Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXXII

Coordinatore: Prof. Salvatore Capozziello

Combined polymer physics and machine learning approach to investigate the chromosome 3D structure

Settore Scientifico Disciplinare FIS/02

Dottorando

Andrea Esposito

Tutore

Prof. Mario Nicodemi

Anni 2017/2020

CONTENTS

Introduction

The spatial organization of the chromatin in the nucleus is known to play an important role in transcriptional regulation of genes in many organisms (Bickmore and Van Steensel 2013; Cremer and Cremer 2001; Dekker, Marti-Renom, and Mirny 2013; Lieberman-Aiden et al. 2009; Misteli 2007; Tanay and Cavalli 2013). However, the comprehension of genome architecture and of the molecular mechanisms shaping its structure, represents a challenging problem which remains not fully understood. During the last two decades, the development of new technologies has allowed to investigate the three-dimensional spatial folding of chromosomes in a quantitative way. Along with high resolution microscopy techniques, e.g. fluorescence in situ hybridization (FISH), which have been a great tool to visualize nuclear organization (Boyle 2001; Tanabe et al. 2002), the majority of recent discoveries are based on Chromosome Conformation Capture (3C) techniques and its derivatives (Dekker, Marti-Renom, and Mirny 2013). These experiments can measure, in a population of cells, the frequency of interaction between pairs of chromatin regions close in the 3D nuclear space, but which may have a large separation along the linear genomic sequence. The method of choice for detecting genome-wide contacts and determining large chromatin structure is Hi-C, which generates contacts maps between all parts of the genome. Thanks to these technologies, we now know that chromosomes are characterized by a complex, non-random, 3D organization occurring at different genomic length scales, through local and long-range interactions. In mammals, different chromosomes occupy distinct territories (Cremer and Cremer 2001) and have preferred positions depending on cell type and transcription activity (Bickmore and Van Steensel 2013; Misteli 2007). At the super-megabase scale, it was established that chromosomes are separated in two types of domains, namely A and B compartments, which tend to interact with each other in a homotypic fashion (Lieberman-Aiden et al. 2009).

Further improvements in the resolution of the 3C-based methods have revealed the partitioning of the genome into domains of preferential chromatin interactions, the topologically associating domains or TADs (Dixon et al. 2012; Nora et al. 2012). TADs extend up to ~3Mb (Mega bases) and are mostly stable between different cell types and across species (S. S. P. P. Rao et al. 2014). The function of TADs is to delimit the genomic regions sampled by each locus in order to correctly direct enhancer-promoter communication and, at the same time, to prevent the activation of promoters by spurious enhancer located in other TADs. TADs are, in turn, only one level of a more complex, hierarchical organization of higher-order domains (metaTADs) extending up to chromosomal scales (Fraser et al. 2015). However, non-trivial patterns are also seen within TADs and these domains, known as sub-TADs, are mostly associated with CTCF (a highly conserved zinc finger protein implicated in diverse regulatory functions) (Phillips-Cremins et al. 2013; Sexton et al. 2012). It is therefore conceivable that every one of these domains provide a different frame in which enhancer and promoter can find each other, or be insulated from each other, ultimately controlling the transcription of genes. Therefore, disruption or alterations of these structures, for instance via genomic mutations, can affect the regular gene activity and produce effects on the phenotype (Lupiáñez et al. 2015; Spielmann and Mundlos 2013). Although the functional significance of the discussed genomic features in gene regulation is better understood, the factors underlying their formation are still to be investigated. In this sense, polymer physics is turning out to be a great tool to understand the molecular mechanisms of the 3D chromatin spatial organization from first principles. So far, different models have been proposed (M. Barbieri et al. 2012; Brackley et al. 2013; Chiariello et al. 2016; Fudenberg et al. 2016; Giorgetti et al. 2014; Jost et al. 2014; Nicodemi and Pombo 2014; Rosa and Everaers 2008; Sachs et al. 1995; Sanborn et al. 2015; Tiana et al. 2016) which try to make sense of the genome-wide contacts data and lay the foundations of an exciting research field where Physics and Biology intermix.

The studies discussed in the present work have been devised in this general framework. They consist of a detailed description of results and conclusions from the projects that we have conducted in the Physics Department of University of Naples Federico II, under the supervision of Professor Mario Nicodemi, in the group of Complex Systems. Many results have been published or are currently in progress in collaboration with the Epigenetic Regulation and Chromatin Architecture group directed by Prof. Ana Pombo, at Max Delbruck Centre For Molecular Medicine (Berlin) and the Development and Disease Group directed by Professor Stefan Mundlos, at Max Planck Institute for Molecular Genetics (Berlin).

The thesis is organized in four chapters. In Chapter 1, we introduce some basic concepts necessary to the comprehension of this research activity and summarize recent results related to the chromatin spatial organization, as the main experimental techniques, the interpretation of the chromosome interaction data and the relationship between spatial organization and cell functionality. Then, we briefly review the polymer models currently proposed to describe the chromosomes three-dimensional organization in the cell nucleus. In Chapter 2, we outline the 'Strings and Binders Switch (SBS)' model, developed in our research group, and we make use of it to quantitatively explain the information contained in the Hi-C interaction data via Molecular Dynamics simulations. We show that the thermodynamic phases envisaged by our mode can be used to explain the long-range contact profile of chromosomes; then we try to schematically model the hierarchical structure of chromatin, and finally we present a theoretical study of the multiple co-localization contact landscape. In Chapter 3, we introduce more sophisticated variant of the SBS polymer model by which we can reconstruct the 3D structure of real genomic region with high accuracy. Next, we employ this model to study the folding mechanisms and the enhancer-promoter communication at some important chromosome loci where

the failure of these mechanisms can lead to severe diseases. Finally, in Chapter 4, we extend our modeling genome-wide, i.e. to the entire set of chromosomes of the mouse genome. The increase in statistics obtained with the genome-wide study, allows us to compare our polymer models with epigenetics factors, known to play an important role in gene regulation. In this way, we can clarify the molecular nature of the binding factors inferred by our model.

Chapter 1 - 3D genome organization in the cell nucleus

The way in which eukaryotic DNA is organized inside the cell nucleus is likely to be of great importance for basic biological processes, such as transcription and gene regulation. Recent developments in molecular biology and novel computational methods, give us the opportunity to explore this interesting problem. In this chapter, far from being exhaustive about this huge topic, we briefly review some recent results which are crucial in this research field and that will help the comprehension of our research activity described in the following chapters. In Section 1.1 and 1.2 we summarize some fundamental concepts of molecular biology and some recent advances about genes regulation and epigenetics. Then in Section 1.3 we discuss the main technologies that allow to quantitatively investigate the architecture of the genome with a focus on the Hi-C experimental technique. Thanks to Hi-C, important results have been obtained about the chromosome spatial organization, and that will be discussed in Section 1.4. Finally, in Section 1.5 we review the most recent polymer models that aim to quantitatively explain and reconstruct the three-dimensional structure of the genome. The results described in this chapter have been introduced and discussed in the following important papers: (Dekker, Marti-Renom, and Mirny 2013; Dixon et al. 2012; Fraser et al. 2015; Lieberman-Aiden et al. 2009; Nora et al. 2012).

1.1 Packaging of the DNA inside the cell nucleus

The cell nucleus of the living organisms can be modelled as a complex system in which multiple interactions between many different components operate to

translate the genetic information into physical processes which ensure the correct development of the organism itself.

The genetic information is written in the **DNA** (deoxyribonucleic acid) molecule, which is a double helix consisting of two coupled polymer chains made of simple units called **nucleotides**. These latter are composed of three elements: a five-carbon sugar, a phosphate group and a nitrogenous base, which may be either adenine (A), cytosine (C), guanine (G), or thymine (T). Hydrogen bonds between specific pairs of nucleotides (A binds T and C binds G) join the opposite strands together. The sequence of these four nucleobases along the polymeric chain encodes the **genetic information**.

In eukaryotes, almost all the DNA is located inside the cell nucleus where it is packed into structural entities called **chromosomes**. In the crowded nuclear environment, each chromosome occupies a specific spatial region referred to as **chromosomal territory**, which is clearly visible in microscopy experiments (**Figure 1.1**). The majority of eukaryotic cells are diploid, i.e. they have two copies of each chromosome. The number of the different chromosomes and the total genomic length, that is the number of base pairs (bp) in the cell nucleus, strongly depends on the considered species. In the human genome, for instance, there are around 3.2×10^9 bp distributed over 24 chromosomes. The importance of the DNA folding is immediately clear if we consider that the linear length of the human genome is about 2m and is constrained in a nucleus of 10÷15μm, that is geometrically equivalent to packing 40km of an extremely fine thread into a tennis ball. This compaction level is achieved through an efficient interaction between DNA and proteins.

Proteins binding the DNA are not only necessary for its spatial structure, but also to exploit many biological purposes as gene expression, DNA replication, DNA repair and DNA recombination. The complex of DNA and proteins is known as

chromatin and represents the real physical structure containing the information to be processed. Chromatin is spatially organized at different genomic length scale and degree of compaction. At the very first level we found the **histones**, around which the double strand of DNA is rolled up in forming the most basic unit of chromosome packing, the **nucleosome**. It consists of a structure of eight histone proteins (two molecules each of histone H2A, H2B, H3 and H4). The length of DNA associated with each nucleosome is 147 bp (about 11nm). Each nucleosome is separated from the next by a filament of linker DNA, which can vary in length from a few nucleotides pairs up to about 80 bp. On average, nucleosomes repeat at intervals of 200 bp. So, since human genome has 6.4×10^9 bp, it consists of about 30×10^6 nucleosomes. This structure is known as "beads on a string" (where the "bead" is the nucleosome and the "string" is linker DNA) organization. As a second level of compaction, the nucleosomes are packed on top of one other, thanks to an additional protein known as histone H1. In this phase, the chromatin fiber has a diameter of about 30nm, which corresponds to a 0.1cm in length for a mean human chromosome, which is too long to fit the cell nucleus. Clearly, there must exist other mechanisms of folding, still largely unknown, that eventually give rise to discrete 1μm wide chromosome territories (Cremer and Cremer 2001).

Different chromosomal regions can be classified into two categories: **heterochromatin** and **euchromatin**. Heterochromatic regions are composed by DNA showing a high degree of compaction. Since the high level of compaction reduces its accessibility, it is less likely that these regions are transcribed. Euchromatin, on the other hand, is a lightly packed form of chromatin, enriched in genes and, often but not always, under active transcription.

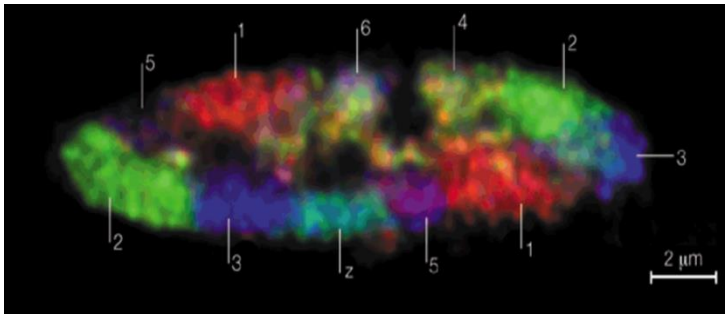


Figure 1.1: Chromosome territories

Optical microscopy image (from chicken fibroblast cells) in which chromosomes have been tagged with different colors. Each chromosome occupies a specific, mutually exclusive, region of the cell nucleus, with homologous chromosomes seen in separate locations. Figure adapted from Cremer and Cremer 2001.

1.2 Transcriptional regulation of gene expression

Genes are crucial genomic sequences of DNA which encodes the synthesis of a gene product, either RNA or protein. Different cells in a multicellular organism may express very different sets of genes, even though they contain the same DNA. The set of genes expressed in a cell determines the set of proteins and functional RNAs it contains, giving it its unique properties. The mechanism of interpretation of genomic information is known as **transcriptional regulation of gene expression** and is mediated by the functionally diversified cis-regulatory elements, such as *promoters*, *enhancers*, *silencers*, and *insulators*. The transcriptional activity is primarily regulated by a control sequence, the gene **promoter**, usually located, in eukaryotes, within 1Kb upstream of the transcription start site (TSS) of the gene. Promoters contain specific DNA sequences such as response elements that provide a secure initial binding site for RNA polymerase and for proteins called **Transcription Factors (TFs)** that recruit RNA polymerase. In general, TFs are proteins that recognize a specific DNA sequence and bind to it to regulate gene expression by promoting or suppressing transcription. **Enhancers**, on the other hand, play a central role in

driving cell-type-specific gene expression and are capable of activating transcription of their target genes at great distances, ranging from several to hundreds, in some cases even thousands, of kilobases (Bulger and Groudine 2011; Calo and Wysocka 2013; Ong and Corces 2011). The mechanism of action of the enhancers, albeit still poorly understood, involves the physical proximity with their target genes. Indeed, extensive intra- and inter-chromosomal interactions between enhancers and promoters are detected at many co-regulated genes during development. This network of contacts among the different regulatory elements (REs) contributes to orchestrate gene expression giving rise to complex spatial conformations.

