and increased roundness of the Xi signals. f, Left panel: scheme of the thresholding-based method for volume quantification. Thirty increasing threshold levels were imposed, starting from the residual grayscale background level surrounding the signal, up to the minimum between the red and green channel grayscale maxima. For each of these thresholds we determined the number of voxels in each channel, where the grayscale intensity was higher than the threshold. Center panel: The fraction of cells where the Xa signal is larger than the Xi is between 60% and 80% in the entire threshold range. Right panel: in a wide range of thresholds, the volume of Xa signals is approximately 25% bigger than Xi signals. Results are shown here for probes a and b; the same holds for probe c (not shown).

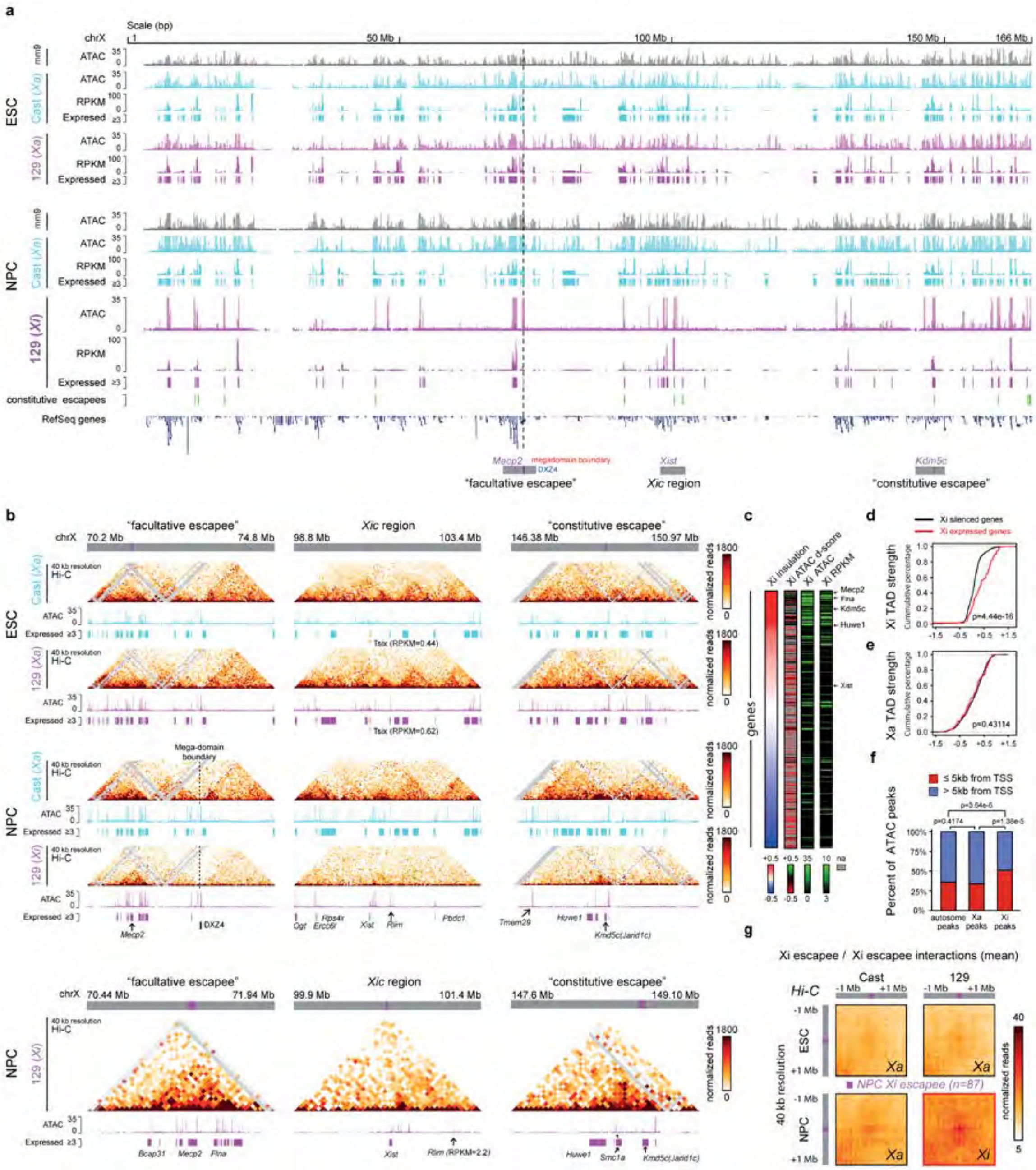


Figure 3.6 | Integrating expression, chromatin accessibility and chromatin conformation along the Xi.

a, X-Chromosome-wide ATAC-Seq and RNA-Seq in ESCs and NPCs. ATAC shows signal for ambiguous, 129- and Cast-specific reads in ESCs and NPCs. RNA-Seq shows total signal as well as expressed gene calls. ATAC-Seq shows global loss of chromatin accessibility and expression on the Xi, except at specific locations (pink boxes) that mostly overlap with escape genes. Dotted line: mega-domain boundary. Bottom: location of regions shown in panel b. Position of constitutive escapees was adapted from ref. [39]. b, The size and strength of residual TAD-like structures on the Xi correlates with the genomic extent of residual transcription and chromatin accessibility, as exemplified by allele-specific Hi-C, RNA-Seq and ATAC-Seq in the *Mecp2*, *Xist* and *Jarid1c (Kdm5c)* regions. Hi-C data are shown at 40-kb resolution. * = Tsix expression in ESC, manually indicated. c, Integrative analysis of Hi-C insulation (TAD structure), ATAC-Seq d-score, ATAC-Seq read counts, and RNA-Seq RPKM. Each row is a gene/promoter. All heatmaps are sorted by insulation score, highest to lowest (strongest-to-weakest TAD signal). Regions with elevated TAD structure harbor promoters that are expressed and accessible on the Xi. ATAC d-scores are calculated by comparing Xi vs. Xa ATAC-Seq peaks within gene promoters. d, The 87 Xi expressed genes (escapees) fall within regions with higher insulation scores on the Xi as compared to the 567 Xi silenced genes (KS test p-value = 4.44e-16). e, The 87 Xi expressed genes (escapees) and the 567 Xi silenced genes have similar insulation scores on the Xa. (KS test p-value = 0.43114). f, ATAC-Seq peaks on the Xi tend to be closer to TSSs (within 5kb) than peaks on autosomes and the Xa. g, Interaction pile up map showing mean interaction signal for all pairwise combinations of the 87 WT NPC Xi Escapees on the Xa and Xi. Escape genes tend to contact one another in 3D space on the Xi.

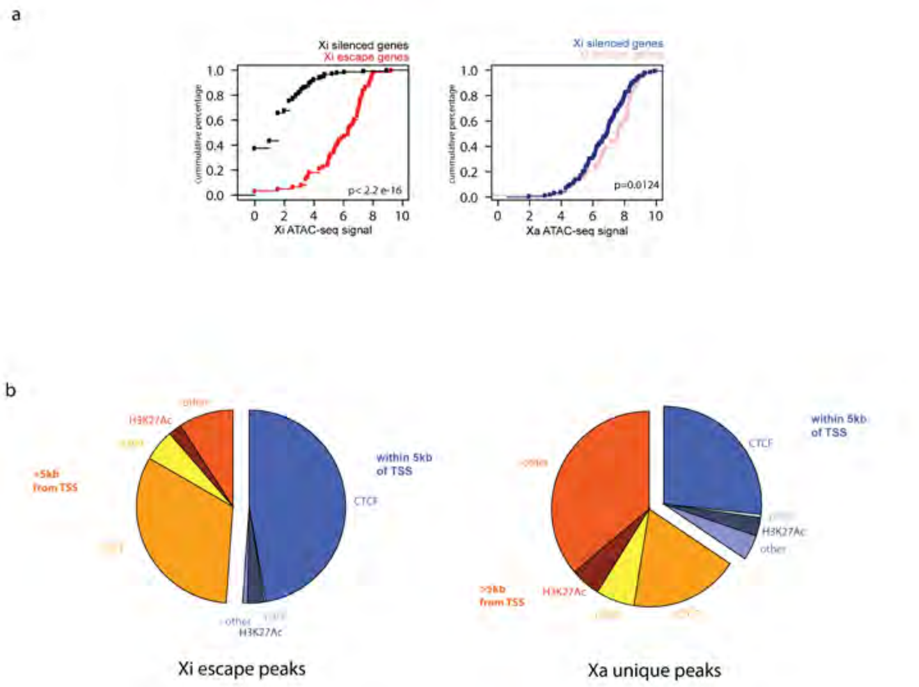


Figure 3.7 | ATAC peaks

a, Escape genes on the Xi (as determined by RNA-Seq) fall within regions with high ATAC-Seq signal (KS test p-value < 2.2 e -16). b, Pie charts showing the distribution of peaks that escape XCI vs. the peaks that are unique to the Xa. Peaks are classified into those that are promoter-proximal (within 5kb of TSS) and distal (>5kb from TSS). Annotations are based on binding sites identified by ChIP-Seq [84], [85].

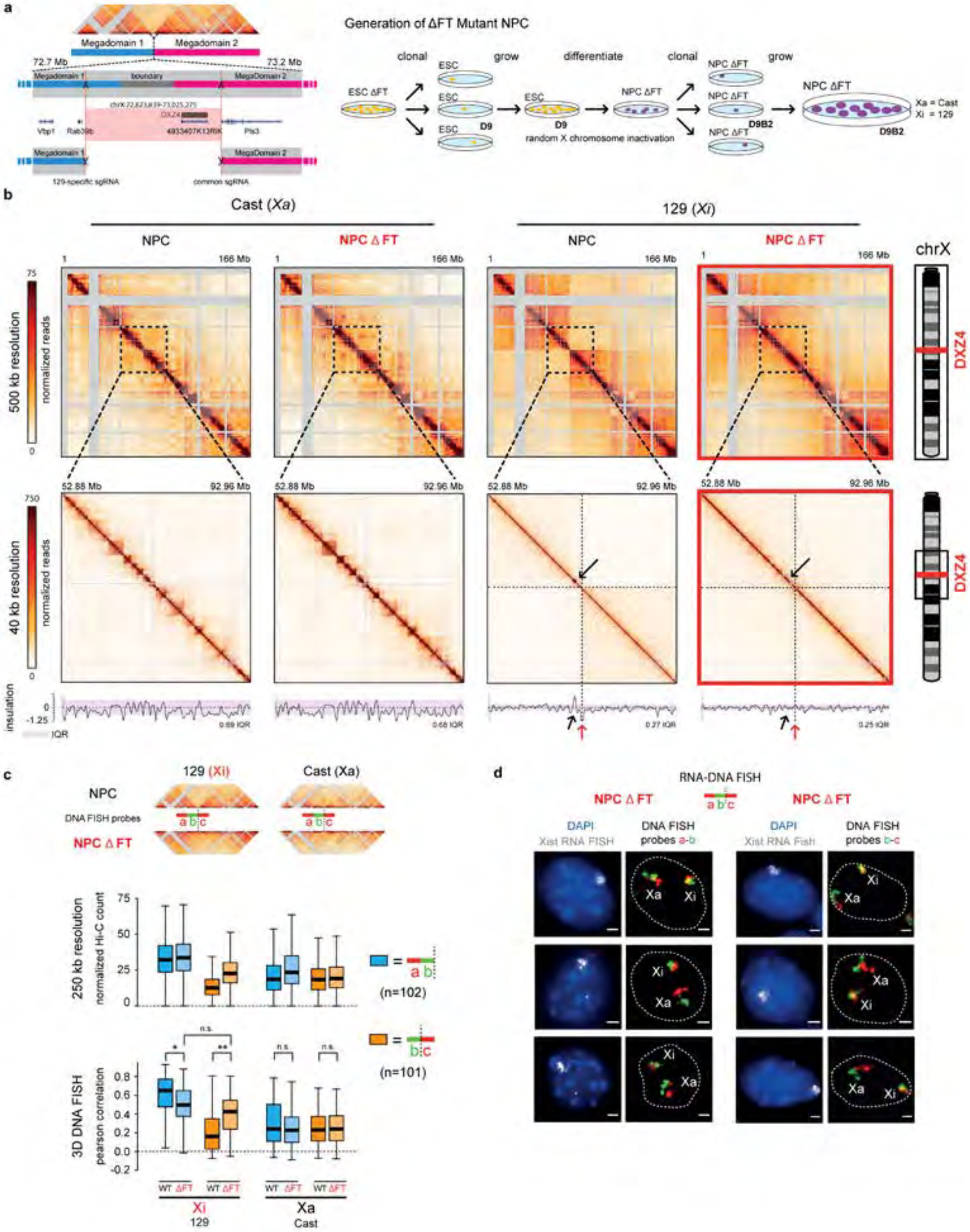


Figure 3.8 | Deletion of the boundary between Xi mega-domains leads to loss of bipartite folding.

a, Scheme of the 200-kb mega-domain frontier deletion (ΔFT) encompassing the DXZ4 macrosatellite. A sgRNA targeting a 129-specific SNP was used to generate a deletion specifically on the 129 X chromosome. Black arrow: position of the residual TAD in the 40-Mb region. Red arrow: position of DXZ4. B, Relative allele-specific Hi-C contact probability maps for chromosome X in wild-type and ΔFT NPCs at 100-kb resolution (top), and for a 40-Mb region centered around the DXZ4 position at 40-kb resolution (bottom). Insulation score is plotted at the bottom of each 40-Mb heatmap. Shaded areas indicate the range of insulation scores along the chromosome. c, Top: Scheme of the DNA FISH probe sets (*a*-*b*: inside the same mega-domain, *b*-*c*: across the boundary). Bottom: Loci detected by probe set b-c are more interacting in the ΔFT than in the wild-type Xi both in Hi-C (left) and in 3D-DNA FISH (right), showing loss of mega-domain boundary. * denotes $p<2e-4$ and ** $p<1e-5$ in a Wilcoxon's rank sum test corrected with Bonferroni for multiple hypothesis testing. d, Sample RNA/DNA FISH images showing that signals from probe set b-c are more overlapping on the ΔFT Xi than on the wild-type Xi.

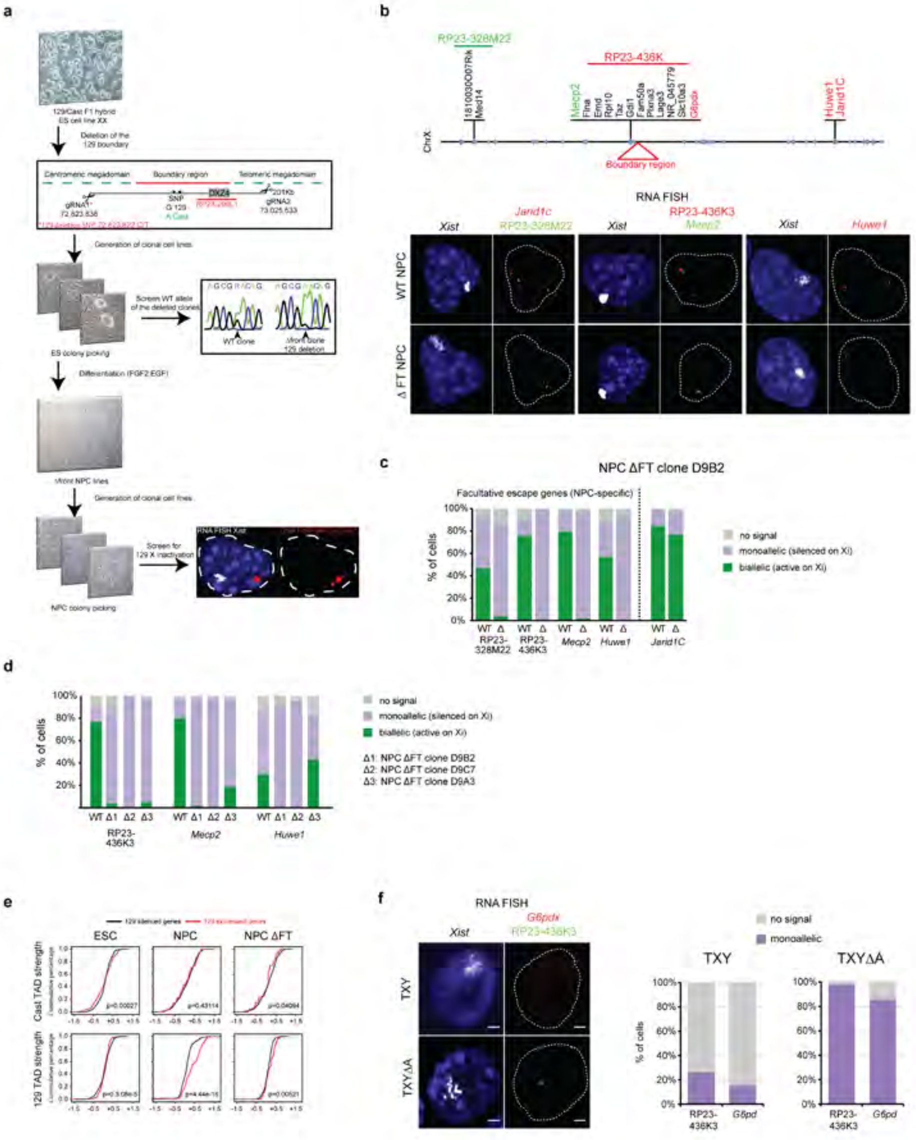


Figure 3.9 | Deletion Strategy and mutant RNA FISH

a, Scheme of the strategy used to delete the mega-domain boundary region in ESCs and to derive ΔFT NPCs. b, RNA FISH against constitutive and facultative escapees confirms RNA-Seq and ATAC-Seq results. Top: The positions of BAC probes (RP23-328M22 and RP23-436K) are shown relative to the escape genes that they span. Colored gene names correspond to transcripts that were detected with specific fosmid probes. Bottom: sample RNA FISH images showing that expression of facultative (*Mecp2* and BAC probes) but not constitutive (*Jarid1c*) escapees is lost on the ΔFT Xi. c, Quantification of the RNA FISH experiment in panel b. d, Quantification of the same RNA FISH experiment as in panel b including two additional mutant NPC clonal cell lines. WT and Δ1 bars represent the same data as in panel c. e, Cumulative plots TAD strength of the expressed versus the silences genes on Cast and 129 chromosomes for all three samples (ESC, NPC, ΔFT). Escapee genes on the Xi chromosomes (NPC 129, ΔFT NPC 129) show higher insulation scores as compared to silenced genes. f, RNA FISH against *G6pdx* and a group of genes recognized by the RP23-436K BAC (see panel b) showing that expression of X-linked genes is lost upon induction of wild-type but not A-repeat mutant *Xist* in male ESCs. TXY and J1:XistΔA were treated with doxycycline for two days.

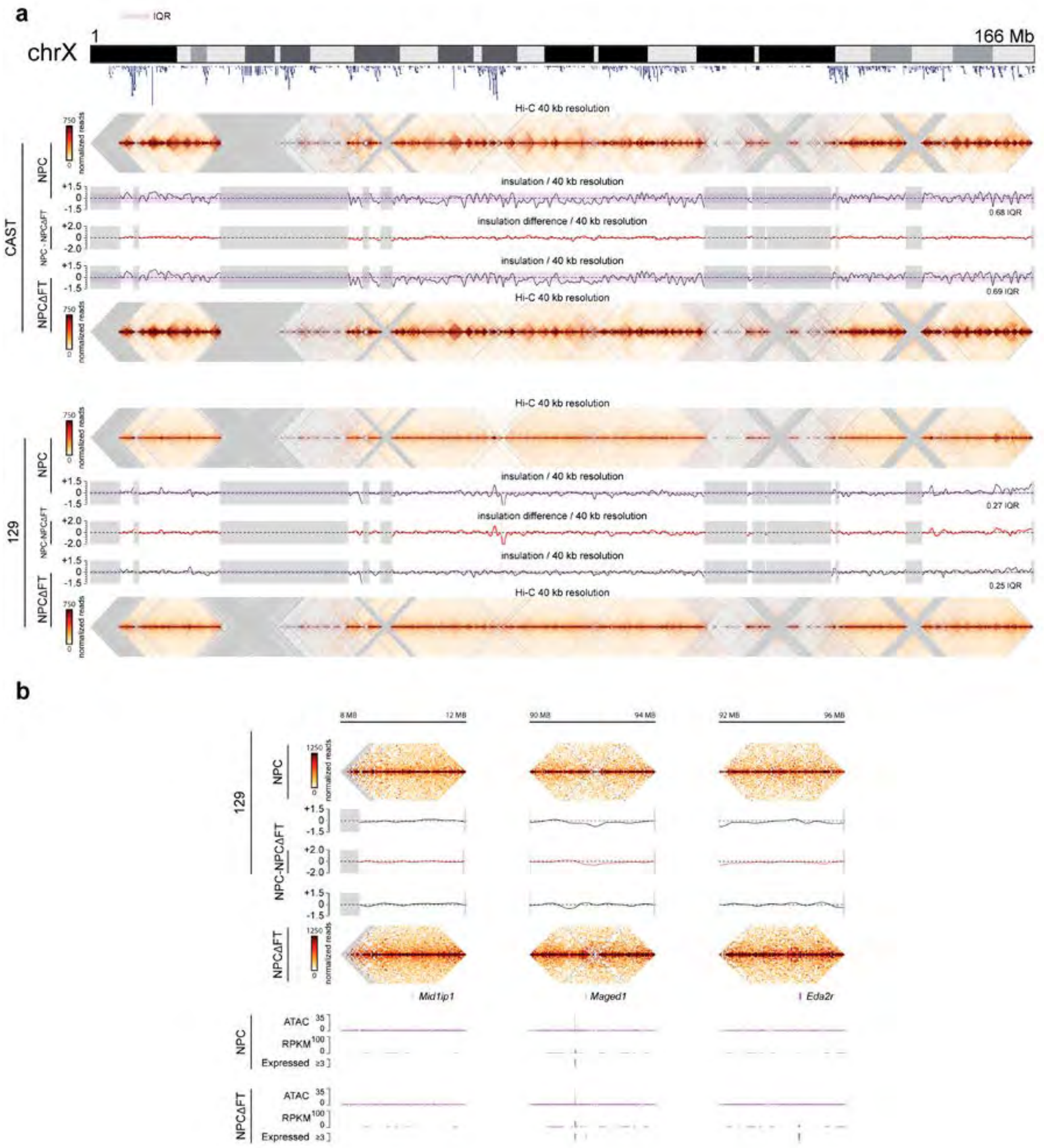


Figure 3.10 | WT and MT insulation

a, Hi-C data, insulation scores, and the difference in insulation scores are shown to compare the WT Xi (NPC 129) and the ΔFT NPC 129). (Top) shows the Cast allele (Xa) for both samples. (Bottom) shows the 129 allele (Xi) for both samples. Large dips in the insulation vector are indicative of TAD boundaries. Peaks in the insulation vector are found towards the center of each TAD. The insulation difference plot highlights areas of differential TAD structure between the WT and ΔFT NPCs. B, Zoom in of 3 regions centered on novel escapees identified on the ΔFT NPC Xi. Left, *Mid1ip1*; Center, *Maged1*; Right, *Eda2r*.

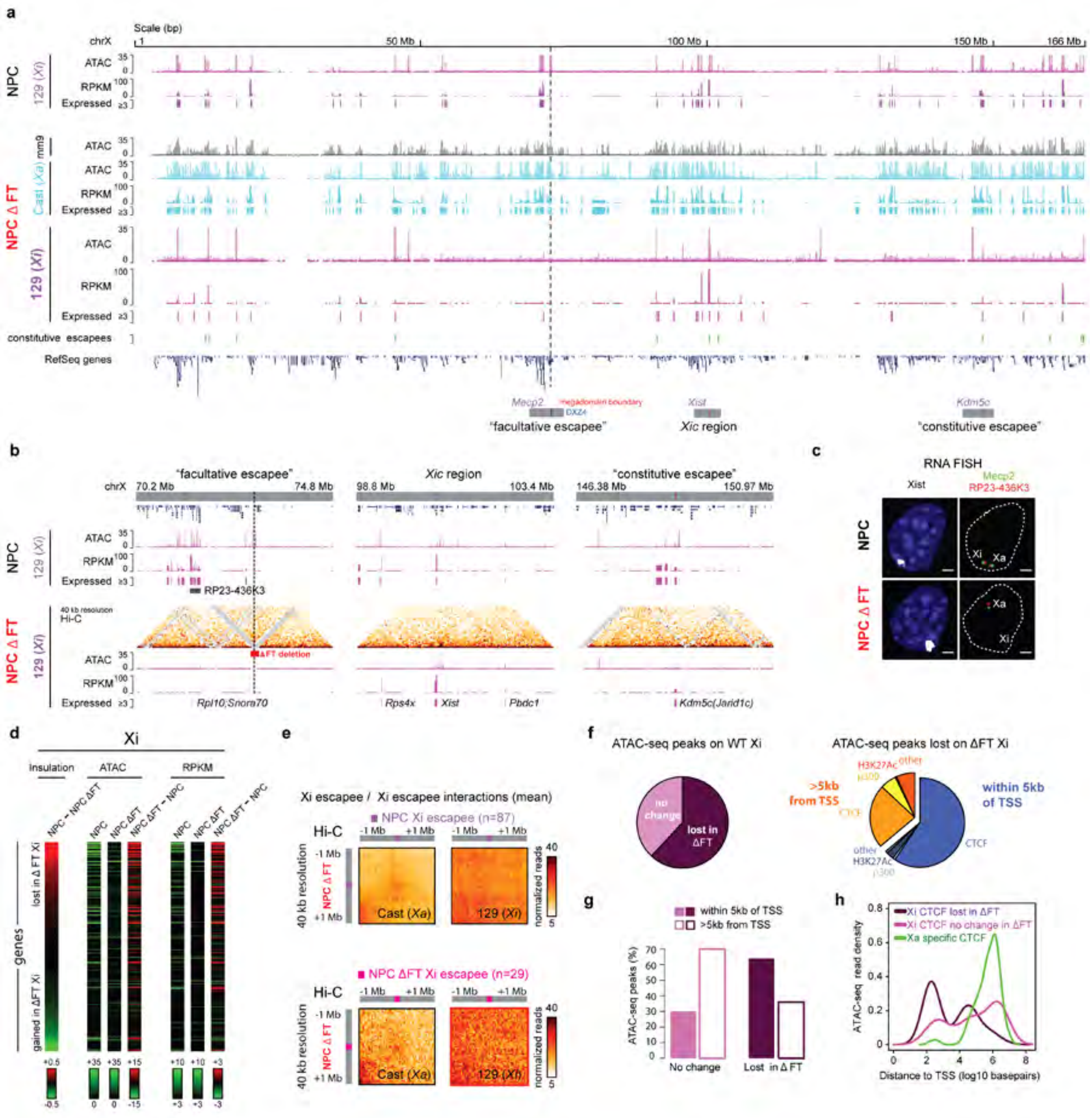


Figure 3.11 | Deletion of the mega-domain boundary region results in altered facultative escape profiles on the Xi.
a, Chromosome-wide ATAC-Seq signal generated with ambiguous, 129- and Cast-specific reads in WT NPC and ΔFT NPC, showing global loss of chromatin accessibility on the ΔFT Xi except at the XIC and constitutive escape genes. b, Zoomed in view of three regions on the ΔFT Xi showing Hi-C interactions, RNA-Seq and ATAC-Seq signal. Regions from left to right show the cluster of escape genes proximal to the deletion region containing *Mecp2*, the Xic, and the region encompassing *Jarid1c.* ATAC-Seq from WT NPCs is included for reference

(previously shown in Figure 2). Also shown in the left panel is the position of the BAC probe (RP23 436K3) used for the RNA FISH experiment in panel c. c, RNA FISH with a fosmid probe hybridizing to *Mecp2* and a BAC probe spanning 13 of its neighboring facultative escape genes (see panel b). d, Integrative analysis shows correlation between loss of TAD structure in ΔFT NPCs with loss of accessibility (ATAC) and loss of expression (RPKM). Plotted is the difference in insulation (NPC - ΔFT NPC) in the 40kb region overlapping the promoter of each gene, NPC ATAC counts, ΔFT NPC ATAC counts, NPC-ΔFT NPC ATAC difference, NPC RPKM, ΔFT NPC RPKM and NPC - ΔFT NPC RPKM difference. ATAC counts are extracted from the promoter of each gene (+/- 500bp from TSS). e, Interaction pile up map showing mean interaction signal in ΔFT NPCs for all pairwise combinations of the (87) WT NPC Xi Escapees and the (29) ΔFT NPC Xi escapees. f, Left panel: Quantification of ATAC-Seq peaks in WT and ΔFT NPCs on the Xi. Of 224 Xi peaks in WT, 139 are lost in the mutant. Right panel: ChIP-Seq annotation of ATAC-Seq peaks lost in ΔFT NPCs. Peaks are divided into those that are within 5kb of a TSS (blues) and those that are >5kb from a TSS (oranges). g, Histogram showing the distance of ATAC-Seq peaks that are lost and those that do not change upon deletion of the mega-domain boundary to TSSs. Peaks are quantified within 5kb of the nearest TSS and >5kb from the nearest TSS. h, Plot showing the distance (log10) of CTCF peaks that are lost in ΔFT NPCs, that do not change in ΔFT NPCs, and those on the Xa from TSSs of escape genes.

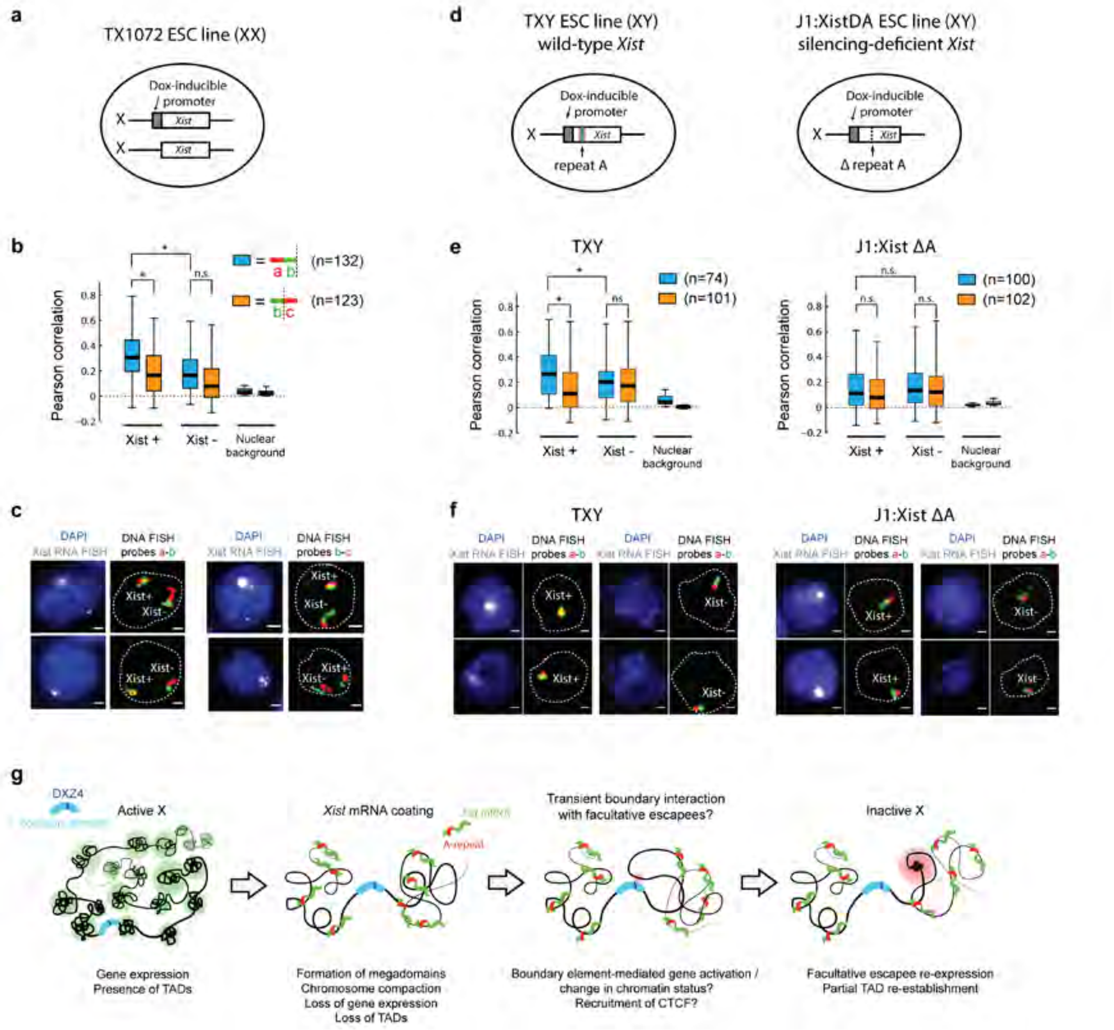


Figure 3.12 | *Xist*-mediated silencing is sufficient to generate a boundary at DXZ4 in undifferentiated ESC.

a, Schematic representation of TX1072 female ESC in which *Xist* expression can be induced via a tetracycline-responsive promoter at the endogenous *Xist* locus. b, RNA/DNA FISH as in Figures 1 and 3 was performed in TX1072 cells treated for three days with doxycycline. Probes a-b overlap more on the *Xist*-coated than on the wild-type X chromosome, whereas signals from b-c show lower overlap and partitioning of the *Xist*-coated chromosome into two separate domains. * denotes p<1e-7 in a Wilcoxon's rank sum test corrected with Bonferroni for multiple hypothesis testing. c, Sample RNA/DNA FISH images from the experiment in panel b. d, Schematic representation of TXY and J1:XistΔA male cell lines, carrying a tetracycline-inducible wild-type and A-repeat mutant *Xist*, respectively, at the endogenous *Xist* locus. e, RNA/DNA FISH shows increased overlap of probes a-b on the *Xist*-coated X chromosome in the wild-type, but not the A-repeat mutant cells. Probes b-c show lower overlap (indistinguishable from the non-*Xist* coated chromosomes in cells where *Xist* expression was not induced upon doxycycline treatment). * denotes $p<0.05$ in a Wilcoxon's rank sum test corrected with Bonferroni for multiple hypothesis testing. f, Sample RNA/DNA FISH images from the experiment in panel e. g, Model of mega-domain boundary-mediated control of chromosome folding and facultative escape. *Xist* coating causes gene silencing and leads to chromosome-wide conformational changes, including formation of mega-domains, overall compaction of chromosome folding, and loss of TADs. Further during differentiation, transient interactions with the mega-domain boundary may occur and result in facultative escape and re-establishment of TADs at facultative escape loci.

Table 3.1 | RNA-Seq Table for all genes (available only via GEO due to size). Table containing all relevant RNA-Seq data for all locations (genes) along the X chromosome. The table columns are as follows: 1, xloc; 2, chr; 3, start; 4, end; 5, gene; 6, B129T3__129S1__category; 7, B129T3__129S1__pval; 8, B129T3__129S1__reads; 9, B129T3__129S1__rpkm; 10, B129T3__129S1__status; 11, B129T3__CAST__category; 12, B129T3__CAST__pval; 13, B129T3__CAST__reads; 14, B129T3__CAST__rpkm; 15, B129T3__CAST__status; 16, GEI.72b__129S1__category; 17, GEI.72b__129S1__pval; 18, GEI.72b__129S1__reads; 19, GEI.72b__129S1__rpkm; 20, GEI.72b__129S1__status; 21, GEI.72b__CAST__category; 22, GEI.72b__CAST__pval; 23, GEI.72b__CAST__reads; 24, GEI.72b__CAST__rpkm; 25, GEI.72b__CAST__status; 26, GUR.2d__129S1__category; 27, GUR.2d__129S1__pval; 28, GUR.2d__129S1__reads; 29, GUR.2d__129S1__rpkm; 30, GUR.2d__129S1__status; 31, GUR.2d__CAST__category; 32, GUR.2d__CAST__pval; 33, GUR.2d__CAST__reads; 34, GUR.2d__CAST__rpkm; 35, GUR.2d__CAST__status. xloc is a numerical ID for each gene location. chr is the chromosome. start is the start position of the gene. end is the end position of the gene. (for positions, start<end, not re-oriented by strand) gene is the gene name. The remaining columns are broken down into groups of 5 per sample, per allele. NNNN is the sample name. ESC = GUR.2d; WT NPC = GEI.72b; ΔFT NPC = B129T3 (D9B2). XXXX is the allele, 129 for the 129S1 allele, CAST for the Cast allele. The five columns are: NNNN__XXXX_category, category assignment of expression (bi,mono,biased,na, see ref 20) NNNN__XXXX__pval, p-value of the allelic assignment. NNNN__XXXX__reads, number of allelic reads. NNNN__XXXX__rpkm, RPKM value for the allelic gene. NNNN__XXXX__status, expression status of the gene, expressed or silenced. We defined expressed as ≥ 3 RPKM.

EXTENDED DATA FIGURES AND TABLES
Figure 1-5
Extended Data Table 1
Extended Data Figures 1-7

AUTHOR INFORMATION

ACCESSION NUMBERS:
Hi-C: Will be made available to referee
ATAC-Seq: GSE71156
RNA-Seq: Will be made available to referee

Acknowledgements

We would like to thank members of the Heard and Dekker labs for their help and critical insights and the PICT-IBiSA@BDD (UMR3215/U934) Imaging facility of the Institut Curie. Supported by grants from the National Institutes of Health (P50-HG007735) and Scleroderma Research Foundation (to H.Y.C.), from the National Human Genome Research Institute (R01 HG003143) to J.D, the Human Frontier Science Project to N.K., an ERC Advanced Investigator award, EU FP7 grants SYBOSS and MODHEP, La Ligue, Labex DEEP (ANR-11-LBX-0044) part of the IDEX Idex PSL (ANR-10-IDEX-0001-02 PSL) to E.H.

Competing financial interests

The authors declare no competing financial interests.

Methods

Cell Culture

The hybrid mouse ES cell line F121.6 (129Sv-Cast/EiJ), a gift from Prof. Joost Gribnau, was grown on mitomycin C-inactivated MEFs in ES cell media containing 15% FBS (Gibco), 10^{-4}M ß-mercaptoethanol (Sigma), 1000U/ml of leukaemia inhibitory factor (LIF, Chemicon).

Boundary Deletion

To generate the boundary region deletion, 5X106 ESCs were transfected with 5µg each of two plasmids (pX459) each expressing Cas9 and a chimeric guide RNA (gRNA1: CATGTTTGAGCATGGAAACCCGG, chrX:72823838-72823860; gRNA2: GGGTTATGGCGGTCGGTTCCTGG, chrX:73025513-73025535). Subcloning of ESC was made by limiting dilution. Cells were treated for 24 hours with puromycin. As soon as visible, single colonies were picked under a microscope to be screened for deletion by PCR (forward primer: CGTAGACGCGGCAGTAGTTT, reverse primer: ACATAAACTCCTTTTCAGGACCA). To identify the targeted allele, we performed a PCR using primers (F:CTGTCCAAATGGAGGTGCTT R:C CTAGGTCCGCTCTCTATCG) that amplify a 203-bp amplicon specifically on the WT allele, which contains a SNP (rs29035891). After amplification, PCR products were gel-purified and sequenced using the forward or reverse primer used for PCR. Clones positive carrying the deletion were expanded and differentiated into NPC as previously described (20) and subcloned by limiting dilution. NPC lines were maintained in N2B27 medium supplement with EGF and

FGF (10ng/ml each), on 0.1% gelatin-coated flasks. Clones carrying the boundary deletion on the inactive X were identified by RNA FISH against Xist with the p510 plasmid probe and DNA FISH with a BAC hybridizing inside the deleted region (RP23-299L1).