REs can often be identified by specific epigenetic marks associated to them. **Epigenetics** is the study of heritable changes in a phenotype arising in the absence of alterations in DNA sequence, such as chemical modification of DNA and interactions with molecular factors. DNA **methylation**, the addition of a methyl group on a substrate, is an important example of epigenetic modification with a known functional role. For instance, methylation of cytosine may occur at CpG sites (regions of the genome where a cytosine is followed by a guanine), which are frequently encountered at gene promoter, to repress gene transcription. The amino acids of the histone tails are also subject to chemical modifications like acetylation, methylation, phosphorylation, and ubiquitination. **Histone modifications** exhibit both a repressive and active function on gene transcription and they can be used to assign a functional role to some genomic regions: for example H3K4me3 (histone H3 lysine 4 trimethylation) is associated with promoter regions, H3K4me1 (histone H3 lysine 4 methylation) and H3K27ac (histone H3 lysine 27 acetylation) characterizes enhancer regions, H3K36me3 (histone H3 lysine 36 trimethylation) marks transcribed regions, H3K27me3 (histone H3 lysine 27 trimethylation) Polycomb-mediated repressed regions and poised enhancers, and H3K9me3 (histone H3 lysine 9 trimethylation) heterochromatin regions (Allis and Jenuwein 2016). Histone modification

patterns can be mapped genome-wide thanks to the development of chromatin immunoprecipitation followed by next-generation sequencing (**ChIP-seq**) technique, a method to detect where a protein is bounded along the DNA, by use of specific antibodies that target the protein of interest.

1.3 Experimental techniques to map chromosome contacts

We have seen that transcriptional control may be mediated through physical contacts between enhancers and their target genes so highlighting the role of the 3D organization of the chromatin in the functioning of the cell. In the last two decades, many efforts have been done in molecular biology to develop a series of experimental approaches to investigate the chromosomes spatial organization with high accuracy. Of particular importance are the methods based on the chromosome conformation capture (3C) protocol, which allow the measuring of the frequency with which any pair of genomic regions (loci) in the genome is in close enough physical proximity (in the range of 10÷100nm) to become crosslinked, that is the pair can be bound by some molecule. The most common methods are the 3C, 4C, 5C and HiC methodologies, all based on the following main steps: cells are crosslinked with formaldehyde to covalently link chromatin segments that are in close spatial proximity; next, chromatin is fragmented by sonication or restriction enzyme digestion; crosslinked fragments are then ligated to form unique hybrid DNA molecules; finally, the DNA is purified and analysed (Dekker, Marti-Renom, and Mirny 2013). The difference among each specific method is how the ligation product is detected and quantified.

3C and 4C generate single interaction profiles for individual loci. More precisely, **3C** (Dekker et al. 2002) yields a long-range interaction profile of a selected gene promoter or other genomic element of interest versus surrounding chromatin (one

versus one) whereas **4C** (Simonis et al. 2006) generates a genome-wide interaction profile for a single locus (one versus all). These data sets can be represented as single tracks that can be plotted along the genome and compared to other genomic features. **5C** (Dostie et al. 2006) and Hi-C methods are not anchored on a single locus of interest but instead generate matrices of interaction frequencies that can be represented as two-dimensional heat maps with genomic positions along the two axes (many versus many).

1.3.1 The Hi-C technique

Hi-C (Lieberman-Aiden et al. 2009) in particular, represents the first genome-wide (all versus all) adaptation of 3C and includes a further step in which, after restriction digestion, the staggered DNA ends are filled in with biotinylated nucleotides. The resulting DNA sample is composed by ligation products of chromatin that were in spatial proximity in the nucleus, with biotin at the ligation junction (**Figure 1.2a**). This facilitates selective purification of ligation junctions that are collected in a Hi-C experiment and then directly sequenced along the genome, producing a list of interacting fragments. In **Figure 1.2b** a typical output of a Hi-C experiment is depicted. Data are organized in a **contact matrix** C_{ij} whose generic bin is the number of ligation products between the locus *i* and the locus *j*. To this aim, the genomic sequence is split into windows whose length is an important parameter and determines the Hi-C data **resolution**. Higher resolution, i.e. a reduced size of the genomic windows, corresponds to matrices with higher sizes. So, the outcome of a Hi-C experiment is a set of interaction matrices associated with each chromosome (*Cis* data) together with interaction data between loci belonging to different chromosomes (*Trans* data), the latter with a much lower frequency. Only *Cis* contact matrices will be analysed throughout this work.

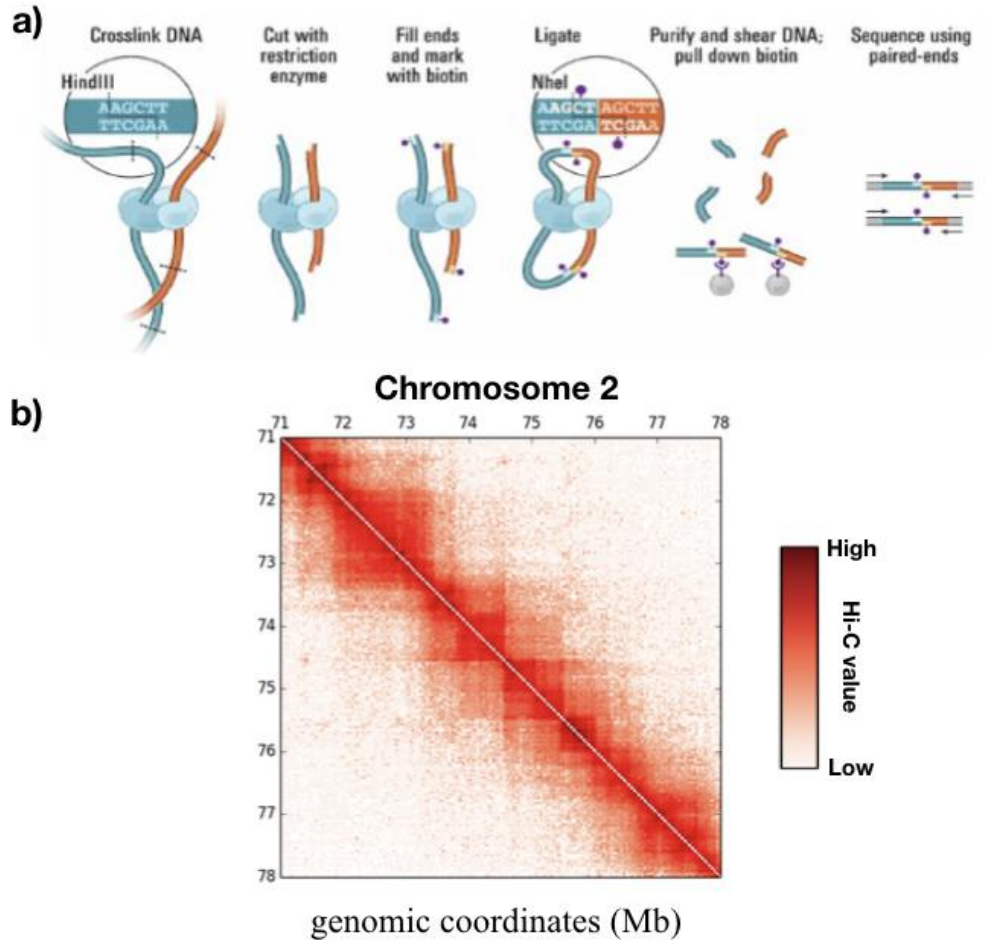


Figure 1.2: Hi-C technique

a) Schematic representation of the Hi-C experimental procedure. Cells are cross-linked with formaldehyde to covalently link spatially adjacent chromatin segments; chromatin is digested with a restriction enzyme and the resulting sticky ends are biotinylated; crosslinked fragments are then ligated to form unique hybrid DNA molecules; DNA is purified and sheared. The labelling with biotin allows to efficiently detect the ligated fragments, which are finally identified by paired-end sequencing. Figure from Lieberman-Aiden et al. 2009 **b)** Example of Hi-C data output collected in a bidimensional heatmap. Data are from a 7 Mb long region of the mouse chr2 at 40Kb of resolution. The color intensity of each pixel is proportional to the number of reads detected in the experiment. Data from Dixon et al. 2012.

It is important to observe that the 3C-based methods describe the relative frequency in a cell population by which two loci are in close spatial proximity, not distinguishing functional from non-functional associations. Indeed, spatial proximity can be the result of direct and specific contacts between two loci, the result of indirect co-localization to the same subnuclear structure (such as the nuclear lamina), or it can be due to random collisions in the crowded nuclear space. The polymer nature of chromosomes also determines the frequency with which pairs of loci interact even in the absence of any specific structure (Fudenberg and Mirny 2012). Finally, the precise conformation of the chromatin fiber is highly variable from cell to cell and is the effect of many different constraints that act on it. Each ligation event represents a contact involving a pair of loci in a single cell of the population. In other words, Hi-C (and all 3C-based) interaction frequency data represent the fraction of cells in which pairs of loci *i* and *j* are in spatial proximity at the time the cells are fixed and the final value contained in the matrix bin C_{ij} represent the sum of interactions over a large cell population. Therefore, the typical maps obtained by these approaches probably represent a superimposition of all possible conformation states of a cell. This has important implications for the biological significance of chromatin contact data.

1.3.2 Ligation-free methods

The just discussed 3C methods are all based on proximity ligation, which creates covalent bonds between regions spatially close. However, these technologies often fail to detect chromatin regions too far apart to directly ligate. Recently, two techniques have been developed which can overcome this problem by using a ligation-free approach: **Genome Architecture Mapping** (GAM) and **Split-Pool Recognition of Interactions by Tag Extension** (SPRITE). Both methods are also capable to detect multi-way contacts such as triplets, quadruplets, and so on, while only pairwise interaction can be detected by 3C-based techniques. GAM (Beagrie et al. 2017) was the first ligation-free approach able to detect contacts at a genome-wide level. Starting from a collection of slices obtained by

cryo-sectioning a population of nuclei in random directions, it is possible to estimate the frequencies of interaction between pairs of loci. In fact, loci which are physically close in the 3D nuclear space, have a high probability to be co-segregated (to fall in the same slice). SPRITE (Quinodoz et al. 2018) shares some of the initial steps with the 3C-based methods, but it does not use the ligation as well as GAM. Precisely, after the crosslinking and the fragmentation processes, the interacting molecules in a cluster are barcoded by using a split-pool strategy. Interactions are identified by matching all the reads having the same barcode via genomic sequencing. The cluster obtained in this way are then converted in contact frequencies by counting all contacts observed in a single cluster and weighting each contact by the total number of molecules contained within the cluster. One of the advantages of this method is that it can also detect, in addition to DNA interaction, higher-order RNA interaction in the nucleus.

1.4 Chromosome structural features from Hi-C data

Genomic compartments

The emergence of whole-genome 3C-based methodologies has allowed the discovery of peculiar structure of chromatin, providing powerful insights into how gene expression relates to chromatin compaction. Application of principal component analysis to Hi-C data revealed a strong segregation of the interactions into two distinct classes, named **A and B compartments** (Lieberman-Aiden et al. 2009; S. S. P. P. Rao et al. 2014), recently confirmed by microscopy experiments (Nir et al. 2018). A compartments preferentially interact with other A compartments throughout the genome. Similarly, B compartments associate with other B compartments (**Figure 1.3**). Compartments are considerably large regions of chromatin, having a characteristic size of 5÷10 Mb, and alternate along the chromosomes. Comparisons with indicators of transcriptional activity such as DNA accessibility, gene density, and several histone marks reveal a strong

relationship between the A compartment and transcriptionally active, open chromatin (euchromatin) and the B compartment with closed chromatin (heterochromatin). Increased depth of Hi-C data sets has allowed smaller sub-compartments to be detected, which capture fine differences in replication timing as well as preferred associations with the nucleolus or the nuclear lamina (S. S. P. P. Rao et al. 2014).

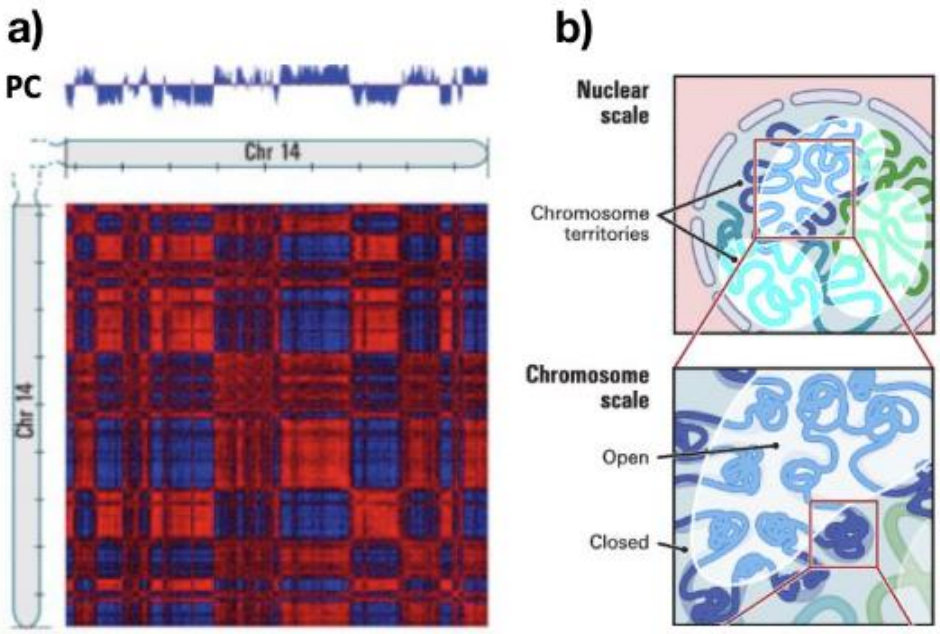


Figure 1.3: A and B compartments

a) Pearson correlation map of a genomic region on the chromosome 14 and the principal component (PC) associated (bar plot on the top). The PC correlates with the plaid pattern in the correlation matrix, defining the compartment A (positive PC values) and B (negative PC values). b) Schematic representation of chromatin organization at nuclear scale, where chromosome territories (hundreds of Mb) occupy distinct regions, and at chromosome scale, where open and closed chromatin regions (5÷10 Mb) alternate. Figure adapted from Lieberman-Aiden et al. 2009.

Topologically associated domains (TADs)

One of the most interesting discoveries was that chromosomes are spatially segregated into sub-mega base scale domains, often called **topologically**

associating domains or, briefly, **TADs** (Dixon et al. 2012; Nora et al. 2012). They typically appear as contiguous square domains along the diagonal of Hi-C or 5C maps (or triangles as represented in **Figure 1.4**), in which regions within the same TAD interact with each other much more frequently than with regions located in adjacent domains. To identify TADs several computational algorithms have been developed (Dixon et al. 2012; Fraser et al. 2015; S. S. P. P. Rao et al. 2014). Although initially mammalian TADs were identified with a median size of ~800 Kb (Dixon et al. 2012), subsequent analysis of higher resolution data Hi-C data suggested a smaller median domain size of ~185 Kb (range 40 Kb–3 Mb). TADs are found to be a universal building blocks of chromosomes, as both mouse and human are composed by more than 2000 domains, covering almost all the genome. The spatial partitioning of the genome into TADs correlates with many linear genomic features and enhancer–promoter interactions seem to be mostly constrained within a TAD (Shen et al. 2012). The mechanism that regulates the formation of TADs is still not completely understood, and polymer models have been proposed to quantitatively describe it (M. Barbieri et al. 2012; Brackley et al. 2013; Chiariello et al. 2016; Fudenberg et al. 2016; Sanborn et al. 2015).

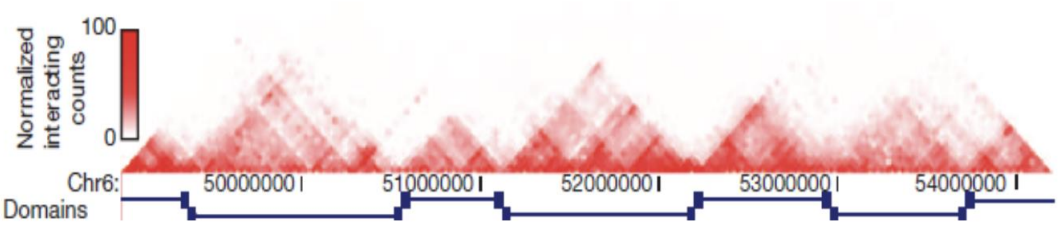


Figure 1.4: Topologically Associated Domains (TADs)

Hi-C data for a region along chromosome 6 in mouse embryonic stem cells (mESC). TAD domains appear as high intensity square blocks along the diagonal of Hi-C matrix (here represented as a upper triangular matrix, since Hi-C is symmetric by construction). Loci belonging to the same TAD interact more frequently than loci in different TADs. (Figure from Dixon et al. 2012)

Further research developments

Beyond TADs and compartments, other important genomic features have been discovered thanks to the development of more sophisticated and refined experiments which resulted in the production of higher quality data. Analysing the interaction among TADs, higher-order chromatin 3D structures were identified through mouse neuronal differentiation. Chromatin was found to be organized in a hierarchy of domains-within-domains, named **metaTADs**, up to chromosomal scales. As for TADs, metaTADs correlate with a variety of epigenetic features, pointing towards a functional role of this organization (Fraser et al. 2015). In addition, analysis of higher resolution contact data (S. S. P. P. Rao et al. 2014) led to identification of **loop domains**, a particular type of TADs exhibiting a point-like interaction peak at its boundaries, known as **chromatin** "**loops**". Loop domains derived from 3C-based technologies often coincide with pairs of convergent CTCF (CCCTC-binding factor) binding sites, indicating that CTCF can contribute to the partition of specific regions of the genome into self-associating domains. Chromatin folding below the scale of TADs has recently been analysed with a novel high resolution, 3C-based, method Micro-C (Hsieh et al. 2019) which showed the existence of finer structures named **microTADs** and **stripes**. Micro-TADs encompass either single genes, multiple genes, or intergenic regions but are much smaller than TADs, while stripes correspond to lines extending from the diagonal in contact maps. Finally, experiments have been performed to evaluate the impact on health of chromatin structure alterations such as **TADs disruption** (Lupiáñez et al. 2015) or **neoTAD** formation (Franke et al. 2016), so demonstrating the deep relationship between chromatin organization and phenotype.

1.5 Polymer physics of chromatin folding

Following the improvements of the experimental technologies, important progresses have been made in the development of theories. Several models have been proposed to understand the molecular mechanisms of chromosome folding and, in this subsection, we will list very briefly some of them, for sake of completeness (a detailed review of the models can be found in (Esposito et al. 2018)). The **String and Binders Switch (SBS)** model (M. Barbieri et al. 2012) will be the fundamental starting point for our considerations in the following chapters. In this model a chromatin fiber is modelled as a bead chain, where some of those (binding sites) can interact with floating particles (binders), and the polymer folds through the interaction between binding sites and binders. Further details on the SBS model will be given in the next chapter. The idea of chromatin interacting with floating particles has been used also in other studies (Brackley et al. 2013; Chiariello et al. 2016). The first proposed model was the **Fractal Globule** (Lieberman-Aiden et al. 2009) in which the polymer condensation is subjected to some topological constraints preventing knotting and slowing down equilibration of the polymer. Another important model is the **Dynamic Loop** model (Bohn and Heermann 2010), where chromatin moves under diffusional motion and when two sites colocalize, they form a loop with a certain probability for a certain lifetime. In (Jost et al. 2014) a model is presented that considers chromatin as a sequence of regions characterized by an epigenetic state with regions in the same state having specific interactions. Other models consider chromatin folding as the result of interaction of boundary elements through dynamic mechanisms of **Loop Extrusion** (Fudenberg et al. 2016; Sanborn et al. 2015). Following this, a Loop Extruding Factor (LEF), progressively extrudes a chromatin loop until it is stalled by a roadblock. In mammals, the Cohesin complex and zinc finger protein CTCF have been identified as the LEF and roadblock factors, respectively. Yet, CTCF binding events are not impermeable

to the extrusion complex, thereby accounting for the formation of chromatin domains within or between TADs (sub-TADs and meta-TADs).

Chapter 2 - Polymer physics explains key features of chromosome organization

In this chapter, we are going to show how it is possible to recover complex features of chromatin organization with polymer physics. In Section 2.1, the Strings and Binders Switch Model (SBS) of chromatin, initially presented in Barbieri et al. 2012, will be discussed and followed by its molecular dynamics implementation (Section 2.2) and its phase diagram (Section 2.3). In Section 2.4 we will show how, with few parameters, we are able to recapitulate the average behaviour of the experimental chromatin contact frequency, as mapped from Hi-C methods, in a large range of genomic lengths going up to chromosomal scales. In Section 2.5, we will use the SBS model to reproduce important features of chromatin organization like TADs and higher order structures. Finally, the theoretical multiple contact profile of genome architecture will be described in Section 2.6.

Most of the material presented in this chapter, including figures, paragraphs and sentences, is adapted or taken literally from the published papers: Annunziatella et al. 2018; Chiariello et al. 2016; Esposito et al. 2019; which I co-authored.

2.1 The Strings & Binders Switch (SBS) model of chromatin

In the **Strings and Binders Switch (SBS)** model (M. Barbieri et al. 2012; Nicodemi and Prisco 2009), a chromatin filament is represented as a self-avoiding-walk (SAW) chain made of consecutive beads linked to each other. A fraction of the beads, named **binding sites**, can interact with molecular particles, the **binders**, dispersed in the surrounding environment with a concentration *c*,

through an attractive potential with interaction energy E_{int}. The interaction between beads and binders drives the folding of the polymer.

A schematic representation of the SBS model is shown in **Figure 2.1**, where the binding sites (green and orange beads) can interact with their cognate binders. For the sake of simplicity, only two type of binders and binding sites (two different 'colors') are showed, yet, to describe more complex situations, different types of interactions can be introduced (Bianco et al. 2018; Chiariello et al. 2016).

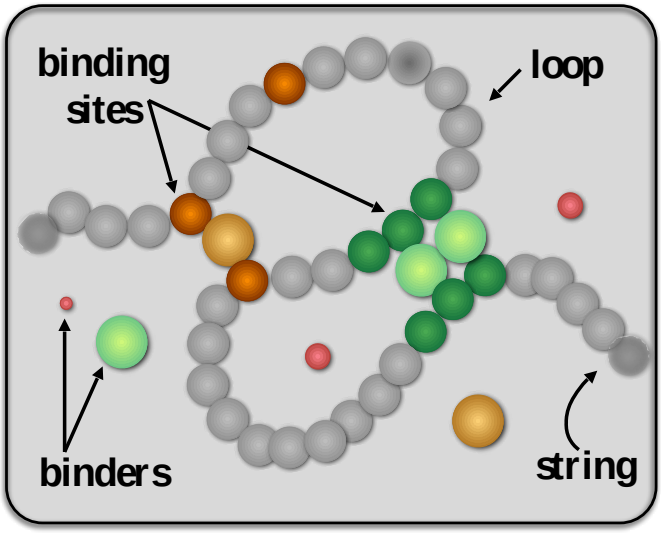


Figure 2.1:The Strings and Binders Switch (SBS) model

The SBS model is a self-avoiding chain of beads (the "string"), a fraction of which (the "binding sites") interact with diffusing molecules (the "binders") having a concentration c and a binding affinity E_{int}. The binders can bridge distant beads looping the polymer. Figure adapted from (Esposito et al. 2019)

2.2 Molecular Dynamics of the SBS model

To study the spatio-temporal evolution of an SBS polymer, we perform **molecular dynamics** (MD) simulations, a standard method to investigate molecular systems. To this aim, the LAMMPS (Large Atomic Molecular Massive Parallel Simulator) software will be adopted, which employs the Verlet algorithm to integrate the equations of motion (Plimpton 1995).

Equations of motion

In our simulations, N beads compose the polymer chain and each bead has a diameter equal to σ. The polymer and its binders are embedded in a surrounding viscous fluid, describing the cell nuclear environment, and undergoes a Brownian motion. Hence, the dynamics of each of the system particles obeys to the Langevin equation (Kremer and Grest 1990):

$$m\frac{d\vec{v}(t)}{dt} = -\zeta\vec{v}(t) + \vec{f}(t) - \nabla V \qquad (1)$$

where m is the mass of the generic particle, $\vec{v}(t)$ the particle velocity, V is the potential acting on the particles and $\vec{f}(t)$ a stochastic random force which takes into account the thermic fluctuation of the environment. The friction coefficient ζ is related to the viscosity of the solvent η from the Stokes relation $\zeta=3\pi\eta\sigma$. As usual in MD simulations, we work in dimensionless units. So, we set the diameter of the polymer bead σ equal to 1 (the same is done for the binder diameter). The diameter fixes our length unit. Analogously, we set the mass of the particle m equal to 1. The energy scales are measured in k_BT, where the Boltzmann constant k_B and the temperature T are both equal to 1. For the dynamics, we set ζ=0.5 (Kremer and Grest 1990; Rosa and Everaers 2008). The simulation box, with boundary periodic conditions, has a linear size D, that is as large as the gyration radius of a SAW with the same number of beads ($D \propto N^{0.588}$). Physical units will be obtained once we fix the length scale and other parameters of the system, as described in the following subsection.

From the MD units to the physical units

The units used in MD simulation, called Lennard-Jones or reduced units, are dimensionless. This means that σ, $\varepsilon = k_bT$ and m are taken as units of length, energy and mass, respectively. The mapping of these units to the physical quantities can be easily obtained with a simple multiplication by a factor

representing the specific unit, linked to the molecular details of the system or to experimental data. For instance, the physical diameter of the bead σ is estimated by imposing that the local chromatin density matches the average nuclear DNA density, i.e., by the relation $\sigma = (s_0/G)^{1/3} D_0$ where D_0 is the nucleus diameter, G is the total genomic content of DNA, and s_0 the number of base pair of each bead of the polymer (M. Barbieri et al. 2012). The molar concentration of binders is calculated using the relation $c = P/VN_A$ where P is the absolute number of binders in solution, V is the box volume and N_A is the Avogadro number. Finally, the time scale τ is fixed by the standard MD relation $\tau = \eta(6\pi\sigma^3/\varepsilon)$. So, by considering $\eta = 0.1$P at room temperature T=300K, we obtain τ =0.03s.