Hi-C Read Mapping / Binning / ICE correction

Hi-C was performed as previously described (13, 15). To obtain allele-specific Hi-C interaction maps in female ESCs ($Xa^{cast}Xa^{129}$) and a derived clonal NPC line ($Xa^{cast}Xi^{129}$) (Methods; **Extended Data Figure 1**) (20), we first constructed an allelic genome using the reference mm9 genome and all 19,722,473 SNPs. The allelic (Cast and 129) genomes were then combined to create a reference diploid genome (consisting of 44 chromosomes; 1-19 X,Y,M). All reads were aligned to the diploid genome (as described in ref. [86]), thus allowing for a competitive mapping strategy between the two alleles. All reads were trimmed to 50bp and then aligned using the novoCraft novoalign (v 3.02.00) software package. Reads were aligned using the following options (-r all 5 -R 30 -q 2 -n 50, minimumReadDistance=5). The best alignment was selected from the list of the top 5 alignments. The alignment was considered unique (allelic), if it's alignment score was ≥ 5 from the 2nd best alignment score (alignment score taken from the ZQ tag). Reads that aligned uniquely to an allele were classified as allelic (either Cast or 129) whereas reads that aligned to both alleles equally (≤ 5 distance) were classified as ambiguous (AMB)

(**Extended Data Figure 1d**). Uniquely aligned Hi-C interactions between loci located on the same chromosome were assigned to a specific parental chromosome *in cis* when at least one of the two reads contained a diagnostic SNP, and the other either contained a SNP from the same allele, or mapped to both alleles [87]. We obtained the following paired-end read counts: For ESC (GUR.2d), a total of 401,684,614 interactions could be aligned, 372,272,389 of which were unique (after PCR duplicate filter), and 95,650,438 of which could be placed to either the Cast or 129 allele (25.69%). For NPC (GEI.72b), a total of 277,440,656 interactions could be aligned, 253,254,798 of which were unique (after PCR duplicate filter), and 82,323,031 of which could be placed to either the Cast or 129 allele (32.51%). For ΔFT NPC (D9B2/B129T3), a total of 229,331,123 interactions could be aligned, 222,941,525 of which were unique (after PCR duplicate filter), and 85,331,870 of which could be placed to either the Cast or 129 allele (38.28%). The difference in percent of reads assignable to either allele is likely due to differences in the percent of cis interactions found in each sample (biological or technical variation). The 82-95 million read depth supported generation of allele-specific chromatin interaction maps at multiple resolutions (10 Mb, 2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, and 40 kb).

Biological replicates were highly correlated. Pearson's correlation coefficients for 500kb data on chrX were as follows: EHSNP-mF1216__R1R2__chrX-129S1, 0.992331; EHSNP-mF1216__R1R2__chrX-cast, 0.990373; EHSNP-mNPe-deltaRF__R1R2__chrX-129S1, 0.976562; EHSNP-mNPe-deltaRF__R1R2__chrX-cast, 0.983614; EHSNP-mNPe__R1R2__chrX-129S1, 0.990976; EHSNP-mNPe__R1R2__chrX-cast, 0.995202; Autosomes showed similar correlation values. Overall these numbers indicate that the produced Hi-C data was of high quality and well correlated between biological replicates. We pooled all biological replicates into a single Hi-C data set per sample and subsequently used the pooled data for all analyses.

Iterative mapping and error filtering/iterative correction (IC) of the chromatin interaction data were performed as previously described [62], [86]. IC was performed on the diploid (44 chromosomes) (replicate pooled) genome-wide matrix for all resolutions.

Hi-C SNP density filter

To remove potential biases in the Hi-C data related to the density of SNPs in each bin, we calculated the number of SNPs residing in each genomic interval (bin) for all Hi-C bins across all bin sizes. We then calculated the median number of SNPs per bin, and produced a minimum required SNP density cutoff defined as the (median - 1.5 * IQR). Any bins with less SNP than the cutoff were removed from all analyses. The SNP density cutoffs used for each bin size were: 40 kb, 43 SNPs; 100 kb, 216 SNPs; 250 kb, 776.5 SNPs; 500kb, 1767.25 SNPs. The non SNP density filtered data was only used for visualization purposes (figure heatmaps). For all intents and purposes, we refer to IC Hi-C as data that has been IC'd and run through the snp-density filter.

Compartment analysis

The presence and location of the A/B-compartments were calculated as previously described (14). Compartments were derived from the 250 kb IC Hi-C data for each chromosome separately using the CIS maps for each sample / allele (Fig 1). The code used to generate the compartments (PC1 from PCA analysis) will be publically available on github following publication (matrix2compartment.pl). Compartments were generated all default options except the (cis alpha) option, set to (-ca 0.005).

Insulation and Boundary calculation

TAD structure (insulation/boundaries) was defined via the insulation method as previously described with minor modifications (14). The code used to calculate the insulation score will be publically available on github following publication (matrix2insulation.pl). Insulation vectors were detected using the following options: (-is 480000 -ids 320000 -im iqrMean -nt 0 -ss 160000 -yb 1.5 -nt 0 -bmoe 0). The output of the insulation script is a vector of insulation scores, and a list of minima along the insulation vector (inferred as TAD boundaries). The TAD boundaries were not used in this study.

Interaction Pile Up Maps

Interaction pile up maps were constructed from all pairwise interactions between either the list of (87) WT NPC Xi Escapees or the (29) ΔFT NPC Xi escapees. Using the 40 kb Hi-C data, a 2 MB window centered around each pairwise interaction (pixel) was taken (25 bins in each direction, yielding 51 x 51 sub-matrix). Any resulting sub-megabase which overlapped the (y=x) diagonal in the matrix was excluded from the analysis (effectively excluding all interactions < 2 MB). All sub-matrices were then averaged to produce the final (mean) pile up map. A strong signal at the center suggests that the elements used tend to contact one another in 3D space.

The Xi is as accessible and detectable in Hi-C as Xa and autosomes

The number of RAW reads observed for both the Xa and Xi were very similar for all chromosomes thus demonstrating that the Xi is simply not less accessible/visible to the Hi-C methodology. ESC-chrX-129S1, 1,118,327; ESC-chrX-Cast, 1,104,709; NPC-chrX-129S1, 1,147,072; NPC-chrX-Cast, 1,148,128; ΔFTNPC-chrX-129S1, 1,314,476; ΔFTNPC-chrX-Cast, 1,288,802. Bias in read directional due to partial digestion is typically observed up to ~10kb. For interactions between fragments separated by over 10kb this bias is negligible, indicating at least one digestion even occurring between them in every cell. This genomic distance is therefore a measure for digestion efficiency (13). For both the Xa and the Xi this genomic distance is ~6-10 kb, indicating that digestion efficiency of chromatin on the Xa and Xi are comparable. Thus, the unique conformation of Xi does not affect Hi-C analysis, as was also found for condensed mitotic chromosomes (15).

RNA and 3D-DNA FISH

FISH was performed as previously described [88]. ESCs and NPCs were cultured on gelatin-coated coverslips #1.5 (1 mm) and fixed in 3% paraformaldehyde for 10 min at RT. Cells were permeabilized on ice for 5 min in 1X PBS, 0.5%Triton X-100 and 2 mM Vanadyl-ribonucleoside complex (VRC,a New England Biolabs), and coverslips were stored in 70% EtOH at −20°C. Prior to FISH, samples were dehydrated through an ethanol series (80%, 95%, 100% twice) and air-dried briefly. For RNA FISH, cells were directly hybridized with denatured probes. For DNA FISH, samples were first denatured in 50% formamide / 2X SSC (pH = 7.3) at 80°C for 37 (ESC) and 35 (NPC) min, immediately placed on ice and washed two times with ice-cold 2X SSC. After overnight hybridization at 37°C for RNA FISH or 42°C for DNA FISH, coverslips were washed at 42°C for RNA or 45°C for DNA, three times for 5 min in 50% formamide / 2X SSC at pH = 7.3 and three times for 5 min in 2X SSC. Nuclei were counterstained with 0.2 mg/ml DAPI (2 mg/ml for structured illumination microscopy), further washed two times for 5 min in 2X SSC at RT and finally mounted with 90% glycerol, 0.1X PBS, 0.1% p-phenylenediamine at pH9 (Sigma).

RNA FISH probes

We used the p510 plasmid coupled with Cy5 to detect Xist. For RNA FISH on escape genes, we used the following BAC and fosmid probes: RP23-436K3, RP23-328M22, RP24-436K3, WI1-1269O10 (*Mecp2*), RP24-157H12 (*Huwe1*), RP23-13D21 (*G6pdx*), RP24-148H21 (*Jarid1c*).

DNA FISH probes

In experiments to detect mega-domain boundary, fluorescent oligonucleotides (average length 45 bp, 5'-modified with Atto 448 or Atto 550, average density: one oligo every 3 kb) were obtained from MYcroarray Inc (Ann Arbor MI, USA). Oligos were designed to tile the following consecutive 18-Mb regions: chrX:35'000'000-53'000'000, chrX:53'000'000- 72'000'000, and chrX:72'000'000-90'000'000. To detect the DXZ4 region we used the RP23-299L1 BAC.

Imaging and quantification of 3D-DNA FISH

Three-dimensional image stacks (200 nm distance between consecutive xy planes) were acquired on a DeltaVision Core wide-field microscope (Applied Precision) equipped with a CoolSNAP HQ2 camera operated at 2X binning, and a 100X PlanApo oil immersion objective (the effective pixel size was 129x129 nm). Xi signals were identified via the presence of an *Xist* mRNA cloud in the far-red channel (p510-Cy5 probe). Pearson correlation between red and green signals was calculated using custom-made ImageJ macros as follows. After subtracting the background from each xy plane (generated by morphological opening the image with a circle of 5 pixels in radius), Pearson correlation between red and green pixel intensities was measured inside a fixed-size region of 40x40x20 pixels (5.16 x 5.16 x 4 μm^3) centered on each FISH signal. Significance of Xi vs Xa differences in correlation was assessed by Wilcoxon's rank sum test. Random nuclear positions were used to estimate the background correlation that could be observed due to non-specific probe hybridization.

The gyration tensor of an image is defined as $S_{ab} = \sum_k I_k (r_a^k - r_a^{CM})(r_b^k - r_b^{CM}) \,/ \sum_k I_k$, where *k* is an index running over voxels, I_k is the grayscale intensity of voxel *k*, and r_a^k and r_a^{CM} are the *a*-th components (x,y, or z) of the xyz position of voxel *k*, and of the center of mass of the image, respectively. The gyration tensor was valuated in a region of interest of 3.8 x 3.8 x 4 μm3 centered on each FISH signal and the gyration radius was calculated as $R_g = \sqrt{\lambda_1 + \lambda_2 + \lambda_3}$ where $\lambda_{1,2,3}$ are the eignevalues of S_{ab}.

RNA-Seq

RNA-Seq data for the ESC (GUR.2d) and NPC (GEI.72b) was obtained from previously published work [76](20). RNA-Seq data for the mutant NPC (D9B2/B129T3) was obtained and processed as previously described (20).

RNA-Seq 'Expressed/Escapee' classification

The allelic RPKM values were derived for each gene by splitting the RPKM value by the 129 ratio. 129 RPKM = (RPKM * 129 ratio); Cast RPKM = (RPKM * (1 - 129 ratio)); Any gene with an allelic RPKM value ≥ 3 RPMK was classified as being expressed. Any gene expressed on the Xi was classified as being an escapee.

ATAC-Seq

ATAC-Seq library preparation was performed exactly as previously described [89]. Sequencing was carried out on an Illumina NextSeq 500 generating 2 x 75 bp paired-end reads. Libraries were sequenced to a depth of 25-35 million reads per sample. Reads were trimmed using CutAdapt and aligned using Bowtie2. Reads were aligned to a custom 129/CastEiJ genome in which SNP sites were replaced by "N." 52-58% of reads per line contained "N"s and were assigned to the 129 or Cast allele based on the identity of the base at that location. Reads containing non-concordant SNPs were rare and were discarded. Reads not containing SNP sites were included in overall peaks but not were excluded from allele-specific tracks. ATAC-Seq Peaks were called using MACS2 with no shifting model.

Assigning allele-specific ATAC-Seq peaks

For each ATAC-Seq peak, all N-containing reads were counted and assigned to 129 or Cast alleles based on SNP at the N-containing position. For each peak, a d-score was calculated as a measure of allelic imbalance [90]. Briefly, for a given peak the d-score was calculated as the ratio of 129 reads to total number of reads -1/2. A peak with a d-score ≥0.3 was assigned as a 129-specific peak. A peak with a d-score ≤-0.3 was assigned as a Cast-specific peak. Any peak with a d-score >-0.3 was assigned as a peak in 129 (monoallelic or biallelic). Any peak with a d-score <+0.3 was assigned as a peak in Cast (monoallelic or biallelic).

Annotating ATAC-Seq peaks using ChIP-Seq data

ATAC-Seq peaks were annotated using existing published ChIP-Seq datasets. CTCF ChIP-Seq came from whole female mouse brain [91]. Called CTCF binding sites were used and extended +/-300 bp before overlapping with ATAC-Seq peaks. H3K27Ac and p300 ChIP-Seq are from mouse NPCs [85]. For H3K27Ac and p300 ChIP-Seq data, peaks were called using MACS2 and then overlapped with ATAC-Seq peak locations.

Integrating Hi-C, ATAC-Seq, and RNA-Seq data

Integrative analysis of Hi-C insulation (TAD structure), ATAC-Seq d-score, ATAC-Seq counts, and RNA-Seq RPKM was performed as follows. A promoter region was defined for each gene as +/- 500 bp from the TSS. ATAC peaks were assigned to a gene if they overlapped with the promoter region. In the event that > 1 ATAC peak overlapped with the promoter, the closer ATAC peak was chosen. An ATAC count of 0 was assigned to each promoter, if it did not contain an ATAC peak. If the ATAC allelic counts overlapping the promoter were < 10, then the ATAC count was set to “NA”. The 40-kb bin overlapping the promoter region was used to display the insulation and insulation-difference value. ATAC d-scores are calculated by comparing Xi vs. Xa ATAC-Seq (-0.5=Xa-specific, 0=biallelic and 0.5=129-specific signals). For ATAC-Seq data, d-score was calculated as described above [90]. Briefly, the d-score for a given peak is calculated as $d = X_i/N_i - \frac{1}{2}$ where X_i is the number of reads coming from the 129 genome and N_i is the total number of reads covering that peak.

CHAPTER IV: The Hitchhiker’s Guide to Hi-C Analysis: Practical guidelines

Preface

This research chapter encompassed work published in Methods by Bryan R. Lajoie, Noam Kaplan and Job Dekker. The publication is entitled, “The Hitchhiker’s guide to Hi-C analysis: practical guidelines”, *Methods*, vol. 72, pp. 65–75, Jan. 2015, entitled [86]

Abstract

Over the last decade, development and application of a set of molecular genomic approaches based on the chromosome conformation capture method (3C), combined with increasingly powerful imaging approaches have enabled high resolution and genome-wide analysis of the spatial organization of chromosomes. The aim of this paper is to provide guidelines for analyzing and interpreting data obtained with genome-wide 3C methods such as Hi-C and 3C-seq that rely on deep sequencing to detect and quantify pairwise chromatin interactions genome-wide.

The 3D genome

The human genome consists of over 3 billion nucleotides and is contained within 23 pairs of chromosomes. If the chromosomes were aligned end to end and the DNA stretched, the genome would measure roughly 2 meters long. Yet the genome functions within a sphere smaller than a tenth of the thickness of a human hair (10 micron). This suggests that the genome does not exist as a simple one-dimensional polymer; instead the genome folds in a complex compact three-dimensional structure.

It is increasingly appreciated that a full understanding of how chromosomes perform their many functions, e.g. express genes, replicate and faithfully segregate during mitosis, requires a detailed knowledge of their spatial organization. For instance, genes can be controlled by regulatory elements such as enhancers that can be located hundreds of Kb from their promoter. It is now understood that such regulation often involves chromatin looping between the enhancer and the promoter [13]–[19]. Further, recent evidence suggests chromosomes appear to be folded as a hierarchy of nested chromosomal domains [1]–[6], and these are also thought to be involved in regulating genes, e.g. by limiting enhancer-promoter interactions to only those that can occur within a single chromosomal domain [7]–[11] .

The chromosome conformation capture methodology (3C) is now widely used to map chromatin interaction within regions of interest and across the genome. Chromatin interaction data can then be interpreted to gain insights into the spatial organization of chromatin, e.g. the presence of chromatin loops and chromosomal domains. The various 3C-based methods have been described extensively before and are not discussed here in detail [92], [93]. We first discuss methods and considerations that are important for using deep sequencing data to build bias-free genome-wide chromatin interaction maps. We then describe several approaches to analyze such maps, including identification of patterns in the data that reflect different types of chromosome structural features and their biological interpretations.

Methods to study the 3D genome

Indiscriminate methods such as microscopy or FISH can study the 3D genome, but have difficulty pinpointing the exact points of contact as well as measuring multiple discrete contacts simultaneously. The Chromosome Conformation Capture (3C) method was the first method to capture and measure all possible contacts of the 3D genome in an unbiased manner [20]. 3C has since been further developed into various other derivatives including 4C [21], [22] and 5C [23]. These methods use 3C as the core methodology by which they capture genomic interactions. They differ in the actual method by which the captured interactions are detected, e.g. by PCR in 3C and by unbiased deep sequencing in Hi-C and 3C-seq. Though the 3C method does capture genome-wide data, it was not until the era of deep sequencing came about that one was able to survey all genome wide interactions in a single experiment, as in Hi-C and 3C-seq.

In 3C, cells are cross-linked using formaldehyde, lysed and the chromatin is then digested with a restriction enzyme of choice (typically HindIII or EcoRI). The chromatin is then extracted and the restriction fragments are ligated under very dilute conditions to favor intra-molecular ligation over inter-molecular ligation. The crosslinks are then reversed, proteins are degraded and DNA is purified. The newly generated chimeric DNA ligation products represent pairwise interactions and can then be analyzed by a variety of down-stream methods.

Currently, there are two 3C-based methods to obtain genome-wide chromatin interaction data: Hi-C and 3C-seq. In the Hi-C protocol one includes a step to introduce biotinylated nucleotides at ligation junctions which enables specific purification of these junctions [1]. This has the important advantage that it prevents sequencing DNA molecules that do not contain such junctions and are thus not informative. In 3C-seq one employs the classical 3C protocol and often a more frequently cutting enzyme (e.g. DpnII) followed by intra-molecular ligation without biotin incorporation [4]. The ligated DNA is then directly sequenced to identify pairwise chromatin interactions genome-wide. The 3C-seq methodology sequences all molecules including un-ligated molecules which can complicate the processing / filtering steps and can reduce the percentage of usable reads. However experimental techniques exist to help minimize uninformative (un-ligated, self-ligated etc.)

Hi-C products

Here we discuss guidelines for analyzing genome-wide chromatin interaction maps generated by Hi-C, but many of these considerations also apply to 3C-seq data. We first discuss the steps required to obtain high-quality unbiased interaction maps. Then, we discuss analysis and interpretation of the interaction maps.

Hi-C data resolution

The space of all possible interactions, which is surveyed by Hi-C experiments, is very large. For example, consider the human genome. Using a 6-bp cutting restriction fragment, there are almost 10^6 restriction fragments, leading to an interaction space on the order of 10^{12} possible pairwise interactions. Thus, achieving maximal resolution is a significant challenge.

In light of this, it is crucial to establish the goals of the experiment, meaning whether one is most interested in either large-scale genomic conformations (e.g. genomic compartments) or specific small-scale interaction patterns (e.g. promoter-enhancer looping).

If the goal is to measure large scale structures, such as genomic compartments, then a lower resolution will often suffice (1MB-10MB). Here, Hi-C using a traditional 6bp-cutting enzyme could be used. However if the goal is to measure at a finer scale the very specific interactions of a small region, e.g. an enhancer of <500bp, then one should choose to use a restriction enzyme that cuts more frequently (e.g. 4bp) and a method that does not measure the entire genome, but instead focuses on exploring only a subset of the genome (i.e. 3C/4C/5C).

In Hi-C the maximum resolution of a dataset is determined by several factors, first and foremost is the sequencing depth. Given increasing amounts of reads, one will cover more of the interaction space and thus improve the resolution.

Library complexity is another factor. Library complexity is defined as the total number of unique interactions that exist in the Hi-C library. A library with a low complexity level (low number of unique interactions) will saturate quickly with increasing sequencing depth e.g. less and less information will be gained from additional sequencing. The saturation curve can be estimated from a dataset by plotting the cumulative number of unique interactions seen versus read depth.

In our experience, given an adequately complex Hi-C dataset for the human genome and roughly 100 million mapped / valid junction reads, one could expect to achieve close to a 40kb data resolution. Data below 40kb may be usable, though it will suffer from a high level of noise.

Computational considerations

Hi-C data produced by deep sequencing is no different than other genome-wide deep sequencing datasets. The data starts out as genomic reads in the traditional FASTQ file format (containing a DNA read string and a phred quality (QV) score string). Hi-C libraries are traditionally sequenced using paired-end technology, where a single read is produced from each end of the molecule. However Hi-C ligation products can also be sequencing using single end reads, assuming reads are sufficiently long to cover both parts of the hybrid molecule and are handled appropriately during the mapping steps.

Processing Hi-C data mainly has requirements in terms of storage and computing power. The data storage requirements for Hi-C datasets are almost solely driven by the sequencing depth needed and the size of the raw FASTQ files. The processed Hi-C data will normally be order(s) of magnitude smaller than the size of the FASTQ files. It is easy to parallelize the steps needed to map the reads to the genome, and thus achieve a significant speedup in the Hi-C processing steps. The necessary Hi-C-specific filtering and processing steps are independent and can therefore also be parallelized.

The average Hi-C datasets produces roughly 100GB in FASTQ files (100-200 million reads), and 50GB in processed data files. The fastQ files take up the bulk of the size. All files can be compressed to save on file size.

Hi-C workflow

Here we describe the major steps needed to process a Hi-C dataset (**Figure 4.1**):

1. Read Mapping
2. Fragment Assignment
3. Fragment Filtering
4. Binning
5. Bin Level Filtering

6. Balancing

Read Mapping

Reads are aligned using standard read alignment software (i.e. bowtie [94]) to the genome of interest. Any aligner can be used for mapping Hi-C reads - the goal is to simply find a unique alignment for each read. Hi-C data is no different than other high-throughput deep sequencing experiments in terms of the mapping logic required. Even though Hi-C data is traditionally sequenced using paired-end reads, the reads are not mapped using the paired-end mode of most aligners. The paired-end mode for most aligners assumes that the ends of a single continuous genomic fragment are being sequenced, and the distance between these two ends is known (following the shearing size distribution). Since the insert size of the Hi-C ligation product can be anywhere from 1bp to hundreds of megabases (in terms of linear genome distance), it is difficult to use most paired-end alignment modes. One straightforward solution is to simply map each side of the paired end read separately/independently using a traditional alignment procedure.

Read Mapping – Iterative Mapping Strategy

Following the Hi-C method, ligation junctions of varying sizes are created (**Figure 4.2 a**). The molecules are then sheared down to the desired size range (normally ~300bp +/- 100bp). Hi-C data is traditionally sequenced using the paired end sequencing approach. Since ‘C’ interactions are simply chimeric ligation products, of two distinct genomic fragments joined at the middle, it makes most sense to sequence the ends of the molecule (to identify the two pairs in the ligation product). However, one could also read the molecule in its entirety and then computationally separate/identify the two distinct genomic fragments (similar to how RNA splicing is processed).

Searching for the actual junction is possible, but the junction site is not guaranteed to fall within the paired end reads. For example, given a 300bp Hi-C ligation product where the junction site is located at position 150 (in the center) of the molecule, if one were to perform a traditional 50 base-pair paired end sequencing, only the 50 bases on each end would be sequenced. No information would be known regarding the 200 internal bases of this molecule. So it would be uninformative to first search for the junction site and then split the reads into two, since the junction site does not exist in the sequencing data. Instead we favor an iterative mapping approach to solve this problem [62] (**Figure 4.2 b**). This approach does not need to explicitly detect the junction site to uniquely map the two sequences in the Hi-C ligation product. The idea is to attempt to map as short a sequence as possible before the sequence reaches the junction site. Reads are first truncated to 25bp starting at the 5’ end and mapped to the genome. Reads that do not uniquely map the genome are extended by an additional 5bp and then re-mapped. This process is repeated until either all reads uniquely map or until the read is extended to its entirety. Only paired end reads in which each side can be uniquely aligned are kept. All other paired end reads are discarded.

Fragment assignment

For each mapped read, the genomic alignment location is assigned to one of the restriction fragments. The mapped read is assigned to a single restriction fragment according to its 5' mapped position. Mapped read positions should fall close to a restriction site, and no further than the maximal molecule length away. Reads that align more than the maximal molecule length away from the closest restriction enzyme are the result of either non-canonical enzyme activity or random physical breakage of the chromatin. It has been shown that these reads produce informative Hi-C interactions, and thus are not discriminated against [62]. Once each read has been assigned to a restriction fragment, filtering must be applied to discard any technical noise in the dataset.

Fragment-level filtering

After assigning each of the paired-end reads to single fragments, it is necessary to perform some basic filtering (**Figure 4.3**). The following scenarios are possible:

1. The read pair falls within the same restriction fragment.
2. The read pair falls within separate restriction fragments.

If the paired reads map to the same fragment, it can represent either an unligated fragment ("dangling end") or a ligated, circularized fragment ("self-circle").

Each of these two cases is considered non-informative, and should therefore be removed.

After removing same-fragment pairs, the remaining pairs are filtered to remove any redundant (identical) PCR artifacts. PCR duplicates can be detected by either sharing the exact same paired-end sequence, or by sharing the exact same 5’ alignment positions of the pair.

Binning

The maximal resolution of a Hi-C dataset is determined by the restriction enzyme used. Normally, a Hi-C dataset is not sequenced deep enough to support this maximal data resolution, as it is not yet cost-effective to obtain a sufficient number of reads. Instead, the data can be binned into various fixed genomic interval sizes, to aggregate data and smooth out noise. Hi-C restriction fragments are assigned to bins by their midpoint coordinate. Binning the Hi-C data reduces the complexity and number of possible genome wide interactions which in turn increases the signal to noise ratio. Data is typically binned into sizes ranging from 40kb to 1MB. All bin-bin interactions are simply aggregated by taking the sum, though one could use other methods to aggregate the signal. A single Hi-C dataset can be binned into multiple bin sizes, as each bin size can be used for different analysis goals. Following the binning, the data can be stored in a fixed-size symmetrical matrix format, though this file format may not be optimal for storing large Hi-C datasets since the number of the matrix entries can be much larger than the number of reads.

Bin-level filtering

Prior to matrix balancing, it is necessary to remove any bins from the dataset that have either very noisy or too low of signal. These bins normally are found in genomic spans with low mappability or high repeat content, such as around telomeres and centromeres. Since these bins suffer from such a high noise level, it is useful to remove them rather than attempting to correct them for technical biases (see below). Various methods can be used to detect these bin outliers. Current methods detect rows/columns with low signal by looking at their sum compared to the sum of all rows/columns. Outliers can be detected by percentile cutoff (e.g. removing the bottom 1% of rows/columns), or by using the variance as a measure of noise. Similarly, outlier pairwise interactions can be detected by a percentile-based filter (such as removing the top 0.5% of data points). In some instances, a single bin-bin point interaction can have a level of reads orders of magnitude higher than one would expect. The outliers can be the result of a strong PCR bias and it is useful to remove them rather than attempting to correct the signal.

Balancing

Hi-C data can contain many different biases, some of known origin and others from an unknown origin. While it can be possible to correct each bias explicitly [95], [96], it can be quite difficult to know each and every bias. We therefore favor an implicit bias correction approach, which we refer to as *balancing* (elsewhere known as iterative correction [62]). The balancing procedure is based on the Sinkhorn-Knopp balancing algorithm [97]. This procedure attempts to balance the matrix by equalizing the sum of every row/column in the matrix. The procedure is based on the assumption that, since we are interrogating the entire interaction space, every fragment/bin should be observed approximately the same number of times in the experiment (interpreted as the sum of the genome-wide row/column in the interaction matrix). The algorithm iteratively alternates between two steps until convergence. First, each row is divided by its mean. Then, each column is divided by its mean. This process is guaranteed to converge. Both explicit bias correction and Sinkhorn-Knopp balancing yield comparable results [62].

Analysis and interpretation of Hi-C data

Following the mapping, filtering and bias-correction of the Hi-C data, we are left with a binned, genome-wide interaction matrix, where each entry reflects an interaction frequency between two genomic loci. The measured interaction frequencies are unscaled, in the sense that they cannot be directly translated into an actual fraction of cells. Extraction of relevant biological knowledge from this interaction matrix is one of the major challenges of Hi-C data analysis. This includes differentiating biological signal from noise, identification of interaction patterns and interpretation of these patterns.

There are a number of factors that complicate this analysis. First, we have to consider the fact that we are measuring interaction frequencies over a population of cells (**Figure 4.4**). This is critical in terms of data interpretation since, when we consider an interaction pattern consisting of multiple pairs of loci, we cannot tell whether such interactions will co-occur simultaneously in a single cell. Accordingly, observing a “smooth” interaction matrix that shows little position-specific structure does not rule out the existence of structure in the underlying genomes - it simply means that if such structures exist, they are not consistent between cells. Second, most of the patterns are given procedural definitions rather than explicit definitions. In other words, rather than formally define what a specific interaction pattern looks like and search for it in the interaction matrix, interaction patterns are defined as the output of some method. As a result, it is difficult to evaluate the validity of a method or compare methods aimed at identifying the same type of interaction pattern. Third, different types of interaction patterns co-exist and overlap each other. Given that in many cases we lack an explicit definition of these patterns, as mentioned above, it can be difficult to disentangle different types of interaction patterns. In practice, many of the current approaches analyze each interaction pattern separately under a simplifying assumption of independence, i.e. by assuming that either the effect of other patterns is negligible or that the other patterns can be normalized out of the data. Finally, we cannot assume *ergodicity* of interaction frequencies. In other words, frequencies in the cell population cannot necessarily be interpreted as frequencies in time (**Figure 4.5**). For example, an interaction which occurs in a small fraction of cells and thus produces weak signal in Hi-C cannot be concluded to necessarily be an unstable interaction. Alternatively, any assumption of ergodicity should be made consciously.

Several different types of interaction patterns have been observed in interaction maps. These patterns vary in scale, from genome-wide patterns to point interactions between loci, and in their ubiquity, from constant between different species to condition-specific. Due to the speculative nature of biological interpretation of interaction patterns and the aforementioned complications, it is often useful to separate the process of pattern identification from the process of pattern interpretation. Here we focus mostly on pattern identification, but also briefly discuss common interpretations of each pattern.

We focus on 5 types of patterns typically observed in mammalian genomes. For each pattern, we discuss how it is defined, how it looks visually in the interaction matrix, how it can be identified computationally and how it can be interpreted.

1. Cis/trans interaction ratio
2. Distance-dependent interaction frequency
3. Genomic compartments
4. Topological domains

5. Point interactions

While we outline possible approaches for independent analysis of each type of pattern, there exist alternative approaches for explicitly considering multiple patterns simultaneously [4]. Finally, as with any approach, we advise not to apply the proposed techniques blindly, but rather critically and always evaluate the data visually. Indeed, other interaction patterns, which we do not discuss here, have also been observed including patterns resulting from circular chromosomes and centromere clustering [98]. Such patterns may require careful consideration and the application of specially-tailored methods. Alternatively, methods can be derived given a specific biological question, for example, whether a given set of genes interact more frequently than expected by random.

Following our discussion of individual patterns, we discuss reconstruction of 3d structures from Hi-C data, application of Hi-C data to problems in genome assembly.

Cis/trans interaction ratio

The strongest interaction patterns which are observed in Hi-C maps are genome-level patterns [1]. By genome-level we mean that the patterns are not locus-specific, but instead reflect average genome-wide trends. Two genome-level patterns have consistently been observed in Hi-C data in various species and cell-types.

The first pattern is a higher interaction frequency, on average, of pairs of loci which reside on the same chromosome (i.e. in cis) than loci which reside on different chromosomes (i.e. in trans). In a genome-wide interaction matrix, this pattern appears as square blocks, of high interaction centered along the diagonal and matching individual chromosomes (**Figure 4.6**). The pattern is likely due, at least in part, to a phenomenon known as *chromosome territories*, where chromosomes are physically separated and occupy a distinct volume in the nucleus. Since this pattern is largely constant across cell types and species, it is typically less useful for studying aspects that are specific to the given biological system. However, this fact makes this pattern a useful proxy for evaluating the quality of the data. If noise in the matrix, due to factors such as random background ligation, is expected to affect both cis and trans interactions similarly, a noisier experiment will result in a lower ratio between cis and trans interactions. Thus, it is common to use this simple statistic (i.e. the ratio between the mean cis interaction frequency and the mean trans interaction frequency) to quantify this pattern. Typical values for the cis/trans ratio in high quality experiments are in the range 40-60. While this interaction pattern is typically dominant, the statistic can be affected by other local large-scale patterns such as inter-chromosomal centromeric interactions, so it is advisable to substitute the mean with robust statistics for estimating the global cis and trans interaction frequencies.

Distance-dependent interaction frequency

The second genome-level interaction pattern is a distance-dependent decay of interaction frequency (**Figure 4.7**). In other words, interaction frequency between loci in cis decreases, on average, as their genomic distance increases. In the interaction matrix this pattern appears as a gradual decrease of interaction frequency the further one moves away from the diagonal. This pattern may be due to random movement of the chromosome, following the intuition that loci which are nearby in the genome will interact frequently if they move randomly in 3D space. The theory underlying this type of intuition is well established in the field of polymer physics [99], [100]. Many basic models of general polymers in polymer physics predict a distance-dependent decay of interaction frequency, where the simplest model, known as the *ideal chain*, is equivalent to a random walk in 3d space. A central aspect of all these models is that they characterize polymers as distributions, rather than single structures, inherently accounting for randomness and structural variability. Specific models are thus characterized by statistical properties such as the mean interaction probability for a pair of loci separated by a given distance. Thus, by estimating the distance-dependent interaction frequency from our data, which is derived from a population of cells, we can ask which polymer models are consistent with the observed pattern. For example, the distance-dependent interaction frequency of an ideal chain is expected to have the form of the power-law decay $p_{interaction}(x,y) = Z^{*}dist(x,y)^{-1.5}$. In fact, this specific decay matches the distance-dependent interaction frequency observed in yeast.

Analysis of distance-dependent interaction frequency is typically performed using one of two methods. The first method is discrete binning. With this method, we bin all interaction frequencies according to their genomic distance, and calculate the average of each bin. The second method is interpolation. With this method, we fit some continuous function to the data and use this function to represent it. In some cases, binning may be used as a preliminary step for fitting a continuous function. Due to the fact that many polymer models predict a power-law decay, it is helpful to plot the resulting decay function on a log-log plot so that power-law decays will appear linear. However, it is important to perform the calculation of the decay function on the initial data, not on the log-transformed data due to theoretical considerations [101]. For related reasons, it is advisable to use logarithmic-sized bins if using the binning scheme, e.g. such that each bin will be double the size of the previous bin.

While it is convenient if the observed distance-dependent interaction frequency matches what is expected by a simple polymer model, this is often not the case. However, it can still be useful to examine a more complicated decay function, since it could provide some insight, such as different regimes of decay at different genomic length scales (**Figure 4.6**). This can, in turn, promote the development of more complex polymer models that reproduce the observed pattern. It is important, though, to realize the limitations of this type of analysis. Hi-C data incorporates several different types of patterns, some of which are locus-specific and will thus not be reproduced by these types of models which do not include locus-specific constraints. Additionally, some of these local patterns could affect the shape of the decay function. Finally, even if a Hi-C map contains no locus-specific interaction patterns and is consistent with some polymer model, it is not sufficient by itself to conclude that the model is correct, since other polymer models could potentially produce the same decay function. Ultimately, what matters is how useful such a model is for gaining biological insight and whether it can produce testable hypotheses.

Genomic compartments

Next, we consider interaction patterns which are position-specific. The largest-scale position-specific interaction pattern is known as *genomic compartments* [1]. This interaction pattern appears on the interaction matrix as a “checker-board”-like pattern consisting of alternating blocks, ~1-10 mb in size, of high and low interaction frequency (**Figure 4.8**). This interaction pattern can be explained by a simple underlying phenomenon where chromosomes are composed of two types of genomic regions that alternate along the length of chromosomes and where the interaction frequencies between two regions of the same type tend to be higher than interaction frequencies between regions of different types. We refer to these two types as A and B compartments [1].

While this interaction pattern is intuitive, its current definition is procedural - the genomic compartments are usually considered to be given by the first principal component of the interaction matrix. The reasoning for this definition is as follows. Imagine each bin in the 1d genome is assigned a number $c(x)$ quantifying whether it belongs to A (positive value) or B (negative value). Now, we decide that the interaction frequency between two loci x,y is $c(x)c(y)$. Note that this formulation is sufficient to reproduce a checkerboard pattern: when the types of x,y are the same, their signs will be the same and will yield a positive interaction frequency, and when their types are different their signs will be different, resulting in a negative interaction frequency. Thus, given an interaction matrix, we are given all interaction frequencies and want to find the compartment $c(x)$ of each position. It turns out that the first principal component found by a Principal Component Analysis can be viewed as finding the optimal values of $c(x)$ such that difference between the observed interaction frequencies and $c(x)c(y)$ is minimal (mean squared error is minimized). Thus, if the compartment pattern is sufficiently strong, this procedure should find it. Alternatively, one could use any standard clustering approach, such as k-means, to cluster the rows of the interaction matrix into two clusters.

Genomic compartments have been found to be correlated with chromatin state, including DNA accessibility, gene density, replication timing, GC content and histone marks [1]. Thus, A-type compartments are interpreted as euchromatic regions while B compartments as heterochromatic regions. Genomic compartments have been found to have high-plasticity, such that they change in different cell-types and biological condition, matching large scale changes in gene activity. Individual compartment blocks tend to be on the order of 1-10 Mb in length, and are thus easy to extract even in experiments with very low sampling. Finally, it is important that while compartment signal is strong and easy to observe in large bins, the interaction frequencies at individual positions that have the same compartment type are quite low. Thus, given that Hi-C measures a population average, it is likely that this pattern reflects a general, highly stochastic, tendency of compartments to interact, rather than a set of deterministic interactions specified by individual loci.

Topological domains

While genomic compartments are useful for understanding general organization principles of the genome, many biological processes occur at a smaller scale. Specifically, enhancer-promoter interactions that underlie gene regulation in metazoans often take place at sub-Mb distances. Recently, 3C-based techniques have revealed the existence of sub-Mb structures that are referred to as *topologically associating domains* or *TADs* [2]–[5]. TADs are contiguous regions in which loci tend to interact much more with each other than with loci outside the region. In the interaction matrix TADs appear as square blocks of elevated interaction frequency centered on the diagonal (**Figure 4.9**). However, the definition of TADs is complicated by the fact that actual interaction patterns are complex and contain multiple hierarchies of overlapping block-like structures, as assessed by visual inspection of chromatin interaction maps. Nonetheless, given some definition of TADs, these domains have been shown to be associated with gene-regulatory features and it is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with enhancers [7], [102], [103]. In addition, TAD-like structures of various sizes have been observed in species ranging from mammals to bacteria [2]–[5], [104].

As hinted above, TADs are also defined procedurally (i.e. as the output of a given method). We outline two such methods for identifying TADs. Both methods take the following approach: First, they summarize the TAD signal using some statistic, such that TAD signal is converted into a 1d profile along the genome. Then, they use the 1d profile to identify potential boundaries between TADs and produce a set of discrete non-overlapping TADs. It is important to note that while these methods provide a useful heuristic for quantifying some of the TAD-level patterns, they do not provide an actual predictive model, or point to physical processes that drive domain formation. Without an explicit definition of TADs, these methods are difficult to compare and evaluate critically. However, it is clear that a discrete set of non-overlapping regions is only a first approximation and likely a significant oversimplification of the interaction patterns which are observed in the data.

An approach by Dixon et al. [2] uses the following statistic: for each bin, we calculate the difference between its average upstream interactions and its average downstream interactions (within some genomic range). This difference is then transformed into a chi-squared statistic and the resulting value is referred to as the directionality index. At the boundaries of TADs, we expect to see a sharp change in the directionality index. Boundaries are then associated with each other using a Hidden Markov Model. Alternatively, others have simply used the ratio between average upstream and average downstream interactions [72].

An alternative approach is to calculate for each bin the average of interaction frequencies crossing over it (within some genomic range). This is referred to as the insulation score. We expect that this value will be lower at TAD boundaries. Then one can use standard techniques to find local minima and use those as boundaries, and define regions between consecutive boundaries to be TADs.

The block-like structure of TADs clearly indicates elevated interaction frequency within a TAD. However, given that we measure a population average and the observed intricate hierarchies of such structures, interpretation of TADs is not straight-forward. It has been proposed that TAD-like structure may be driven at least in part by looping interactions between loci located within them [105] or by supercoiled plectonemes [104], [106]. Additionally, some genomic features such as CTCF and cohesin binding have been shown to be enriched at TAD boundaries [2], [11]. It remains unclear what physical structures TADs exactly represent and how they are specified in the genome.

Point interactions

The final type of interaction pattern we discuss is point interactions. While TADs may be relevant for constraining promoter-enhancer interactions, the actual regulatory interactions are probably of much smaller scale. Ultimately, protein-mediated interactions of two localized genomic elements, e.g. enhancers and promoters, which are typically up to a kb in length, can activate the expression of a gene. Given sufficient resolution, we expect such point interactions to appear as a local enrichment in contact probability.

As with some of the other interaction patterns, current approaches for finding point interactions do not provide an explicit model of what a point interaction should look like. Instead, these approaches try to find outliers which show higher interaction frequency than expected, where the background model may consist of other previously mentioned interaction patterns [13], [14], [107]. Typically, the background model consists only of the strongest signal, namely the distance-decay function, but other patterns such as TADs can be incorporated as well. Given a background model, we can then test the significance of individual pairwise interactions. The resulting set of significant high outliers would then need to be corrected for multiple testing. It is important to note that without an explicit model of point interactions, it may be difficult to distinguish between real point interactions and experimental noise. Thus, it may be helpful to provide additional evidence including analysis of biological replicates, and from alternative methods as to the validity of such interactions (e.g. by showing enrichment for enhancers and promoters).

While the biological interpretation of point interactions seems to be straightforward, it is important to consider what such methods find. If we look for interactions that have a higher interaction frequency then what is expected given their distance, we are not evaluating their absolute interaction frequency. For example, consider two loci which are nearby in the genomic sequence, and are thus expected to interact very frequently. Such interactions may be functional and biologically important, but they may not be have a much higher interaction frequency than expected by distance, and thus may not be found to be point interactions. Similarly, the expected interaction frequency for loci that are separated by large genomic distances is very low. As a result even a small increase in their interaction frequency can make their interaction statistically significant even though their absolute interaction frequency is still low, implying it occurs in only few cells. Thus, careful biological evaluation is always in order for interpreting any statistical approach to identifying point interactions.

Structure reconstruction and polymer modeling

Given that Hi-C measures an aspect of the 3D structure of the genome, it is natural to ask whether we can use Hi-C data to infer the underlying 3D structures. In fact, Hi-C maps are reminiscent of 2D NMR spectrum maps used to infer 3D protein structure with great accuracy. However it is important to realize that there are important differences between protein structure and genome structure that dramatically complicate inference of the genome structure. First, inference of protein structures incorporates knowledge of protein physics and the underlying sequence. There are strong constraints on what conformations are physically possible and there is a relatively good understanding of the physics of various intramolecular interactions. On the other hand, knowledge of chromatin physics is limited and chromatin structure is much less constrained than protein structure. Second, chromatin fibers are much longer than proteins, in the sense the length of a chromosome may be as much as 10^5-10^6 times larger than the smallest structures of interest in the chromosome. Thirdly and most importantly, chromatin structure is much more variable than protein structure, yet we observe only the population average. In fact, it is debatable whether it is even useful to infer a single average “consensus structure”, given the highly-stochastic nature of the genome structure.

With these limitations in mind, we consider 2 general approaches to structure inference from Hi-C data:

1) Consensus structure. These methods essentially ignore the fact that structure is variable across the population and try to find a 3D structure that is as consistent as possible with the 3d interaction matrix [98], [108]–[112]. Most methods follow some form of multidimensional scaling, formalized as seeking 3D coordinates for all loci such that their pairwise distances are as consistent as possible with the observed interaction frequencies. These approaches require making assumptions on how interaction frequency of loci is related to their spatial distance.
2) Ensemble of structures. These methods typically try to create a set of structures such that the either the average distances or the contact

probability between every two loci are consistent with the observed interaction frequencies [105], [109], [113]. While this approach resembles the actual biology more closely, allowing for multiple structures makes the problem even less constrained. In other words, there are likely many different ensembles of structures that could explain a given interaction Hi-C matrix. Additionally, such an ensemble of structures may be difficult to interpret.

Once again, the utility of such models will be measured by whether they can give biological insight and make useful predictions.

Genome rearrangements and genome assembly

Typically, Hi-C data is mapped to a known high-quality genome sequence and is used to answer questions regarding the 3D organization of genomes. However, it has recently been shown in a number of studies that Hi-C data can be useful to learn about the 1D arrangement of the genome sequence and thus solve a number of outstanding problems in the field of genome assembly [87], [114]–[117]. Ironically, the recent major advancement of DNA-sequencing technologies has caused a decrease in the quality of genome assemblies due to the use of short reads. Thus, genomes assembled from short-read data consist of huge sets of contigs (~100000 contigs for Gb-scale genomes), which cannot be grouped and ordered with this type of data. However, by mapping Hi-C data to a set of contigs, we gain interaction frequency data over very large genomic distances. We can then exploit a number of universal principles relating 1d structure to 3d structure in order to associate and order contigs in linear genome. We refer to this set of approaches as *DNA triangulation*, due to their use of multiple lines of long-range evidence (i.e. Hi-C interactions) to resolve genomic positions.

We list these principles and how they can be used:

1. Interactions of loci located in different nuclei are less frequent than those in the same nucleus. This principle seems obvious, but has important implications. In microbiome studies, which analyze large mixed populations of different species, high-throughput sequencing typically yields a large set of contigs, yet it is difficult to establish which contigs belong to the same genome. Using Hi-C data, we can determine that if two contigs interact frequently in 3D they are likely to belong to the same genome with high probability [116], [117].
2. Interactions of loci located on different chromosomes are less frequent than those in the same chromosome. As discussed above, this pattern is both strong and ubiquitous. When performing de novo genome scaffolding, we can thus use Hi-C data to determine that contigs that interact frequently are likely to belong to the same chromosome [114], [115]. Additionally, since homologous chromosomes are also separated into distinct territories, this principle can be used to perform haplotype phasing. A Hi-C paired-end read

that maps to one SNP on each side is much more likely to come from the same chromosome than from the homologous chromosome [87].

3. Interactions of loci located far from each other along a chromosome are less frequent than loci that are near each other. Using Hi-C data, we can arrange contigs which belong to the same chromosome such that strongly interaction contigs are positioned next to each other [114], [115].

While the goal of these techniques is not necessarily to learn about the 3D structure of the genome, it is clear that they are widely useful. When indeed such techniques will be adopted, they may offer large amounts of Hi-C data as an important side benefit. However, if one's goal is to use Hi-C for *DNA triangulation*, it could be useful to carefully consider some of the experimental design and analysis choices. For example, locus-specific interaction patterns are important for studying the biology of genome structure but could pose problems for *DNA triangulation*. Pooling different cell types, computationally or experimentally, could average out some cell-specific interaction patterns.

Acknowledgements

We thank all members of the Dekker lab and the Mirny lab, especially Maxim Imakaev and Geoffrey Fudenberg, for contributing methods and thoughts outlined in this article. Research in the Dekker lab is funded by the National Human Genome Research Institute (HG003143, HG007010 to JD) and the Human Frontier Science Program (to JD and LT000706/2012 to NK).

Figures

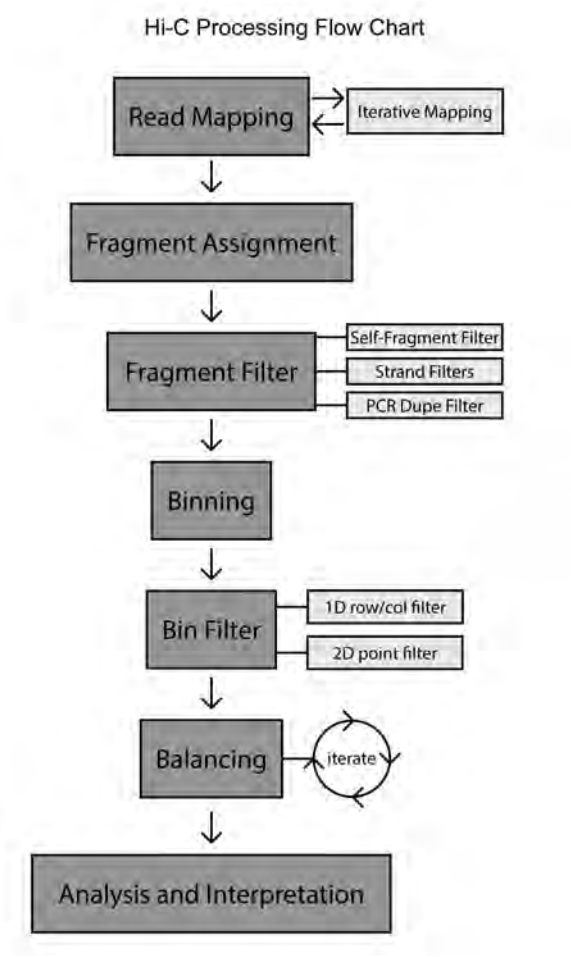
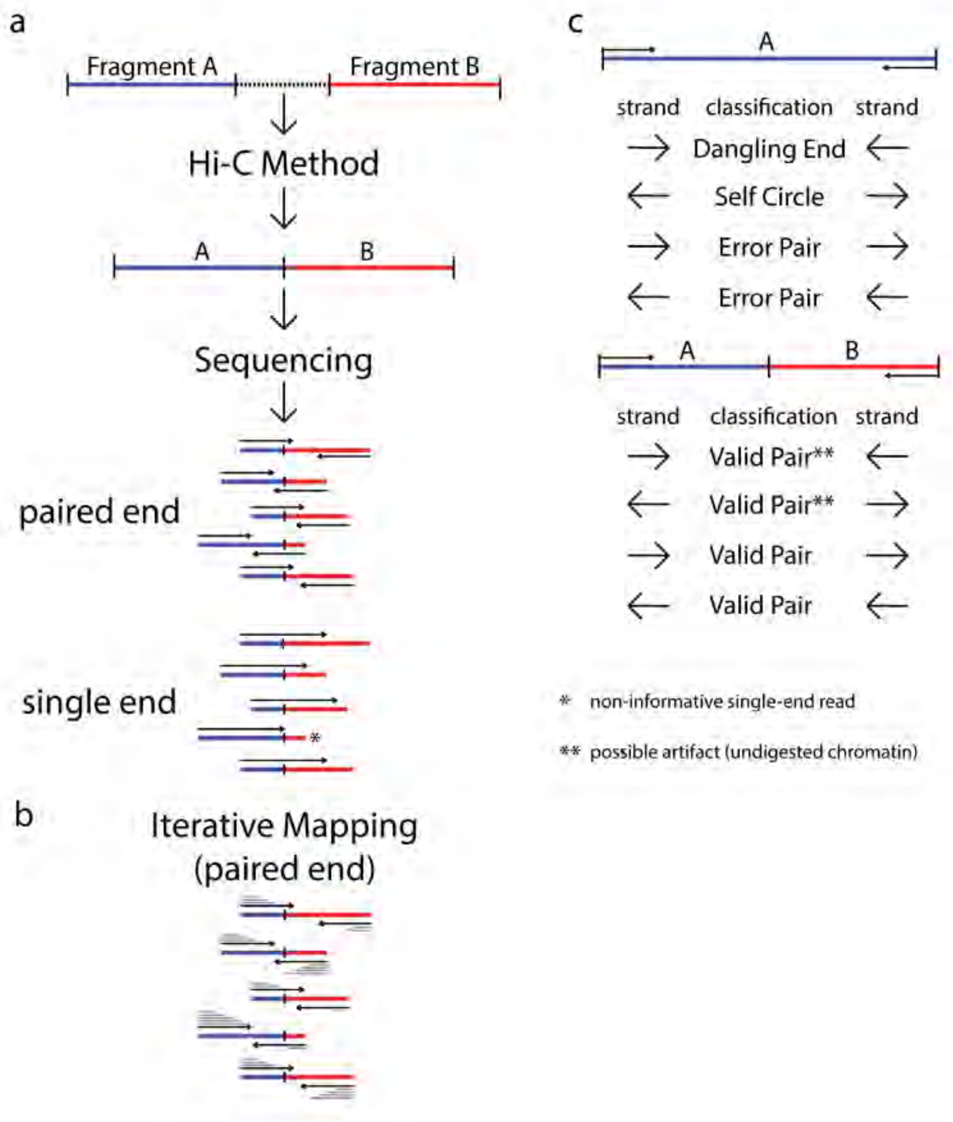


Figure 4.1 | Flow chart for processing Hi-C Data.
Reads are first mapped using the iterative mapping approach for paired end reads. Only paired end reads where both ends map uniquely are kept, all others are discarded. Mapped reads are then assigned to a restriction fragment, and fragment-fragment interactions are assembled. Fragment level filtering is applied. Un-ligated fragments and self-ligated fragments are removed. Optional strand-specific filters are applied. PCR duplicates are removed. Data is then binned. Bin-level filtering is then applied. Outlier bin-bin point interactions (2D) are removed. Outlier bins (1D row/cols) are removed.



* non-informative single-end read

** possible artifact (undigested chromatin)

Figure 4.2 | Mapping and filtering

a, Following the Hi-C method, fragments are ligated. Hi-C junctions are then sheared and sequenced. Hi-C junctions can be sequenced by using either paired-end sequencing or single-end sequencing. * - Here a Hi-C junction is incapable of being sequenced by a 100bp single end run, as the read does not extend past the junction into the second fragment. Should the read length increase, then the sequenced read would cross the junction. b, Iterative mapping approach for aligning paired-end Hi-C reads. In gray, from top to bottom above/below each read, the mapping iterations are shown as the read is extended and re-mapped. Iterative mapping concludes when either the read is uniquely aligned, or the maximal read length is reached. The number of iterations is a factor of mappability and the location of the junction. c, After mapping, the paired reads can either map to a single fragment, or to different fragments. Reads mapping to a single fragment are considered uninformative. Self-ligations and un-ligated fragments are classified by the read strand. Inward pointing reads are considered un-ligated fragments (“dangling ends”). Outward pointing reads are classified as self-ligated fragments (“self-circles”) as they form circular products. Same-strand reads are classified as “error pairs” as these products are a result of either a mis-mapping, random break, or an incorrect genome assembly. Reads mapping to different fragments are used to assemble the Hi-C dataset. All strand combinations are possible and are expected to be observed in equal proportions (25% per combination). However, inward and outward pairs could be the result of un-digested restriction sites, and then processed as either self-ligated or un-ligated products. Imbalance in the relative proportions of the strand combinations, could suggest the need for additional filtering.

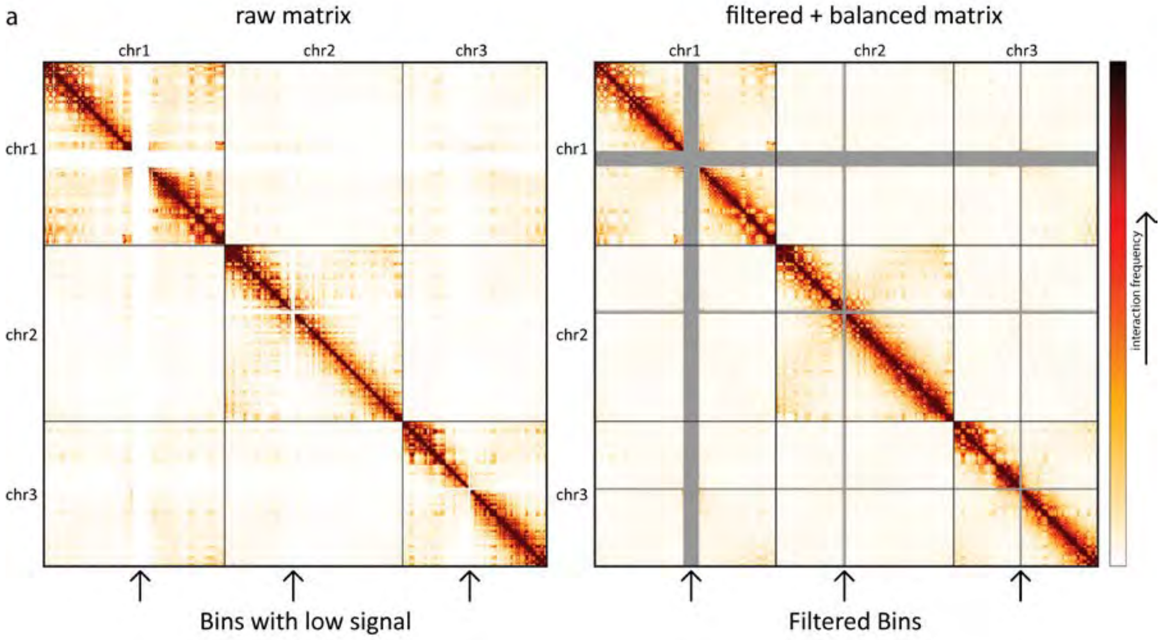


Figure 4.3 | Hi-C interaction matrix for 3 chromosomes.
On the left, raw Hi-C data. On the right, filtered and balanced Hi-C data. The arrows below the heatmaps mark bins (rows/cols) that are filtered. Following the balancing procedure, the sum of each row/col is equal. This results in an overall smoother heatmap.

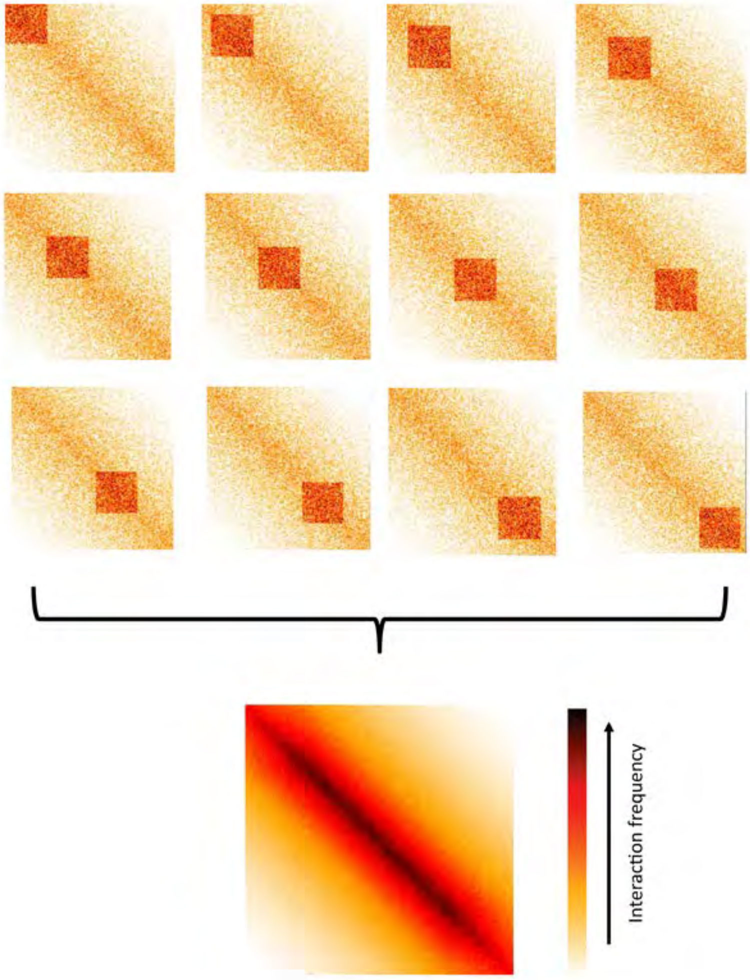


Figure 4.4 | Averaging effects in Hi-C data.
In this toy example, a square interaction pattern is apparent in the top interaction matrices representing subpopulations, yet its location varies. The final Hi-C interaction matrix, which consists of the average of all subpopulations, does not show the square interaction pattern, and shows a pattern that is not present in individual subpopulations.

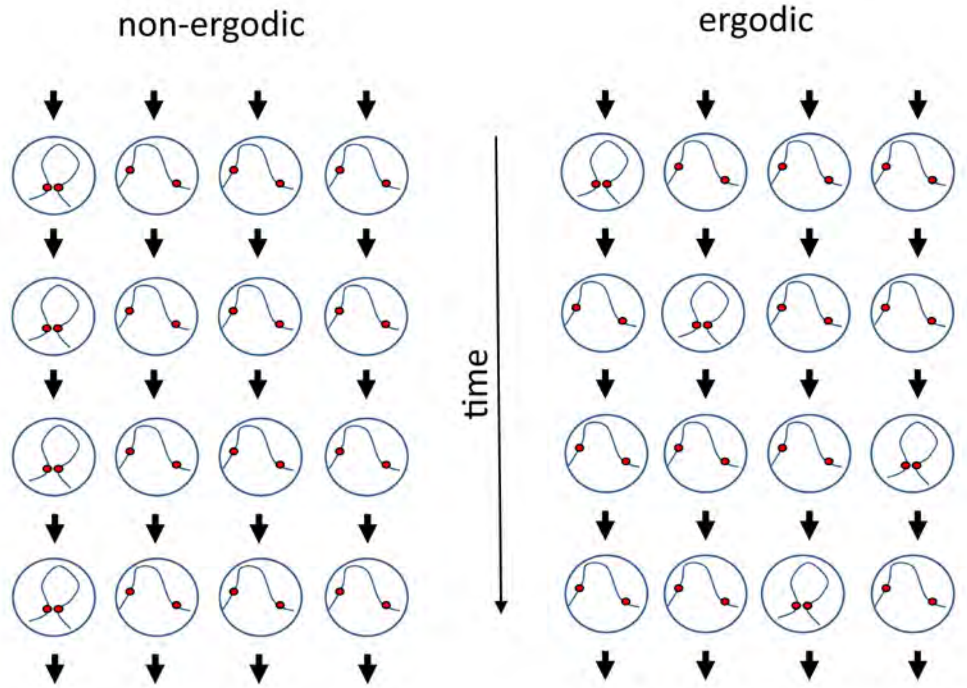


Figure 4.5 | Ergodicity in Hi-C.
This toy example follows, over time, the interaction of two loci in a population of 4 cells. Each row represents a time point and each column represents a cell. In the non-ergodic population (left), the interaction is maintained in the same cell over all time points. In the ergodic population (right), the interaction appears in different cells, such that its frequency in time is equal to its frequency in the population (both are 0.25). In Hi-C, which measures a single time point (i.e. a row) in a population of cells, the ergodic and non-ergodic cases are indistinguishable.

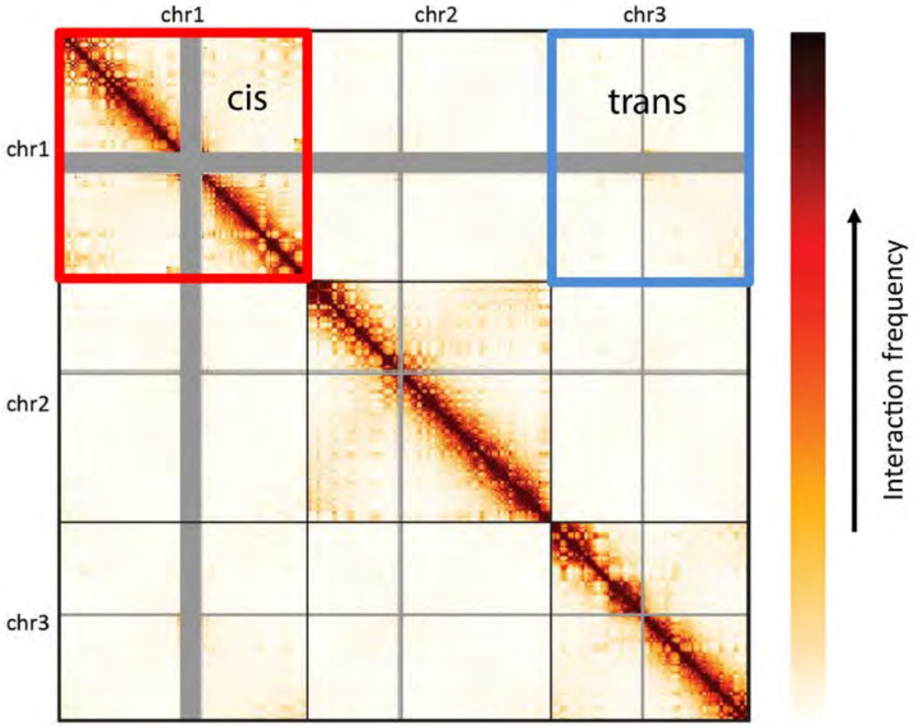


Figure 4.6 | CIS / TRANS ratio.
A Hi-C interaction matrix (shown on 3 chromosomes for simplicity). Sample cis (intra-chromosome) and trans (inter-chromosome) regions are highlighted.

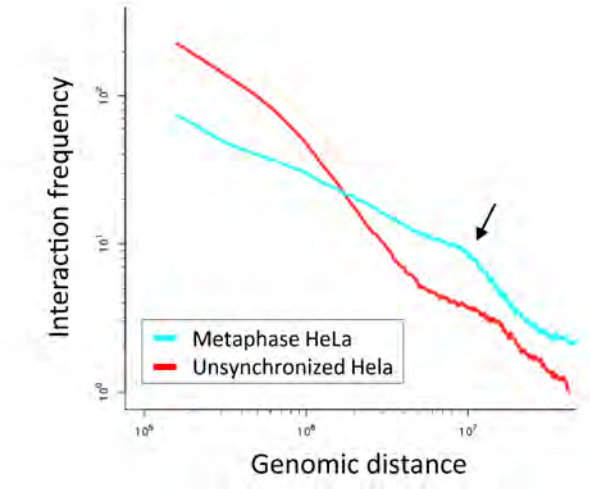


Figure 4.7 | Distance-dependent interaction frequency.
Shown are distance-dependent interaction frequency curves for metaphase and unsynchronized HeLa Hi-C from [35]. Note the slope change in the metaphase data which occurs at 10 Mb (indicated by the black arrow). Thus, loci separated by fewer than 10 Mb interact frequently, whereas loci separated by more than 10 Mb rarely interact. This information has been incorporated into polymer models of mitotic chromosomes.

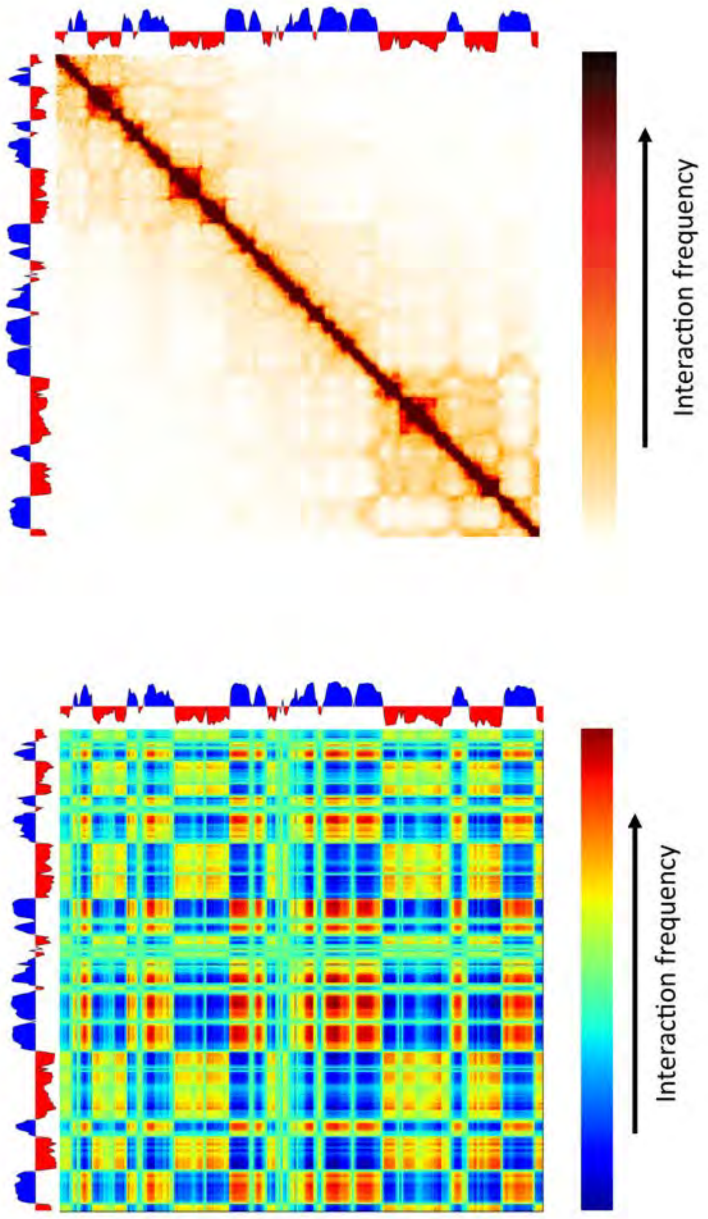


Figure 4.8 | Genomic compartments.
Top: Hi-C interaction matrix (shown on 3 chromosomes for simplicity) along with the calculated compartment value (first principal component; shown as alternating red-blue track next to the matrix). Below: outer product of the first

principal component with itself yields a rank-1 reconstruction of the interaction matrix.

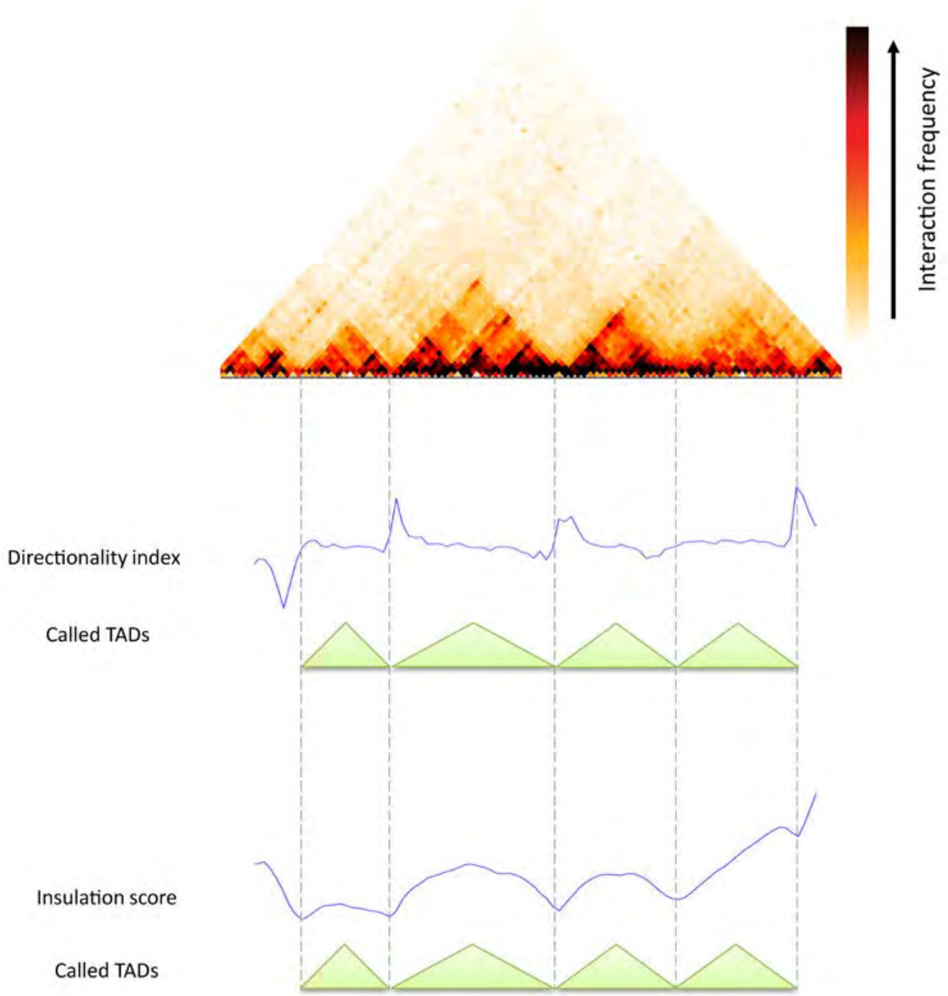


Figure 4.9 | Topologically associating domains (TADs).
A 45-degree rotated interaction matrix shows TAD patterns in a 4 Mb region. Below, the directionality index and insulation score are shown together with the called non-overlapping set of TADs. Data was taken from Dixon et al. [2].

CHAPTER V: The long-range interaction landscape of gene promoters

Preface

This research chapter encompassed work published in Nature by Amartya Sanyal, Bryan Lajoie, Gaurav Jain, Job Dekker. The publication is entitled, “The long-range interaction landscape of gene promoters”, Nature, vol. 489, no. 7414, pp. 109–13, Sep. 2012.

Introduction

The vast non-coding portion of the human genome is awash in functional elements and disease-causing regulatory variants. The relationships between the genomic positions and order of regulatory elements and their impact on distal target genes remain unknown. Genes and distal elements can come together through looping to form higher order chromatin structures involved in gene regulation [24]. Mapping of these structures allows placing loci in three-dimensional context to reveal long-range and possibly functional relationships. Here we have applied chromosome conformation capture carbon copy, 5C [23], to comprehensively interrogate interactions between transcription start sites (TSSs) and distal elements in 1% of the human genome representing the ENCODE Pilot regions [25]. 5C maps were generated for GM12878, K562, HeLa-S3 and H1-hES cells and results were integrated with other data from the ENCODE consortium (NCP0004) [26]. We discovered >1,000 long-range interactions in each cell line. In differentiated cells, interactions occurred preferentially between active promoters and distal elements that are enriched for chromatin features that are hallmarks of regulatory elements. In contrast, in H1-hES cells looping was not correlated with gene expression and often involved elements resembling poised enhancers. Looping interactions are related to the relative genomic positions of the elements and display directionality. First, TSSs interact more frequently with enhancer-like and CTCF-bound elements located upstream than downstream, with a pronounced preference for elements located 100-200 Kb upstream. Second, only ~8% of interactions are with the nearest gene, and some skip as many as 20 genes. Third, in contrast to current insulator models, CTCF-bound elements do not block long-range interactions, implying that many of these sites do not demarcate physically insulated gene domains. Finally, interactions form complex long-range interaction networks. These analyses provide new insights into the links between linear genome sequence, three-dimensional chromatin architecture and gene regulation.

Spatial proximity and specific long-range interactions between genomic elements can be detected using 3C-based methods [20]. Previous studies have been limited to analysis of single loci [20]–[22], [118], to interactions that involve a single protein of interest [119] or to analysis of genome-wide folding of chromosomes at a resolution that cannot detect specific looping interactions between genes and functional elements [1]. To overcome these limitations we had developed 3C-Carbon Copy technology (5C) [23]. 5C is a high-throughput adaption of 3C and employs pools of Reverse and Forward 5C primers to detect long-range interactions between two targeted sets of genomic loci, e.g. promoters and distal gene regulatory elements. By targeting a specific part of the genome 5C facilitates detection of interactions at single restriction fragment resolution.

Results

To start to define principles of long-range gene regulation in the human genome we have employed 5C to systematically map interactions between promoters and distal elements throughout the 44 ENCODE pilot project regions representing 1% (30 Mb, Supplementary Table 1) of the genome in four cell lines (**Figure 5.1 a**). The ENCODE regions, ranging in size from 500 Kb to 1.9 Mb, were selected for comprehensive annotation by the ENCODE pilot project [120]. Here we analyzed interactions between 628 TSS-containing restriction fragments and 4,535 “distal” restriction fragments covering the ENCODE pilot regions (**Figure 5.1 a**; Supplementary Tables 2 and 3, see supplementary methods).

5C libraries were generated for 2 biological replicates of GM12878, K562, HeLa-S3 and H1-hES cells (Supplemental Table 4-6). These cell lines are extensively annotated by the ENCODE consortium [25]. 5C interaction frequencies measured between ENCODE regions located on different chromosomes were used to quantify minor variations in interaction detection efficiencies, due to technical biases related to 5C primer efficiency, restriction fragment length and digestion efficiency. 5C interaction frequencies were then corrected for these biases (Supplementary Methods; Supplementary Data File).

An example of a 5C long-range interaction map representing TSS-distal fragment interactions along and between 14 ENCODE pilot regions (ENm001-ENm014) (**Figure 5.1 b**). 5C detects known general features of spatial chromatin organization. First, interactions within the same ENCODE region are more frequent than those between different ENCODE regions. Within one ENCODE region interaction frequencies are generally higher for pairs of loci located closer together in the linear genome, as is apparent from strong signals along the diagonal of the heatmap (**Figure 5.1 c**). This inverse relationship between genomic distance and interaction frequency is as expected for a flexible chromatin fiber [20], [121]. Second, interactions between ENCODE regions that are located on the same chromosome are more frequent than interactions between regions located on different chromosomes (**arrow in Figure 5.1 b**). This is consistent with 4C and Hi-C analyses [21], [35] and is due to the formation of spatially separated chromosome territories.

5C datasets were analyzed to identify TSS-distal fragment pairs that interact more frequently than expected indicating they are relatively close in space. For each dataset we determined the average relationship between interaction frequency and genomic distance (**solid red line in Figure 5.1 d**). We defined this as the expected interaction frequency. Next we identified interactions that occur significantly more frequent than expected for loci separated by a corresponding genomic distance by transforming 5C signals into a z-score (FDR=1%, Supplementary Methods).

Our analysis correctly identified known interactions between TSSs and their cognate distal regulatory elements, providing validation of the approach. As an example, (**Figure 5.1 d**) shows the 5C interaction profile in K562 cells for a TSS located in the beta-globin locus. We previously found that this TSS, for an intergenic transcript located just downstream of the gamma-globin genes [122], displayed prominent looping interactions with the distal Locus Control Region in K562 cells [23]. Our analysis accurately detected these looping interactions (HS3, 4, 5). We detected additional known long-range interactions with DNAse I hypersensitive sites (DHSs) near distal CTCF-bound elements (3'HS1 and HS-111) [19], [23], [123]. In K562 cells we also detected the known interactions between the alpha-globin genes and three distal regulatory elements including the alpha-globin enhancer HS40, and two CTCF-bound elements (HS46 and HS10), located 40, 46 and 10 Kb upstream of the genes respectively (**Figure 5.2**, [18], [124]). The importance of these distal elements in regulating globin gene expression through looping has been extensively documented. None of these looping interactions in the globin loci were detected in GM12878, HeLa-S3 or H1-hES cells indicating that 5C reliably identifies known cell-type specific functional interactions between TSSs and their distal regulatory elements. Furthermore, our set of significant long-range 5C interactions is correlated with TSS-distal DHS pairs predicted to be functionally connected based on their highly correlated activity across a large panel of cell lines (P < 10-13, one-sided Mann-Whitney test [125]), providing independent validation of their biological significance.

In each cell line we identified large numbers of statistically significant TSS-distal fragment interactions, of which 50-60% were observed in only one of the four cell lines (**Figure 5.3 a**). These data point to intricate cell type specific three-dimensional folding of chromatin. Many previous studies have shown that 3C-based assays detect specific and functional interactions between TSSs and their distal gene regulatory elements [118]. In addition the assay will detect “structural” interactions, e.g. close spatial proximity as a result of other nearby specific looping interactions (bystander interactions) or overall higher order folding of the chromatin fiber. To determine which looping interactions involved distal sites that displayed specific chromatin features associated with functional elements we compared our data with datasets generated by the ENCODE consortium (**Figure 5.3 b**; Supplementary Table 7). We find that looping interactions in all four cell lines are significantly enriched for distal fragments that are bound by CTCF, a protein known to mediate DNA looping [126], contain open chromatin (as determined by FAIRE [127] or DHS mapping, and/or histones with modifications associated with active functional elements (H3K4me1, H3K4me2, H3K4me3). In GM12878, K562 and HeLa-S3 cells interactions are also enriched for H3K9ac and H3K27ac, but are not enriched or significantly depleted for H3K27me3, a mark typically associated with inactive or closed chromatin. Interestingly, the opposite pattern was observed in H1-hES cells where there is no significant enrichment in H3K9ac or H3K27ac, but rather enrichment for H3K27me3. These combinations of marks are associated with different types of regulatory elements, e.g. active enhancers and poised enhancers respectively [128], [129], which suggests that different classes of elements are involved in long-range interactions in H1-hES cells as compared to the three other cell types.