Potentials

Each particle of the system has a potential energy V($\vec{x}$) consisting of the three different components listed below (see **Figure 2.2**):

1) Between any two consecutive beads of the polymer chain there is a potential that models a finitely extensible non-linear elastic spring, the FENE potential (see Kremer and Grest 1990). We set the FENE length constant R_0(the maximum extension of the spring) equal to 1.6σ and K (the strength of the spring) equal to $30k_BT/\sigma^2$ (Brackley et al. 2013; Kremer and Grest 1990);

2) To account for excluded volume effects between any two particles there is also a purely repulsive, shifted Lennard-Jones (LJ) potential:

$$V_{hard}(r) = \begin{cases} 4\left[\left(\frac{\sigma_{b-b}}{r}\right)^{12} - \left(\frac{\sigma_{b-b}}{r}\right)^{6} - \left(\frac{\sigma_{b-b}}{1.12}\right)^{12} + \left(\frac{\sigma_{b-b}}{1.12}\right)^{6}\right] & r < 1.12 \\ 0 & otherwise \end{cases} \quad (2)$$

σ_{b-b} being the distance between any two particles when they are close in space.

3) Finally, there is the bead-binder potential, between each bead and its cognate binders, modelled as an attractive LJ potential with a cut-off:

$$V_{int}(r) = \begin{cases} 4\epsilon_{int}\left[\left(\frac{\sigma_{b-b}}{r}\right)^{12} - \left(\frac{\sigma_{b-b}}{r}\right)^{6} - \left(\frac{\sigma_{b-b}}{r_{int}}\right)^{12} + \left(\frac{\sigma_{b-b}}{r_{int}}\right)^{6}\right] & r < r_{int} \\ 0 & otherwise \end{cases} \quad (3)$$

where r_{int} is the cut-off distance which regulates the interaction range, ϵ_{int}, expressed in k_BT units, is the parameter controlling the strength of the interaction, and σ_{b-b} is the bead-binder distance when they are close in space (i.e. the sum of their radii which in our case is 1σ). For our simulation, we set $r_{int} = 1.3\sigma$, unless otherwise stated.

The absolute value of the minimum of the interaction potential, V_{int}, is taken as the energy of the interaction between beads and binders and it is proportional to ϵ_{int} through the relation:

$$E_{int} = \left|4\epsilon_{int}\left[\left(\frac{\sigma_{b-b}}{r_{int}}\right)^{6} - \left(\frac{\sigma_{b-b}}{r_{int}}\right)^{12} - \frac{1}{4}\right]\right| \quad (4)$$

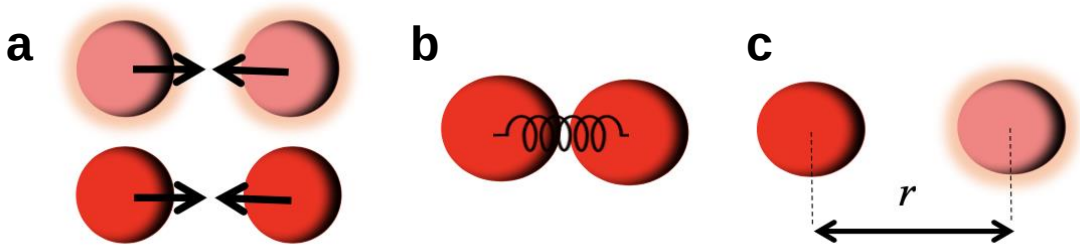


Figure 2.2: Scheme of the three potentials used in the SBS model

a) Any pair of the system (bead-bead, bead-binder, and binder-binder) is subjected to a repulsive LJ potential (equation 2), which models the excluded volume effect. **b)** The bond between two consecutive beads is a finitely extensible non-linear elastic spring (FENE). **c)** Beads and binders interact via an attractive, shifted LJ potential, as in equation (3).

Initial states of the system

All the initial states of our simulations are self-avoiding-walk (SAW) configurations, obtained by a standard approach described in Kremer and Grest 1990. First, we generate a random walk chain, where the distance between two consecutive beads is equal to the average length of an equilibrium SAW chain under the FENE potential above described. The random overlaps between beads and binders are then removed by replacing the hard-core repulsive LJ potential with a soft potential letting the system equilibrate for some timesteps (Brackley et al. 2013; Kremer and Grest 1990). The scaling properties of the polymer are then measured to check that the SAW state is attained. In particular, one of the physical quantities taken into account is the **gyration radius R_g** of the polymer, defined as:

$$R_g^2 = \frac{1}{M}\sum_{i=1}^{N} m_i\,(r_i - r_{CM})^2 \qquad (5)$$

where M is the total mass of the polymer, m_i and r_i are the mass and the position of the i-th bead (respectively), and r_{CM} is the position of the center of mass of the polymer. When the polymer is at an equilibrium SAW state, the gyration radius as a function of time reaches a plateau. Furthermore, the scaling properties of R_g can be studied to check the quality of the SAW state. Indeed, as known from polymer physics (De Gennes 1979), $R_g^2(t)$ exhibits a power-law behaviour as a function of the polymer length $R_g^2 \propto N^{2\nu}$ with the scaling exponent $\nu = 0.588$.

The MD simulation proceeds by randomly introducing, in the simulation box containing the SAW initial state, the binders at a concentration *c*. The system evolves towards its thermodynamics equilibrium state which, once again, can be monitored by looking at the plateauing of the gyration radius.

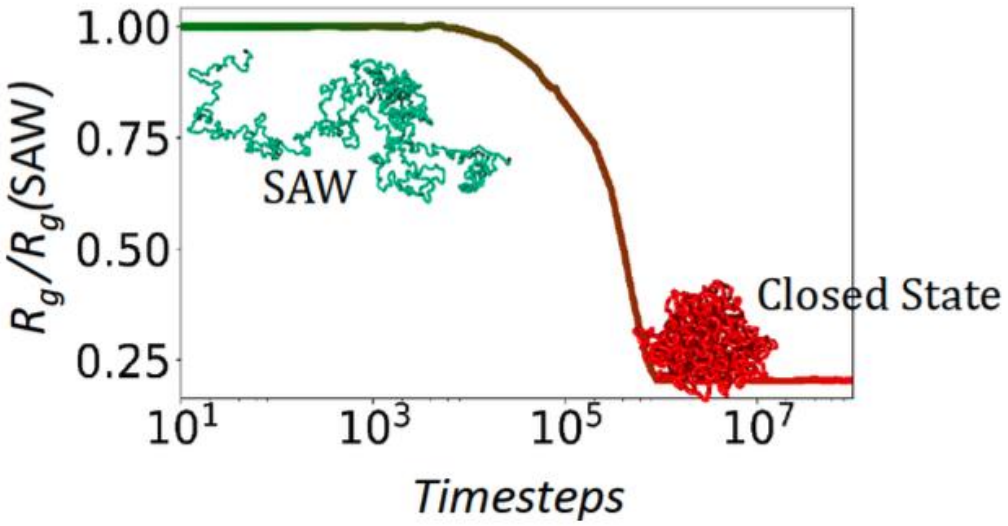


Figure 2.3: Plateauing of the gyration radius

The gyration radius R_g (relative to its initial value) of an SBS polymer as a function of MD time steps. As the system evolves, the polymer folds as showed by the decreasing of its R_g, which, at equilibrium, reaches a plateau. (Figure adapted from Annunziatella et al. 2018)

2.3 Conformational classes in the SBS homo-polymer system

We now want to investigate the average behaviour of a generic region of the genome (i.e. a chromosome) with an SBS polymer (see Section 2.1) made of N = 1000 beads. Each bead of the polymer will correspond to a precise number of DNA base pairs. Indeed, if L is the length of the genomic region to be modelled and N is the number of beads forming the chain, s_0 = L/N is the number of base pairs per bead and is named genomic content. Since for a typical mammalian chromosome L = 100Mb, each bead will contain s_0 = 100Kb. In this section, we focus on the simplest SBS model, that is a self-avoiding chain of identical (equal-colored) beads all interacting with the binders (homo-polymer). In the

next sections, we will extend the model to accommodate different types of binding sites along the chain and their specific cognate binders.

Phase diagram

The control parameters of the system are the bead-binder interaction energy E_{int} and the binder concentration c. As known from polymer physics, there is a **coil-globule** folding transition, highlighted by a sharp drop of the gyration radius, the order parameter for this transition, when crossing the theta point in the phase diagram (in **Figure 2.4**). The coil state is characterized by small values of E_{int} and c, i.e. when the binders not succeed in forming stable loops and the polymer remains open as in a SAW (**Figure 2.4** and **Figure 2.5b** violet box). On the other hand, in the globular state the polymer is in a compact configuration, occupying a very small fraction of the open state volume (**Figure 2.4** and **Figure 2.5b** red box). We also identify a new phase transition, occurring in the polymer globular phase, where the binders undergo an **order-disorder** transition, although they do not interact directly with each other. At low energies or concentrations, the binders form a disordered aggregate attached to the polymer chain, while at high energies, with a sufficiently high concentration, they form an ordered aggregate. Such thermodynamic stable states are expected to play an important role in the chromatin organization as discussed in the following chapters.

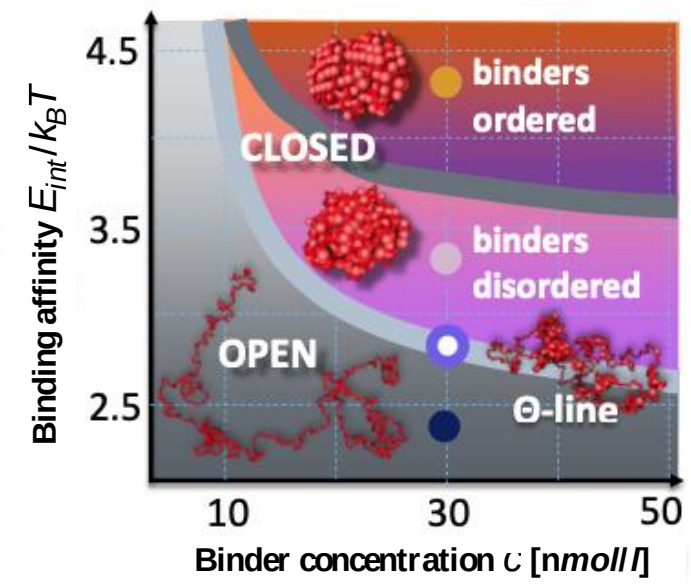


Figure 2.4: The phase diagram

The equilibrium architectural classes correspond to the different phases of its phase diagram: at low binding affinity or concentration, the polymer is open and randomly folded in its coil phase; above its Θ-point transition, in the globule phase, it is closed in more compact conformation. In the closed state, at higher values of E_{int} or c, its binders undergo a transition from a disordered to an ordered arrangement. (Figure adapted from Chiariello et al. 2016).

The order parameters

The transition lines in the phase diagram of **Figure 2.4** are identified as follows. The coil-globule transition is found by measuring, at equilibrium, the collapse of the gyration radius defined in equation (5), which is essentially a measure of the average linear size of the polymer. $\boldsymbol{R_g}$ has the predicted value of a SAW when the polymer is in the open state, while it jumps to a much lower value in the compact state (**Figure 2.5a**). The binder order-disorder transition is captured by two structural quantities associated to the spatial configurations of the binders bound to the polymer: their **pair distribution function** *g(r)* and the **structure factor** *S(k)*, defined as:

$$g(r) = \frac{1}{\rho N_b}\langle \sum_i \sum_{i \neq j} \delta(r - r_{ij}) \rangle \quad (6)$$

$$S(k) = 1 + 4\pi\rho \int_0^\infty r^2 \frac{\sin kr}{kr} g(r) dr \quad (7)$$

where *i*, *j*, label the different binders, r_{ij} is the distance between a pair of them, $\rho = N_b/V$ is the concentration of the binders attached to the polymer and δ is the Dirac delta function. The structure factor is almost flat in the disordered binder state, while it has sharp peaks in the binder ordered state. The transition order parameter is the ratio $S(k^*)/S_{max}$ where k^* is the value of *k* corresponding to the second peak in the *S(k)* function and S_{max} is a normalization coefficient taken to be equal to the maximum value of $S(k^*)$ across the different considered cases. Such an order parameter has a sharp jump at the order-disorder transition (**Figure 2.5c**). Analogous results are obtained if other peaks (for instance the first or the third peak) of *S(k)* are taken.