To gain more insights into the types of elements present in the distal looping fragments we made use of genome-wide and cell-line specific segmentation analyses that identified seven distinct chromatin states based on histone modifications, the presence of DHSs and the localization of proteins such as RNA polymerase II and CTCF ([130]; **Figure 5.3 b**). These states are 1) “Enhancer” (E), 2) “Weak Enhancer” (WE), 3) “TSS”, 4) “Predicted Promoter Flanking regions” (PF), 5) “Insulator element” (CTCF), 6) “Predicted Repressed region” (R) and 7) “Predicted Transcribed region” (T)). The ENCODE consortium tested sets of the E elements in enhancer assays and confirmed that >50% display enhancer activity [26]. However, it is important to point out that the segmentation analysis only identifies elements that resemble states associated with enhancers and promoter proximal elements, but do not unequivocally identify the function of these sites, or identify all functional elements. We find that looping interactions are significantly enriched for distal fragments that contain E, WE and CTCF elements, and the actively transcribed chromatin state (“T”), but are depleted for the repressed chromatin state (“R”). We note that some distal looping fragments contain elements classified as “TSS” or “PF”, even though they do not contain TSSs as defined by the GENCODE v7 annotation [131]. Possibly, these chromatin states are not can be found at some distance from TSSs, or merely resemble promoters in chromatin state. We conclude that 5C identified significant looping interactions between TSSs and distal elements that display hallmarks of functional elements.

Next, we used the 7-way segmentation data to categorize looping interactions into four broader functional groups (**Figure 5.3 c, Figure 5.4, Supplementary Data File**): those that involve a distal fragment that contains a putative enhancer (“E”: E or WE); an element associated with promoters (“P”: TSS or PF), or a CTCF-bound element (CTCF). The final class contains interactions with distal fragments that do not contain any of these three types of elements (“U”: unclassified), although they often do contain individual features such as DHSs.

We find that TSS-E and TSS-P interactions are more cell type specific than TSS-CTCF interactions: in case of the former two categories the ratio of interactions that is seen in only one cell line vs more than one cell line is ~4:1, whereas it is ~1:1 for the latter (**Figure 5.4**). Next, we determined whether looping of a TSS to any of the four categories of chromatin states is correlated with transcription. We used CAGE expression data for the four cell lines [132] to assign an expression level to each TSS. In GM12878, K562 and HeLa-3 cells we find that looping interactions with elements containing enhancer-like E elements are significantly enriched for those that involve expressed TSSs (**Figure 5.5**). Similarly, the set of TSSs that interact with fragments containing E-elements were significantly more highly expressed compared to TSSs that do not interact with E-elements (**Figure 5.3 d**). Interestingly, this is not the case for H1-hES cells where no significant correlation with expression status or level was observed (**Figure 5.5**). We note that the E-class of elements in H1-hES cells differs from the E-class in the other three cells lines: it is characterized by high levels of the repressive mark H3K27me3 [130], as has been observed for poised enhancers. Therefore, it is possible that the E-class in H1-hES cells represents at least in part poised enhancers that do not yet activate gene expression [128], [129]. Interactions with other classes of elements (CTCF, P, and Unclassified) are in some cell lines, but not all, significantly enriched for actively expressed genes (**Figure 5.5**).

Our comprehensive dataset allowed us to determine the distribution of up- and downstream looping interactions. We aligned all TSSs and calculated the average number of interactions that a TSS has with each class of distal element at increasing genomic distances up and downstream of the TSS. (**Figure 5.6 a**) shows the resulting average long-range interaction profile across all four cell lines (similar results were obtained when each of the cell lines was analyzed separately (**Figure 5.7**). Several striking results are obtained. First, we find larger numbers of looping interactions with E, P and CTCF-bound elements upstream of the TSS as compared to downstream (bias of 4:1, up to 20:1). The bias for upstream interactions reveals an unanticipated directionality in long-range interactions with TSSs. This may indicate the presence of topological constraints imposed by the mechanism by which such interactions regulate target promoters. No such bias was observed for the set of unclassified elements, or for the complete set of interrogated interactions (**Figure 5.6 a**). Again, H1-hES cells were different and no directional bias was observed for interactions with any class of long-range interactions. Second, in all four cell lines the highest density of long-range interactions with E, P and CTCF-bound elements was observed with elements located 100-200 Kb upstream of the TSS. Interestingly, previous analyses showed that conserved non-coding elements are also often found within similar distances of target genes [133]. Third, when we analyzed expressed TSSs and non-expressed TSSs separately we find that both have a similar interaction profile but that expressed TSSs tend to have more interactions, especially with the E, P and CTCF classes. We cannot rule out that some TSSs classified as non-expressed based on the absence of CAGE tags are actually expressed at low levels. In addition, expressed and non-expressed TSSs differ in their overall 5C profile. 5C signals are lower around active TSSs as compared to non-expressed TSSs (**Figure 5.6 b**). This has been observed before for the FMR1 gene [121]. One interpretation is that there is a different topological chromatin conformation so that expressed TSSs interact less with their flanking chromatin and instead associate more with distal regulatory elements located farther away. Consistently, we find that expressed TSSs not only display more looping interactions (**Figure 5.6 a**), but that longer-range looping interactions also occur more frequently in the population as indicated by higher 5C scores as compared to non-expressed TSS (**Figure 5.6 b**).

Next we explored whether the relative order of elements in the genome affects which long-range interactions occur. It is often assumed that distal elements such as enhancers target the nearest TSS. We find that only ~8% of the looping interactions are between an element and the nearest TSS (**Figure 5.6 c**). This number goes up to 24% when only active TSSs are included. Similarly, 27% of the distal elements have an interaction with the nearest TSS, and 48% of elements have interactions with the nearest expressed TSS. Thus, when predicting TSS-distal element interactions, picking the nearest (active) gene is often not correct.

It has been suggested that CTCF sites located between an enhancer and a TSS may prevent enhancer-promoter interactions [126], [134]. To specifically address this question we determined how frequently we identified long-range interactions between a TSS and a distal element that skip over a site bound by CTCF. We find that ~80% of long-range interactions are unimpeded by the presence of one or more CTCF-bound sites in the corresponding cell line (**Figure 5.6 d**). Thus the mere presence of a CTCF-bound site does not block physical long-range interactions. Possibly, additional factors need to be recruited to CTCF-bound sites to obtain insulator activity, as has been shown in Drosophila [135].

The large numbers of long-range interactions we discovered suggest that distal elements and TSSs are each engaged in multiple long-range interactions. To characterize this phenomenon in more detail we determined the degree of TSS and distal fragments. We find that ~50% of TSSs display one or more long-range interactions with some interacting with as many as 20 distal fragments (**Figure 5.8 a**). Expressed TSSs interact with slightly more elements as compared to non-expressed TSSs (for GM12878 mean is 1.84 vs 1.35; or 3.79 vs 3.20 when including only those TSS with at least one interaction). Again, H1-hES cells are the exception where non-expressed TSSs displayed a slightly higher number of long-range interactions (**Figure 5.9**). Out of all distal fragments interrogated, ~10% interact with one or more TSS, with some interacting with more than 10 (mean 0.22; or 2.14 when including only those distal fragments with at least one interaction). The degree distribution of the four categories of distal elements were very similar, although we note that E-elements had a slightly lower degree (interacted with fewer TSSs on average) as compared to CTCF-bound elements (**Figure 5.9**).

(**Figure 5.8 b and c**) show examples of the complex long-range interaction networks formed by TSSs and distal elements. It is unlikely that these interactions can all occur at the same time in the same cell, which implies significant cell-to-cell variation. The network of long-range interactions in the HoxA locus (ENm010) in H1-hES cells is particularly interesting (**Figure 5.8 c**). We find that many of the HoxA genes interact with an upstream distal fragment that contains a CTCF-bound element and two E-elements. These E-elements contain peaks in binding of Suz12, which is a component of the repressive PRC2 complex that binds genes poised for activation [136]. Thus, these interactions appear to be between mostly inactive HoxA genes and distal elements that are bound by silencing complexes. This example reinforces our observations in H1-hES cells that long-range interactions are enriched for H3K27me3 (**Figure 5.3 b**), and that looping to E elements is not correlated with expression (**Figure 5.5**). Thus, it appears that H1-hES cells display a unique category of long-range interactions between inactive genes and distal poised or silencing elements.

Conclusions

Overall, our data provide new insights into the landscape of chromatin looping that bring genes and distant elements in close spatial proximity. Besides generating a rich dataset reflecting specific gene-element associations, the average interaction profile of TSSs with surrounding chromatin reveals several general principles regarding the asymmetric relationships between genomic distance, the order of elements, and the formation of looping interactions. The bias for upstream interactions may indicate that the protein complexes on many TSSs may be asymmetric and may preferentially interact on one side with enhancer-protein complexes approaching along the chromatin fiber, as would be proposed by the enhancer tracking model [136]. Furthermore, while these average looping profiles may facilitate computational prediction of long-range interactions throughout the genome, the fact that interactions skip genes and CTCF sites suggests that additional mechanisms for target selection and gene insulation exist.

With further 3C technology development and increases in sequencing capacity, similar high-resolution studies should become feasible to map specific long-range interactions throughout the genome, which may uncover additional principles that guide chromatin looping. Such insights will also be critical for interpreting genome-wide association studies that often identify regions with regulatory elements but not their distally located target genes.

Figures

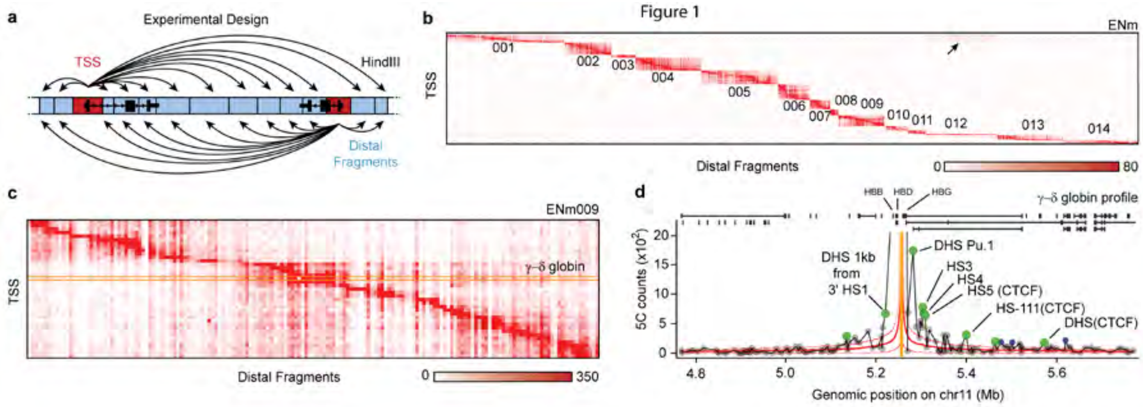


Figure 5.1 | A synopsis of 5C approach to identify long-range looping interactions in ENCODE Pilot regions.

a, 5C design. To interrogate long-range interactions between TSSs and distal elements we used the my5C toolbox 32 to design Reverse primers for HindIII restriction fragments in the ENCODE regions that contains a TSS (red fragments; according the Gencode-v7 annotation (GRCP01123) and Forward primers for all other 'distal' restriction fragments (blue fragments). b, Heatmap of all interrogated TSS-distal fragment interactions in 14 ENCODE regions (ENm001-014) in K562 cells. Fragments are displayed in their genomic order. Each dark rectangular area in the heatmap denotes interactions within a single ENCODE region while remaining areas denote interactions between regions. ENCODE regions that are near each other on the same chromosome show a higher interaction frequency (arrow) than regions that were on different chromosomes. c, Detailed heatmap of interactions of a single ENCODE region (ENm009: β-globin) from b. Interaction frequencies are generally higher for fragments that are located near each other in the genome (strong signal along the diagonal of the heatmap). The orange rectangle shows the 5C interaction profile of a single TSS (γ-δ globin) across the ENm009 region. d, Interaction profile of γ-δ globin (vertical orange bar) across ENm009 (hg19; chr11:4774421-5776011) based on 5C signal illustrating the peak calling method for 5C data. The solid red line shows the expected interaction level (LOWESS line, Supplemental Methods) along the genomic coordinates and dashed red lines above and below indicate LOWESS ± 1 standard deviation. The expected interaction profile demonstrates that contact probability decreases with genomic distance. 5C signals that are significantly higher than expected in both biological replicates (green circles, False Discovery Rate = 1%) are considered as looping interactions between the γ-δ globin and the corresponding distal fragment. Interactions that are significantly higher than expected in only one replicate (blue circles) are not considered as 5C looping interaction. 5C peak calling accurately detects the known long-range interactions of γ-δ globin to HS-3,4,5 and -111 and several additional DHS and CTCF sites (labeled).

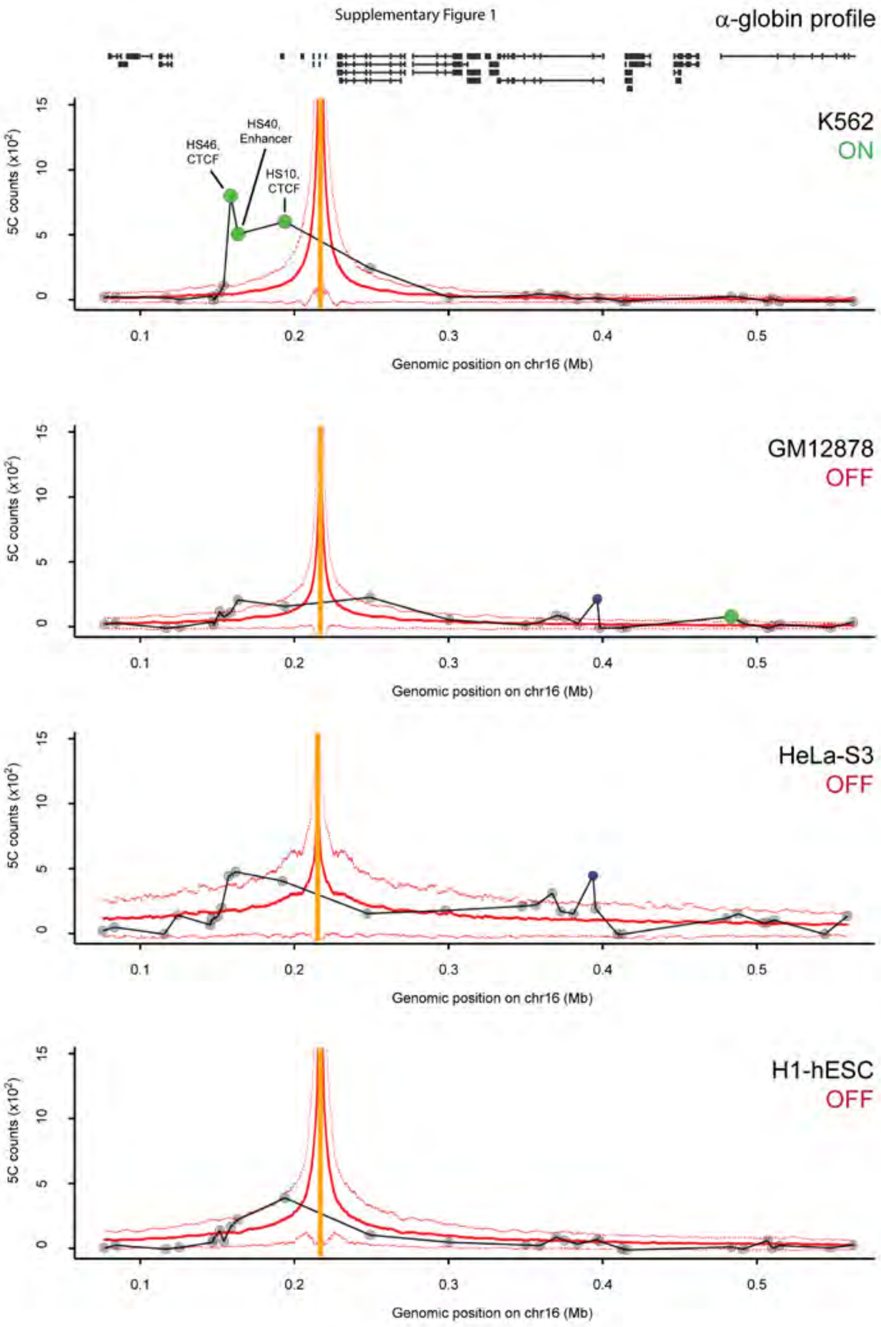


Figure 5.2 | Interaction profile of α-globin genes across ENm008 region in α-globin expressing (K562) and non-expressing cells.
5C interaction profile of reverse fragment (vertical orange bar) containing TSS of α- globin genes (HBA1, HBA2, HBM) vs interrogated distal fragments in ENm008 (hg19; chr16:60002-559999) region. The solid red line shows the expected

interaction profile (LOWESS line) along ENm008 genomic coordinates and dashed red lines above and below indicates LOWESS ± 1 standard deviation. The 5C signals that are significantly higher than expected in two biological replicates (green circles) are considered as long range looping interactions between α- globin and the corresponding distal fragments. The blue circles denote interactions higher than expected in only one replicate (not considered as looping interactions). In α-globin expressing K562 cells (ON), our 5C peak calling method accurately detects the known long-range interactions between the α-globin and its enhancer HS40 and the CTCF-containing HS46 and HS10 hypersensitive sites (indicated in top panel). These interactions are absent in cells (GM12878, HeLa-S3 and H1-hESC) where α-globin is not expressed (OFF).

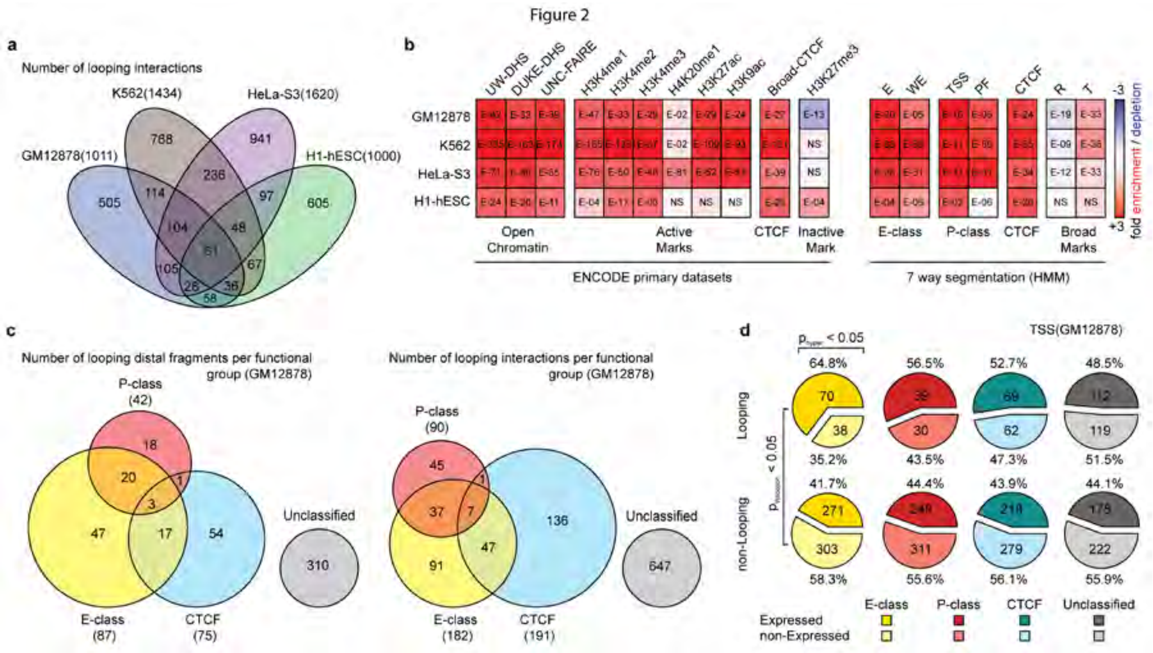


Figure 5.3 | Distribution of looping interactions across cell types and their relationship with chromatin features and gene expression.

a, Venn diagram showing the number of unique and overlapping looping interactions across four cell types (GM12878, K562, HeLa-S3 and H1-hESC). b, Heatmap showing the enrichment/depletion of different chromatin/histone marks and features in looping fragments compared to all interrogated fragments based on genome-wide datasets from ENCODE consortium (Supplemental Table 7) in four cell types. Various genome tracks include Open Chromatin: UW DHS, Duke DHS and UNC-FAIRE (formaldehyde assisted identification of regulatory elements); Active Marks: Broad Histone H3K4me1/2/3, H4K20me1, H3K27ac, H3K9ac; CTCF: Broad CTCF ChIP peaks; Inactive Mark: Broad Histone H3K27me3 and; 7 way segmentation: categories based on HMM prediction analysis for indicated cells - E (predicted enhancer), WE (predicted weak enhancer or open chromatin cis regulatory element), TSS (predicted promoter region including TSS), PF (predicted promoter flanking region), CTCF (CTCF-enriched element), R (predicted repressed or low activity region) and T (predicted transcribed region). We further grouped segmentation categories E and WE into "E-class", TSS and PF into "P-class", and R and T into "Broad Marks (spread at Kb length scale)". The color scale represents the fold enrichment (red) or depletion (blue). The numbers listed inside each box represent p-values of the significant enrichment/depletion for that mark while NS denotes p-values that are not significant. The p-values are calculated based on two-tailed hypergeometric test and corrected for multiple testing using Bonferroni. c, Venn diagram showing the number of unique and overlapping looping distal fragments (left) and looping interactions (right) among 4 functional groups in GM12878 cells. Distal fragments are classified into 4 non-exclusive groups based on the 7-way segmentation. Similarly, TSS - distal fragment interactions are classified based on the functional grouping of the distal fragments to which TSS is looping. The four functional groups are E-class (yellow; E + WE), P-class (magenta; TSS + PF), CTCF (cyan; CTCF enriched elements) and Unclassified (grey; interactions that do not belong to E-, P- or CTCF groups). d, Relationship between looping interactions of a particular group and gene expression in GM12878 cells. Pie charts showing percentages and numbers of expressed/non-expressed TSSs looping or not looping to a particular group (E-, P-, CTCF or Unclassified; colored as in c) of distal fragments (top panel). TSSs with a CAGE value greater that zero are deemed expressed. Significant enrichment for expressed TSSs in the looping or non-looping categories are indicated on top (hypergeometric test; phyper<0.05). Significant differences in expression levels between TSS in the looping vs the non-looping category is indicated on the left (Wilcoxon signed-rank test; pWilcoxon<0.05).

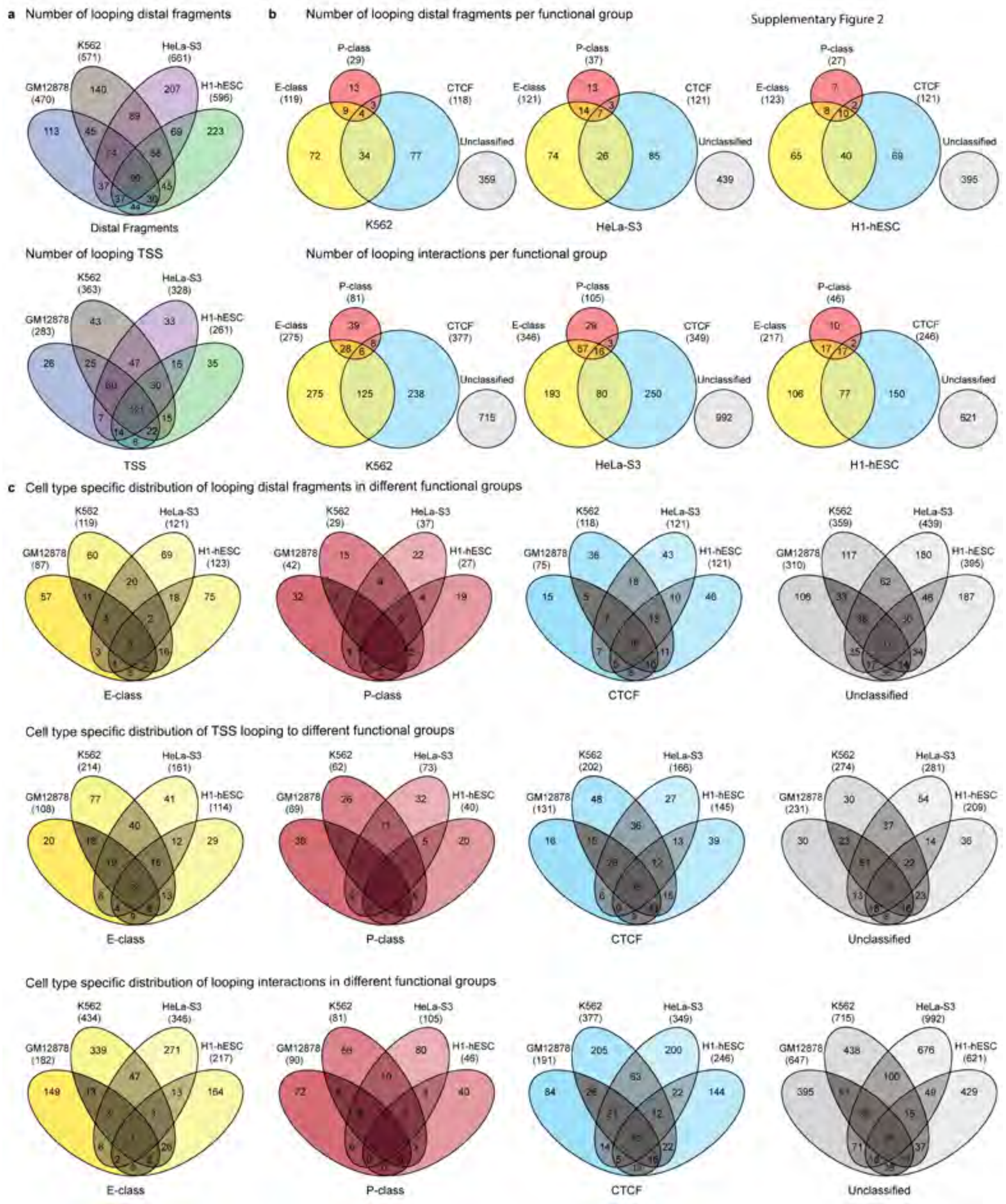


Figure 5.4 | Distribution of looping interactions across cell types and functional groups.

a, Venn diagrams showing the unique and overlapping looping distal fragments (top) and looping TSSs (bottom) across four cell types (GM12878, K562, HeLa-S3 and H1-hESC). b, As described in figure 2c, looping interactions are classified into E-class (yellow), P-class (light magenta), CTCF (cyan) and Unclassified (grey) groups. Venn diagrams showing the distribution of looping distal fragments (above) and looping interactions (below) among the four groups in

K562, HeLa-S3 and H1-hES cells. c, Venn diagrams showing the distributions of looping distal fragments (top), TSSs (TSS) (middle) and looping interactions (bottom) across different cell types in each of the E-class, P-class, CTCF and Unclassified groups.

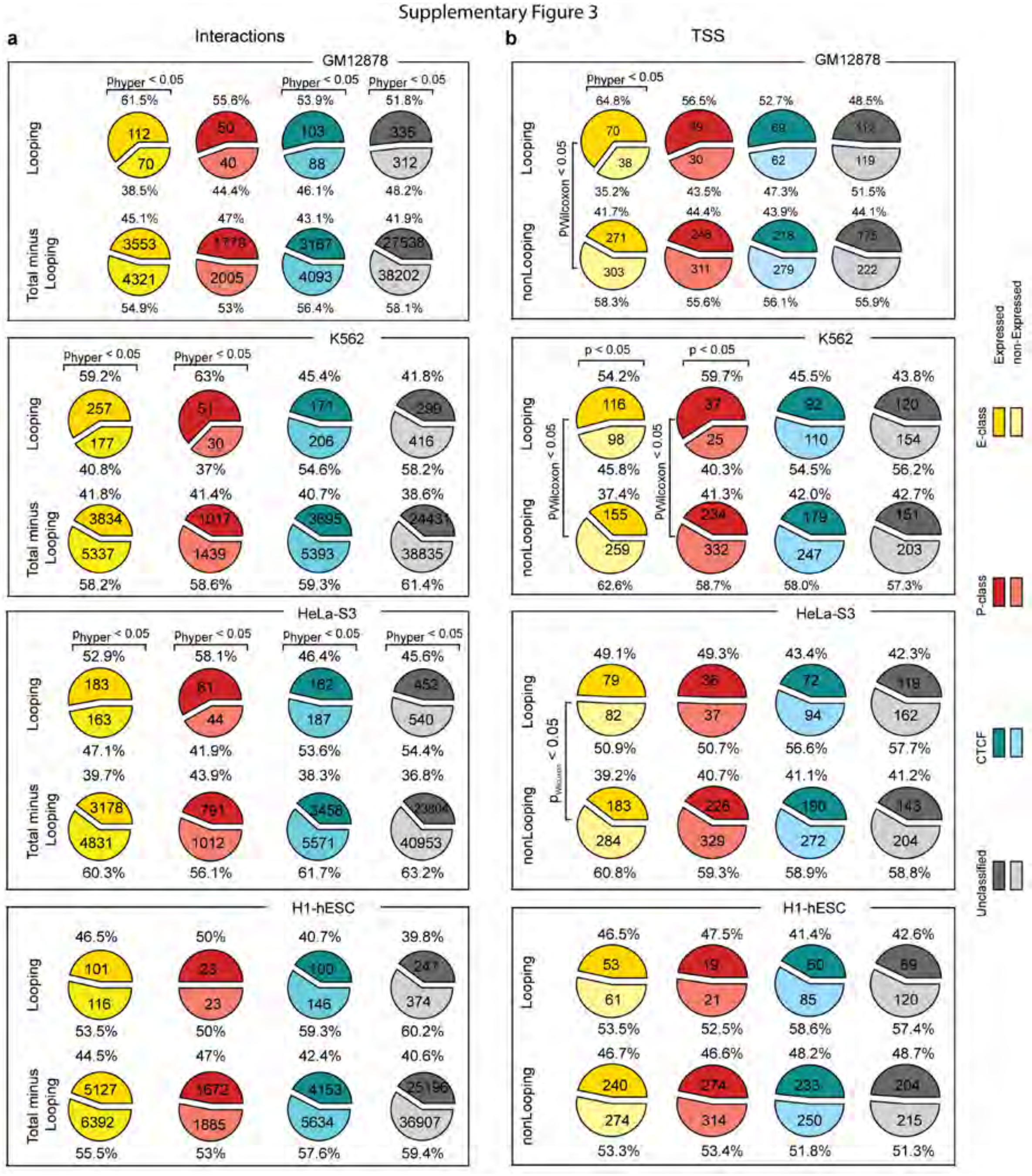


Figure 5.5 | Correlation between looping interactions to a particular groups and gene expression in different cell types.

As in figure 2d, CAGE expression data are used to assign expressions for each TSS in K562, HeLa-S3 and H1-hES cells. TSS with RIKEN CAGE value >0 is considered as expressed. Different groups are represented as: E-class (yellow), P-class (magenta), CTCF (cyan) and Unclassified (grey). a, The top row of pie charts in each panel under “interactions” indicates the percentages/numbers of TSS- distal fragment interactions of a particular functional group (E-, P-, CTCF or Unclassified) in which looping TSSs are expressed (dark color) or not expressed (light color). The bottom row in each panel of pie charts shows the percentages/numbers of expressed (dark color) and not expressed (light color) TSSs for all interrogated but non-looping TSS-distal fragment interactions of a particular group. Significant enrichment for expressed TSSs in the looping or non-looping categories are indicated on top (hypergeometric test; phyper<0.05). b, The top row in each panel of pie charts under “TSS” indicates percentages and numbers of expressed/non-expressed TSSs looping or not looping to a particular group (E-, P-, CTCF or Unclassified; colored as in c) of distal fragments (top panel). TSSs with a CAGE value greater that zero are deemed expressed. Significant enrichment for expressed TSSs in the looping or non-looping categories are indicated on top (hypergeometric test; phyper<0.05). Significant differences in expression levels between TSS in the looping vs the non-looping category is indicated on the left (Wilcoxon signed-rank test; pWilcoxon<0.05).

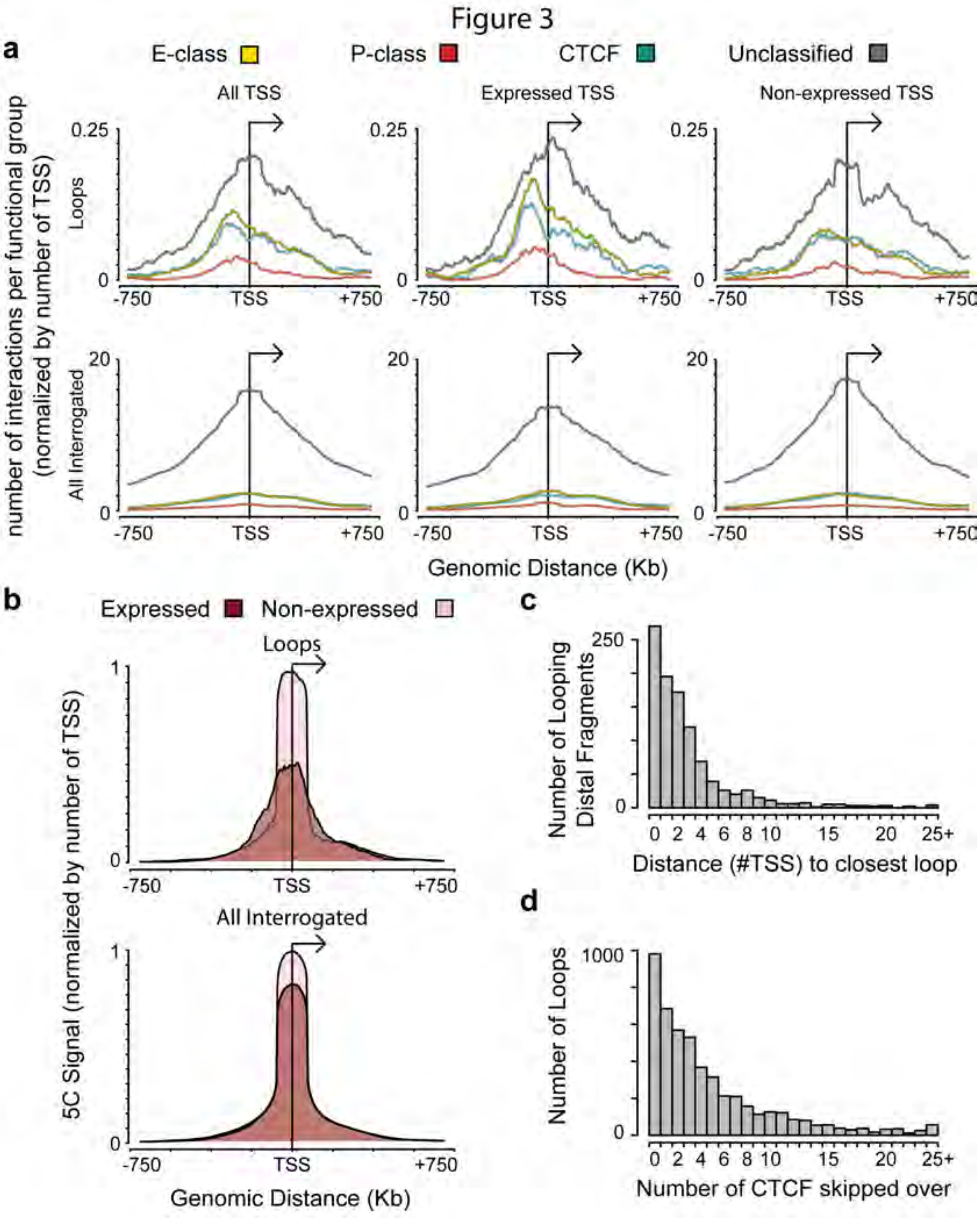


Figure 5.6 | Average looping landscape of TSSs to distal fragments.
a, Composite profile of average number of group-specific looping interactions upstream and downstream of TSSs based on combined 5C interaction data from the four cell lines. Each group is represented by different colors. The top panel shows the average looping profiles of all TSSs (left), of expressed TSSs (CAGE

value of >0, middle) and of non-expressed TSSs (CAGE value = 0; right). The bottom set of plots shows the corresponding profiles of all interrogated TSS-distal element interactions (left), of expressed TSSs (middle) and of non-expressed TSSs (right). All the interaction data for a particular group for all four cell lines are binned with a sliding window of 150 Kb with step size of 5 Kb and the interactions values normalized by the number of TSSs. b, Composite profile of average 5C looping signal upstream and downstream of TSS based on combined 5C looping signals from four cell lines. The signals were normalized by the number of TSSs. The top panel represents the average 5C signal for statistically significant loops for TSSs that are expressed (dark red) and not expressed (light red). The bottom panel represents the average 5C signal for all interrogated interactions for TSSs that are expressed (dark red) and not expressed (light red). c, Histogram showing the number of distal fragments that are involved in looping with their target promoters skipping 0,1,2,…, 25 (and above) TSSs (data for all four cell line combined). All the values above 24 in the x-axis are added and grouped as 25+. d, Histogram showing the number of looping interactions that skip over 0, 1, 2,…, 25 (and above) CTCF-bound elements (based on 7 way segmentation) between the distal fragments and their target TSS (data for all four cell line combined). All the values above 24 in the x-axis are added and grouped as 25+.

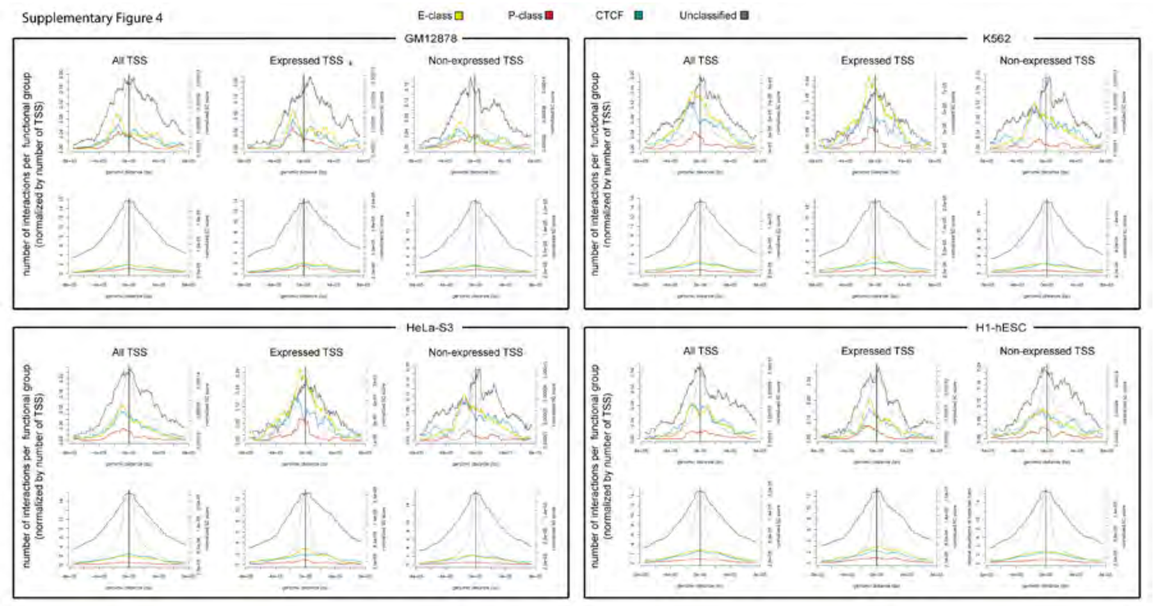


Figure 5.7 | Average TSS-distal fragment looping landscape in different cell lines.

Composite profiles of average number of group-specific looping interactions upstream and downstream of TSSs for each of the four cell lines. Each group is represented by different colors: E-class – yellow, P-class – magenta, CTCF – cyan and Unclassified – grey. In each panel the top row shows the average looping profiles of all TSSs (left), of expressed TSSs (CAGE value of >0, middle) and of non-expressed TSSs (CAGE value = 0; right) with each of the four groups of distal elements. The bottom row of each panel shows plots with the corresponding profiles of all interrogated TSS-distal element interactions (left), of expressed TSSs (middle) and of non-expressed TSSs (right). All the interaction data for a particular group is binned with a sliding window of 150 Kb with step size of 5 Kb and the interactions values normalized by the number of TSSs.