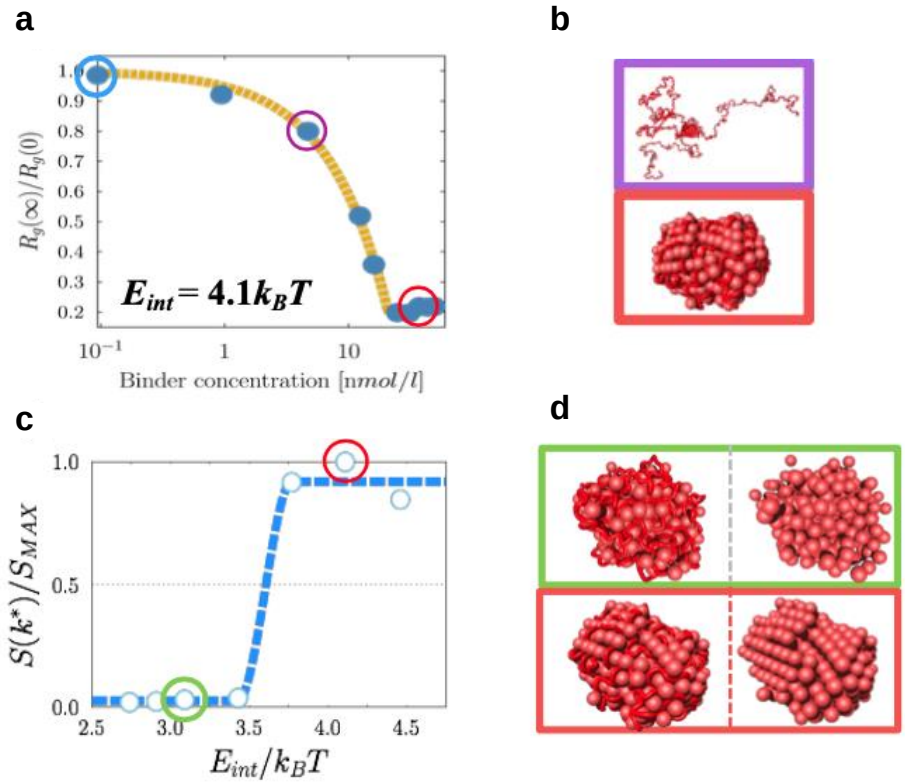


Figure 2.5: The order parameters of the transitions

a) The relative gyration radius of the SBS polymer as a function of the binders concentration. **b)** Two different spatial configurations at different concentrations. **c)** The structure factor sharp increase signals the order-disorder transition in the arrangement of the binders. **d)** The binders in the order (red) and disordered (green) configuration. (Figure adapted from Chiariello et al. 2016).

Contact probabilities

The conformational classes can be characterized by studying the **contact probability** $P_c(s)$ of a bead pair separated by a given genomic distance s. In the coil phase, the probability decreases as a power law, $P_c(s) \sim s^{-\alpha}$, with an exponent $\alpha \sim 2.1$, as predicted by polymer physics (De Gennes 1979). At the theta point, the exponent becomes 1.5, while in the closed state the probability has different shapes depending on whether the system is in the disordered or in the ordered state. In the former, it has an asymptotic plateau, with $\alpha = 0$, in the latter it decreases with $\alpha = 1$ (Chiariello et al. 2016). These properties are general features of this kind of systems. In the following chapters, we will consider more complex polymer models, which, by use of appropriate binding sites positioning, can describe the detailed three-dimensional structure of specific genomic regions. The finer details of the polymer configurations depend anyway on other aspects, like the position of the binding sites on the chain, the presence of 'inert' neutral sites and confinement. Furthermore, off-equilibrium, unstable conformations are also expected to be encountered in real chromosomal regions, for example during changes in the folding state.

2.4 Chromatin is a mixture of regions in different thermodynamics states

To check the ability of our model to recapitulate the average properties of chromosome folding, we fitted the experimental contact probability $P_c(s)$

obtained from Hi-C data with a linear combination of the different contact probabilities of each conformational class. Indeed, a single chromosome is likely to be a mixture of differently folded regions, with some regions more compact than others, and, at first approximation, the conformation of each region must belong to the stable thermodynamic states previously identified (**Figure 2.4**). In a simple coarse- grained model where chromatin is approximately a homo-polymer, the $P_c(s)$ is simply a linear combination of the different contact probability profiles. This combination depends on the relative abundances of the states in the mixture and on a scale factor necessary to map the bead size into genomic distances. The fit of genome-wide Hi-C average pairwise contact data as a function of the pairwise genomic separation is done by use of the Least Square Method (LSM). We compute the model predicted contact probability of a mixture of open and closed states by using the independently derived corresponding contact probabilities from the MD simulations of the homo-polymer chain. Then, we find the composition of the mixture of open and closed states that minimizes the distance between the predicted and experimental $P_c(s)$. We find that the model can fit the experimental contact probability data over very large length scales, from the sub-mega base scale up to the whole chromosome length, in both genome-wide averaged data and single chromosomes data (**Figure 2.6b-c**). Furthermore, we use data obtained from different experimental techniques (Hi-C, TCC and in-situ Hi-C), and the results are similar. From the data fit we obtain the percentages of open and closed state that best describe the chromatin in a cell type (averaged over all the chromosomes), or the percentage that best describe the chromatin for a fixed chromosome. We find different results depending on the cell type: in the human embryonic stem cells, the open state is approximately 75%, while in the differentiated cells as IMR90 fibroblast, this value is approximately 50% (**Figure 2.6d**), in agreement with expectations. If we consider contact probabilities extracted from data obtained from different experimental techniques, the fit gives similar results, with a closed ordered state of 40%, but a slightly different balance between the other states. For a fixed cell

type, we find a wide variability of these fractions among the different chromosomes, as shown in **Figure 2.6e** for IMR90 cell type. Generally, the percentage of open state decreases with the chromosome size, while the closed disordered phase increases, even though it represents a very small fraction.

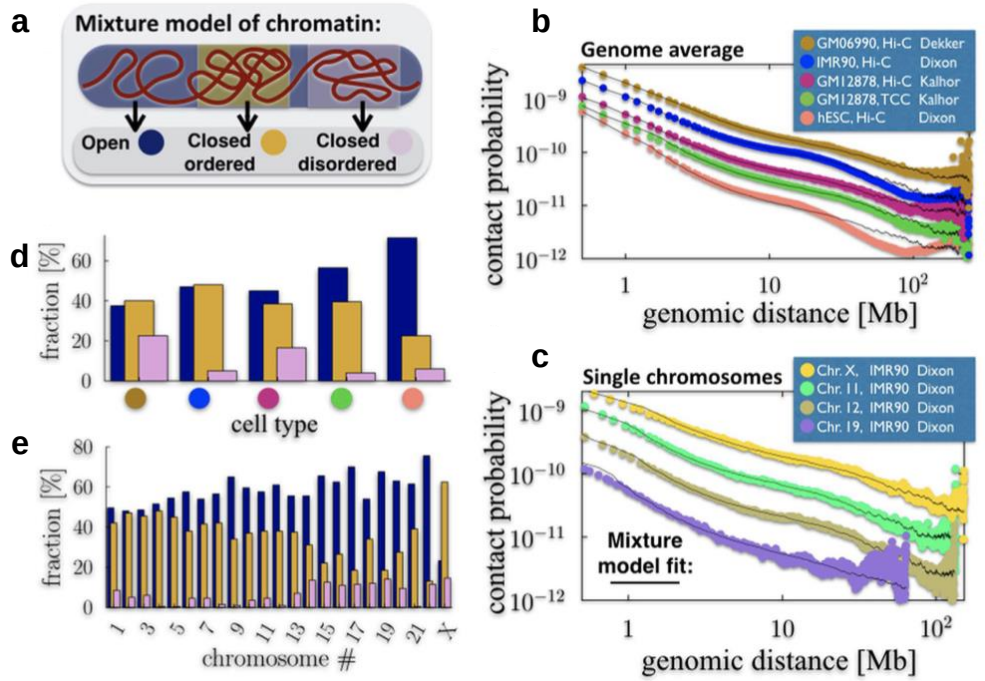


Figure 2.6: Chromatin is a mixture of regions folded in different thermodynamic states

a) We model a chromatin filament as a mixture of differently folded regions, each belonging to one of the stable conformational classes. **b)** Genome-wide average contact frequencies across human cell types, obtained from various experimental techniques, can be fitted from the sub-Mb to chromosomal scales by such a mixture model. **c)**. Single chromosome data (here IMR90 cells) can be similarly explained. **d)** Different cell types have a different chromatin composition (where blue indicates the open state, yellow and pink the closed ordered and closed disordered states, respectively), with hESC (orange circle) more open than differentiated cells, such as IMR90 (blue circle). **e)** Within a given cell type (here IMR90 cells) distinct chromosomes have also a different composition, with chromosome X formed mostly of closed regions, whereas gene rich chromosomes, e.g., chr.19, are up to 70% open. (Figure from Chiariello et al. 2016)

2.5 Chromatin architectural features are reproduced by the SBS model

2.5.1 Block-copolymer model

Although a simple SBS polymer as the one used in the previous section recapitulates the average properties of chromosome folding, more complex, local structures arise from Hi-C data (see Section 1.4). To correctly reproduce these structures, it is necessary to complicate the polymer chain by introducing more types of interaction. Therefore, we now consider a **block-copolymer**, with two types of beads (visually represented in red and green in **Figure 2.7**), that can interact only with its cognate kind of binder (red and green, respectively).

We begin with a 2-block copolymer where each block is made of 500 beads, one red and one green, and the entire polymer is made of 1000 beads in total (**Figure 2.7a**). Since we want to reproduce specific genomic structures, we are considering scales one order of magnitude lower than the chromosome modeling, which are the typical genomic lengths where chromatin is known to be subjected to compartmentalization (Lieberman-Aiden et al. 2009). So, here we suppose that the region is 10Mb long. To estimate the length scale, we proceed as before and we find that the bead has a diameter $\sigma = 64$nm and the time step results to be 0.003s. The concentrations and interaction energies are sampled so to cover the three thermodynamic stable states identified in the homo-polymer study. When equilibrium is reached, each block folds in the configurations discussed in the previous subsection, and two stable globular domains are formed (**Figure 2.7b**), that can be interpreted as TADs. In fact, these objects correspond to enriched interaction squares along the diagonal of the contact matrix, whose calculation details are explained in the subsection 2.5.3 (**Figure 2.7c**).

In the second block copolymer considered, the distribution of the colors along the polymer consists of four consecutive blocks (red-green-red-green, **Figure 2.7d**), each block 250 beads long. As before, each block can fold in the stable configuration and it forms, at the beginning of the dynamic process, a lower level

structure, resembling a TAD sequence. (**Figure 2.7d**, central matrix). When equilibrium is completely reached, the blocks of the same color interact, and the result is a hierarchical organization of higher-order structures, which is known to be a feature of the mammalian genome (Fraser et al. 2015). In the contact matrix (**Figure 2.7d**, right matrix), such organization is represented by a chessboard-like pattern. In the framework of our model, such structural features naturally emerge by specialization of the involved molecular factors under the laws of polymer physics.

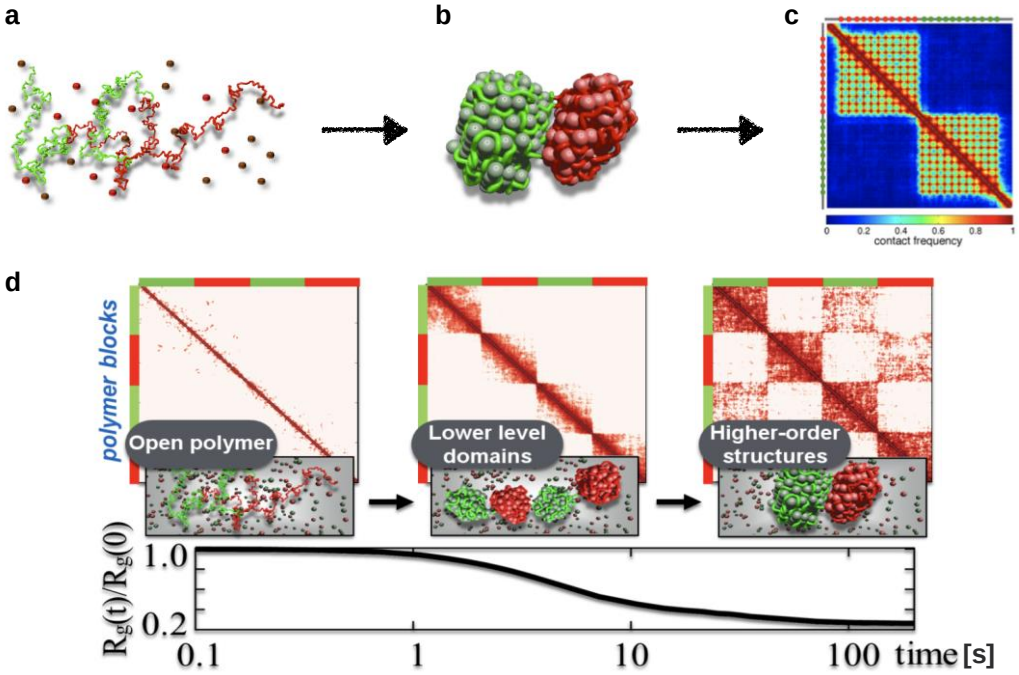


Figure 2.7: Formation of chromatin domains

a-c) Block-copolymer model made of two types of beads (red and green) interacting with two types of binders. At equilibrium, two chromatin domains arise and the contact map of such a system shows a TAD-like pattern. (Figure adapted from Barbieri et al. 2012) **d)** Dynamics of a block co-polymer model made of four consecutive blocks alternating their color (green- red-green-red). The gyration radius of the system decreases in time and a hierarchical self-assembly of domains spontaneously occurs. (Figure adapted from Chiariello et al. 2016)

The self-assembly of domains led to a symmetry breaking mechanism occurring in the spatial organization of the loci. Since TAD boundaries have been associated to an insulating role in the cell functionality, we consider the effect of the domains on the physical distance between pairs of sites having the same genomic distance (that is the contour distance along the polymer chain), but differently located with respect to the domain itself. In particular we focus on two cases where the sites can be symmetrically or asymmetrically located with respect to the boundary of the domains (see **Figure 2.8** bottom panel). In the closed state, we find that the spatial distance between the sites is larger in the symmetric case, while in the open state no difference is observed.