Figure 4

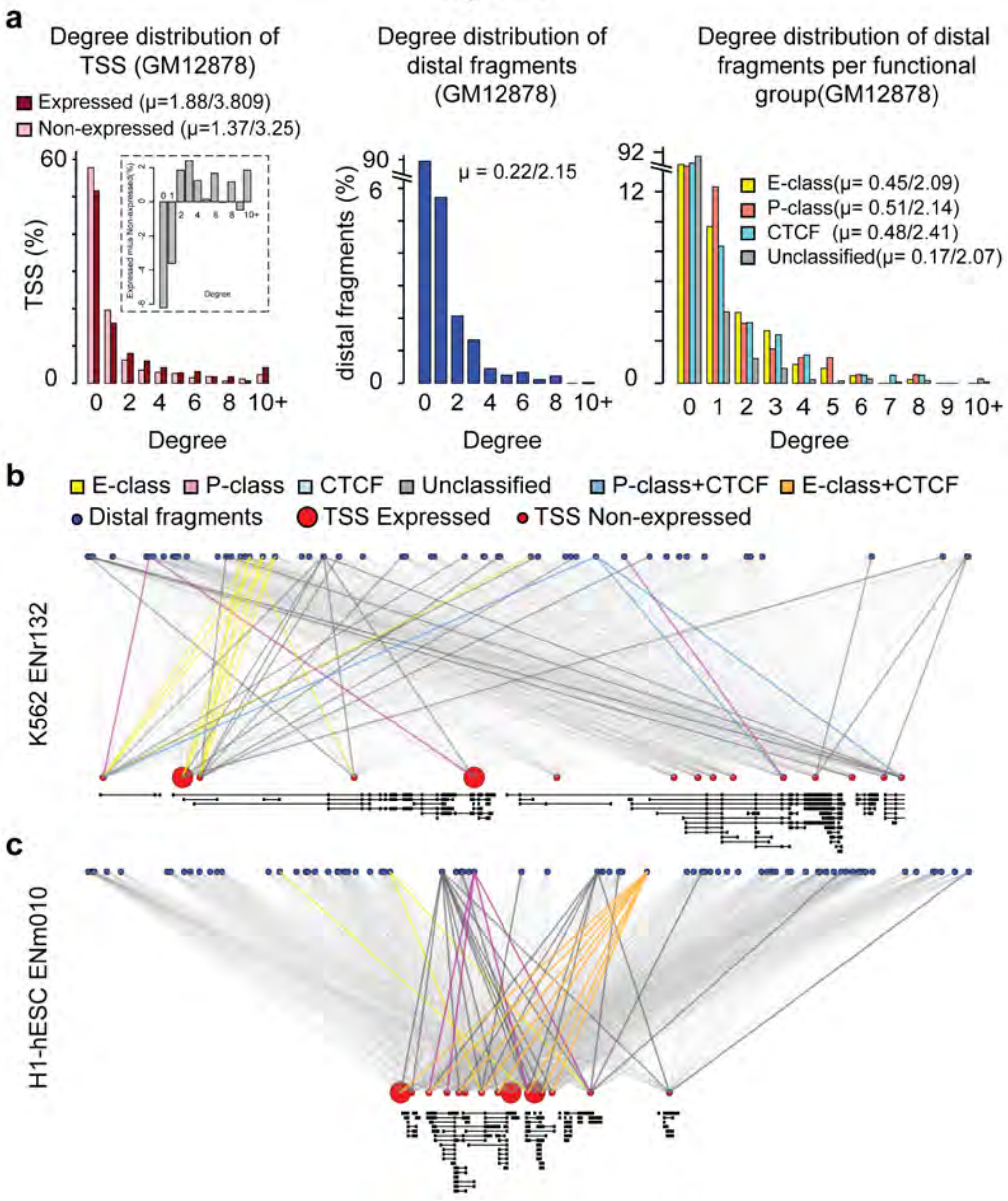


Figure 5.8 | Degree distribution and networks of TSS-distal fragments looping interactions.

a, Histogram showing the number of TSSs (left, red) or distal fragments (middle, blue) in percentages that are involved in 0, 1, 2,…., 10 (and above) looping interactions (degree, x-axis) with distal fragments and TSSs respectively in GM12878 cells. All the values for degrees that are >9 are added and grouped under degree 10+. The dark red bars represent the percentages of looping TSSs that are expressed (CAGE expression value >0) while light red bars represent the percentages of looping TSSs that are not expressed. Inset: the difference in percentages between looping TSSs that are expressed and not expressed for each degree is shown. The right panel: degree distribution for each functional group of distal fragments. The average (mean, μ) degrees for TSSs and distal fragments are indicated. The first value is the mean degree considering all the TSS/distal fragments (looping + non-looping) while the second value is the mean degree of looping TSS/distal fragments (excluding degree = zero). b, Webplot showing the long-range looping interactions in ENr132 region in K562 cells. The interrogated distal fragments (blue circle) and the TSS (red circle) are positioned according to genomic coordinates and the Gencode v7 gene annotation is indicated. The size of the red circles denotes if that TSS is expressed (big circle, CAGE value >0) or not expressed (small circle). The thin grey lines show all the possible interactions that were interrogated. The colored lines show significant looping interactions between TSSs and distal fragments of a particular group. c, Webplot showing the looping interactions of the HOXA (ENm010) region in H1-hES cells.

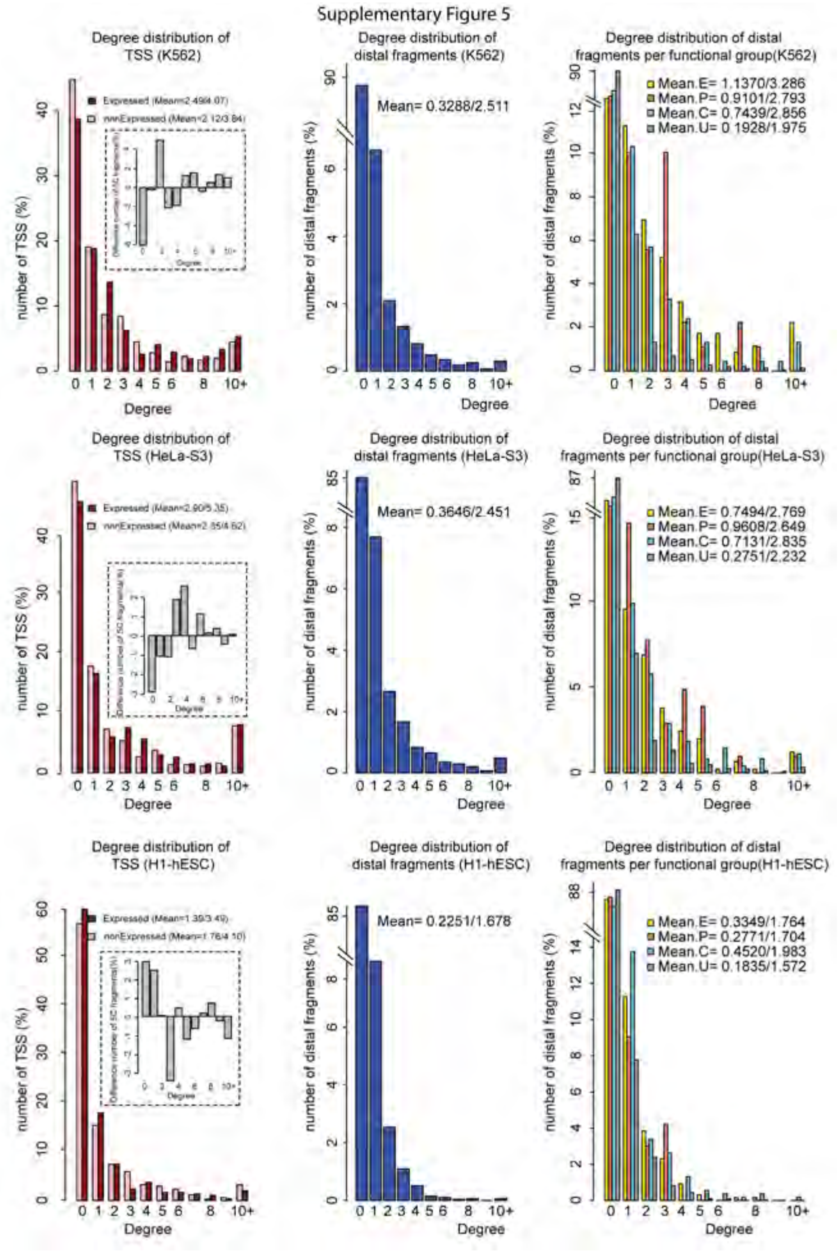


Figure 5.9 | Degree distribution of looping interactions of TSS and distal fragments in K562, HeLa-S3 and H1-hESC.
Histogram showing the number of TSSs (left, red) or distal fragments (middle, blue) in percentages that are involved in 0, 1, 2,..., 10 (and above) number of looping interactions (degree, x-axis) with distal fragments and TSSs respectively in K562 (top panel), HeLa-S3 (middle panel) and H1-hES cells (bottom panel). All the values in degrees that are >9 are grouped and included in the category with

degree 10+. The red bars represent the percentages of looping TSSs that are expressed (CAGE expression value >0) while light red bars represent the percentages of looping TSSs that are not expressed in the corresponding cell line. The difference of percentages between looping TSSs that are expressed and not expressed (red bar minus light red bar) for each degree is shown (inset). The right panel shows the degree distribution for each group of distal fragments. The average (mean, μ) degree for TSSs and distal fragments are indicated. The first value is the mean degree considering all the TSS/distal fragments (looping + non-looping) while the second value is the mean degree of looping TSS/distal fragments (degree greater than zero).

Methods

5C was performed using two pools of 5C primers, one for ENm001 through ENm014 and ENr313, and one pool for all 30 randomly picked ENCODE regions (ENr111-ENr334) 11 (Supplementary Table 2 and 3). 5C libraries (two biological replicates per cell line) were sequenced on an Illumina GAII platform and sequence reads were mapped using Novoalign (http://www.novocraft.com), as described [124]. Raw mapped reads for each experiment will be submitted to GEO. Statistically significant pair-wise interactions were identified (Supplementary Methods) by converting each 5C signal into a z-score using the average 5C signal distribution versus genomic distance as a background estimate. Significant interactions (1%FDR) observed in both replicates were considered looping interactions. 5C looping interactions were compared to a variety of genome-wide data sets generated by the ENCODE consortium (Supplementary Table 7).

Tissue culture

GM12878 lymphoblastoid cells were procured from Coriell Cell Repositories and grown in RPMI 1640 medium supplemented with 2mM L-glutamine, 15% fetal bovine serum (FBS) and antibiotic (1% Pen-Strep). K562 (CCL-243), a CML cell line and HeLa-S3 (CCL2.2), a cervical carcinoma cell line were obtained from American Type Culture Collection (ATCC). K562 cells were cultured in similar media as GM12878 except with 10% FBS while HeLa-S3 cells were maintained in ATCC recommended F-12K Medium (Kaighn's Modification of Ham's F-12 Medium) with 10% FBS and 1% Pen-Strep. H1-hES cells are human male embryonic stem cells and formaldehyde crosslinked cells were distributed by Cellular Dynamics International, Madison WI to individual labs that were part of the ENCODE project consortium. The culture densities and conditions were maintained as per recommendations of the repositories.

Formaldehyde crosslinking

For suspension cells (GM12878, K562) cells a total of 1X10^8 freshly growing cells were centrifuged at 100Xg for 5 minutes. Cell pellets were resuspended in 45 mL of respective growth medium in a 50 mL Falcon tube. Cells were fixed by addition of 1.25 mL of 37% formaldehyde (final concentration of formaldehyde 1%). The cell suspension was gently mixed by inverting the tube up and down 4-6 times at room temperature and the tubes was rotated on an end-to-end shaker for exactly 10 minutes. Crosslinking was stopped by addition of 3M glycine (final concentration 125 mM) and cell suspensions were incubated at room temp for 15 minutes using an end-to-end shaker. The crosslinked cells were then pelleted at 100Xg for 5 minutes and the cell pellet was stored at -80°C. For HeLa-S3, the adherent cells were first trypsinized and then the crosslinking was performed as described above. Cross-linked H1-hES cells were obtained from Cellular Dynamics International.

5C analysis

5C analysis was carried out as previously described [1,2] for the 44 ENCODE Pilot regions (ENCODE Manual – ENm and ENCODE Random – Enr). The chromosomal position and coordinates of the regions as per the Feb 2009 GRCh37/hg19 human genome assembly are enlisted in (supplemental table S1). The 5C experiment is designed to interrogate looping interactions between *Hin*dIII fragments containing transcription start sites (TSS) and any other *Hin*dIII restriction fragment ("distal fragments") in the ENCODE Pilot regions.

5C primer design

5C primers were designed at *Hin*dIII restriction sites (AAGCTT) using 5C primer design tools previously developed and made available online at My5C website (http://my5C.umassmed.edu) [3]. Reverse 5C primers were designed for *Hin*dIII restriction fragments overlapping a known TSS from GENCODE transcripts, or overlapping a start site as experimentally determined by CAGE Tag data of the ENCODE pilot project (Supplemental Table ST2). Forward 5C primers were designed for remaining of the *Hin*dIII restriction fragments (Supplemental Table ST3). For ENCODE regions that do not contain any TSS (ENr112, ENr113, ENr311 and ENr313) we employed an alternative primer design. For these regions an alternating design of forward and reverse 5C primers was used in which forward and reverse primers are designed for alternating restriction fragments [1]. Primers were excluded for highly repetitive sequences that prevented the design of a sufficiently unique 5C primer. Primers settings were as described before [2]: U-BLAST: 3; S-BLAST: 130: 15-MER: 1320; MIN_FSIZE: 40; MAX_FSIZE: 50000; OPT_TM: 65; OPT_PSIZE: 40. The 5C primers contained up to 40 bases that were specific for the corresponding restriction fragment. If a shorter sequence was sufficient to obtain a predicted annealing temperature of 65°C, that shorter sequence was used, and random sequence was added to make a total of 40 bases. All the 5C primers have an extension of universal tail sequences, at the 5' end for Forward 5C primers, and at the 3' end of Reverse 5C primers. DNA sequence of the universal tails of forward primers was 5′-CCTCTCTATGGGCAGTCGGTGAT-3′; DNA sequence for the universal tails of reverse primers was 5′-AGAGAATGAGGAACCCGGGGCAG-3′. A six base barcode was included between the specific sequence of the primers and the universal tail to aid in mapping of the high throughput short sequencing reads. The length of each primer was 69 bases. In total, 981 reverse primers and 5,321 forward primers were designed (corresponding to ~77.1% (6,302/8,174) of all *Hin*dIII fragments in the 44 ENCODE regions).

Generation of 5C libraries

3C was performed with *Hin*dIII restriction enzyme as previously described [2,4] for GM12878, K562, HeLa-S3 and H1-hES cells separately with two biological replicates for each cell line. The 3C libraries were then interrogated by 5C. The

44 ENCODE regions were analyzed in two groups using two separate 5C primer pools. The first group (ENm) contained the manually picked ENCODE regions ENm001-014, and ENr313. The second group (ENr) contained the 30 randomly picked ENCODE regions. The two 5C primer pools were made by pooling 5C primers for interrogating long-range interactions in the two groups of ENCODE regions. In these pools each primer was present at a final concentration of 0.5fmol/µL.

The primer pool for the ENm group contained a total of 3,150 primers (476 reverse 5C primers and 2674 forward 5C primers). This primer pool allows interrogation of a total of 1,272,824 interactions. Of these, 83,427 interactions were between fragments that were both located in the same ENCODE region. The primer pool for the ENr group contained a total of 3,152 primers (505 reverse 5C primers and 2647 forward 5C primers). This primer pool allows interrogation of a total of 1,336,735 interactions. Of these, 34,859 interactions were between fragments that were both located in the same ENCODE region.

5C was performed in 10-15 reactions each containing an amount of 3C library that represents 200,000 genome equivalents and 0.5 fmol of each primer. The multiplex annealing reaction was performed overnight at 55 °C. Pairs of annealed 5C primers were ligated at the same temperature using Taq DNA ligase for 1 hour. Ligated 5C primer pairs, which represent a specific ligation junction in the 3C library and thus a long-range interaction between the two corresponding loci, were then amplified using 28 cycles of PCR with universal tail primers that recognize the common tails of the 5C forward and reverse primers. At least four separate amplification reactions were carried out for each 10-15 annealing reactions described above and all the PCR products were pooled together. This pool constitutes the 5C library. The libraries were concentrated using Qiaquick PCR purification kit and 3′-A tailing reaction was done using dATP and Taq DNA polymerase in presence of 1X standard Taq buffer (NEB) at 72ºC for 30 minutes.

To facilitate Illumina paired end DNA sequence analysis of 5C libraries, Illumina paired end adapter oligos (Illumina, San Diego, CA) were ligated to the 5C library using the Illumina PE protocol. The linkered 5C library was then amplified by PCR (17 or 18 cycles, with Phusion High Fidelity DNA polymerase) using Illumina PCR primer PE 1.0 and 2.0. The 5C library gel purified and sequenced on the Illumina GA2 platform generating 36 base paired end reads.

5C read mapping

Sequencing data was obtained from an Illumina GAIIx machine and was processed through a custom pipeline to map and assemble 5C interactions. We used thirty six (36) base-pair paired end reads to sequence all 5C libraries. Due to sequencing efficiency some 5C libraries were re-sequenced as many as 10 times to obtain the required read depth for our analysis.

The fastQ files were taken directly from the Illumina GAIIx and fed into our in-house 5C mapping pipeline. Each side of the paired end read was independently mapped to a pseudo-genome of all possible 5C primer sequences using the novoalign mapping algorithm (V2.05 http://novocraft.com). The default alignment settings for novoalign were used. After mapping, if both of the paired end reads could be uniquely mapped to a 5C primer, a 5C interaction was assembled. Invalid interactions between the same primer or between primers of the same type were removed as these would represent a mapping artifact or an issue with the 5C technique. The number of invalid interactions detected across all libraries was < 0.01%, which would be expected if solely due to random mapping errors.

Statistics regarding the 5C library quality, mapping efficiency etc. can be found in (supplemental table ST4). Since it is only necessary to map the paired end reads to the list of all possible 5C primers rather than to the entire genome, a higher percentage of mapped/usable reads can be achieved. We find that > 90% of all paired reads (after Illumina chastity filtering) can be uniquely mapped to a single 5C interaction. For libraries where more than one lane was used to achieve adequate sequence depth, the interactions from each lane were summed to produce the complete 5C interaction dataset. A table summarizing the read depth of each 5C library can be found in (supplemental table ST5). Pearson correlation coefficients between the biological replicates can be found in (supplemental table ST6).

Detection bias correction

5C experiments involve a number of steps that can locally differ in efficiency, thereby introducing biases in efficiency of detection of pairs of interactions. These biases could affect the efficiency of cross-linking, the efficiency of restriction digestion (related to cross-linking efficiency), the efficiency of ligation (related to fragment size), the efficiency of 5C primers (related to annealing and PCR amplification) and finally the efficiency of DNA sequencing (related to base composition). All these potential biases, several of which are common to other approaches such as chromatin immunoprecipitation (e.g. cross-linking efficiency, PCR amplification, base-composition dependent sequencing efficiency), will impact the overall efficiency with which long-range interactions for a given locus (restriction fragment) can be detected. To determine this overall efficiency of interaction detection we have developed the following general strategy. To determine overall interaction detection efficiency for a given restriction fragment we analyzed the large set of inter-chromosomal interactions that are detected for each fragment. We then defined the overall efficiency of inter-chromosomal interaction detection for a given fragment as the ratio of the average inter-chromosomal signal obtained with that fragment and the average inter-chromosomal signal of all fragments. We then corrected the frequency of each interrogated long-range intra-chromosomal interaction using a correction factor that is the product of the overall efficiency of inter-chromosomal interaction detection for the two interacting fragments.

This procedure will correct for any of the biases in detectability of interactions for a given locus, as listed above, and will also adjust for copy number variation of a locus, which can vary in transformed cell lines such as K562 and HeLa S3 cells, as these factors will also affect the level of inter-chromosomal interactions.

Detailed Primer Filtering

To approximate the relative 5C signal of each restriction fragment interrogated in the experiment we first calculated the average 5C signal for all trans interactions (interactions between different chromosomes). To remove any extreme outliers from the mean calculation, e.g. due to primer failure, we first filtered down the distribution of 5C signals in trans for each restriction fragment by removing all signals beyond the mean +/- three (3) standard deviations. After calculating the filtered mean for each restriction fragment in trans, we calculated the global mean of all inter-chromosomal interaction frequencies. We then calculated a correction factor for each restriction fragment that would normalize its set of trans interactions to the entire set. Once the correction factors were calculated, we then calculated the mean and standard deviation correction factor and flagged any restriction fragments requiring a correction value beyond the mean +/- 1.654 standard deviations. Fragments with a correction factor outside of this limit were flagged for removal since their trans signal is too above/below the expected signal by chance. Here, we assume that any variation in 5C signals detected within the trans space is due to experimental factors, differing primer efficiencies, ligation efficiencies etc.

Detailed Primer Correction

Once the outlier fragments are removed from the 5C dataset, we repeat the above described steps to calculate the primer correction values required to normalize the 5C signals for the remaining restriction fragments. Then, for each 5C interaction within an ENCODE region in the dataset, we use the product of the correction factors from the two restrictions fragments involved in the interaction as the final correction factor to apply to the 5C signal. 5C signals are then either increased or decreased by the correction factor to correct for varying signals from the fragments visibility in the trans interaction space.

Peak calling

To detect significant looping interactions from background looping interactions we developed an in-house “5C peak calling” algorithm. We chose to call peaks in each 5C biological replicate separately and then take only the peaks that intersect across replicates as our final list of significant looping interactions.

5C signals represent the three-dimensional contact probabilities between pairs of loci. This relationship inversely scaled with genomic distance. To properly control for the varying genomic distances tested in the 5C dataset, we first determined the relationship of 5C signals over genomic distance. Using a LOWESS smoothing algorithm we find the weighted average and weighted standard deviation of all 5C signals across the range of all interrogated genomic distances. We used the traditional tri-cubic weighting function and an alpha parameter of (0.01) to average the closest 1% of the 5C signals around each genomic distance. We assume the large majority of interactions are not significant looping interactions and thus we interpret this weighted average as the expected 5C signal for any given genomic distance. The 5C signals are then transformed into a z-score by calculating the (obs-exp/stdev). Where the (obs) value is the detected 5C signal for a specific interaction, (exp) is the calculated weighted average of 5C signals for a specific genomic distance and (stdev) is the calculated weighted standard deviation of 5C signals for a specific genomic distance. Once the z-scores have been calculated, the distribution of z-scores are fit to a Weibull distribution. We find that the distribution of z-scores fits to the Weibull distribution with a R2 value of > 0.939 for all cell-lines. P-values can then be mapped to each z-score and then also transformed into q-values for FDR analysis. The ‘qvalue’ package from R (qvalue.cal [siggenes]) was used to compute the q-values for the given set of p-values determined from the fit to the Weibull distribution. Using an FDR cutoff of 1%, we select all 5C interactions with a q-value <= 0.01. We then take the intersection of all significant looping interactions across the two biological replicates as our final list of 5C looping interactions.

Fragment Annotation

To annotate the interrogated restriction fragments, a variety of ENCODE datasets were used to check for overlap with our list of restriction fragments. A list of all utilized ENCODE datasets can be found in (supplemental table ST7).

Acknowledgements

We thank the University of Massachusetts Medical School Deep Sequencing core for sequencing 5C libraries, and Robert Thurman and John Stamatoyannopoulos for discussion and help with peak calling analysis. We thank Marian Walhout and members of the Dekker lab for discussions. This work was supported by grants from the National Institutes of Health, National Human Genome Research Institute (HG003143 and HG003143-06S1) and a W.M Keck Foundation distinguished young scholar in medical research grant to JD.

Author Contributions

JD conceived of the project. AS performed all 5C experiments. BRL designed 5C experiments, and built the data analysis and visualization pipelines. BRL, AS, GJ and JD analyzed the data and wrote the paper.

Author information

Author Information All data are publicly available at GEO (accession number GSE39510). 5C data has also been deposited in the public UCSC ENCODE database (http://encodeproject.org/ENCODE/). 5C data can be found at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUmassDekker5C/.

The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D. (job.dekker@umassmed.edu).

CHAPTER VI: cWorld – a toolbox for manipulating genome structure data

Preface

This chapter contains a description of unpublished work encompassing a git repositories containing code written in python, perl, R which has been used to process, filter, normalize, analyze, visualize, integrate and manipulate all 3C, 5C and Hi-C data contained in this thesis. This toolbox is a result of countless collaborations and discussions with others. The code has grown exponentially in both its usefulness and robustness over the years. My goal is to first describe the toolbox in terms of its design and expected usage and then describe the various algorithms and methods that each particular function utilizes.

Introduction

cWorld first started as a collection of perl scripts, used to quickly manipulate genome structure data (e.g. 5C). The code was published and made publically available in the form of a web tool (http://my5C.umassmed.ed) [137]. It soon grew to include various visualization methods, analysis methods (peak calling, normalization, etc.) and integration methods (UCSC, bed files etc.) and as such outgrew its implementation as a web tool.

cWorld at its core contains a perl module (cWorld.pm) and a collection of action-based scripts that each perform a distinct manipulation of the data.

cWorld has grown to also include a python module and a series of scripts to perform specific functions. cWorld is slowly beginning to transition to a python environment to make use of both a hd5f file format implementation and to make use of numpy, a scientific computing package that is specialized for N-dimensional matrix operations.

cWorld file format

Since the introduction of 3C and 5C a new datatype was created. In order to standardize the file format to contain this 2D interaction data, cWorld and my5C introduced a standard in the field as to how to represent and store 2D genome structure data. This standard has been adopted by others and has helped to increase the usability of the datatype and development of additional tools and analysis packages. In the basic form, the 2D interaction data can be represented by a text formatted, tab delimited file (tsv). The key here is to add the specific row and column headers which represent the genomic loci of each particular row/column. This file format is visualized in (**Figure 6.1**). The headers of each specific row/col must be in a specific format that cWorld can read and understand. The basic structure of each header is as follows:

lociName|assembly|coordinates

lociName is any name that can describe the genomic loci encompassed by the row/col. **assembly** is a UCSC formatted assembly name (e.g. hg18, hg19, mm9, sacCer3, dm3 etc.). **coordinates** is a UCSC formatted genomic coordinate string that encompasses the genomic interval of the row/col (e.g. chr4:50000-150000, chrX:72000000-74000000, etc.). Each header must be unique from one another, except in the case of a symmetrical matrix. In this case, the header may be repeated on both the X and Y axis. A further level of details allows one to encode additional information into the headers, such as information relative to a 5C primer, or a specific allele of a chromosome. For instance the following 5C row/col header, **5C_195_ENm009_REV_249|hg18|chr11:5598207-5605246** encodes information about the 5C design, 5C region, 5C primer type, 5C primer number, assembly, and genomic coordinates of the primer/fragment location. Should the **lociName** contain 4 underscores and start with "5C", then the cWorld module will use the second field as the 5C specific region ID (**195**), the third field as the 5C specific region name (**ENm009**), the fourth field as the 5C specific primer type (**REV**) and the fifth field as the 5C specific primer/fragment number (**249**). This additional information can have implications in specific tools and the user can select specific ways to handle this additional information (e.g. such as defining cis/trans as either on different physical chromosomes, or between different 5C regions). To encode allele information into the chromosome, one can add the following token to the coordinates field, e.g. 62232|mm9-cast-129s1|**chrX-129S1:16640001-16680000**. In this header, the lociName is **62232**, an index relative to the genome-wide bin number. The assembly is **mm9-cast-129s1** signifying that the genome is mm9 (mouse) but also a diploid genome of the cast and 129s1 strains. The chromosome location is **chrX-**

129S1:16640001-16680000. Here the allele (parental genome) is encoded into the chromosome name. chrX-129S1 signifies that this bin, row/col represents data from the 129S1 allele on chrX at genome interval 16640001-16680000. By leveraging allele information into the headers of the cWorld matrix, various analyses can now be performed in an allele specific manner.

cWorld functions

cWorld.pm contains 82 private and public sub routines. These sub-routines are utilized by all of the action-based scripts and together represent the API that one can use to interface with the library.

cWorld logic flow

The normal logical flow is as follows: A user starts with an interaction matrix in the cWorld tsv file format, denoted $inputMatrix. First the user would pass the $inputMatrix to the getMatrixObject() function, this function first validates that the matrix is intact and in the correct file format, then it returns various objects such as a list of the headers, missing row/cols, number of NANs in the matrix, number of 0S, whether the matrix is symmetrical or not, the number of contigs/chromomes, indices into the matrix/headers for all chr/contig/region breakpoints and so on and so forth. All of these data structures are encompassed within the $matrixObject data structure. The reason getMatrixObject validates and returns this many objects is to pre-process and memoize as many useful metrics about the matrix as possible. This memoization helps to speed up later steps and allow assumptions about data integrity and matrix structure.

After the user has received a $matrixObject representing their matrix file, various other functions can be called with the $matrixObject and the $inputMatrix, such as getData() which returns a 2dimensional hash object containing the signal within each i,j pixel within the matrix, normally denoted $matrix. cWorld utilizes a spare data storage implementation, meaning the signal for every pixel, (i,j coordinate) is not stored. During the initial getMatrixObject() call, cWorld make a decision whether it is more efficient to omit the 0s or the NaNs from an input matrix. Depending on the type of data (5C, HiC), binned or not binned, raw or iced etc., the number of NaNs and 0s can different and potentially represent over 50% of the entire matrix. By omitting to store these values, memory usage can be dramatically reduced and these omitted values can then be inferred later when performing analysis or writing output data.

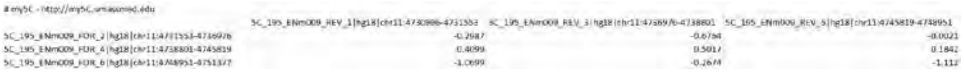
With both the $matrixObject and $matrix in hand most all of the analysis can be performed. Existing data structures may be created and memoized to reduce redundant computation and can be added to the $matrixObject for continued processing.

cWorld scripts

The cWorld toolbox contains over 43 different scripts which can perform a set of standard function calls in a particular sequence in order to perform a specific analysis or transformation of the input data. The collection of scripts is hosted on github (https://github.com/blajoie/cWorld-dekker). These scripts can perform functions such as transforming a matrix file into a heatmap png image, binning an interaction matrix into fixed size genomic intervals, subset a matrix by genomic coordinates or by a user-defined element (BED) file and so on and so forth. These scripts are built in a modular nature and specialize in a singular transformation of the data. By creating a ‘pipeline’ that calls multiples scripts in a specific sequential order, a user can perform complicated multi-stage analyses. Complicated procedures such as peak calling, significant difference detection between N samples, outlier filtering and data normalization can be performed in the above manner. I will now describe the various scripts, their intended usage and expected output.

addMatrixHeaders.pl - add headers to a matrix txt file

This script can add headers to a matrix txt file. This is useful for interaction matrices that are produced without embedded row/col headers and or converting from a python 2D list/np.array data structure.



my5C - http://my5C.umassmed.edu

	5C_195_ENm009_REV_1|hg18|chr11:4730986-4731553	5C_195_ENm009_REV_3|hg18|chr11:4736976-4738801	5C_195_ENm009_REV_5|hg18|chr11:4745819-4748951
5C_195_ENm009_FOR_2|hg18|chr11:4731553-4736976	-0.2987	-0.6754	-0.0021
5C_195_ENm009_FOR_4|hg18|chr11:4738801-4745819	0.4099	0.5017	0.1842
5C_195_ENm009_FOR_6|hg18|chr11:4748951-4751377	-1.0699	-0.2674	-1.112

Figure 6.1 | Schematic depicting cWorld tsv file format
y-axis headers, x-axis header and the data matrix can be seen. Lines that start with a # are comment lines and are ignored by cWorld.

aggregateBED.pl - sliding window aggregate of BED5 data

This script can perform a sliding window genomic interval binning on an input bed/bedGraph file. For example, if a user supplies a read-level signal file for a chip-Seq experiment, this script can bin/summarize the data into fixed size genomic intervals that the user defined via the --wsize,--wstep and --wmode options. The user can also select the signal aggregation method (mean, min, max, sum, median, iqrMean etc.). The user can also make choices regarding whether to include/exclude 0s single. This script is useful for binning external 1D data tracks into the same intervals as 5C or Hi-C data.

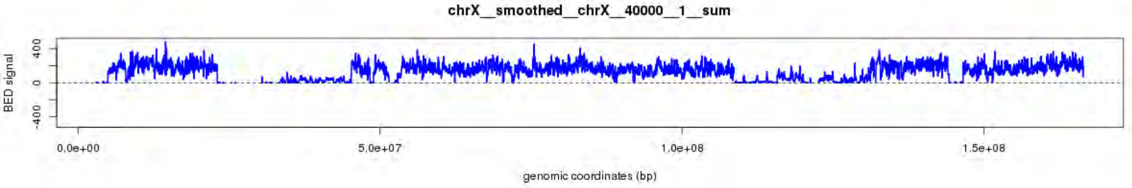


Figure 6.2 | Depiction of the aggregateBed method
Above, a tag level ATAC-Seq bedGraph track for chrX in mm9 is binned into fixed sized non-overlapping genomic intervals using the sum aggregation. Binsize=40000, binstep=1, binmode=sum.

anchorPurge.pl - filters out row/col from C data matrix file

This script can detect and remove outlier row/cols from a matrix. The outliers can be defined by various use-selectable methods. In the simplest case, outlier row/cols are detected by the sum of each row/col. Row/cols (anchors) can be removed if their sum is on the tails of the distribution of all row/col sums. This in effect would remove row/cols with either too low or too high of signal captured across the entire matrix. The tails of the distribution can be defined by various means but again in the simplest form the script will utilize a percentile based threshold. The user can also opt for a IQR based outlier threshold, which defaults to Q1 - 1.5 x (IQR), or above Q3 + 1.5 x (IQR). This metric is favored as it is more robust to various distribution shapes. After the row/col are detected, all signal within the row/col is set to “NA”, the string used to signify missing or unavailable data within the matrix.

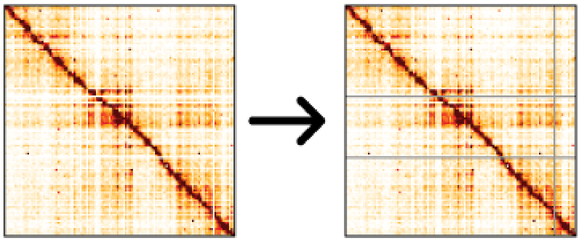


Figure 6.3 | Depiction of the anchorPurge method
The input matrix (raw 5C matrix) is on the left, and the resulting matrix from (anchorPurge.pl) is on the right. Here outlier row/cols are detected and then removed (removed row/cols show in gray).

applyCorrection.pl - apply correction to factor - external factors

This script can apply a set of externally-calculated correction factors to a given matrix. The correction factors can be applied by either of the two methods described in **correctMatrix.pl.**

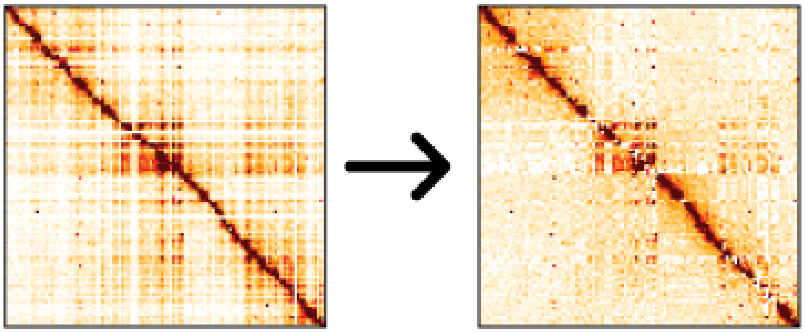


Figure 6.4 | Depiction of the apply correction method
The input matrix (raw 5C matrix) is on the left, and the resulting matrix from (applyCorrection.pl) is on the right. Here a set of externally calculation correction factors for each row/col were supplied and the pixel (i,j) adjustment was applied using the ‘zscore’ method.

binMatrix.pl - bin/aggregate matrix into fixed sized intervals

This script can bin a supplied matrix into fixed size genomic intervals. This script is useful for binning fragment-level data (3C, 5C, Hi-C) into larger genomic intervals to reduce noise and increase signal by reducing the resolution of the data. The user can also select the signal aggregation method (mean, min, max, sum, median, iqrMean etc.). The available aggregation methods are described in the function **listStats()**. The user can also make choices regarding whether to include/exclude 0s single. Normally if the number of 0s in a given matrix is > 50%, then one should bin the data into large intervals with the goal of reducing the percentage of '0' interactions.

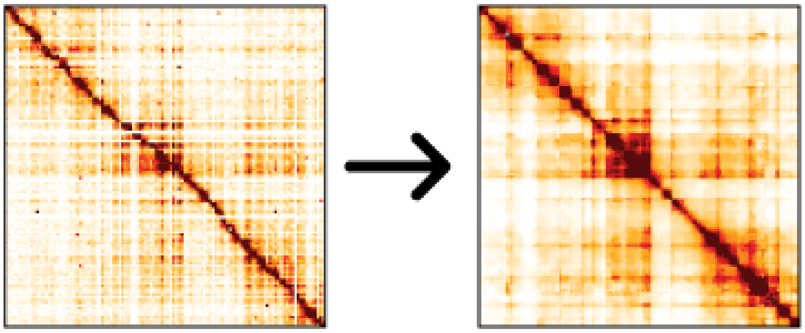


Figure 6.5 | Depiction of the binMatrix method
The input matrix (raw 5C matrix) is on the left, and the resulting matrix from (binMatrix.pl) is on the right. Here the fragment-level 5C matrix was binned using binsize=30000, binstep=10, binmode=median.

changeMatrixHeaders.pl - replace matrix row/col headers

This script can alter the headers of a supplied matrix file. This is can be useful to embed additional information (such as allele or tad definitions) into a matrix file. This script expected a *map file, which is a two column tsv file containing in column 1 the original header (found in the matrix) and in column 2 the new header that is to be replaced.

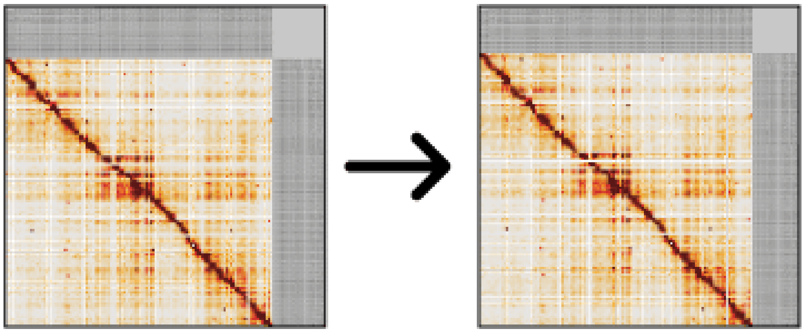


Figure 6.6 | Depiction of the changeMatrixHeaders method
The input matrix (raw 5C matrix) is on the left, and the resulting matrix from (changeMatrixHeaders.pl) is on the right. Here the headers of each row/col have been changed, since the headers have only changed by name, the resulting data matrix is unaltered.

collapseMatrix.pl - collapse matrix by (chr,name,group), sum signal

This script can collapse and aggregate the signal of a matrix by various means. For instance, if the user select the collapse by chromosome, then all signal within that chromosome can be aggregated (e.g. summed). If a genome wide matrix binned at 500kb is supplied and the user chooses to collapse by chromosome using the sum aggregation option, then all signal between each chrxchr combination will be summed. The resulting matrix will then be a NxN matrix where N is the number of chromosomes and each i,j pixel in the matrix represents the sum of all signal between chromosome.i and chromosome.j. This visualization is useful for looking at global chromosome-chromosome associations (e.g. the chromosomes containing the rDNA).

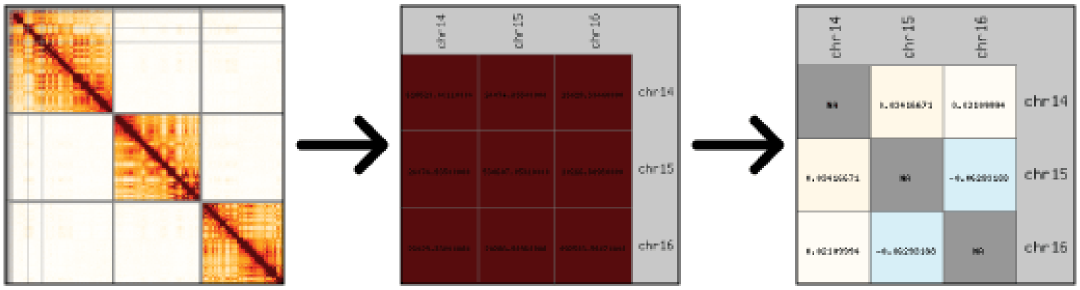


Figure 6.7 | Depiction of the collapseMatrix method
The input matrix (3 x 3 Hi-C matrix) is on the left, and the resulting matrix(s) from (collapseMatrix.pl) follow. Here a 500kb binned matrix, depicting chr14, chr15 and chr16 is used as input to the collapseMatrix.pl method. CollapseMatrix.pl first sums all signal per each of the 9 possible chr x chr cells, and then performs

a normalization for chromosome length. The final heatmap on the right shows the enrichment/depletion of interaction between each of the chromosomes.

column2matrix.pl - turn list (3 tab) file into matrix

This script takes as input a 3 column tab delimited text file (tsv) and turns it into an interaction matrix. The format is as follows: column1 = row header name. column2 = col header name. column3 = signal value. The matrix can then be constructed by first assembling a N1 x N2 matrix where N1 = number of distance row headers, and N2 = number of distinct col headers. Then the signal value (column 3) will be placed into the matrix at the combined of header.y and header.x. There are additional options to choose to include 0s or NaNs, or whether or not the matrix should be constructed in a symmetrical manner. If symmetrical is selected, the rows and cols will consist of the union of all distinct headers found in columns 1 and 2 of the input tsv fie.

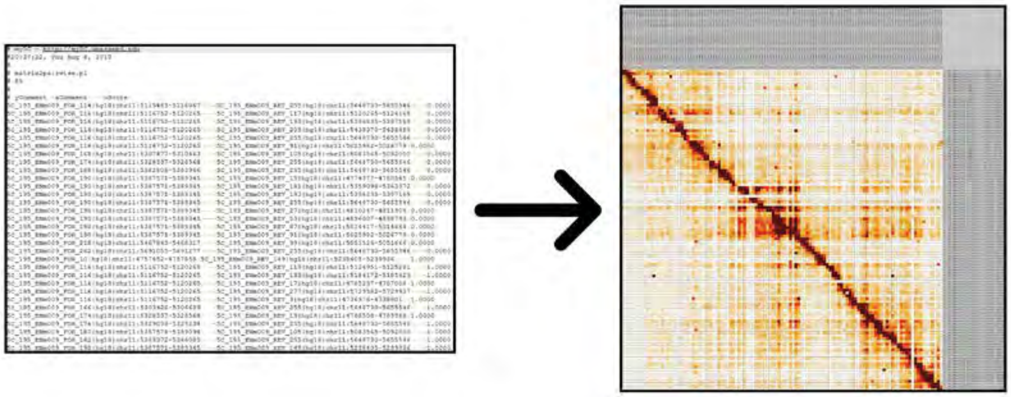


Figure 6.8 | Depiction of the column2matrix method.
The input 'pairwise' file (3 column tsv file) is on the left, and the resulting matrix from (column2matrix) is on the right. Here the matrix and resulting heatmap is displayed after the 'pairwise' (3 column tsv file) is transformed into a cWorld formatted matrix file.

combineMatrices.pl - combine matrices [sum,mean,median,min,max]

This script can combine an N number of input matrices. This can be useful for combined interaction matrices from biological replicates or samples of similar type, or even for various other analyses goals (pooling samples for average/consensus structure). Similar to other scripts, the user can specify the aggregation method as described in **listStats().**

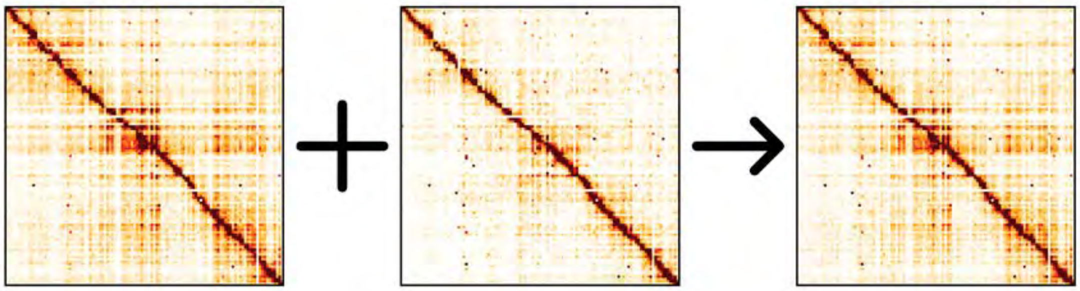


Figure 6.9 | Depiction of the combineMatrices method
Two input matrices are displayed on the left (between the '+' symbol). These two matrices are summed via the combineMatrices.pl method and the resulting summed matrix is displayed on the right. Here combineMode=sum is used.

compareInsulation.pl - compare insulation vector - calculate difference

This script can compare insulation vectors calculated from the **matrix2insulation.pl** script. By comparing insulation vectors one can deduce and assess structural changes between two samples. Since insulation signal is already in log-space, this script simply calculated the subtraction between the two vectors and outputs by data files and files formatted into bed and bedGraph format for visualization.

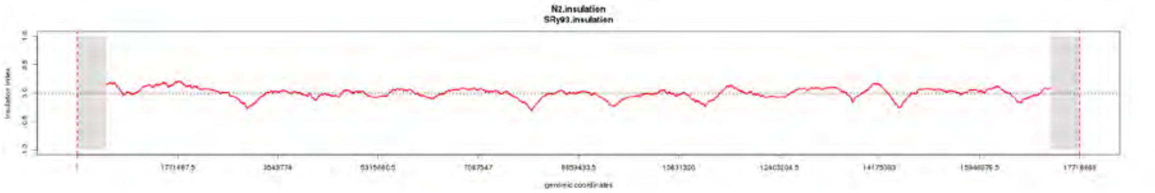


Figure 6.10 | Depiction of the compareInsulation method
Here two insulation vectors (not shown) are being compared via the compareInsulation.pl script. The resulting difference vector (red) displays area where the two insulation vector differ. The minima along this vector signify lost or weakened boundaries in one of the input insulation vectors compared to the other.

compareMatrices.pl - performs comparison between two matrices

This script can perform a comparison between any two matrices. The user can select the comparison method, such as [log2ratio, add, sum, mean, subtract, divide, multiply, min, max, deconvolve]. Log2ratio is the most common method to use when comparing two matrices as it can highly signal that is either higher or lower in the either sample. This script can properly handle 0s and NaNs that may differ between the two matrices. The deconvolved method is a useful mode that can be used to sample from a single matrix, and then subtract away the sampled signal. For instance, given two samples A and B. If A is an untreated control sample, and B is a treated sample within a 30% treatment efficiency. To create a matrix that contains only the sample that results from the treatment, one could very simply subtract away 70% of the A sample from the B sample. This could then create a new matrix that contains only the signal from the treatment in B. Of course this method is only an approximation but it can still serve a useful purpose during initial data exploration.

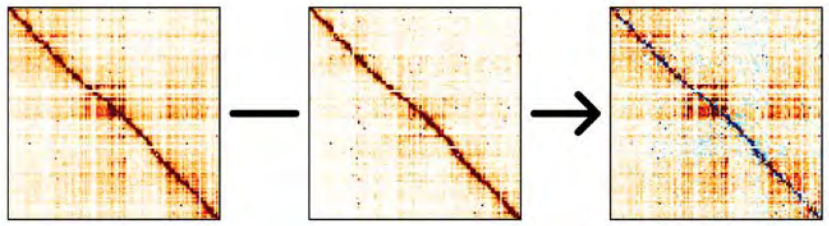


Figure 6.11 | Depiction of the compareMatrices method.
Two input matrices are displayed on the left (between the '-' symbol). The second matrix is subtracted away from the first matrix, and the resulting 'difference' matrix is displayed on the right. Here compareMode=subtract. The blue pixels

represent interactions that are now negative in value, signifying that the interaction score was higher in the second input matrix (middle).

correlateMatrices.pl - performs correlation between two matrices

This script can perform a correlation analysis between any two matrices. Only the i,j pixels that both contain valid signal as defined by the user are used to calculate the pearson/spearman correlation R value. The user can choose to subset the cis, trans, by genomic distance, exclude 0s etc. The output of this script is a plot generated by R which shows the scatter of signals between matrix_1 and matrix_2. A linear regression line is drawn through the scatter and the correlation value is printed on the top of the plot. The user can choose correlation either by the Pearson or Spearman method.

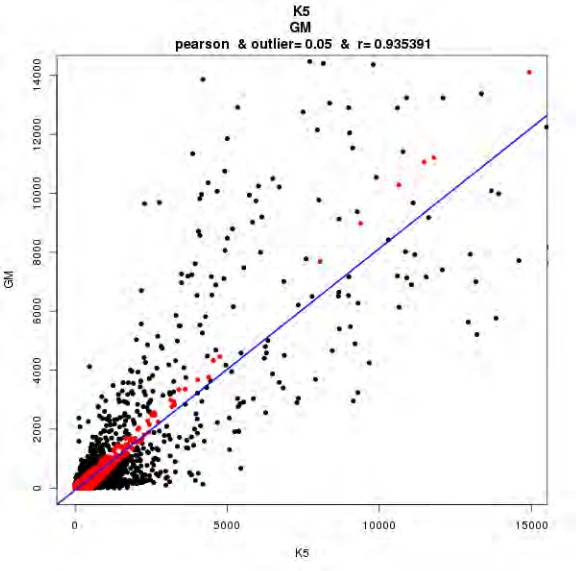


Figure 6.12 | Depiction of the correlateMatrices method
Here, two 5C matrices are being correlated. Each dot represents a pixel position (i,j) within the matrix. For each interaction, the interaction score from inputMatrix-1 is plotted on the X axis (labeled K5) and the interaction score from inputMatrix-2 is plotted on the Y axis (labeled GM). This scatter plot represents the relationship between the two variables. A linear regression is performed and the resulting fit is shown by the blue line. An outlier removal step (0.05 percentile removal) is selected and the black data points are flagged as outliers and ignored

from all analyses (linear fit and correlation analysis). Only the red data points are used. The resulting Pearson's R value is show on top of the plot (0.935).

coverageCorrect.pl - can perform coverage correction on matrix [balancing]

This script can perform a row/col balancing on a supplied matrix. Unlike the Hi-C balancing method which leverages genome-wide interaction data to apply the Sinkhorn-Knopp iterative balancing procedure, this script can perform more complicated balancing procedures which are normally reserved for non-genome wide or 5C datasets. This script has two main usage modes defined by the –cm option (correction mode). The user can choose to either use the CIS or the TRANS data to infer a visibility/performance score for each row/col. If the CIS mode is used, then matrix is first transformed into a z-score relative to an expected matrix calculated by the **matrix2loess.pl** script, and then the average of all z-score for each row/col is calculated. If the TRANS mode is used, then the average signal is calculated for each row/col across the entirety of the TRANS space. This average score is used a measure of how visible or how well each row/col performs in the experiment. If a specific row has an average ‘high’ score (compared to the entire row/col distribution), then one can assume that that specific row/col has a technical bias which causes it to have an elevated signal within the experiment.. If a specific row has an average ‘low’ score (compared to the entire row/col distribution), then one can assume that that specific row/col has a technical bias which causes it to have an low signal within the experiment. The goal of this balancing procedure is to normalize and equalize the signal for each row/col within the experiment. Each i,j interaction is corrected by the product of both the row and column factor (col.f * row.f). This method is iterative in nature and repeats the above steps until the procure either converges or meets a user-defined convergence limit. Of course this method makes a similar assumption to the Hi-C balancing procedure (ICE) in that each row/col should be equally visible in the experiment given that region is large enough and that a single biological interaction can and will not alter the region wide signal. Normally this assumption holds true as long as the region that one is normalizing is at least 1-2MB in size. For any smaller regions, any specific row/col can have elevated signal across the entire region as a result of a true biological signal (looping interaction). In most case, it is preferred to design experiments to sample from at least 1-2MB of the genome and then apply the Hi-C style balancing procedure.

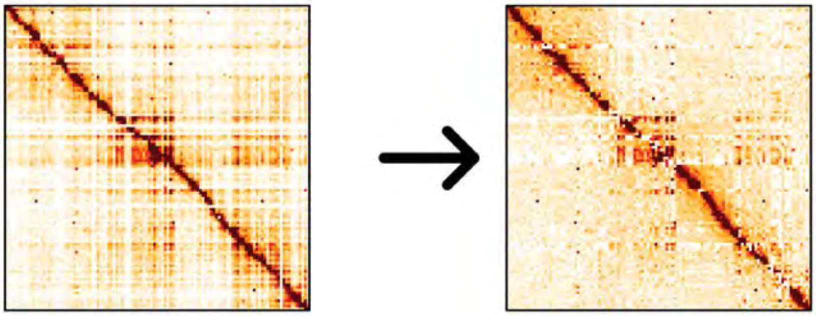


Figure 6.13 | Depiction of the coverageCorrect method.
Here, a 5C matrix is used as input to the coverageCorrect.pl script. Correction is performed using correctionMode=CIS and factorMode=ZSCORE. The resulting matrix has converged where the average z-score per row/col is < the convergence threshold (convergenceThreshold=0.05).

digitizePicture.pl - digitize picture into my5C matrix format

This script can digitize any PNG image into a cWorld formatted matrix. This script can be useful for restoring an interaction matrix from a heatmap image. The signal of each cell, can be calculated as either the sum or mean of either specific color values in the RGB range or the sum, mean of all colors.

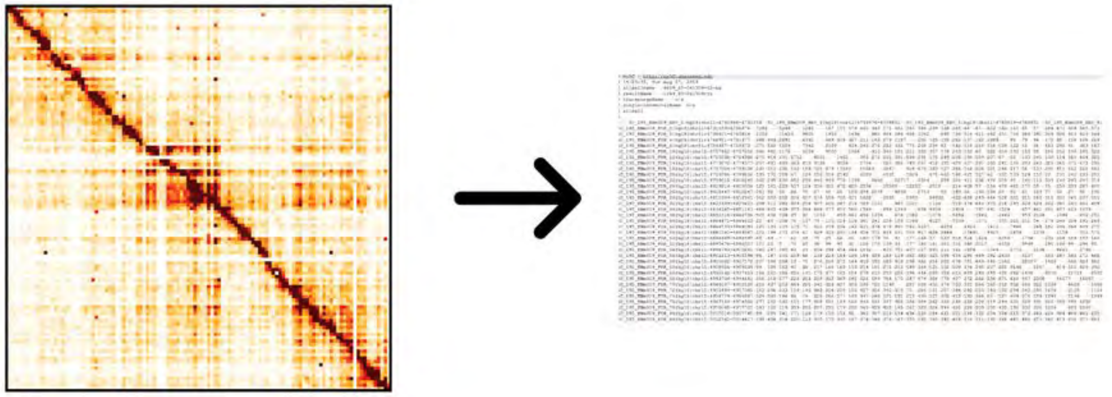


Figure 6.14 | Depiction of the digitizePicture method
Here, a 5C matrix is used as input to the digitizePicture.pl script. Here colorMode=mean and thus the average R,G,B values for each pixel is calculated and used as the relative interaction score. The produced cWorld tsv matrix file is shown on the right.

elementPileUp.pl - pile up cData around specified list of 'elements'

This script can aggregate the interaction data around a set of genomic elements. This script can be useful to determine whether or not the genomic structure is conserved around a set of genomic elements. For instance, if one were to assume that each bound CTCF protein in the genome caused a topologically association domain (TAD) to be formed, then if one were to aggregate all signal around each CTCF site, the resulting consensus structure should show two TAD structure on either side of the CTCF site. This script can be very useful during the initial data exploration phase. Once can 'pileup' the signal around any set of user specified elements. The script can also apply distance limits or change how the signal is aggregated to help reduce artifacts and noise.

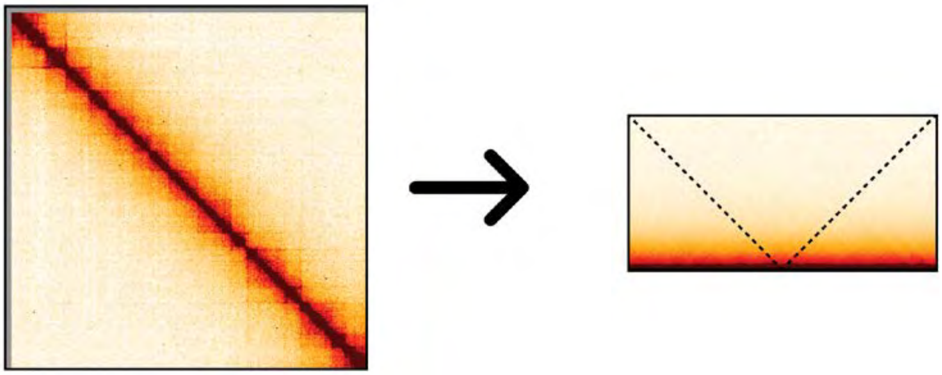


Figure 6.15 | Depiction of the elementPileUp method
Here, a Hi-C matrix is used as input to the elementPileUp.pl script along with a bed file containing the binding location of a specific protein / element. The Hi-C data around is element is gather and aggregated into the resulting matrix on the right.

extractSubMatrices.pl - extract sub matrices

This script can extract sub-matrices from a supplied matrix files. Sub-matrices can be defined either by chromosome, group (allele specification) or by name (5C region specification). For instance given a supplied matrix consisting of all interactions between chr1, chr2 and chr3, if the user were to selected extraction by chr, then 9 sub-matrix files would be created. 3 CIS matrices, consisting of chr1xchr1, chr2xchr2 and chr3xchr3 and 6 TRANS matrices, chr1xchr2, chr1xchr3,chr2xchr1, chr2xchr3, chr3xchr1 and chr3xchr2. The first denoted chromosome is plotted on the Y axis and the second denoted chromosome is denoted on the X axis. If the user selected the –eco (extract cis only) option, then only the CIS matrices will be produced. One can also subset the selection by genomic coordinates or by a list of genomic intervals. This script is efficient in terms of both memory and speed. Rows of the matrix are first written as ‘chunks’ (only those chunks that satisfy the user’s selection are written) Each chunk is then transposed and the procedure is repeated to achieve the desired result.

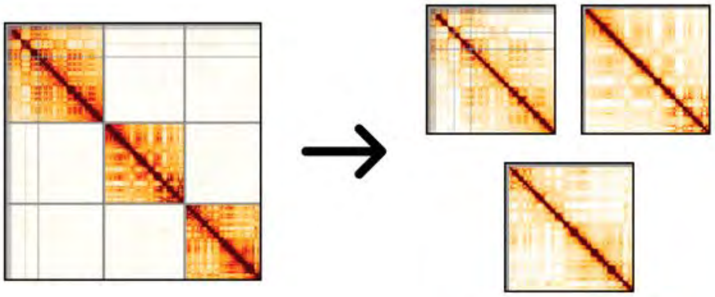


Figure 6.16 | Depiction of the extractSubMatrix method
Here, a 3 x 3 Hi-C matrix consisting of chr14, chr15 and chr16 is used as input to the extractSubMatrit.pl script. By select the –eco (extract cis only) option, only the CIS sub-matrices are extracted as seen on the right.

fillMissingData.pl - replace NAs with expected signals

This script can replace missing data (NaNs) with data samples from an expected distribution. The expected data distribution is calculated for each distinct genomic distance (distance between any two genomic loci in CIS). A random drawing from the expected distribution is used to replace each NaN value. This script can be useful to visualization purposes or various other custom analyses goals.

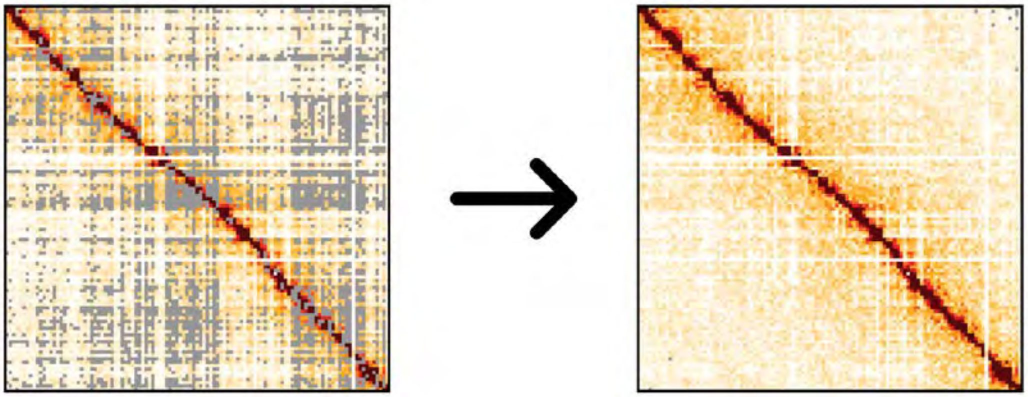


Figure 6.17 | Depiction of the fillMatrix method
Here, a 5C matrix is used as input to the fillMatrix.pl script. The NaNs are filled from a random sampling of the distribution calculated from the LOWESS function.

generateBins.pl - create my5C formatted headers

This script can generate a list of bins or genomic intervals from a list of fragment-level cWorld matrix headers.

heatmap.pl - draws heatmap PNG of matrix file

This script can transform an interaction matrix into a heatmap image representation. Each interaction score is linearly translated into a pixel color which represents the strength of the interaction. This script has multiple options that can be used to fine tune the resulting heatmap image. The –sfs (scale fragment size) option can be used to scale the pixel size by the row/col genomic loci size. In the case of binned data (fixed-size intervals), this option has no effect. In the case of fragment-level interaction data, where each row/col (header) corresponds to a different sized genomic interval representing the restriction fragment, this option would scale the pixel size in the heatmap image by the fragment size. Since each pixel is the intersection of the 'fragment' on the Y axis and the 'fragment' on the X axis, the pixel would be a rectangle with Y1 as the length and X2 as the width, where Y1 is a factor which represents the row (y-axis) fragment size and X2 is a factor which represents the column (x-axis) fragment size. The –dt (drawTriangle) option can output a rotated and cropped 'triangle' heatmap. The –dd (drawDiamond) option can output a rotated 'diamond' heatmap. The –em (embed meta data) option can embed metadata related to the input matrix file into the resulting heatmap image. The –dpb (drawPixelBorder) option can draw borders around every pixel in the heatmap. The –dl (drawLabel) option can write the headers for every row and column on the right and top of the heatmap image. The –ocb (omitContigBorder) option can omit lines drawn between all contigs/chromosomes/regions in the heatmap image. The –ds (drawScore) option can write the interaction score value within each pixel of the matrix. The –ps (pixelSize) option can control the x/y size of every cell in the heatmap image in terms of number of pixels. The –yps (y-pixelSize) option can control the y size of every cell in the heatmap image in terms of the number of pixels. The –xps (x-pixelSize) option can control the x size of every cell in the heatmap image in terms of the number of pixels. The –lt (logTransform) option can log transform the data before plotting in the heatmap, a user specified base is supplied after the –lt flag, e.g. –lt 2 is a log2 transformation. The –start (startColor) and –end (endColor) options are used to specify the range of colors (absolute value) that are to be visualized on the heatmap. A **start** of 0 and an **end** of 100 would color all score between 0 and 100, any scores outside of this range would receive either the lower bound or upper bound specified colors. The –startTile (start tile) and –endTile (end tile) options are used to specify the range of colors (relative value) that are to be visualized on the heatmap. A **start** of 0.25 and an **end** of 0.75 would color all score between the 25th percentile and the 75th percentile, any scores outside of this range would receive either the lower bound or upper bound specified colors. The –ebf (elementBedFile) option is a useful option to highlight specific row/cols that overlap a list of user specified elements. The –sm (scale mode) option is used to set the auto color scaling options, --sm combined pools all CIS and TRANS data before determine the color scale bar, --sm separate colors the CIS and TRANS data separately using two distinct color scale bars. The –pc (positive color) option controls the colors to use for the positive values in the matrix. –pc **white,red,blue** would color all positive colors from white to red to blue from the user specified **start** and **end** values. Named colors as well as RGBA codes can be supplied. For example **white,255.0.0.0,blue** would again color all positive colors from white to red to blue from the user specified **start** and **end** values. The –nc (negative color) option does the same as the above –pc option but for the negative values of the interaction matrix. The –mc (missing color) option controls the color to use for all missing data (e.g. NaNs). The –t (transparency) option controls the transparency to use for all colors in the heatmap [0-255].

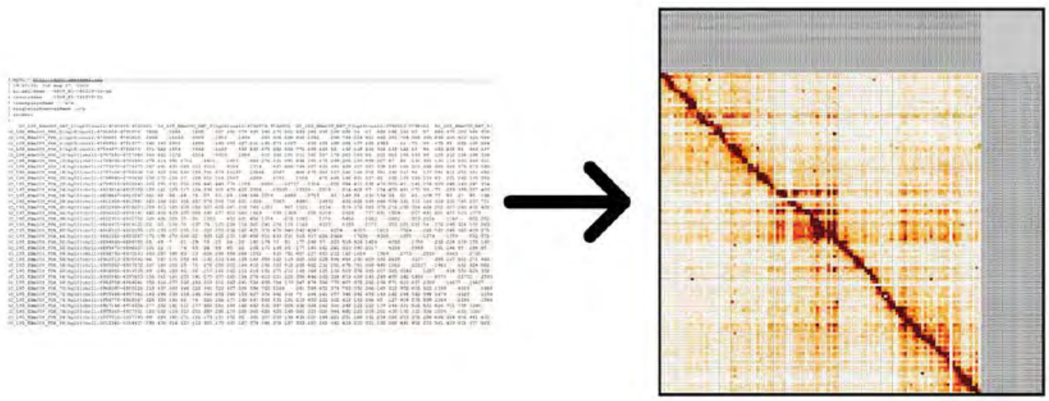


Figure 6.18 | Depiction of the heatmap method
Here, an example cWorld tsv matrix file is converted into a heatmap image, depicted on the right.

insulation2tads.pl - create tad specific headers

This script can in an insulation vector and a list of called boundaries calculated in the **matrix2insulation.pl** and translate them into a list of consecutive TADs. Consecutive TADs are defined as the space between any two called TAD boundaries. There are additional options once can employ to either limit the set of boundaries to use or to filter out genomic spans that contain missing data in the insulation vector. The strength of each TAD is defined as the difference between the **abs(max(insulation) – min(insulation))** for all insulation values within the TAD region (between two boundaries).

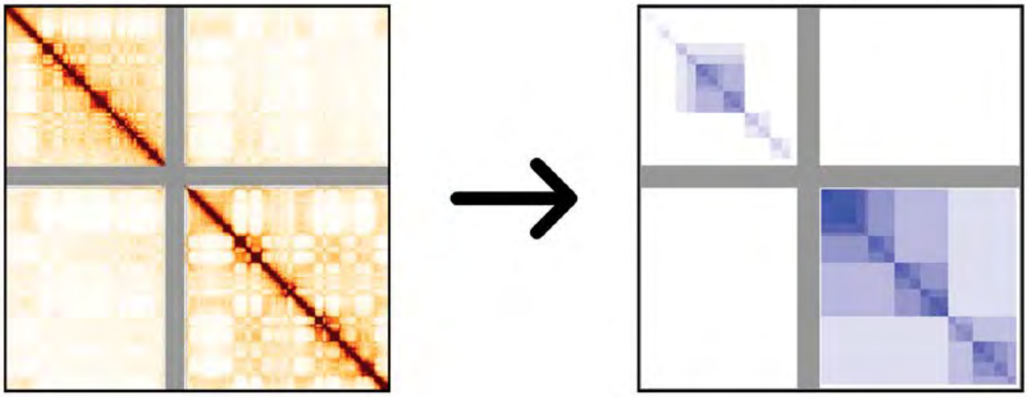


Figure 6.19 | Depiction of the insulation2tads method
Here, an example Hi-C matrix is used to first calculate both an insulation vector and a list of minima (boundaries). The insulation vector and boundaries as used as input to the insulation2tads.pl script and a set of nested TADs is inferred.

interactionPileUp.pl - pile up cData around specified list of 'elements'

This script can ‘pile up’ or aggregate interaction data between a set of element:element interactions. This is useful to determine whether or not a set of elements have a tendency to interact in 3D space.

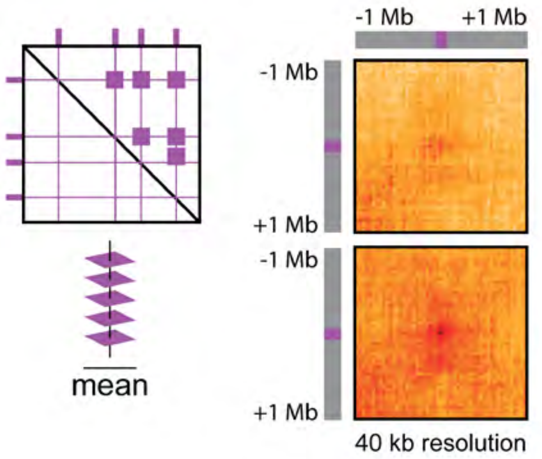


Figure 6.20 | Depiction of the interactionPileUp method
Here, the area surrounding a set of element:element interactions is extracted from an input matrix, aggregated and visualized in the two heatmaps on the right. The top heatmaps shows little interaction (clustering) between the elements, whereas the bottom heatmaps shows a much stronger (clustering) of the elements.

matrix2anchorPlot.pl - transform each row/col into 4C style 'anchor' plot.

This script can transform an interaction matrix into 3C/4C style ‘anchor’ plots. Since each row/col specifies all interactions within a specific genomic interval (e.g. the anchor), this script can create a plot per every row/col. This script can plot the expected signal per distance and the observed signal for every row/col.

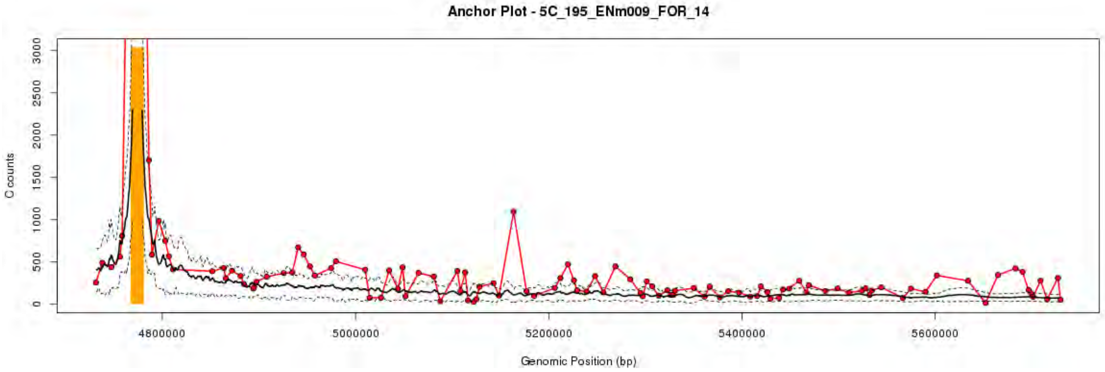


Figure 6.21 | Depiction of the matrix2anchorPlot method
The anchor bin/fragment is shown in orange. The red line shows the observed data extracted from the input matrix. The solid black line and dotted black line are from the LOWESS calculation run on the input matrix, the solid black line is the LOWESS mean (expected mean) and the dotted black line is the LOWESS stdev (expected stdev).

matrix2bed12.pl - transform matrix into bed12 format (track per row)

This script can transform an interaction matrix into a bed12 format which is useful for visualizing a set of interactions in the UCSC genome browser. It is useful to first subset or call significant interactions in the matrix before attempting to visualize all possible interactions in the genome browser.

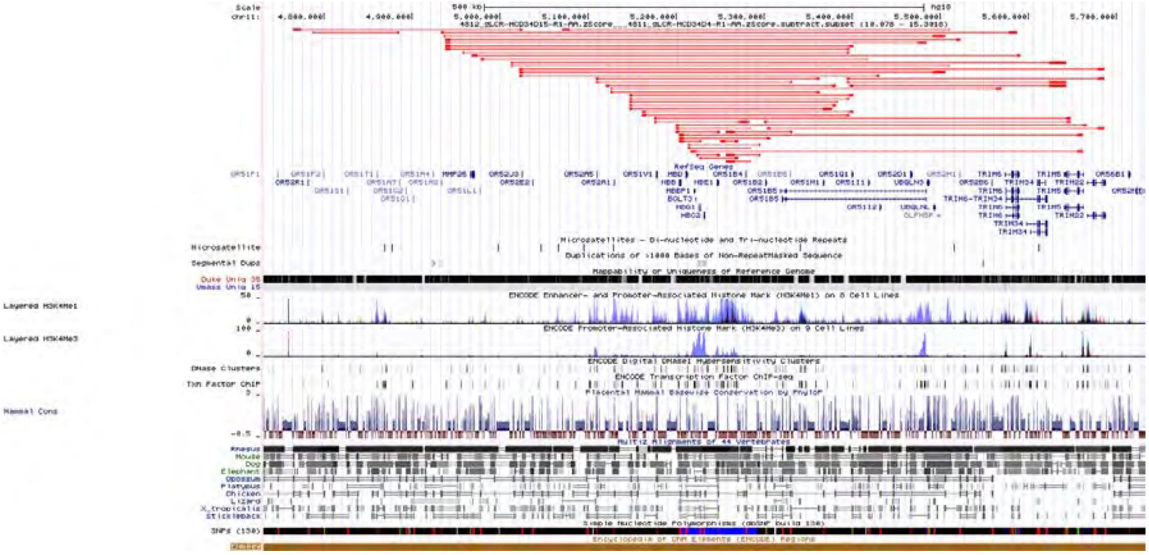


Figure 6.22 | Depiction of the matrix2bed12 method
In red, a set of 5C interactions (peak called) are visualized.

matrix2compartment.pl - perform PCA on input matrix

This script can perform a PCA analysis on a supplied input matrix. The PCA analysis is primary used as a proxy measure of the compartment signal found in almost all interaction matrices. Prior to running the PCA analysis, the input matrix is first transformed into a z-score matrix (via **matrix2loess.pl**). Then the z-score matrix is transformed into a correlation matrix (via **matrix2correlation.py**). The correlation matrix is then used as input to the sklearn.decomposition.PCA function and N components are calculated. The explained variance ratio of each component is output in a plot, and the eigenvalues of each bin along eigenvector 1 is used a measure of A or B compartment signal. Eigenvector1 is filliped so that the most gene rich compartment (positive or negative values) is positive. This ensures that the A compartment is always detected as the positive eigenvalues. The eigenvalues for eigenvectors 1 – 3 are plotted and the eigenvalues for eigenvector 1 are output in a bedGraph file for visualization in the UCSC genome browser. The positive values are colored red, and represent bins that are a member of A compartment (the active genomic compartment). The negative values are colored blue, and represent bins that are a member of the B compartment (the inactive compartment). Careful consideration must be applied this method and is described in detail in Chapter 4 of this thesis. PCA analysis is not guaranteed to detected and describe the active and inactive compartments. PCA analysis will only detect the source of the most variation in the matrix, which sometimes can the two arms of the chromosome.

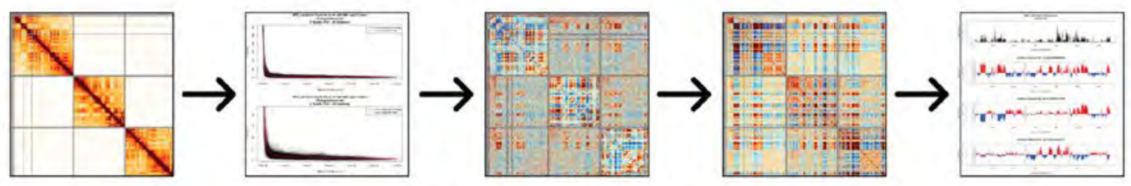


Figure 6.23 | Depiction of the matrix2compartment method
First (image 1), the input 3 x 3 Hi-C matrix is shown on the left. Next (image 2) depicts the LOWESS calculation run on the input matrix. Next (image 3) shows the z-score transformation of the input matrix. Next (image 4) shows the correlation matrix of the z-score matrix. Finally (image 5) shows the PCA eigenvector decomposition of the correlation matrix.

matrix2direction.pl - calculate directionality [tads] on matrix

This script can detect the directionality of every row/col (bin) within a matrix. The directionality measure, described previously [2], is a useful metric for detecting and summarizing TADs. The directionality score is defined as the log2 ration between the mean signal upstream of the bin, and the mean signal downstream of the bin. Calculating the directionality index for every bin along the chromosome creates a directionality vector. As one travels through the directionality vector and approaches a TAD boundary, the directionality index will rapidly shift from very positive to very negative, this transition point can be detected and inferred as a TAD boundary. The amount of the shift can be used as a proxy measure for the boundary 'strength'.

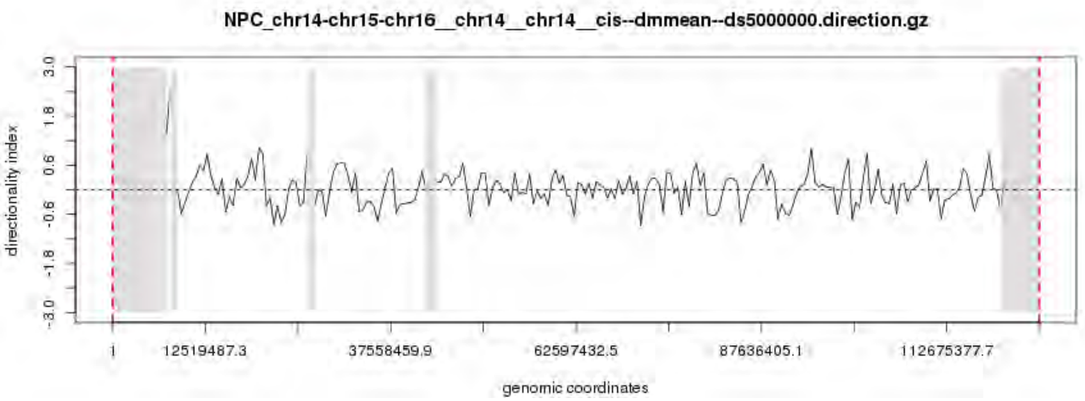


Figure 6.24 | Depiction of the matrix2direction method
A directionality index is calculated for every bin and visualized in the above vector (colored black).

matrix2distance.pl - cumulative reads versus distance

This script can translate a matrix into a pairwise tsv file with the following format. Column 1 = y-axis header, column 2 = x-axis header, column 3 = interaction distance between the two interacting loci. Interaction distance is defined as -1 if the interaction is between two different chromosomes (TRANS) or the midpoint between the two genomic intervals if the intervals are binned (fixed sized intervals) or the closest distance if the intervals vary in size (fragment level). This script also plots a useful metric, which is the cumulative signal per genomic distance. This plot can be used to infer quality of a interaction matrix (or experiment).

matrix2headerBed.pl - dump matrix headers as BED file

This script output a list of all matrix headers in BED format – useful for overlapping with other genomic element/signal tracks or integrating into the UCSC genome browser.