2.5.2 Distance distribution calculation

To measure the physical distances between two sites in the block co-polymer model, we consider two loci A and B, belonging to different blocks (A in the red block and B in the green block). In both cases, their contour distance is $d = 125\sigma$. In the symmetric case, they are equally distant from the boundary of the domain, while in the asymmetric case the site A is located at distance of 5σ from the domain boundary, and consequently the site B is 120σ from the boundary (so it is well inside the domain).

2.5.3 Contact matrices calculation

To obtain the pairwise contact matrices of the polymer models, we proceed in this way. We fix a contact threshold distance $k\sigma$, where σ is the length unit, and k is a dimensionless constant threshold, which we set to k=3.5. For a given spatial conformation of the polymer chain, we consider the distance r_{ij} between each bead pair i and j, ($i \neq j$, where i and j are bead indices along the chain). If $r_{ij} < k\sigma$, then we count a contact between the beads i and j. We then compute the average of these matrices across the different configurations in the considered polymer state.

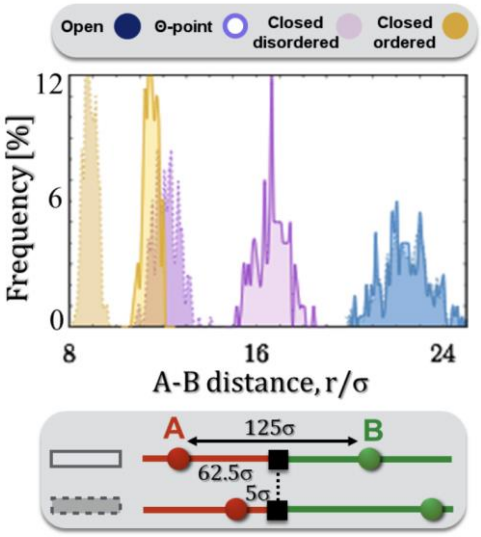


Figure 2.8: Symmetry-breaking mechanism

Pairs of sites with the same contour separation (here 125σ), differently positioned across a block boundary (see bottom panel), have the same average physical distances, r, in the open phase. Yet, in the closed states, the symmetry is broken by their different position relative to the boundary as the two pairs have a different physical distance, as seen from the corresponding distributions of *r*.

2.6 Multiple contacts landscape

In this section, we are going to discuss the probability of co-localization events of multiple sites, or many-body contact probability, which is a generalization of the pairwise contact probability previously defined. We first start looking at the probability of triple contact events $P_c(s_1, s_2)$ where the three beads are separated by different genomic separations s_1 and s_2. Then we compute the frequency of observing n>3 sites in physical proximity, and we do this in the three thermodynamic states previously identified. We find that in the closed states many-body contacts are exponentially more frequent than in the open state, as a function of the genomic distance.

To estimate the average number of many-body contacts involving simultaneous interactions of *n* beads occurring in a given polymer conformation, we count the

number of beads m_i that are in contact with the *i*-th bead within the fixed threshold *k* (here we use as above *k*=3.5) and the number of possible combinations of *n* simultaneous contacts that contain the *i*-th bead,$\binom{m_i}{n-1}$. We average that number over all the beads in the polymer. As normalization factor, we consider the number of total possible many-body contacts of *n* particles with the *i*-th bead, $\binom{N}{n-1}$. In **Figure 2.9**, we show the value of this frequency as a function of the multiplet complexity n, computed in the homopolymer case.

Although multiple interactions cannot be detected by Hi-C, our model highlights that they are likely to be an abundant structural component of chromatin, as is emerging from new researches in the field (Beagrie et al. 2017; Olivares-Chauvet et al. 2016; Quinodoz et al. 2018). That hints towards an important functional role of chromatin domains where multiple regulatory regions (like enhancers) can loop simultaneously onto a given target (gene promoter) with a much higher probability than in open regions.

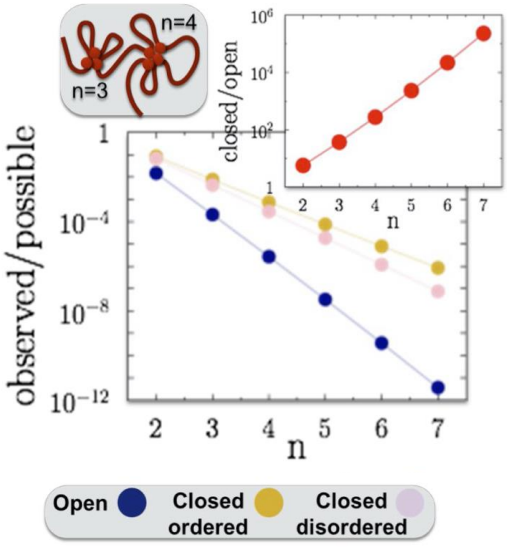


Figure 2.9: The multiple contact profile

The plot shows the frequency of observing n sites in simultaneous physical contact (normalized by the number of possible combinations of n sites) along the SBS homo-polymer discussed previously. The inset shows the ratio in the compact-disordered and open states. (Figure adapted from Chiariello et al. 2016).

Chapter 3 - Polymer physics investigation of real genomic regions from pairwise contact data

In the previous chapter, we have seen how the simple homo-polymer SBS model is able to recapitulate with a good degree of accuracy the average behaviour of the chromosome structure in a wide range of genomic lengths (from the sub-Mb scale up to the whole chromosome scale). The introduction of a second type of interaction (the red-green model) allowed us to explain other aspects of the chromatin architecture and to highlight mechanisms that could possibly have important functional roles: the existence of domains, the symmetry-breaking in the distance distribution, and the hierarchical structure of the experimental Hi-C contact matrices. In this chapter, we generalize our SBS model by introducing a multicolour polymer, where each color interact only with its cognate type of binder. We will show that, with this generalized model, the 3D folding of real genomic regions can be explained. The specialization of the generalized model to each of the different cases here presented will be obtained with the application of the PRISMR inference method (described in Section 3.1), which aims to infer the minimal factors that shape the folding of a chromatin locus and its equilibrium 3D structure under the laws of physics, without a-priori assumptions. The PRISMR algorithm takes as input an experimental contact matrix and gives as output the optimal SBS polymer model of the corresponding genomic locus. A brief description of the PRISMR algorithm will be given in Section 3.1. In the subsequent sections, we show that, albeit our polymer models are derived from pairwise contact matrices, they can be used to derive any further aspect of folding, such as the ensemble of its 3D conformations, not directly accessible from the interaction data. A first application of PRISMR will be shown in Section 3.2, where we use our generalized SBS model to investigate the role of the folding on enhancer-promoter communication in developing limbs at the *Pitx1* locus (Kragesteen et al. 2018). In Section 3.3 we will presents the results about

the modeling of the Sox9 locus, containing a very important gene for the cell functionality (Franke et al. 2016). Then, in Section 3.4 we will model the Xist locus, which is another very important region (Giorgetti et al. 2014; Nora et al. 2012), and we will apply the model to predict the effect of a deletion variant. Finally, in Section 3.5, the mechanism of enhancer-promoter interaction will be further investigated (together with the role of the structural protein CTCF) by a recent application of our model to the *Shh* locus, containing a gene crucial in posterior limb development (Paliou et al. 2019).

Most of the material presented in this Chapter, including figures, paragraphs and sentences, is adapted or taken literally from the following papers, which I co-authored: (Chiariello et al. 2016; Kragesteen et al. 2018; Paliou et al. 2019).

3.1 The PRISMR algorithm

In the generalized SBS model (**Figure 3.1**), the different types of interacting polymer beads within the chain are identified by different colors, while 'gray' marks inert beads in our notation, i.e., sites that do no interact with any binder except for the excluded volume effects. The number of different colors allowed in the model will be denoted by *n*. A given SBS polymer is then completely characterized by the number of the different binding sites and by their arrangement along the polymer chain. To determine the configuration of the system related to a real genomic region, here we use the **PRISMR** (Polymer-based Recursive Statistical Inference Method) algorithm, which aims to find the minimal number and types ("colors") of binding sites in a SBS polymer chain, and their position along the chain, that best reproduce the input contact matrix of a given chromosomal locus (**Figure 3.1**). Although here we focus on the SBS polymer model to describe a chromatin filament, the PRISMR algorithm can be easily generalized to different models. A detailed description of PRIMSR can be

found in (Bianco et al. 2018), here we just summarize the key points of the algorithm.

An SBS polymer model of a genomic region is composed by *L* beads, depending on the resolution of the input contact matrix of the region, such as a Hi-C pairwise matrix. For instance, a 10Mb locus at 10kb resolution will be partitioned in *L*=1000 bins. Furthermore, we split each bin in *r* different sub-units, considering that a single DNA bin along the locus could include many binding sites and could interact with different molecular factors. In this way, we can resolve finer details such as the different binding sites located within a bin. The minimal value of r required to explain the input data is one of the outputs of PRISMR. However, in the practical cases discussed in this chapter, the number of binding sites within a bead is typically smaller than the total number of different types of binders, hence $r = n$ is a safe assumption and will be adopted in the description of the algorithm.

To find the SBS model which, at equilibrium, best describes the input contact matrix, PRISMR minimizes a given cost function, *H*, which includes two terms accounting for two main requirements, the necessity to fit well the input data and, at the same time, the attempt to avoid overfitting. The first term, H_0, considers the distance between the input contact matrix, $C_{exp}(i,j)$, from the one predicted by polymer physics thermodynamics, $C(i,j)$. H_0 is normalized to the average contact frequency and to the total number of sites. A constant scale factor F, whose value is returned by the optimization algorithm itself, is used to map the total counts in experimental matrix data onto the derived physical contact frequencies of loci in our 3D models. The second part of the cost function is a Bayesian term (a chemical potential in Statistical Mechanics) which penalizes the addition of new interacting beads, and it is indicated with H_λ. It is proportional to the total number of colored sites of the polymer through a parameter λ and it is normalized to the total number of beads of the polymer chain, N.

For a given value of the parameters *n* and λ, PRISMR samples the huge space of all allowed color arrangements of the chain (which has (n+1)N elements) in order to find the one which minimizes the above cost function. To this aim we employ a standard Simulated Annealing (SA) iterative procedure (Kirkpatrick, Gelatt, and Vecchi 1983; Salamon, Sibani, and Frost 2002). Schematically, each SA step consists in randomly changing the color of a polymer bead, compute the average contact matrix of the new polymer, evaluate the new cost function, compare it with the cost function in the previous step and accept or reject the color change on the basis of the Metropolis algorithm. These steps are iteratively repeated until convergence (Bianco et al. 2018). The procedure is repeated to search for the minimal allowed value of *n* and then for the maximum of λ required to fit the data within a predefined accuracy. The best color arrangement is the final output of the algorithm, returning the minimal required number of binding domains and their best positioning along the SBS polymer to explain the input data within a given accuracy.

The calculation of C(*i*,*j*) is a computationally demanding step of PRISMR and can be achieved, for instance, either by Molecular Dynamics computer simulations, which may require huge computational efforts, or by enhanced folding algorithms (see, e.g., MELD MacCallum, Perez, and Dill 2015), which albeit approximate can be much faster. Here, to speed up computation and to make our procedure feasible over genomic scales, we considered an approximation typical of mean-field methods of Statistical Mechanics. In our approach, the average contact frequency over the thermodynamics ensemble of allowed 3D conformations of the polymer, is estimated from the average contact frequency between two sites at the same genomic separation in a homopolymer SBS model of N beads (for a detailed description see Bianco et al. 2018). The advantage of the mean-field approximation is that the contact matrices can be calculated in an easy way and used throughout the SA Monte Carlo procedure to

make computation times feasible. Next, we can run MD simulations of the optimal model found by the SA procedure to derive its contact matrix without any approximation.