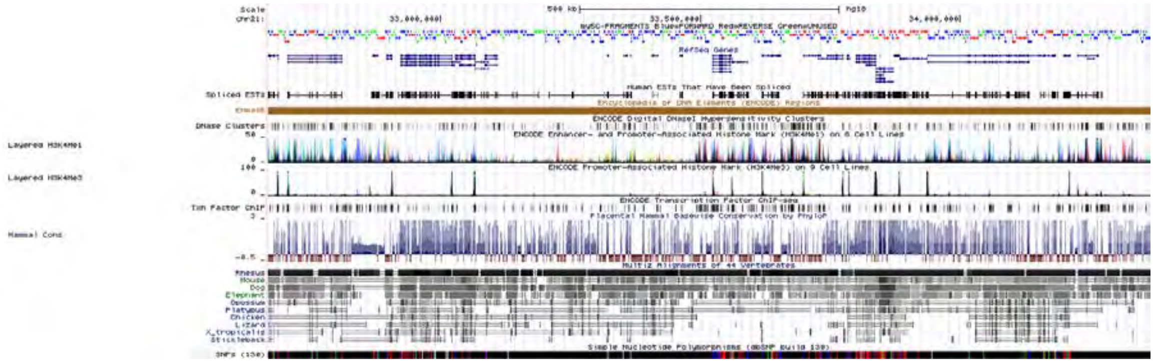

Scale
chr21:
500 kb
hg18
33,000,000
33,500,000
34,000,000
RefSeq Genes
Spliced ESTs
Human ESTs That Have Been Spliced
Encyclopedia of DNA Elements (ENCODE) Regions
DNase Clusters
ENCODE Digital DNaseI Hypersensitivity Clusters
Layered H3K4Me1
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
Layered H3K4Me3
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
Txn Factor ChIP
ENCODE Transcription Factor ChIP-seq
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
SNPs (130)
Simple Nucleotide Polymorphisms (dbSNP build 130)

Figure 6.25 | Depiction of the matrix2headerBed method
On top (red, green and blue) is a BED track visualizing the 5C primers designed for an example region.

matrix2info.pl - get matrix info

This script can be used to quickly asses various metrics contained within a supplied interaction matrix such as: number of contigs, percent of cis data, percent of trans data, sum of matrix etc.

matrix2insulation.pl - calculate insulation index (TADs) of supplied matrix

This script is used to calculate the insulation index of every row/col (bin) in a given interaction matrix. This script has been previously described in detail [46]. Briefly, a square is slid along the diagonal of the matrix. The size of this square is defined by the –is (insulation square size) option. The aggregate signal is then calculated according to the –im (insulation mode) option. This in essences assigns a singular value to each bin (row/col) within the matrix. The average of all insulation signals is calculated, and then each insulation signal is translated into a log2ratio (log2(insulation.i/mean(insulation)) where insulation.i is the insulation value for each bin and mean(insulation) is the mean signal for all insulation values. The resulting normalized insulation values are then plotted as a QC metric. A proper insulation vector should look smooth and contain large valleys followed by peaks throughout the entire chromosome. Valleys represent bins that have low interactions occurring across them. Peaks represent bins that have high interactions occurring across them. Valleys are inferred as TAD boundaries and peaks are inferred as the interior of a TAD. To detect peaks and valleys (minima and maxima) a method similar to the zero-derivative procedure is used. Briefly, the slope of the insulation vector (+/- --ids (insulation delta span)) is calculated. Every zero crossing of the slop represents either a PEAK or a VALLEY. Valleys are then detected and thus the TAD boundaries are called.

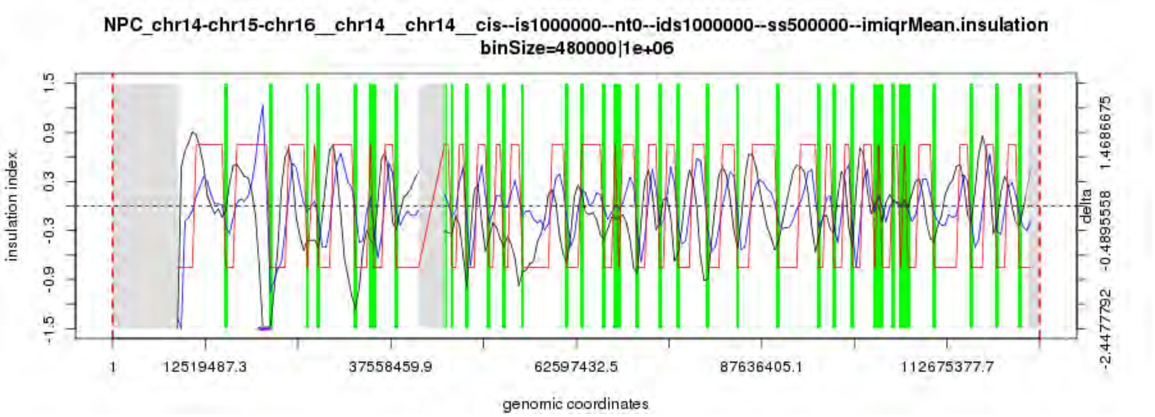


Figure 6.26 | Depiction of the matrix2insulation method
An example insulation plot is shown. In Black is the insulation vector. In Blue is the first derivative of the insulation vector. In red is a visualization of all zero-crossings of the blue line (derivative). In green are all detect minima (boundaries) of the insulation vector. Gray vertical bars represent areas with no data (NaN).

matrix2insulationRange.pl - calculate insulation index over range of square sizes

This script can calculate a series of insulation vectors for a range of –is (insulation square size). For instance, the insulation vector for all squares sized from 40,000 bp to 4,000,000 bp. This script is a useful metric for quantifying structure across various distance regimes.

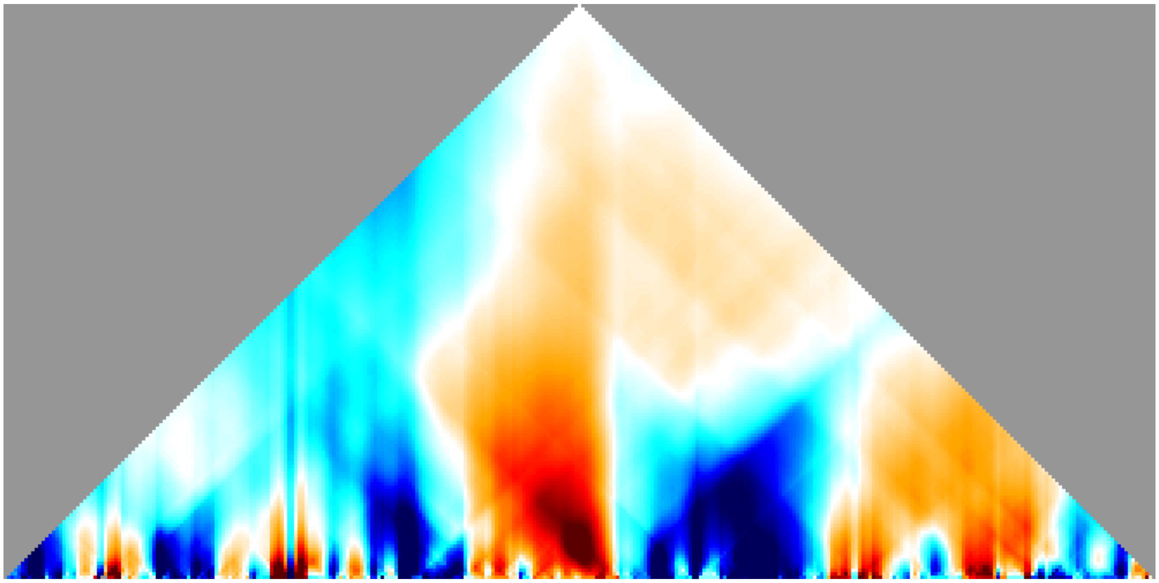


Figure 6.27 | Depiction of the matrix2insulationRange method
Here all possible insulation square sizes are calculated for an example matrix. Each row in the above heatmap correlated with an increasing insulation square size, starting at the bottom at 40kb, to the top at the maximal distance in the input matrix. Blue regions signify minima in the insulation vectors and can be inferred as regions of high insulation (boundaries). Red regions signify regions with high interaction or high local compaction (areas with low insulation).

matrix2loess.pl - calculate the loess (expected/stdev/zScore) for a given matrix

This script can calculate the expected and standard deviation signal for each distinct genomic distance contained within the input data. Rather than representing the expected signal per genomic distance as a mathematical model/function of genomic distance x genomic signal, I instead chose to use LOWESS to estimate the relationship between distance and signal. The LOWESS method [138] (Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting) is a "LOcal regrESSion" technique that utilized linear least squares regression analysis. LOWESS has one option which can drastically alter the performance of the fit, namely the ‘alpha’ parameter. This alpha parameter controls the amount of N closest data points to the anchor when performing the linear regression. For example, given a X/Y relationship within X being equal to the genomic distance and Y being equal to the observed interaction signal and the desire to calculate an average (or expected) Y for every X one can utilize the LOWESS method in the following manner. Using a an alpha parameter of 0.05 (or 5% of the total data points) would cause LOWESS to calculate a locally weighted linear regression of the 5% closest data points along the x axis for every distinct X value. Traditional the tri-cubic weighing function is used to determine weights for every data point, however one could alter the weight function as needed. LOWESS in its simplest form is quite sensitive to outliers, in order to remove outliers and produce a more representative fit of the input data, I adapted the original LOWESS method by adding an two-step procedure which includes an IQR outlier filter. This procedure encompassed two passes of the LOWESS algorithm through the entirety of data. The first pass works as previously described, producing both a regression value (weighted mean and a weighted stdev) for every X (expected mean signal and expected stdev signal for every genomic distance). During the second pass, only those **Y** points that fall within Q_1 - **1.5** * (**IQR**) > **Y** < Q_3 + **1.5** * (IQR) are used in the second pass linear regression. This additional steps creates a far more robust estimation of the relationship between the two variables. Once the modified robust LOWESS procedure is completed, various transformations of the input data can be calculated, such as a z-score transformation. A z-score is defined as $z = \frac{x - \mu}{\sigma}$ where x = observed data, μ = LOWESS mean of a specified genomic distance (X), and σ = LOWESS stdev of a specific genomic distance. This z-score transformation in effect normalizes out the distance dependency for every interaction. Various other metrics can be calculated such as the log2(observed/expected) or observed-expected. These transformations are calculated and each one is output in a separate matrix file for further use. One area of the LOWESS calculation that can be altered for reduce computation time is the fact that not every distinct X (genomic distance) needs to be represented in the final distribution. By leveraging the –caf (cis approximate factor) one can control the precision of the LOESS calculation by either setting the –caf = 1, which will calculate an expected signal for every distinct X, or by setting the –caf = 1000, which divides every genomic distance by 1000 and takes the floor. This in essence bins or clusters the data into discrete distance bins which in effect reduces the number of distance X that must be used in the LOWESS calculation.

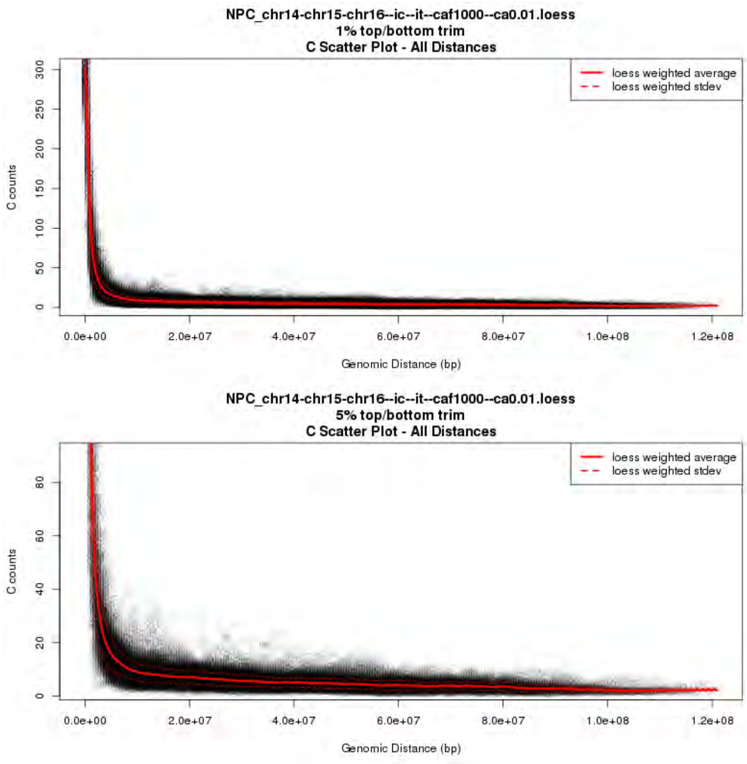


Figure 6.28 | Depiction of the matrix2loess method
Here an example scatter plot of interaction signal (labeled C counts) on the Y axis, and genomic distance on the X axis. The solid red line is the LOWESS expected value (weighted mean), the dotted red line above and below the solid line signify the mean +/- 1 stdev. Each black dot represent an interaction (pixel) in the input matrix, it’s location along the X axis signifies its genomic distance between the two genomic loci and its location along the Y axis signifies the observed interaction count.

matrix2pairwise.pl - transform tsv matrix into 3 column tsv file

This script can translate a matrix into a pairwise tsv file with the following format. Column 1 = y-axis header, column 2 = x-axis header, column 3 = interaction score between the two interacting loci. This transformation and resulting file format can be useful for integrating with other data types or interacting with various plotting methods.

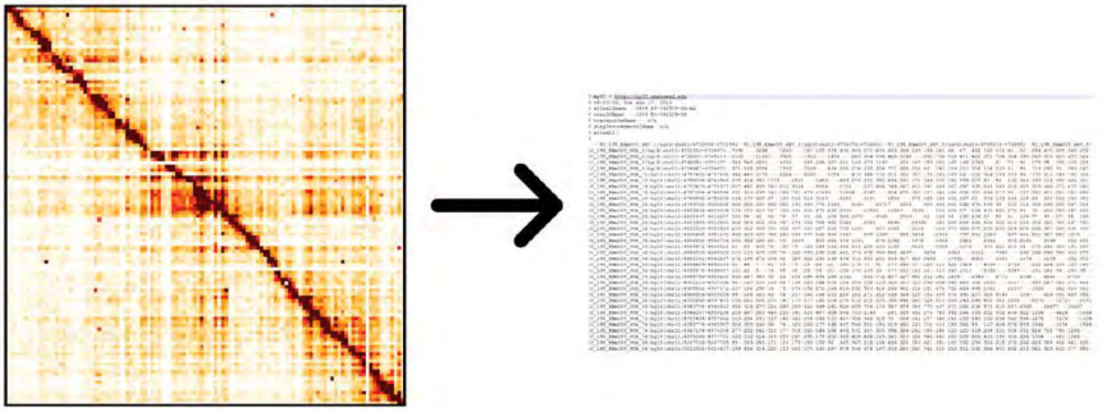


Figure 6.29 | Depiction of the matrix2pairwise method
Above a transformation from a cWorld tsv matrix file to a ‘pairwise’ (3 column tsv) file is visualized.

matrix2scaling.pl - transform matrix into scaling (polymer) plot

This script can transform and summarize a matrix into a ‘scaling plot’. This plotting technique is useful for inferring the state of the polymer as the log-transformed relationship between genomic distance and signal. The slope of line represents various theoretical polymer states and can be used to infer various biological models. The shape of this line can also be a useful metric in evaluating quality of a given experiment or similarity between biological replicates. This script can take as input N matrices and each matrix can be summarized into a single ‘scaling’ line and plotted together for visual comparison. The (**matrix2loess.pl**) and LOWESS method are used to estimate the relationship between distance and signal.

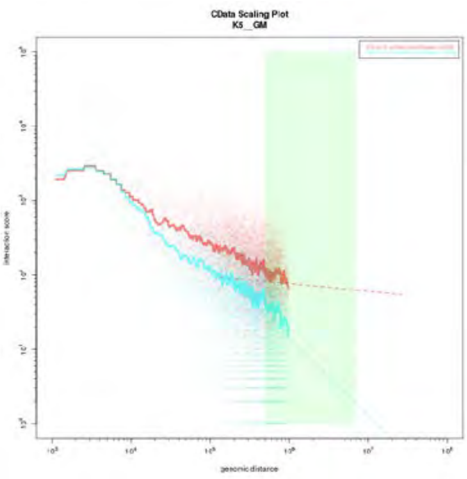


Figure 6.30 | Depiction of the matrix2scaling method
Two input matrices are transformed into ‘scaling plots’ above. One input matrix is show in red, the second is show in blue. The green vertical box represents the regime from 500kb – 7.5MB. The slope of each line is calculated for all points within the regime (green box) and is shown on the plot.

matrix2stacked.pl - transform matrix into stacked anchor matrix

This script transforms a matrix into a ‘stacked’ matrix. A stacked matrix is composed of linear stacks of each row centered on the diagonal bin. This van be visualize as taking a 1 x 21 row from the matrix, centered on each diagonal bin. These 1x21 rows are then stacked and output in a matrix format. This transformation is useful for visualizing structure relative to genomic distance, e.g. performing this transformation on all bins/rows that contain a gene.

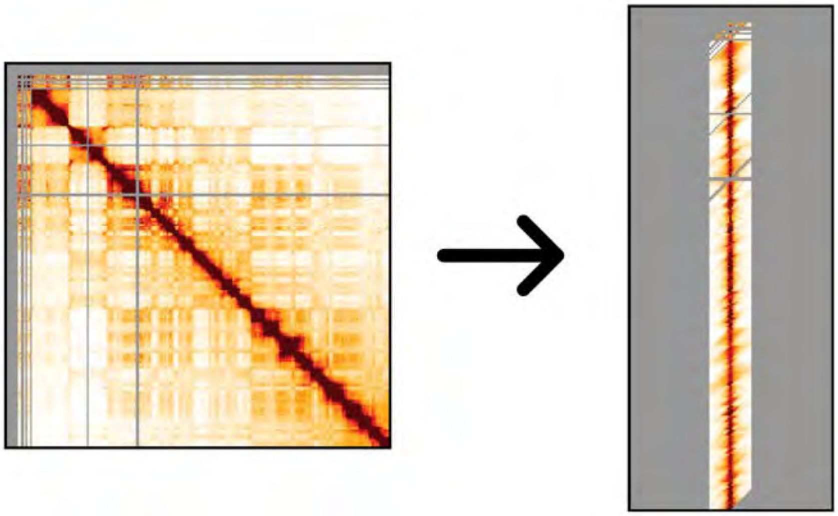


Figure 6.31 | Depiction of the matrix2stacked method
Here the matrix2stacked transformation is visualized with the input matrix on the left, and the stacked matrix on the right. Here only the first 2MB of interactions are show on the stacked matrix.

matrix2symmetrical.pl - transform rectangular matrix into symmetrical matrix

This script can transform a matrix into a symmetrical matrix. This operation is described as taking the union of all distinct row/col headers, producing a new matrix, and then filling in all available interaction data. Interactions between row/col headers that were not defined in the original matrix are replaced with a NaN signal. This transformation can be useful when a symmetrical matrix is required for specific analyses (e.g. Sinkhorn-Knopp).

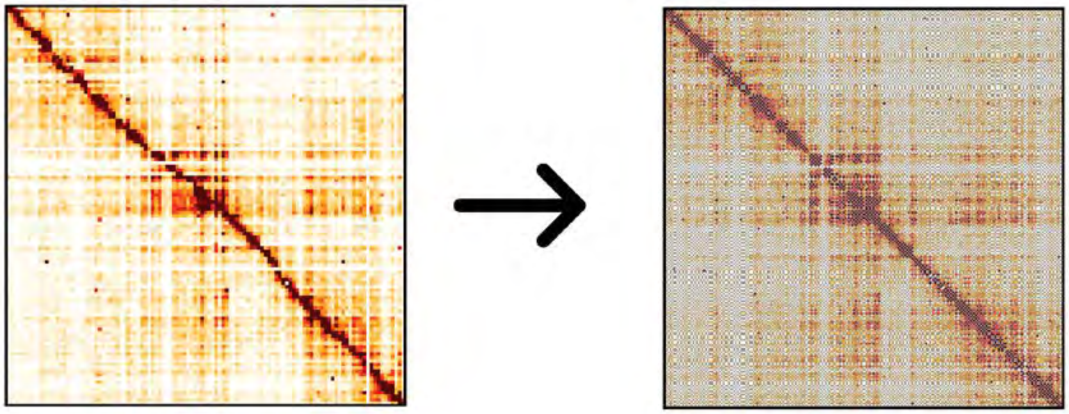


Figure 6.32 | Depiction of the matrix2symmetrical method
Here, the matrix2symmetrical transformation is visualized. The input non-symmetrical 5C matrix is show on the left, and the symmetrical form of this matrix is shown on the right. For the 5C input matrix, all FOR primers are show on the Y axis and all REV primers are shown on the X axis. In the symmetrical matrix, the union of all FOR/REV primers are shown on both axes. All signal between FOR:FOR and REV:REV are inferred as NaN (gray pixels).

matrix2webplot.pl - draws 'web-plot' of matrix file

This script can transform a matrix into a ‘webplot’ visualization. A webplot consists as two vectors which represent the row and col headers in the supplied matrix. The row header vector is color blue, and plotted at the top of the diagram from left to right, the column header vector is colored red and is plotted at the bottom of the diagonal from left to right. Interactions are then visualized as lines between any two points along the top (row) and bottom (column) vectors. The color of the line and or thickness of the line can be used as a measurement of signal intensity. This can serve as an important visualization technique to describe genomic interactions. Depending on how the input matrix is pre-processed (e.g. binarized to include only the significant interactions), various results/biological meanings can be visualized in the webplot.

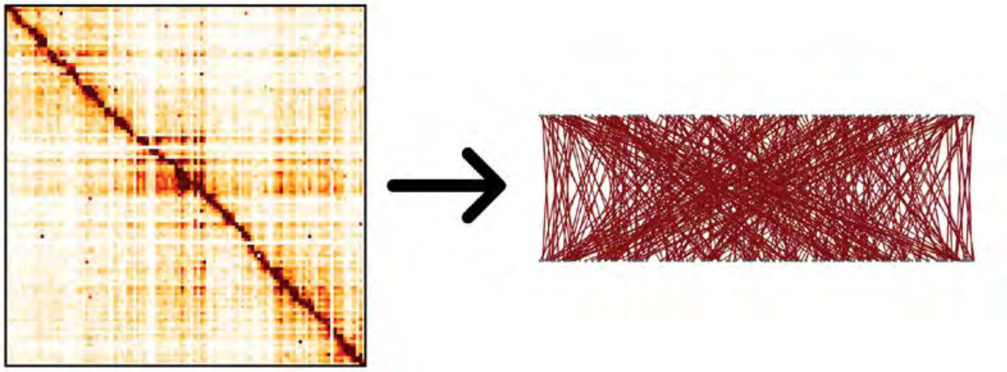


Figure 6.33 | Depiction of the matrix2webplot method
Here, the matrix2webplot transformation is visualized. All y-axis FOR primers (rows) are shown on the top of the webplot, visualized as small circles. All x-axis REV primers (cols) are shown on the bottom of the webplot, visualized as small circles. All interactions (FOR:REV) with an interaction score >= 50 are shown on the webplot as a solid red line.

primer2plates.pl - layout primers in 96-well plate format

This script can quickly transform a tsv list of 5C primers (to be ordered) into an Invitrogen formatted plate specification order form.

reOrderMatrix.pl - re-order matrix by list of headers

This script can re-order a matrix by a user specific list of ordered headers. This can be useful for clustering, or sorting an interaction matrix by any means desired. This script can also be used to remove or add specific row/cols. A user supplied –yohl (y ordered header list) file is used to set the ordering and composition of the row (y-axis) headers. A user supplied –xohl (x ordered header list) file is used to set the ordering and composition of the column (x-axis) headers.

scaleMatrix.pl - normalizes matrix sum - scales to 10^6

This script can scale a matrix to have a desired sum. By setting the –st (scaleTo) option, this script will scale all signal within the matrix to have the desired sum. This transformation can be used to normalize matrices for read depth before further analyses/comparison.

singletonRemoval.pl - detect and remove singleton outliers

This script can detect and remove singletons (single pixels) within a matrix that have a higher than expected signal. In the case of 5C experiments, it is impossible to detect and remove PCR blowouts during the mapping steps (since every chimeric ligation product is a combination of two 5C primers). The sequence of these chimeric ligation products is identical in the case of distinct molecules or in the case of PCR duplications from a single molecule. These PCR duplicates come through into the interaction matrix and can be seen as pixels with a higher than normal signal. To detect and remove these interactions, the matrix is first transformed into a z-score matrix via the (**matrix2loess.pl**) script, and then all pixels with a z-score >= SZT are remove, where SZT is the user specified singleton z-score threshold. Separate singleton z-score thresholds can be set for both the CIS and TRANS data to produce the desired filtering. It can be extremely useful to remove all singletons early in the data processing steps to avoid adding noise which can obscure later downstream analyses.

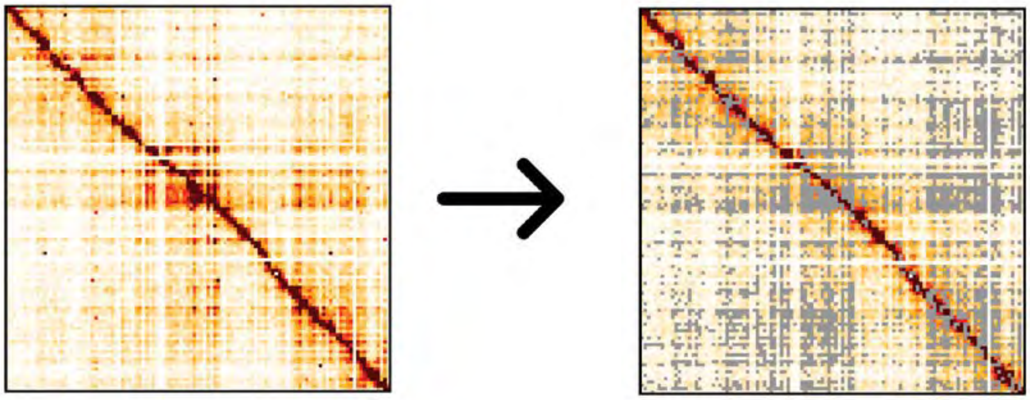


Figure 6.34 | Depiction of the singletonRemoval method
Here, the singletonRemoval method is visualized. First the input matrix is transformed into a z-score matrix (via matrix2loess), then any pixel with a z-score >=3 is removed and set to NaN (gray pixel).

subsetMatrix.pl - subset matrix by distance, or by BED file (bin overlap)

This script can subset an input matrix by various means. --minDist and –maxDist control the distances that are to be included in the subset matrix. Any interactions with distance x which fail to satisfy **minDist > x < maxDist** will be set to NaN. --lowerScore, --upperScore and --scoreSubsetMode control which interactions will be included in the subset matrix. When –scoreSubsetMode is set to outer, then scores which fail to satisfy **y < lowerScore or y > upperScore** will be set to NaN. When –scoreSubsetMode is set to inner, then scores which fail to satisfy **lowerScore > y < upperScore** will be set to NaN. –ec (excludeCis) can be used to exclude all CIS data. –et (excludeTrans) can be used to exclude all TRANS data. –ebf (elementBedFile) can be used to include only those bins (row/col) which directly overlap the list of elements found in the supplied element bed file. Multiple element bed files can be supplied to produce the desired effect. --yebf and --xebf can be used to subset the rows (y) and columns (x) separately by different element bed files. –z (zoomCoordinate) can be used to subset only those bins (row/col) which overlap genomic interval. Multiple genomic intervals can be supplied to produce the desired effect. --yz and --xa can be used to subset the rows (y) and columns (x) separately by different genomic intervals. By leveraging a variety of these options, users can subset input matrices by an almost unlimited means.

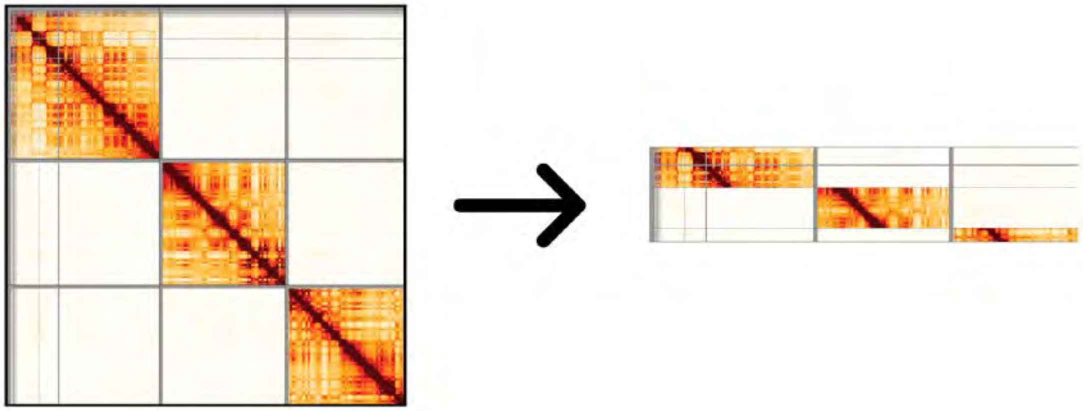


Figure 6.35 | Depiction of the subsetMatrix method
Here an assortment of multiple selections via genomic coordinate and bed file overlap are visualized. The input matrix is seen on the left and the final (subset) matrix is shown on the right.

symmetrical2seperate.pl - transform symmetrical matrix into non-symmetrical

This script performs the inverse of (**matrix2symmetrical.pl**), assuming the transformation was originally performed on a 5C cWorld matrix file. This script will place all FOR primers on the y-axis (rows) and all REV primers on the x-axis (cols).

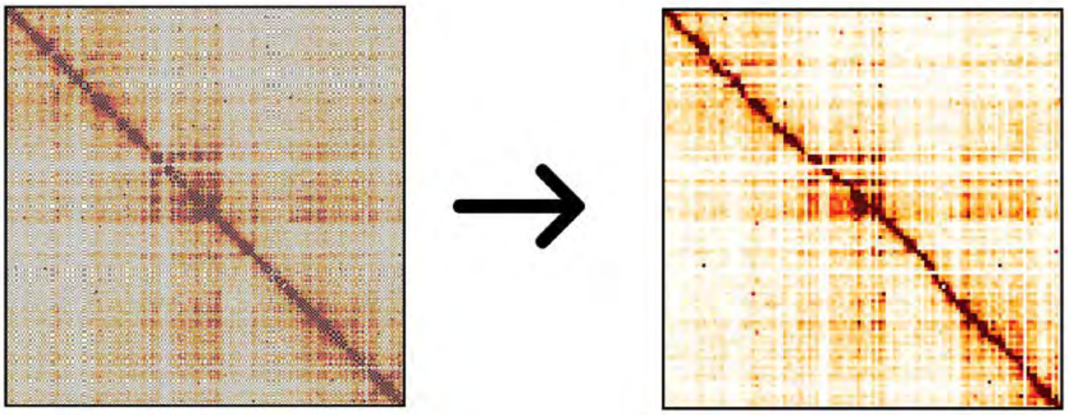


Figure 6.36 | Depiction of the symmetrical2seperate method
Here, the matrix2symmetrical transformation is visualized. This transformation is the inverse of matrix2symmetrical.

Conclusions

cWorld has become quite robust over the past few years and will only continue to improve with continued use and development. cWorld has now been used in several high impact publications. As further analyses and tools are added and become automated, the speed at which a 3C, 5C or Hi-C experiment can be processed, analyzed and biological significance inferred will continue to improve. To continue to expand and improve the cWorld toolkit not only must additional tools be added, but specific protocols must be improved and adapted to new file formats and advanced processing techniques.

CHAPTER VII: Conclusions and future directions

Preface

This conclusions chapter is partially adapted from a review written by Noam Kaplan, Job Dekker and myself entitled “The Hitchhiker’s Guide to Hi-C Analysis: Practical guidelines” [86], as well as the discussions sections of Chapter II, II, IV, V and VI.

Introduction

The genome structure field has grown exponentially over the past few years, mainly due to the increased availability of NGS. Prior to NGS, less discriminate and more specific methods such a microscopy, FISH or even PCR-based 3C were used to gain insights into the organization and structure of genomes. In only a short period of time, NGS technologies have grown from yielding a few hundred thousand short reads to being able to produce billions of long reads in less time and for less money. This advancement has unlocked cutting edge research to thousands of scientists worldwide and has driven the development of hundreds of genome-wide functional assays and thousands of computational methods to analyze this new and exciting data.

This thesis first introduced 3C based methods (3C-Seq, 5C, Hi-C etc), and then applied the methods to gain insights into the relationship between genome structure and function in the context of two dosage compensation systems.

Then, insights into the long range interaction landscape of genes and enhancers across a panel of ENCODE cell lines was gained by leveraging a targeted 5C based study. These data demonstrated and characterized a set of significant looping interactions between genes and enhancers, constructed a network of regulatory elements and provided insights into the roles of insulator proteins (e.g. CTCF) in the context of controlling gene expression. During this process, multiple novel processing, analysis and visualization methods have been developed and published. These methods aim to lower the bar needed for researchers to be able to perform, analyze and interpret genome structure data when applied to specific biological contexts. This thesis has discussed the necessary considerations one should make when processing, analyzing and interoperating genome structural data. This thesis has also introduced and discussed a set of tools for processing, manipulating, analyzing and visualizing genome structural data. Taken together, the insights gained from the work described in this thesis have made a significant contribution to the complex relationship between genome structure and genome function.

Worm Dosage Compensation

The results of Chapter II support the model that TAD structure on the X chromosome mediated by DCC binding to rex sites creates a 3D topology that acts chromosome-wide to repress gene expression. Given that changes in TAD boundaries occur locally, while changes in gene expression occur chromosome-wide, a parsimonious model posits that DCC-dependent changes in X chromosome structure imposed by rex–rex interactions drive the chromosome-wide reduction in gene expression. Potential DCC-dependent nuclear positioning of the X chromosome might also affect gene expression, as speculated by others [54].

In summary, DCC-induced formation of TAD structure on the X chromosome demonstrates a striking remodeling of chromosome topology that reveals a central role for condensin in shaping the 3D landscape of interphase chromosomes. Not only does condensin compact and resolve mitotic and meiotic chromosomes, it acts as a key structural element to regulate gene expression. No other molecular complex or set of DNA binding sites is yet known to cause comparably strong effects on megabase-scale TAD structure in higher eukaryotes [55]–[57]. The new understanding of the topology of dosage-compensated chromosomes provides fertile ground to decipher the detailed mechanistic relationship between higher-order chromosome structure and chromosome-wide regulation of gene expression.

Mouse Dosage Compensation

The study described in Chapter III reveals that the inactive X chromosome is a surprisingly elaborate entity, with a global partitioning into two mega-domains and loss of TAD organization, except at clusters of genes that are still expressed from the otherwise silent Xi. TADs were previously thought to be highly stable across cell generations and differentiation [2], [3], and their presence or maintenance not to require transcription in general. However, our study demonstrates that 1) TADs can indeed be lost in some contexts (as also observed on mitotic chromosomes [72], although in the case of the Xi, TAD loss is not a transient state but is stably transmitted through cell division) and that 2) gene expression and/or binding of factors such as CTCF can enable their maintenance and/or *de novo* re-creation. The findings show that gene silencing and loss of accessibility is accompanied by loss of structure, but that *de novo* gain of escape corresponds to re-creation of local structure, and further that transcription at clusters of genes coincides with TAD formation. Together these findings suggest that gene expression and DNA binding factors may be the driving forces of TAD organization in the context of the inactive X, which is otherwise devoid of TADs. The Xi may therefore represent a sequence-independent chromosome state at the structural level, from which sequence specific TADs can arise.

The reduced level of facultative escape in cells where the mega-domain has been deleted is intriguing. Although escape can be quite variable even in normal cells, three NPC clones derived from the D9 ΔFT mutant ESC line showed reduced escape by RNA FISH. These results suggest that during XCI the mega-domain boundary and the bipartite folding of the Xi that it induces, may modulate or affect the process leading to facultative escape. Constitutive escapees are much less affected by the boundary deletion and presumably have an intrinsic capacity to override the XCI process [80]. Facultative escapees on the other hand are first silenced during XCI and then re-expressed ([81], [82] and unpublished data). Although the mega-domain boundary region does not appear to interact with escapee regions in NPCs and is transcriptionally silent in NPCs, this region is transcribed and possibly euchromatic at the onset of XCI (MA and EH, unpublished observations). Transient interaction of this region with facultative escape loci during differentiation may thus occur and may be sufficient to regulate the local amount of escape and/or re-establish TADs at escape loci due to its unusual chromatin status and atypical enrichment in CTCF binding [83]. An additional, but not mutually exclusive, model is that the boundary region helps position the Xi in a particular sub-nuclear location during or after XCI, that facilitates the establishment of a given escape pattern. These results establish the Xi as a powerful model system for studying the mechanistic interrelationships between chromosome conformation and gene regulation, and point to a key role for gene activity in the establishment of chromosome structure at the level of TADs in the context of facultative heterochromatin.

Landscape of gene promoters

The data in Chapter V provide new insights into the landscape of chromatin looping. Here, the results demonstrate that physical chromatin looping can bring genes and distant elements into close spatial proximity. Besides generating a rich dataset reflecting specific gene-element associations, the average interaction profile of TSSs with surrounding chromatin reveals several general principles regarding the asymmetric relationships between genomic distance, the order of elements, and the formation of looping interactions. The bias for upstream interactions may indicate that the protein complexes on many TSSs may be asymmetric and may preferentially interact on one side with enhancer-protein complexes approaching along the chromatin fiber, as would be proposed by the enhancer tracking model [136]. Furthermore, while these average looping profiles may facilitate computational prediction of long-range interactions throughout the genome, the fact that interactions skip genes and CTCF sites suggests that additional mechanisms for target selection and gene insulation exist.

With further 3C technology development and increases in sequencing capacity, similar high-resolution studies should become feasible to map specific long-range interactions throughout the genome, which may uncover additional principles that guide chromatin looping. Such insights will also be critical for interpreting genome-wide association studies that often identify regions with regulatory elements but not their distally located target genes.

Practical Guidelines

As discussed in chapter IV of this thesis there are many considerations regarding the design and analysis of genome structural projects. Given the only very recent development of methods to probe the three-dimensional genome, many new analysis tools and methods will be developed and enhanced in the coming years, however these guidelines and principles should apply to even variations of the original Hi-C method and may be applicable to other similar methods (such as ChIA-PET).

Before starting a genome structure experiment, it is important to first, carefully consider the desired resolution of the data. Depending on the experimental goals, one must carefully choose between either a 5C or Hi-C (or a hybrid capture / targeted enrichment strategy). The space of all possible interactions, which is surveyed by Hi-C experiments, is very large. For example, consider the human genome. Using a 6-bp cutting restriction fragment, there are almost 10^6 restriction fragments, leading to an interaction space on the order of 10^{12} possible pairwise interactions. Thus, achieving maximal resolution is a significant challenge without adequate sequencing depth. To adequately cover a genome-wide Hi-C experiment at high resolution (5 kb) one may require billions if not tens of billions of mapped reads. However if one is interested in only a specific loci of the genome (say 1 MB in length) and given the same requirement of high resolution (5 kb), one may only require tens of millions of reads.

In light of this, it is crucial to establish the goals of the experiment, meaning whether one is most interested in either large-scale genomic conformations (e.g. genomic compartments) or specific small-scale interaction patterns (e.g. promoter-enhancer looping). If the goal is to measure large scale structures, such as genomic compartments, then a lower resolution will often suffice (1MB-10MB). Here, Hi-C using a traditional 6bp-cutting enzyme could be used. However, if the goal is to measure at a finer scale the very specific interactions of a small region, e.g. an enhancer of <500bp, then one should choose to use a restriction enzyme that cuts more frequently (e.g. 4bp) and a method that does not measure the entire genome, but instead focuses on exploring only a subset of the genome (i.e. 3C/4C/5C).

In Hi-C the maximum resolution of a dataset is determined by several factors, first and foremost is the sequencing depth. Given increasing amounts of reads, one will cover more of the interaction space and thus improve the resolution.

Library complexity is another factor. Library complexity is defined as the total number of unique interactions that exist in the Hi-C library. A library with a low complexity level (low number of unique interactions) will saturate quickly with increasing sequencing depth e.g. less and less information will be gained from additional sequencing. The saturation curve can be estimated from a dataset by plotting the cumulative number of unique interactions seen versus read depth.