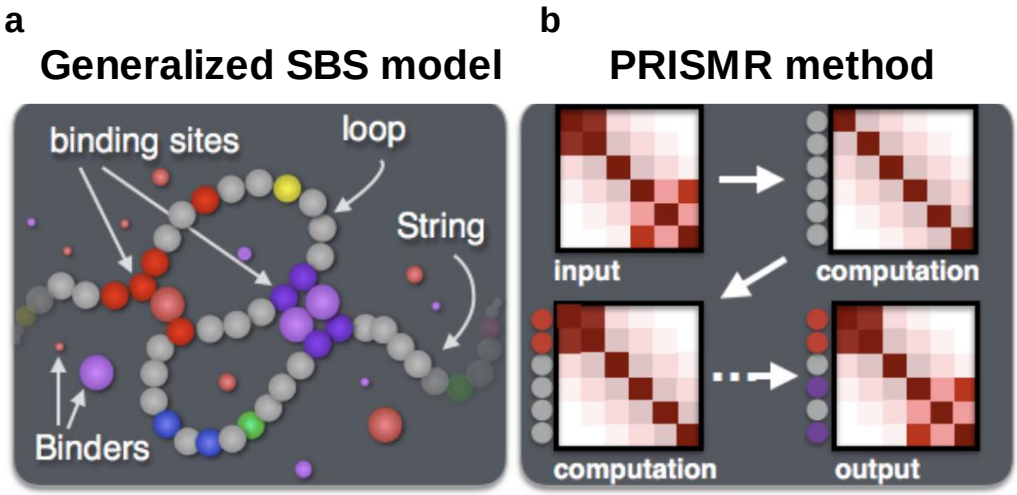


Figure 3.1: The PRISMR method

a) The generalized SBS model, with the different types of binding sites (and their cognate bridging molecules) sowed in different colors. **b)** PRISMR samples the possible states of a given SBS polymer model to find the one which best describes a input contact matrix. Figure adapted from (Bianco et al. 2018)

3.2 Modeling of the *Pitx1* locus in mouse limb cells

As discussed in the first chapter, the mechanism of gene regulation involves enhancers, which are short stretches of DNA that drive gene expression over long distances by physically contacting their target gene promoter. In this section, we employ the polymer physics models just described above to dissect the basic mechanisms of enhancer-promoter interaction and their role in gene regulation at the **mouse *Pitx1* locus**, a gene which is critically required for establishing the identity and differentiation (DeLaurier, Schweitzer, and Logan 2006) of **hindlimbs** (posterior limb). Specifically, during limb development, the *Pitx1* gene is only expressed in hindlimbs, but not expressed in **forelimbs** (anterior

limb). However, we demonstrate that *Pitx1* is regulated by an enhancer (named *Pen*) which displays activity in both tissues. The restriction of this activity to the hindlimb is associated with tissue-specific differences in three-dimensional chromatin structure enabling the *Pen* enhancer to control *Pitx1* transcription in hindlimbs only. This study has been developed in collaboration with Prof. Stefan Mundlos' research group at Max Plank Institute, Berlin, who performed all the experimental part (Kragesteen et al. 2018).

3.2.1 The regulatory landscape of the *Pitx1* gene

In the mouse genome, the *Pitx1* gene is located on the chromosome 13 (chr13). To determine the position of regulatory elements controlling *Pitx1* in hindlimbs, we examined the Capture-C (a 4C technique) interaction profiles and chromatin immunoprecipitation sequencing (ChIP-seq) for the enhancer mark H3K27ac (see Section 1.2) in hindlimbs, at mouse embryonic day (E) 10.5 (**Figure 3.2**). The chromatin interaction profile shows that the *Pitx1* regulatory landscape extends over 400kb and forms several chromatin loops corresponding to H3K27ac peaks, termed regulatory anchors (RAs) 1–5. In the next, we will specifically focus our analysis on **RA2**, the *Pitx1* promoter, and **RA5**, which marks *Pen* (pan-limb enhancer), an enhancer showing activity in both forelimb and hindlimb buds.

3.2.2 Polymer modeling of Capture Hi-C (CHi-C) data highlights a switch in 3D chromatin architecture

In order to characterize 3D chromatin folding with a higher definition, we analysed CHi-C interaction maps encompassing a 3Mb long region around *Pitx1* in mouse forelimbs and hindlimbs at E11.5. We found that the locus is divided into subdomains separated by the previously characterized regulatory anchors and that multiple interactions occur between the various RAs (**Figure 3.3c, d** top triangles).

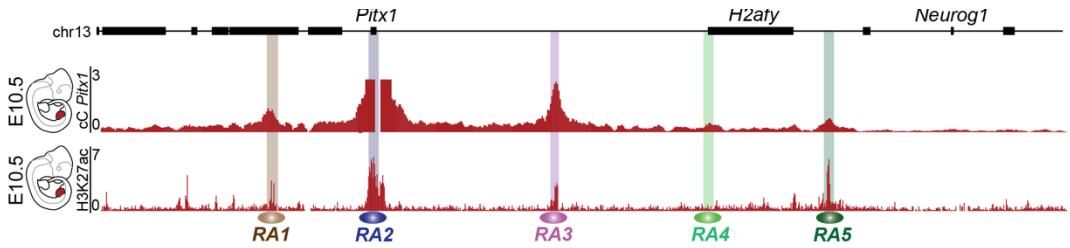


Figure 3.2: *Pitx1* regulatory landscape includes a pan-limb region

The upper track shows the Capture-C data in E10.5 hindlimbs using the Pitx1 promoter as the viewpoint demonstrating chromatin interactions with regulatory anchors (RA1–RA5). The bottom track displays H3K27ac ChIP-seq enrichment profile in E10.5 hindlimbs. (Figure adapted from Kragesteen et al. 2018)

The *Pitx1* regulatory landscape displays extensive differences between the tissues, as highlighted by the subtraction of forelimb and hindlimb CHi-C contact maps (**Figure 3.4a**). On the one hand, we could recapitulate the forelimb-specific repressive interaction between *Pitx1* and the functionally unrelated *Neurog1* gene (blue arrow in **Figure 3.4a**), while, on the other hand, hindlimb-specific interactions between *Pitx1* and RA1, RA3, and *Pen* occur (red arrows in **Figure 3.4a**). To understand how these differences are translated into the locus 3D conformation, we employed our polymer physics approach.

Based on the CHi-C interaction data at 10kb resolution, our PRISMR inference procedure gives a total of $n=14$ different types of binding sites, whose position and abundance along the genome is showed by the histograms in **Figure 3.3a, b**, where a different color is associated with each type of binding site. The derived ensembles of polymer structures allowed us to visualize the conformational changes in the 3D space and to perform quantitative measures, such as physical distances among regions of interest (**Figure 3.6**), helping to provide a clearer biological interpretation. As shown in **Figure 3.3e**, in forelimbs the locus segregates into two chromatin hubs, containing (1) *Pitx1*, RA3 and *Neurog1* (blue, pink and red spheres, respectively) and (2) *Pen* and RA4 (dark green and light green spheres, respectively). This spatial conformation is such that *Pen* and

Pitx1 are separated from each other and the repressed gene *Neurog1* is close to *Pitx1*, so preventing its activation. Conversely, the hindlimb 3D structure is partitioned in three major hubs (**Figure 3.3f**), one containing RA1 only, another containing *Pitx1* and RA3, and the last one RA4, *Pen* and *Neurog1*. Here, the physical proximity between *Pitx1* and its enhancer *Pen* ensures a correct regulation of the gene. Therefore, chromatin spatial configuration restricts the activity of *Pen* to the hindlimb tissue by separating the enhancer from its promoter in forelimbs.

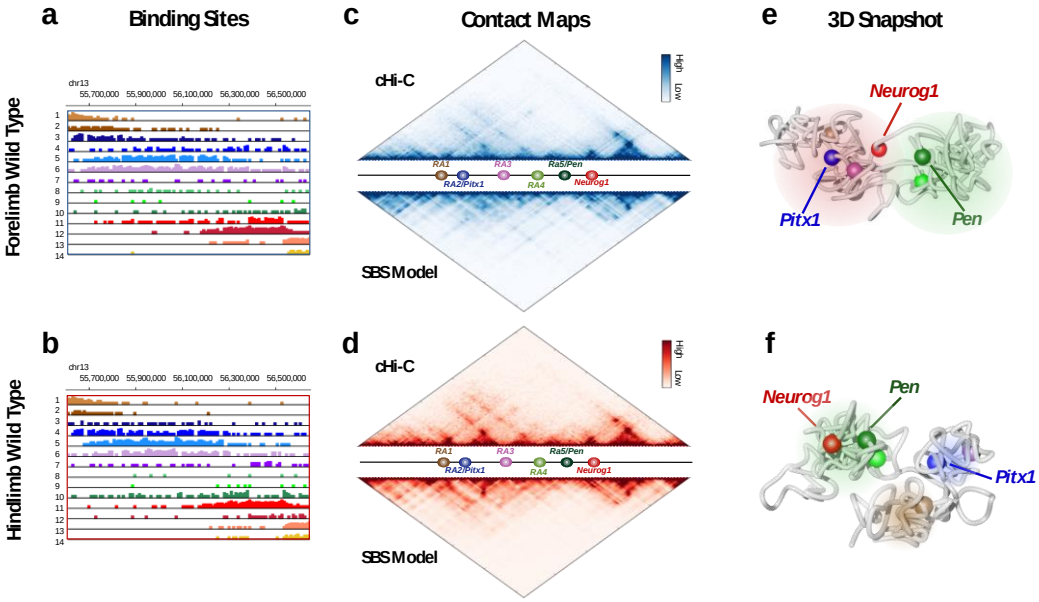


Figure 3.3: Tissue-specific 3D conformation reconstructed by the SBS model

a-b) Histograms displaying the position and abundance of the 14 different types of binding sites along the genome, in forelimbs (top) and hindlimbs (bottom) as derived from the E11.5 CHi-C data. Each binding site is displayed with a different colour. **c-d)** CHi-C (above) and model (below) derived contacts maps display high similarities. The Pearson correlation, r, and the distance corrected Pearson correlation, r´, between the CHi-C and SBS matrices are r=0.98 and r'=0.84 in forelimb, r=0.98 and r'=0.82 in hindlimb. **e-f)** A representative 3D-structure of the locus in forelimb (top) and hindlimb (bottom), selected from the ensemble of 'single-cell' model derived conformations. (Figure adapted from Kragesteen et al. 2018)

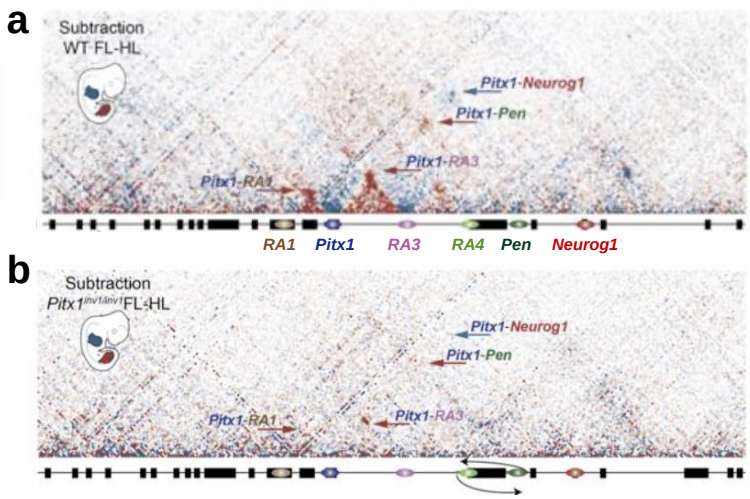


Figure 3.4: *Inv1* inversion induces a spatial reorganization in forelimb

a) CHi-C subtraction between forelimb and hindlimb Wild Type (WT). Chromatin interactions more prevalent in forelimb or hindlimb tissues are shown in blue or red, respectively. The interaction between *Pitx1* and *Neurog1* that is more prevalent in forelimbs is indicated with a blue arrow. Chromatin interactions between *Pitx1* and *Pen* that are more prevalent in hindlimbs are indicated with red arrows. **b)** Subtraction maps of $Pitx1^{inv1}$ forelimbs and WT hindlimbs. Note the high similarity of 3D chromatin structure between both tissues in comparison to WT animals showed in panel a. (Figure adapted from Kragesteen et al. 2018)

3.2.3 The forelimb *Inv1* inversion effects are well captured by our polymer modeling

One of the major differences between forelimbs and hindlimbs is the interaction of *Pitx1* with the repressed gene *Neurog1*. To investigate if the inactive forelimb configuration can be converted to the active hindlimb state and induce *Pitx1* transcription, we perturbed the regulatory landscape of the locus by inverting, in the forelimb buds, a 113kb fragment containing *Pen* and RA4 (**Figure 3.5b**, horizontal bar). In contrast to the healthy Wild Type (WT) tissues, the chromatin organization in the forelimb *Inv1* locus, indicated as $Pitx1^{Inv1}$, was nearly identical to the WT hindlimbs, as showed in the CHi-C subtraction matrices of **Figure 3.4b**. Subtraction between WT and $Pitx1^{Inv1}$ virtual Capture-C (obtained by considering the column in the contact matrix corresponding to the considered

viewpoint) from the *Pitx1* viewpoint showed several hallmarks of WT hindlimb architecture in *Pitx1*^{Inv1} forelimbs: a gain of interaction between *Pitx1*, RA3, and *Pen*, as well as a diminished interaction with *Neurog1* (not shown here, see Kragesteen et al. 2018). Interestingly, 3D modeling of *Pitx1*^{Inv1} showed a hindlimb-like spatial conformation, with the formation of three chromatin hubs and a closer proximity between *Pen* and *Pitx1* compared to forelimb wild type (**Figure 3.5c**). Because of this increased proximity, the *Pitx1* gene in the *Pitx1*^{Inv1} tissue displayed a 44-fold increase in expression and limb malformation was detected on these mice, that displayed a partial arm-to-leg transformation. As a control, a slightly smaller genomic region has been inverted (99kb, indicated by *Pitx1*^{Inv2}) which leaves *Pen* at its original location. *Pitx1*^{Inv2} embryos had a normal skeleton (not shown here, see Kragesteen et al. 2018) and did not show ectopic expression, thus confirming the direct effect of the *Pen* enhancer and its role on the mis-expression of *Pitx1* in *Pitx1*^{Inv1}.