Chapter IV has also touched upon interpretations of the data type and methods by which one can extract biological information. A key measurement of the quality of a Hi-C experiment is the percent of cis reads (e.g. the number of interactions between the same chromosome). Normally the percent of reads which are CIS is between 60–80. A high CIS percent is normally correlated with a high(er) quality Hi-C datasets. This is for obvious reasons, assuming no crosslinking and only random ligation, one would expect most of the reads to fall in TRANS. This is for the sample reason that at least for human/mouse, the number of possible interactions in TRANS is much larger than the number of possible interactions in CIS. With random ligation, the TRANS space should contain more reads. Therefore in a way, the percent of reads in CIS is a proxy measurement for the perceived percent of random ligation occurring in an experiment.

The second feature of the data is the distance-dependent decay of interaction frequency. In other words, interaction frequency between loci in cis decreases, on average, as their genomic distance increases. In the interaction matrix this pattern appears as a gradual decrease of interaction frequency the further one moves away from the diagonal. This pattern may be due to random movement of the chromosome, following the intuition that loci which are nearby in the genome will interact frequently if they move randomly in 3D space. The theory underlying this type of intuition is well established in the field of polymer physics [99], [100]. Depending on the experimental goals, one may wish to remove or normalize away this distance dependent decay, to then better highlight interactions that may be significantly higher or lower than their expected signal given their genomic distance.

A third feature of the data is the *genomic compartments* [1]. This interaction pattern appears on the interaction matrix as a “checker-board”-like pattern consisting of alternating blocks, ~1-10 mb in size, of high and low interaction frequency. This interaction pattern can be explained by a simple underlying phenomenon where chromosomes are composed of two types of genomic regions that alternate along the length of chromosomes and where the interaction frequencies between two regions of the same type tend to be higher than interaction frequencies between regions of different types. We refer to these two types as A and B compartments [1].

Analysis and characterization of the genomic compartments can highlight regions of either an active or inactive state. Changes in the compartments between two samples can uncover large scale differences between the expression state of the samples. One can use this very simple comparison to quickly pinpoint regions that may show quite different expression levels between the genes contains within the regions of compartment difference.

A fourth feature of the data are the topologically associations domains (TADs) While genomic compartments are useful for understanding general organization principles of the genome, many biological processes occur at a smaller scale. Specifically, enhancer-promoter interactions that underlie gene regulation in metazoans often take place at sub-Mb distances. Recently, 3C-based techniques have revealed the existence of sub-Mb structures that are referred to as *topologically associating domains* or *TADs* [2]–[5]. TADs are contiguous regions in which loci tend to interact much more with each other than with loci outside the region. In the interaction matrix TADs appear as square blocks of elevated interaction frequency centered on the diagonal. These domains have been shown to be associated with gene-regulatory features and it is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with enhancers [7], [102], [103]. In addition, TAD-like structures of various sizes have been observed in species ranging from mammals to bacteria [2]–[5], [104].

The block-like structure of TADs clearly indicates elevated interaction frequency within a TAD. However, given that we measure a population average and the observed intricate hierarchies of such structures, interpretation of TADs is not straight-forward. It has been proposed that TAD-like structure may be driven at least in part by looping interactions between loci located within them [105] or by supercoiled plectonemes [104], [106]. Additionally, some genomic features such as CTCF and cohesin binding have been shown to be enriched at TAD boundaries [2], [11]. It remains unclear what physical structures TADs exactly represent and how they are specified in the genome.

The fifth and final feature of the interaction matrix is a point interaction. While TADs may be relevant for constraining promoter-enhancer interactions, the actual regulatory interactions are probably of much smaller scale. Ultimately, protein-mediated interactions of two localized genomic elements, e.g. enhancers and promoters, which are typically up to a kb in length, can activate the expression of a gene. Given sufficient resolution, we expect such point interactions to appear as a local enrichment in contact probability. Point interactions have been discussed extensively in Chapter V of this thesis. Hi-C methods currently do not support a high enough resolution (< 1 kb) that would be necessary to detect and annotation functional point interactions between say a promoter and an enhancer. With added sequencing depth coupled with enrichment strategies (5C, hybrid capture), one can focus the sequencing power on a subset of the genome and thus increase the resolution.

cWorld

cWorld has become quite robust over the past few years and will only continue to improve with continued use and development. cWorld has now been used in several high impact publications. As further analyses and tools are added and become automated, the speed at which a 3C, 5C or Hi-C experiment can be processed, analyzed and biological significance inferred will continue to improve. To continue to expand and improve the cWorld toolkit not only must additional tools be added, but specific algorithms must be improved and adapted to new file formats and advanced processing techniques. One specific focus is on the hdf5 file format that has been recently adapted to store Hi-C data. This file format features a hierarchical chunked storage scheme which allows rapid retrieval of specific data chunks. The user has control of the chunking strategy and can be fined tuned to fulfill the user's exact needs. All chunks are stored on disk and the entire matrix is never required to be loaded into memory. Interaction with the matrix object is completely abstract to the user via the h5py library in python. However, one can take advantage of the internal file structure to process the matrix chunk by chunk to limit the amount of time needed reading the disk. Each chunk can be loaded to memory, or even an entire stripe (multiple rows) can be loaded into memory and processed at the same time. A further feature of the hdf5 file format and the h5py library is its multi-read, single write implementation. This means that the hdf5 file can be accessed in parallel and computation can be sped up by using multiple cores/threads. Quite a few scripts have now been re-worked and re-factored in python with knowledge of the hdf5 data format to achieve incredibly efficient time and space cost. As cWorld continues to grow, it will need to move most of its heavy computation to a python environment with access to the hdf5 data file. Specific aspects of cWorld will not benefit largely from moving to python or having access to the hdf5 file, as these transformations are simple in nature and would not necessarily benefit from or be capable of being parallelized.

cWorld can handle matrices up to 30000x30000 in size. Since cWorld utilizes a dynamic sparse matrix storage format, the memory footprint does not necessarily scale with matrix size, instead it scales with the number of observed data points (excluding 0s or NaNs). With additional memory, cWorld could grow to handle matrices up to 100,000 x 100,000 in size. When necessary, cWorld will avoid storing the entire matrix file in memory. Instead only sub-sections of the matrices, or slices along the diagonal can be extracted and loaded into memory. With further development of the hdf5 file format, one could imagine chunking and storing the matrix not only by square n x n chunks, but instead in row, or column or even diagonal ‘chunks’. This could give the user rapid access into even more complicate transformations or slices from the supplied matrix file/object.

cWorld has served mainly as a platform to rapidly prototype and develop new methods aimed to infer biological function from genome structure data, usually stored in a matrix format. Even though cWorld has been designed in a modular fashion and abstraction has been a general theme throughout the entire code base, it could benefit from a major re-working or even a migration to python.

TADs and gene expression

Throughout my thesis work, my research has provided insights into the mechanisms of dosage compensation across two species. It has also become more clear that the specific structure of a chromosome is highly correlated the functional output of that chromosome. However, this connection is only a correlation, it has not yet been demonstrated whether TAD structure can cause function or function can create TAD structure. Experiments to further elucidate this relationship are underway now. By manipulating the genome elements that control and define a TAD and or experimenting with ways to control or shut down expression, insights into the relationship between structure and function can be gained.

In the case of the worm dosage compensation, the two hermaphrodite X chromosomes which are down-regulated by one half, seem to be highly structured and packaged into multiple MB sizes TADs along the entire length of the chromosome. When binding and the resulting function of the DCC complex is disrupted, the two hermaphrodite X chromosomes regain 1 full dose of X chromosome genes each (increase expression by 1 fold) and lose their prominent TAD structure. These results suggests that TADs must act as a sort of local insulator, insulating a gene to only the enhancers that are contained with the gene's TAD. If this process is tightly controlled, this could have the net effect of lowering a genes total output/expression. If a gene is allowed to sample all enhancers within the genome, then proper tight regulation of that gene may be distributed. Having a tightly regulated micro neighborhood of enhancers for each gene could facilitate tighter control of gene expression.

In the case of the mouse dosage compensation, one of the X chromosomes is inactivated and packaged into heterochromatin, whereas the other X chromosomes remains active and packaged just as the other autosomes are. The active X chromosome still shows prominent TAD structure and strong compartment signal, suggestive of normal gene expression and regulation. The inactive X chromosome is packaged into two massive domains (~90 Mb). Why the X chromosome is compacted into two chromatin domains instead of one is unknown. The two domains could be the result of a tethering of the boundary region to the nuclear periphery; however, further experimental evidence would be required to either rule this hypothesis as correct or false.

Future Directions

In the future, I predict that genome structure experiments will become just as common as RNA-Seq or Chip-Seq is today. In fact, given how rapidly new methods such as ATAC-Seq have grown in popularity, I suspect that this may happen sooner than most would think. Using ATAC-Seq as an example, ATAC-Seq signal contains protein/TF specific footprint patterns. This means that given enough depth, from a single ATAC-Seq experiment, one can detect all accessible regions of the DNA, and from the footprint patterns, one could infer physical binding of a multitude of proteins/TFs. This sort of assay which can infer multiple layers of additional information from a single experiment will be the future. Hi-C could serve as such an assay. Hi-C can be used to better assemble genomes. From ordering contigs, scaffolding, detecting translocations or breakpoints, measuring copy number variations or structural variants, detecting chromosome territories, measuring genome wide active and inactive compartments, detecting sets of nested TAD structures, detecting co-expressed clusters of genes or transcription factories, characterizing gene – enhancer looping interactions and so on and so forth. From a single Hi-C experiment, currently a wealth of information can be extracted. However, one must not forget that the genome structure field is still in it’s infancy, the amount of data that can be extracted or inferred will continue to grow and multiply as the years pass. Hi-C may become the go-to method for assembling cancer patient genomes. One could also envision using Hi-C data to infer gene expression. Given adequate high-resolution (< 1kb), one could imagine that expressed genes may have a unique topology or structure compared to inactive genes. From this observed structural difference, one could infer expression status. Or the same could hold true for protein / TF binding. Do specific proteins or TFs have a unique local topology or organization relative to other TFs? Is the neighboring DNA altered in any way? If so – this information could ultimately be used to infer TF binding. Once again, HI-C allows us to extract a wealth of information from a single genome wide experiment.

Given the advancement and availability of longer and longer reads (now up to 30kb via PacBio Sciences) or even up to 100kb via virtual long read technologies such as those offered by 10X Genomics or Illumina's Moleculo technology, one could envision developing a Hi-C variant which aims to capture multiple interactions in a single molecule. From this molecule, one could then detect a set of hundreds or even possibly thousands of DNA fragments that from a single cell, were all co-localized, co-occurring interactions. This added layer of information could be used to detect mutually exclusive or co-occurring events, both of which are currently masked given Hi-C's population average data.

One could also imagine devising a Hi-C variant which could track the progression of interactions. By this I mean given a single loci in the genome, measuring all other loci that it samples (in 3D space) through time. By leveraging clever, unique barcode delivery techniques over time, and constructing large concatemers of ligations junctions, it may be possible to capture interaction events through time. This would mean that one can now measure the dynamics of interactions through various biological processes such as mitosis, differentiation, gene expression, external stimuli and so on and so forth.

The Hi-C method could also benefit from mixing data from many different variations of the method. For instance, normally one would use a 6-bp cutting restriction enzyme to fragment the DNA prior to ligation to capture interacting DNA fragments. When instead a 4-bp cutting enzyme is used, a marked increase in the number of local (< 1MB) interactions is observed. Does the same hold true for other digestion strategies? Can the depth of data per genomic distance be altered simply by varying the digestion? If so, one could imagine performing multiple Hi-C experiments across a range of digestion levels and then computationally combining the data. This could potentially reduce the amount of reads needs for maximal resolution of a given datasets. To explain further, a 4-bp cutting enzyme may have 80% of all sequencing reads for all interactions <= 5MB. A 6-bp cutting enzyme may have 80% of all sequencing reads for all interactions <= 50MB. A 8-bp cutting enzyme may have a 80% of all sequencing reads for all interactions <= 200MB. By leveraging this observation and pursuing this further, one can imagine focusing the sequencing power on a specific distance regime. Then, depending on the goals of the experiment, the researcher could optimally target the regions of highest interest.

Hi-C can also benefit from additional controls or spike-ins to better measure the efficiencies of cross-linking, digestion, ligation etc. Without the proper controls, it is difficult to conclude a biological mechanism for the observed data over a simple technical artifact.

Conclusion

In conclusion, this thesis has attempted to make the case for the usefulness of studying the genome structure in the context of many different cellular functions. As the genome structure field continues to mature and grow, additional insights will be gained and this information will be leveraged to better understand and tease apart complex biological systems. This thesis has also attempted to outline and make clear a set of practical guidelines that one should follow when working with genome structure data. This thesis has also introduced the cWorld toolkit, a set of computation tools that implement the aforementioned guidelines and give users a set of powerful computational methods to process, analyze, visualize and infer biological meaning from genome structure data. Given the modular design of the cWorld toolkit, it is difficult to describe all possible workflows and or techniques one could employ to process genome structure data. By leveraging both the set of guidelines and the very modular / abstract design of the cWorld toolkit, custom analyses can be performed, which can help to demonstrate the usefulness of this new datatype.

And finally, as the cost of NGS continues to decrease, the ability to obtain billions, if not tens, or even hundreds of billions of paired end sequencing reads will become possible. This new level of sequencing depth, will unlock the ability to have extremely high resolution (possibly down to the single base pair) which I expect, will reveal an even further layer of genome organization and it’s implications on genome function.

BIBLIOGRAPHY

[1] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome.,” *Science*, vol. 326, no. 5950, pp. 289–93, 2009.

[2] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, “Topological domains in mammalian genomes identified by analysis of chromatin interactions,” *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[3] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, “Spatial partitioning of the regulatory landscape of the X-inactivation centre.,” *Nature*, vol. 485, no. 7398, pp. 381–5, May 2012.

[4] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, “Three-dimensional folding and functional organization principles of the Drosophila genome,” *Cell*, vol. 148, no. 3, pp. 458–472, 2012.

[5] C. Hou, L. Li, Z. Qin, and V. Corces, “Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains,” *Mol. Cell*, 2012.

[6] W. a Bickmore and B. van Steensel, “Genome architecture: domain organization of interphase chromosomes.,” *Cell*, vol. 152, no. 6, pp. 1270–84, 2013.

[7] J. Gibcus and J. Dekker, “The hierarchy of the 3D genome,” *Mol. Cell*, 2013.

[8] W. de Laat and D. Duboule, “Topology of mammalian developmental enhancers and their regulatory landscapes.,” *Nature*, vol. 502, no. 7472, pp. 499–506, 2013.

[9] W. Schwarzer and F. Spitz, “The architecture of gene expression: integrating dispersed cis-regulatory modules into coherent regulatory domains,” *Curr. Opin. Genet. Dev.*, 2014.

[10] D. Gorkin, D. Leung, and B. Ren, “The 3D genome in transcriptional regulation and pluripotency,” *Cell Stem Cell*, 2014.

[11] K. Van Bortle, M. Nichols, L. Li, and C. Ong, “Insulator function and topological domain border strength scale with architectural protein occupancy,” *Genome Biol*, 2014.

[12] W. Bickmore, “The spatial organization of the human genome,” *Annu. Rev. Genomics Hum. Genet.*, 2013.

[13] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, “The long-range interaction landscape of gene promoters.,” *Nature*, vol. 489, no. 7414, pp. 109–13, Sep. 2012.

[14] F. Jin, Y. Li, J. Dixon, S. Selvaraj, Z. Ye, and A. Lee, “A high-resolution map of the three-dimensional chromatin interactome in human cells,” *Nature*, 2013.

[15] W. Deng, J. Lee, H. Wang, J. Miller, and A. Reik, “Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor,” *Cell*, 2012.

[16] I. Krivega and A. Dean, “Enhancer and promoter interactions-long distance calls.,” *Curr. Opin. Genet. Dev.*, vol. 22, no. 2, pp. 79–85, Apr. 2012.

[17] S. Razin, A. Gavrilov, E. Ioudinkova, and O. Iarovaia, “Communication of genome regulatory elements in a folded chromosome,” *FEBS Lett.*, 2013.

[18] D. Vernimmen and M. De Gobbi, “Long -range chromo
regulate the timing of the transition between poised and active gene expression,” *EMBO J.*, 2007.

[19] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat, “Looping and interaction between hypersensitive sites in the active beta-globin locus.,” *Mol. Cell*, vol. 10, no. 6, pp. 1453–65, Dec. 2002.

[20] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing chromosome conformation.,” *Science*, vol. 295, no. 5558, pp. 1306–11, 2002.

[21] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).,” *Nat. Genet.*, vol. 38, no. 11, pp. 1348–54, Nov. 2006.

[22] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson, “Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.,” *Nat. Genet.*, vol. 38, no. 11, pp. 1341–7, Nov. 2006.

[23] J. Dostie, T. a Richmond, R. a Arnaout, R. R. Selzer, W. L. Lee, T. a

Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, “Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.,” *Genome Res.*, vol. 16, no. 10, pp. 1299–309, Oct. 2006.

[24] J. Dekker, “Gene regulation in the third dimension,” *Science (80-.).*, 2008.

[25] T. E. P. Consortium, “A user’s guide to the encyclopedia of DNA elements (ENCODE).,” *PLoS Biol.*, vol. 9, no. 4, p. e1001046, Apr. 2011.

[26] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, “An integrated encyclopedia of DNA elements in the human genome.,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[27] B. J. Meyer, “X-Chromosome dosage compensation.,” *WormBook*, pp. 1–14, 2005.

[28] N. Brockdorff and B. Turner, “Dosage compensation in mammals,” *Epigenetics*, 2007.

[29] C. Grimaud and P. Becker, “Form and function of dosage-compensated chromosomes–a chicken-and-egg relationship,” *Bioessays*, 2010.

[30] J. Jans, J. M. Gladden, E. J. Ralston, C. S. Pickle, A. H. Michel, R. R. Pferdehirt, M. B. Eisen, and B. J. Meyer, “A condensin-like dosage compensation complex acts at a distance to control expression throughout the genome.,” *Genes Dev.*, vol. 23, no. 5, pp. 602–18, 2009.

[31] R. R. Pferdehirt, W. S. Kruesi, and B. J. Meyer, “An MLL/COMPASS subunit functions in the C. elegans dosage compensation complex to target X chromosomes for transcriptional regulation of gene expression,” *Genes Dev.*, vol. 25, pp. 499–515, 2011.

[32] G. Csankovszki, K. Collette, K. Spahl, J. Carey, M. Snyder, E. Petty, U. Patel, T. Tabuchi, H. Liu, I. McLeod, J. Thompson, A. Sarkesik, J. Yates, B. J. Meyer, and K. Hagstrom, “Three Distinct Condensin Complexes Control C. elegans Chromosome Dynamics,” *Curr. Biol.*, vol. 19, no. 1, pp. 9–19, 2009.

[33] D. G. Mets and B. J. Meyer, “Condensins Regulate Meiotic DNA Break Distribution, thus Crossover Frequency, by Controlling Chromosome Structure,” *Cell*, vol. 139, no. 1, pp. 73–86, 2009.

[34] B. J. Meyer, “Targeting X chromosomes for repression,” *Curr. Opin. Genet. Dev.*, vol. 20, no. 2, pp. 179–189, 2010.

[35] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J.

Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome.,” *Science*, vol. 326, no. 5950, pp. 289–93, Oct. 2009.

[36] H. E. Dawes, D. S. Berlin, D. M. Lapidus, C. Nusbaum, T. L. Davis, and B. J. Meyer, “Dosage compensation proteins targeted to X chromosomes by a determinant of hermaphrodite fate.,” *Science*, vol. 284, no. 5421, pp. 1800–1804, 1999.

[37] A. Wutz, T. P. Rasmussen, and R. Jaenisch, “Chromosomal silencing and localization are mediated by different domains of Xist RNA.,” *Nat. Genet.*, vol. 30, no. 2, pp. 167–174, 2002.

[38] C. Chu, Q. C. Zhang, S. T. da Rocha, R. A. Flynn, M. Bharadwaj, J. M. Calabrese, T. Magnuson, E. Heard, and H. Y. Chang, “Systematic Discovery of Xist RNA Binding Proteins,” *Cell*, vol. 161, no. 2, pp. 404–416, 2015.

[39] S. B. Peeters, A. M. Cotton, and C. J. Brown, “Variable escape from X-chromosome inactivation: identifying factors that tip the scales towards expression.,” *Bioessays*, vol. 36, no. 8, pp. 746–56, 2014.

[40] J. Chaumeil, I. Okamoto, and E. Heard, “X-chromosome inactivation in mouse embryonic stem cells: analysis of histone modifications and transcriptional activity using immunofluorescence and FISH.,” *Methods Enzymol.*, vol. 376, pp. 405–19, Jan. 2004.

[41] E. Splinter, E. de Wit, and E. Nora, “The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA,” *Genes Dev.*, 2011.

[42] K. Teller, D. Illner, S. Thamm, C. S. Casas-Delucchi, R. Versteeg, M. Indemans, T. Cremer, and M. Cremer, “A top-down analysis of Xa- and Xi-territories reveals differences of higher order structure at ≥ 20 Mb genomic length scales,” *Nucleus*, vol. 2. pp. 465–477, 2011.

[43] C. Naughton, D. Sproul, C. Hamilton, and N. Gilbert, “Analysis of Active and Inactive X Chromosome Architecture Reveals the Independent Organization of 30 nm and Large-Scale Chromatin Structures,” *Mol. Cell*, vol. 40, no. 3, pp. 397–409, 2010.

[44] R. Eils, S. Dietzel, E. Bertin, E. Schröck, M. R. Speicher, T. Ried, M. Robert-Nicoud, C. Cremer, and T. Cremer, “Three-dimensional reconstruction of painted human interphase chromosomes: Active and inactive X chromosome territories have similar volumes but differ in shape and surface structure,” *J. Cell Biol.*, vol. 135, no. 6, pp. 1427–1440, 1996.

[45] S. Rao, M. Huntley, and N. Durand, “A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping,” *Cell*, 2014.

[46] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer, “Condensin-driven remodelling of X chromosome topology during dosage compensation,” *Nature*, vol. 523, no. 7559, pp. 240–244, Jun. 2015.

[47] Y. Zhang, R. P. McCord, Y.-J. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, and J. Dekker, “Spatial organization of the mouse genome and its role in recurrent chromosomal translocations.,” *Cell*, vol. 148, no. 5, pp. 908–21, Mar. 2012.

[48] B. D. Towbin, C. González-Aguilera, R. Sack, D. Gaidatzis, V. Kalck, P. Meister, P. Askjaer, and S. M. Gasser, “Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery,” *Cell*, vol. 150, no. 5, pp. 934–947, 2012.

[49] K. Ikegami, T. a Egelhofer, S. Strome, and J. D. Lieb, “Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2.,” *Genome Biol.*, vol. 11, no. 12, p. R120, 2010.

[50] T. Liu, A. Rechtsteiner, T. A. Egelhofer, A. Vielle, I. Latorre, M.-S. Cheung, S. Ercan, K. Ikegami, M. Jensen, P. Kolasinska-Zwierz, H. Rosenbaum, H. Shin, S. Taing, T. Takasaki, A. L. Iniguez, A. Desai, A. F. Dernburg, H. Kimura, J. D. Lieb, J. Ahringer, S. Strome, and X. S. Liu, “Broad chromosomal domains of histone modification patterns in C. elegans.,” *Genome Res.*, vol. 21, no. 2, pp. 227–36, 2011.

[51] P. McDonel, J. Jans, B. K. Peterson, and B. J. Meyer, “Clustered DNA motifs mark X chromosomes for repression by a dosage compensation complex.,” *Nature*, vol. 444, no. 7119, pp. 614–618, 2006.

[52] R. a J. Chen, P. Stempor, T. a. Down, E. Zeiser, S. K. Feuer, and J. Ahringer, “Extreme HOT regions are CpG-dense promoters in C. elegans and humans,” *Genome Res.*, vol. 24, no. 7, pp. 1138–1146, 2014.

[53] W. S. Kruesi, L. J. Core, C. T. Waters, J. T. Lis, and B. J. Meyer, “Condensin controls recruitment of RNA polymerase ii to achieve nematode X-chromosome dosage compensation,” *Elife*, vol. 2013, no. Pol II, pp. 1–31, 2013.

[54] R. Sharma, D. Jost, J. Kind, G. Gómez-Saldivar, B. van Steensel, P. Askjaer, C. Vaillant, and P. Meister, “Differential spatial and structural organization of the X chromosome underlies dosage compensation in C. elegans.,” *Genes Dev.*, vol. 28, no. 23, pp. 2591–6, 2014.

[55] V. C. Seitan, A. J. Faure, Y. Zhan, R. P. McCord, B. R. Lajoie, E. Ing-Simmons, B. Lenhard, L. Giorgetti, E. Heard, A. G. Fisher, P. Flicek, J. Dekker, and M. Merkenschlager, “Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments.,” *Genome Res.*, vol. 23, no. 12, pp. 2066–77, Dec. 2013.

[56] S. Sofueva, E. Yaffe, and W. Chan, “Cohesin -mediated interactio organize chromosomal domain architecture,” *EMBO J.*, 2013.

[57] J. Zuin, J. R. Dixon, M. I. J. a van der Reijden, Z. Ye, P. Kolovos, R. W. W. Brouwer, M. P. C. van de Corput, H. J. G. van de Werken, T. a Knoch, W. F. J. van IJcken, F. G. Grosveld, B. Ren, and K. S. Wendt, “Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 3, pp. 996–1001, 2014.

[58] R. S. Kamath and J. Ahringer, “Genome-wide RNAi screening in Caenorhabditis elegans,” *Methods*, vol. 30, no. 4, pp. 313–321, 2003.

[59] P. T. Chuang, D. G. Albertson, and B. J. Meyer, “DPY-27: A chromosome condensation protein homolog that regulates C. elegans dosage compensation through association with the X chromosome,” *Cell*, vol. 79, pp. 459–474, 1994.

[60] N. van Berkum and J. Dekker, “Determining spatial chromatin organization of large genomic regions using 5C technology,” *Chromatin Immunoprecipitation Assays*, 2009.

[61] J. Belton, R. McCord, and J. Gibcus, “Hi–C: a comprehensive technique to capture the conformation of genomes,” *Methods*, 2012.

[62] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, “Iterative correction of Hi-C data reveals hallmarks of chromosome organization.,” *Nat. Methods*, vol. 9, no. 10, pp. 999–1003, Oct. 2012.

[63] A. Sanyal, B. Lajoie, G. Jain, and J. Dekker, “The long-range interaction landscape of gene promoters,” *Nature*, 2012.

[64] H. Chen, D. D. Hughes, T. a Chan, J. W. Sedat, and D. a Agard, “IVE (Image Visualization Environment): a software platform for all three-dimensional microscopy applications.,” *J. Struct. Biol.*, vol. 116, no. 1, pp. 56–60, 1996.

[65] A. E. Friedland, Y. B. Tzur, K. M. Esvelt, M. P. Colaiácovo, G. M. Church, and J. a Calarco, “Heritable genome editing in C. elegans via a CRISPR-Cas9 system.,” *Nat. Methods*, vol. 10, no. 8, pp. 741–3, 2013.

[66] B. Chen, L. a Gilbert, B. a Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, and B. Huang, “Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system.,” *Cell*, vol. 155, no. 7, pp. 1479–91, 2013.

[67] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. a Hutchison, and H. O. Smith, “Enzymatic assembly of DNA molecules up to several hundred kilobases.,” *Nat. Methods*, vol. 6, no. 5, pp. 343–5, 2009.

[68] T.-W. Lo, C. S. Pickle, S. Lin, E. J. Ralston, M. Gurling, C. M. Schartner, Q. Bian, J. A. Doudna, and B. J. Meyer, “Precise and heritable genome editing in evolutionarily diverse nematodes using TALENs and CRISPR/Cas9 to engineer insertions and deletions.,” *Genetics*, vol. 195, no. 2, pp. 331–48, 2013.

[69] L. R. Baugh, J. Demodena, and P. W. Sternberg, “RNA Pol II accumulates at promoters of growth genes during developmental arrest.,” *Science*, vol. 324, no. 5923, pp. 92–94, 2009.

[70] T. D. Wu and S. Nacu, “Fast and SNP-tolerant detection of complex variants and splicing in short reads.,” *Bioinformatics*, vol. 26, no. 7, pp. 873–81, 2010.

[71] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome biol*, 2010.

[72] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker, “Organization of the mitotic chromosome.,” *Science*, vol. 342, no. 6161, pp. 948–53, Nov. 2013.

[73] J. Giacalone, J. Friedes, and U. Francke, “A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes,” *Nat. Genet.*, vol. 1, no. 2, pp. 137–143, 1992.

[74] B. P. Chadwick, “DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts,” *Genome Res.*, vol. 18, pp. 1259–1269, 2008.

[75] A. Minajigi, J. E. Froberg, C. Wei, H. Sunwoo, B. Kesner, D. Colognori, D. Lessing, B. Payer, M. Boukhali, W. Haas, and J. T. Lee, “A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation,” *Science (80-.).*, no. June, pp. 1–19, 2015.

[76] A.-V. Gendrel, M. Attia, C.-J. Chen, P. Diabangouaya, N. Servant, E. Barillot, and E. Heard, “Developmental Dynamics and Disease Potential of Random Monoallelic Gene Expression,” *Dev. Cell*, vol. 28, no. 4, pp. 366–380, 2014.

[77] M. Vietri Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur, “Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture,” *Cell Rep.*, vol. 10, no. 8, pp. 1297–309, 2015.

[78] E. G. Schulz, J. Meisig, T. Nakamura, I. Okamoto, A. Sieber, C. Picard, M. Borensztein, M. Saitou, N. Blüthgen, and E. Heard, “The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating the ESC signaling network.,” *Cell Stem Cell*, vol. 14, no. 2, pp. 203–16, 2014.

[79] A. Wutz and R. Jaenisch, “A Shift from Reversible to Irreversible X Inactivation Is Triggered during ES Cell Differentiation,” *Mol. Cell*, vol. 5, no. 4, pp. 695–705, 2000.

[80] J. Xu, X. Deng, and C. M. Disteche, “Sex-specific expression of the X-linked histone demethylase gene Jarid1c in brain,” *PLoS One*, vol. 3, no. 7, pp. 1–6, 2008.

[81] C. Corbel, P. Diabangouaya, A.-V. Gendrel, J. C. Chow, and E. Heard, “Unusual chromatin status and organization of the inactive X chromosome in murine trophoblast giant cells.,” *Development*, vol. 140, pp. 861–72, 2013.

[82] C. Patrat, I. Okamoto, P. Diabangouaya, V. Vialon, P. Le Baccon, J. Chow, and E. Heard, “Dynamic changes in paternal X-chromosome activity during imprinted X-chromosome inactivation in mice.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 13, pp. 5198–5203, 2009.

[83] A. H. Horakova, J. M. Calabrese, C. R. McLaughlin, D. C. Tremblay, T. Magnuson, and B. P. Chadwick, “The mouse DXZ4 homolog retains Ctcf binding and proximity to Pls3 despite substantial organizational differences compared to the primate macrosatellite,” *Genome Biol.*, vol. 13, no. 8, p. R70, 2012.

[84] C. A. McHugh, C.-K. Chen, A. Chow, C. F. Surka, C. Tran, P. McDonel, A. Pandya-Jones, M. Blanco, C. Burghard, A. Moradian, M. J. Sweredoski, A. A. Shishkin, J. Su, E. S. Lander, S. Hess, K. Plath, and M. Guttman, “The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3,” *Nature*, vol. 521, no. 7551, pp. 232–236, 2015.

[85] B. Martynoga, J. L. Mateo, B. Zhou, J. Andersen, A. Achimastou, N. Urbán, D. van den Berg, D. Georgopoulou, S. Hadjur, J. Wittbrodt, L. Ettwiller, M. Piper, R. M. Gronostajski, and F. Guillemot, “Epigenomic enhancer annotation reveals a key role for NFIX in neural stem cell quiescence,” *Genes Dev.*, vol. 27, no. 16, pp. 1769–1786, 2013.

[86] B. R. Lajoie, J. Dekker, and N. Kaplan, “The Hitchhiker’s guide to Hi-C analysis: practical guidelines.,” *Methods*, vol. 72, pp. 65–75, Jan. 2015.

[87] S. Selvaraj, J. Dixon, V. Bansal, and B. Ren, “Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing,” *Nat. Biotechnol.*, 2013.

[88] J. Chaumeil, S. Augui, J. C. Chow, and E. Heard, “Combined immunofluorescence, RNA fluorescent in situ hybridization, and DNA fluorescent in situ hybridization to study chromatin changes, transcriptional activity, nuclear organization, and X-chromosome inactivation,” *Methods Mol. Biol.*, vol. 463, pp. 297–308, 2008.

[89] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.,” *Nat. Methods*, vol. 10, no. 12, pp. 1213–8, 2013.

[90] M. A. Eckersley-Maslin, D. Thybert, J. H. Bergmann, J. C. Marioni, P. Flicek, and D. L. Spector, “Random monoallelic gene expression increases upon embryonic stem cell differentiation.,” *Dev. Cell*, vol. 28, no. 4, pp. 351–65, Feb. 2014.

[91] J. B. Berletch, W. Ma, F. Yang, J. Shendure, W. S. Noble, C. M. Disteche, and X. Deng, “Escape from X Inactivation Varies in Mouse Tissues,” *PLOS Genet.*, vol. 11, no. 3, p. e1005079, 2015.

[92] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, “Hi-C: A comprehensive technique to capture the conformation of genomes.,” *Methods San Diego Calif*, pp. 1–9, 2012.

[93] N. Naumova, E. Smith, Y. Zhan, and J. Dekker, “Analysis of long-range chromatin interactions using Chromosome Conformation Capture,” *Methods*, 2012.

[94] B. Langmead and S. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, 2012.

[95] E. Yaffe and A. Tanay, “Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture,” *Nat. Genet.*, 2011.

[96] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. Liu, “HiCNorm: removing biases in Hi-C data via Poisson regression,” *Bioinformatics*, 2012.

[97] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices,” *Pacific J. Math.*, 1967.

[98] Z. Duan, M. Andronescu, K. Schutz, and S. McIlwain, “A three-dimensional model of the yeast genome,” *Nature*, 2010.

[99] P. de Gennes, “Scaling Concepts in Polymer Physics,” 1979.

[100] G. Fudenberg and L. Mirny, “Higher-order chromatin structure: bridging physics and biology,” *Curr. Opin. Genet. Dev.*, 2012.

[101] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions inEmpirical Data,” *SIAM Rev.*, vol. 51, pp. 661–703, 2009.

[102] Y. Shen, F. Yue, D. McCleary, Z. Ye, and L. Edsall, “A map of the cis-regulatory sequences in the mouse genome,” *Nature*, 2012.

[103] O. Symmons, V. Uslu, T. Tsujimura, and S. Ruf, “Functional and topological characteristics of mammalian regulatory domains,” *Genome Res.*, 2014.

[104] T. Le, M. Imakaev, L. Mirny, and M. Laub, “High-resolution mapping of the spatial organization of a bacterial chromosome,” *Science (80-.).*, 2013.

[105] L. Giorgetti, R. Galupa, E. Nora, and T. Piolot, “Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription,” *Cell*, 2014.

[106] F. Benedetti, J. Dorier, Y. Burnier, and A. Stasiak, “Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes.,” *Nucleic Acids Res.*, pp. 1–8, 2013.

[107] F. Ay, T. L. Bailey, and W. S. Noble, “Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.,” *Genome Res.*, pp. 1–23, 2014.

[108] M. Rousseau, J. Fraser, and M. Ferraiuolo, “Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling,” *BMC Bioinformatics*, 2011.

[109] M. Hu, K. Deng, Z. Qin, J. Dixon, and S. Selvaraj, “Bayesian inference of spatial organizations of chromosomes,” *PLoS Comput Biol*, 2013.

[110] Z. Zhang, G. Li, K. Toh, and W. Sung, “3D chromosome modeling with semi-definite programming and Hi-C data,” *J. Comput. Biol.*, 2013.

[111] F. Ay, E. Bunnik, N. Varoquaux, and S. Bol, “Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene,” *Genome Res.*, 2014.

[112] N. Varoquaux, F. Ay, W. Noble, and J. Vert, “A statistical approach for inferring the 3D structure of the genome,” *Bioinformatics*, 2014.

[113] M. Marti-Renom and L. Mirny, “Bridging the resolution gap in structural modeling of 3D genome organization,” *PLoS Comput Biol*, 2011.

[114] N. Kaplan and J. Dekker, “High-throughput genome scaffolding from in vivo DNA interaction frequency,” *Nat. Biotechnol.*, 2013.

[115] J. Burton, A. Adey, R. Patwardhan, and R. Qiu, “Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions,” *Nat. Biotechnol.*, 2013.

[116] J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure, “Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps.,” *G3 (Bethesda).*, May 2014.

[117] C. Beitel, L. Froenicke, and J. Lang, “Strain-and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products,” *PeerJ*, 2014.

[118] A. Miele and J. Dekker, “Long-range chromosomal interactions and gene regulation,” *Mol. Biosyst.*, 2008.

[119] M. Fullwood, M. Liu, Y. Pan, and J. Liu, “An oestrogen-receptor-&agr;-bound human chromatin interactome,” *Nature*, 2009.

[120] E. Birney, others, J. a Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. a Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. a Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. a Hirsch, E. a Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J.

Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaöz, A. Siepel, J. Taylor, L. a Liefer, K. a Wetterstrand, P. J. Good, E. a Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. a Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. a Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. a Singer, T. a Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. a Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. X. Zhang, M. Xu, J. N. S. Haidar, Y. Yu, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. a Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. W. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrímsdóttir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. a Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. a Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.

[121] N. Gheldof, T. Tabuchi, and J. Dekker, "The active FMR1 promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications," *Proc. Natl. Acad. …*, 2006.

[122] J. Gribnau, K. Diderich, S. Pruzina, R. Calzolari, and P. Fraser, "Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus.," *Mol. Cell*, vol. 5, no. 2, pp. 377–386, 2000.

[123] R. Palstra, B. Tolhuis, and E. Splinter, “The β-globin nuclear compartment in development and erythroid differentiation,” *Nat. Genet.*, 2003.

[124] D. Baù, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom, “The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules.,” *Nat. Struct. Mol. Biol.*, vol. 18, no. 1, pp. 107–14, Jan. 2011.

[125] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos, “The accessible chromatin landscape of the human genome.,” *Nature*, vol. 489, no. 7414, pp. 75–82, Sep. 2012.

[126] J. E. Phillips and V. G. Corces, “CTCF: Master Weaver of the Genome,” *Cell*, vol. 137, no. 7. pp. 1194–1211, 2009.

[127] L. Song, Z. Zhang, L. Grasfeder, and A. Boyle, “Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity,” *Genome Res.*, 2011.

[128] M. Creyghton, A. Cheng, and G. Welstead, “Histone H3K27ac separates active from poised enhancers and predicts developmental state,” *Proc. …*, 2010.

[129] A. Rada-Iglesias, R. Bajpai, and T. Swigut, “A unique chromatin signature uncovers early developmental enhancers in humans,” *Nature*, 2011.

[130] M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, and W. S. Noble, “Integrative annotation of chromatin elements from ENCODE data,” *Nucleic Acids Res*, vol. 41, no. 2, pp. 827–841, 2013.

[131] J. Harrow, A. Frankish, and J. Gonzalez, “GENCODE: the reference human genome annotation for The ENCODE Project,” *Genome Res.*, 2012.

[132] X. Dong, M. Greven, A. Kundaje, and S. Djebali, “Correlating histone

modifications and gene expression," *Genome Biol*, 2012.

[133] T. Vavouri, G. McEwen, A. Woolfe, W. Gilks, and G. Elgar, "Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key," *Trends Genet.*, 2006.

[134] J. Wallace and G. Felsenfeld, "We gather together: insulators and genome organization," *Curr. Opin. Genet. Dev.*, 2007.

[135] A. Wood, K. Van Bortle, and E. Ramos, "Regulation of chromatin organization and inducible gene expression by a Drosophila insulator," *Mol. Cell*, 2011.

[136] T. I. Lee, R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young, "Control of developmental regulators by Polycomb in human embryonic stem cells.," *Cell*, vol. 125, no. 2, pp. 301–13, 2006.

[137] B. Lajoie, N. van Berkum, A. Sanyal, and J. Dekker, "My5C: webtools for chromosome conformation capture studies," *Nat. Methods*, 2009.

[138] W. Cleveland and S. Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *J. Am. Stat. Assoc.*, vol. 83, no. 403, pp. 596–610, 1988.