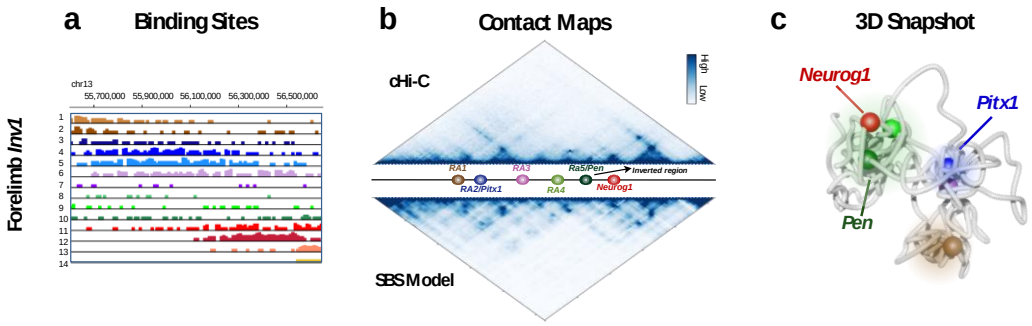


Figure 3.5: The SBS model correctly describes the architectural changes caused by the *Inv1* genomic inversion

a) Histograms displaying the position and abundance of the 14 different types of binding sites along the genome in *Pitx1*^{Inv1} forelimbs. **b)** Comparison of CHi-C (above) against SBS model (below) derived contacts maps in *Pitx1*^{Inv1} forelimbs show high similarities. The Pearson correlation, r, and the genomic distance corrected Pearson correlation, r´, between the matrices are r=0.97 and r´=0.74. **c)** Representative 3D structure of the locus in *Pitx1*^{Inv1} forelimbs, selected from the ensemble of 'single-cell' model derived conformations. Note the similarity between this conformation and the one from WT hindlimbs in Figure 3.3 panel f. (Figure adapted from Kragesteen et al. 2018)

Summarizing, the application of the SBS model to the *Pitx1* gene region allowed us to translate the pairwise information in the ensemble of 3D conformations of the region under study, obtaining significant insights into the spatial organization of the locus, as the hindlimb-specific three hubs and forelimb-specific two hubs organization, not directly accessible from the experimental contact data. Additionally, physical distances among any region of the locus can be computed from the ensemble of polymers, confirming the greater proximity between *Pen* and *Pitx1* in hindlimbs and between *Neurog1* and *Pitx1* in forelimbs with respect to the other tissue.

3.2.4 Simulation details of the *Pitx1* polymer model

We applied our SBS model in forelimbs, hindlimbs, and *Inv1*-forelimbs to a broad genomic sequence encompassing the mouse *Pitx1* regulatory landscape to avoid boundary effects and focused on chr13:55,600,000–56,650,000 (mm9). To derive an ensemble of the model equilibrium 3D conformations we implemented Molecular Dynamics (MD) computer simulations (see Section 2.2). Specifically, we used a polymer chain of N=1785 beads, so the elementary bead of the polymer is approximately 17nm. The molar concentration of binders is c = 135 nmol/l and the scale of the bead-binder interaction energy is E_{int}=1.0 k_BT and E_{int}=8.1 k_BT, corresponding to the coil and globule conformational state of the polymer, respectively. The dimensionless friction coefficient is set to ζ=0.5 (Chiariello et al. 2016; Kremer and Grest 1990). The MD integration time step is Δt=0.012 (Rosa and Everaers 2008) and we let the system evolve up to $5x10^8$ time steps, to reach stationarity. An ensemble of at least 10^2 different equilibrium configurations is derived by MD for each of the considered case.

To test our models against the experiments, we compared the CHi-C data with the average contact matrix obtained from the ensemble of 3D model

conformations derived via MD. The contact maps were computed following the approach described in Section 2.5.3 using, as parameter for the interaction threshold k=8. To take into account the effects of cell population heterogeneity, that is, the possibility that the locus could be in different states (coil/globule) in different cells, we considered the contact matrix of the coil/globule mixture which maximized the Pearson's correlation coefficient, r, with the CHi-C data (Bianco et al. 2018). An 80% (coil) – 20% (globule) mixture well describes all cases. To account for the effects of genomic proximity beyond Pearson's r between model-predicted and CHi-C contact matrices, we also computed the **distance-corrected Pearson's correlation coefficient r´**, that is the correlation between the two matrices where the average contact frequency at each genomic distance has been previously subtracted. The MD model versus the CHi-C Pearson's r is 0.98 in WT forelimb, 0.98 in WT hindlimb, and 0.97 in the *Inv1*-forelimb (**Figure 3.3c, d** and **Figure 3.5b**); the distance-corrected correlation r´ is 0.84 in WT forelimb, 0.82 in WT hindlimb, and 0.74 in the *Inv1*-forelimb (here, strong outliers above the 90th percentile were excluded).

Finally, to capture the structural differences between forelimb and hindlimb, we measured the physical distances among all the regions of interest. The relative distance changes shown in **Figure 3.6** represent to the ratio $(d_{FL}- d_{HL})/d_{FL}$ of the distances in forelimbs and hindlimbs (respectively d_{FL} and d_{HL}) among its key regulatory regions averaged over the discussed state mixture; this confirms the above described scenario.

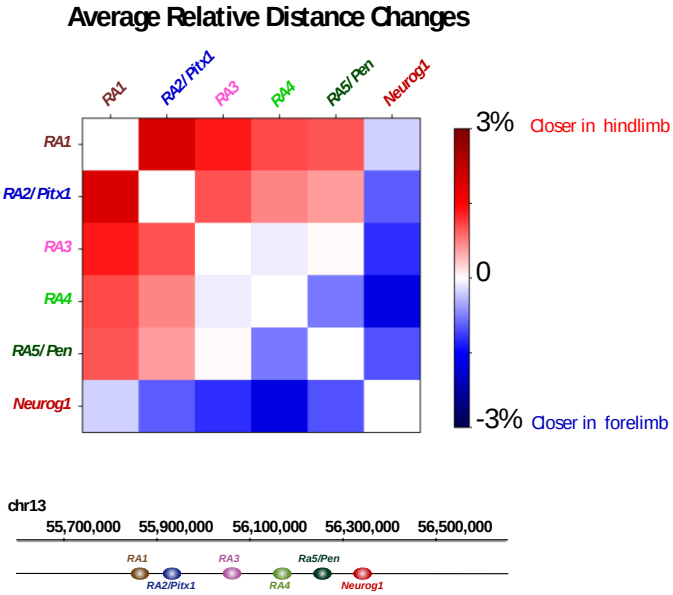


Figure 3.6: Measuring physical distance changes among the regions of interest

Heatmap showing relative changes in physical distances between forelimb and hindlimb 3D structure as measured by the polymer model. Blue squares indicate that the corresponding regulatory elements are closer in forelimbs, while red squares indicate that they are closer in hindlimbs. (Figure adapted from Kragesteen et al. 2018)

3.3 Polymer physics investigation of the *Sox9* genomic region in mouse embryonic stem cells (mESC)

To test the general applicability of the SBS model and the PRISMR method in reconstruct and visualize the 3D architecture of real loci, we modelled another important genomic region containing *Sox9*, a gene which is crucial in sex development and whose mutations are linked to severe congenital diseases (Franke et al. 2016). In particular, we focused on the 6Mb long region chr11:109000000-115000000 (mm9), which includes both gene rich and gene desert regions, as shown in **Figure 3.7a**. The Hi-C data set used to infer the model comes from mouse embryonic stem cell (mESC) line at 40kb resolution (published in Dixon et al. 2012) and is normalized as described in Yaffe and Tanay 2011 (**Figure 3.7c**). The optimal number of binding domains (colors) returned by our inference procedure PRISMR is, in this case, $n = 15$ and the

polymer is made of *N*=2250 beads. So, the elementary bead of the polymer has a genomic content of L/N=2.67Kb and the size of the bead is 26nm, as follows from the calculation described in Section 2.2. The parameters used in the MD simulations are the same used for the modeling of the *Pitx1* locus (see subsection 3.2.4).

As depicted in **Figure 3.7a** (bottom panel), the binding domains tend to overlap with the different TADs existing in the locus, but they also overlap with each other, producing interactions between TADs that result in the hierarchical structure (metaTADs) visible in the original experimental matrix. Once obtained the optimal arrangement of the binding sites along the polymer, we performed MD simulations to reconstruct the 3D structure of the region. Then we computed, from the ensemble of configurations, the model-derived contact maps and compare it with the experimental data. As in the previously discussed *Pitx1* case, we considered separately the open phase (i.e. the SAW conformational class) and the closed phase (i.e. the equilibrium phase after the complete folding of the polymer) and we selected the open-closed mixture that maximizes the Pearson's correlation coefficient between model inferred and Hi-C data. The contact matrix returned by our model is very similar to the experimental one, as the value of their Pearson's correlation coefficient r=0.95 (**Figure 3.7c**). **Figure 3.7b** shows a single typical configuration of the locus in the closed state, with the relative positioning of *Sox9* and other important genes, which are *Kcnj2* and *Slc39a11*. Interesting features can be obtained by the conformational ensemble. For instance, we can consider the transcription starting sites (TSS) of the three nearby genes in the locus and compute their physical distances. We found that the *Sox9* and *Kcnj2* TSS, having a genomic separation s =1.72Mb, have an average physical distance d =1190 nm, while the *Sox9* and *Slc39a11*, having a genomic separation of s = 0.46Mb (four times smaller) have a spatial distance d = 590 nm, so the two pairs are proportionally closer, as they belong to consecutive

regional areas. The Sox9 locus is marked by many-body contacts which are exponentially more abundant than expected in a randomly folded conformation (**Figure 3.7d**, error bars within symbol size). The self-assembly of the locus spatial structure starts from a totally random SAW initial state and proceeds hierarchically, passing through early local domains folding into larger and larger domains that cover the whole locus (**Figure 3.7e**).

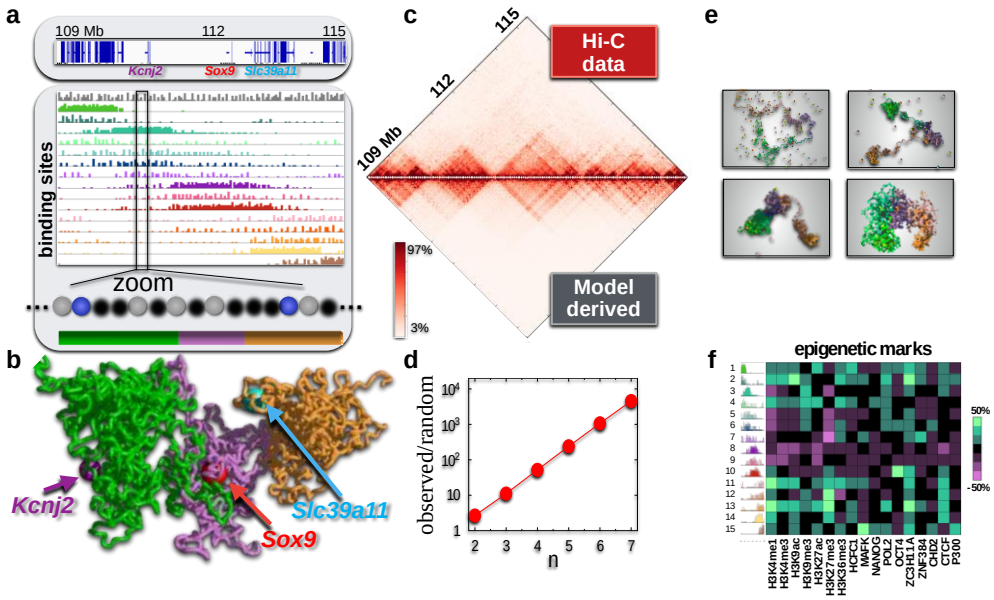


Figure 3.7: Polymer physics captures the folding of the Sox9 locus

a) Top panel: the considered region in mESC cells, with some important genes marked. Bottom panel: distribution of the binding domains (different colors) of the polymer model best explaining the Hi-C contact data; their abundance is represented as a histogram over the genomic sequence. The bar at the bottom highlights three main regional areas to help 3D visualization. **b)** A snapshot of the *Sox9* locus in its closed state as derived by the polymer model, with the position of some key genes highlighted. **c)** The model derived pairwise contact frequency matrix (bottom) has a 95% Pearson correlation with Hi-C experimental data (top). **d)** Many-body contacts of *n* sites result to be exponentially more abundant than in random SAW conformations. This could help the simultaneous colocalization of multiple functional regulatory regions. **e)** The Sox9 genomic region dynamics starts from an initially open conformation and self-assemble hierarchically in higher-order structures in approximatively 20s. **f)** Heat map showing the Pearson's correlations between the binding domains and some chromatin features. Single colours correlate with a combination of different marks. (Figure adapted from Chiariello et al. 2016)

Next, to investigate the molecular nature of the inferred binding domains, we compared the information of their positioning with epigenetic data available in mESC (see next chapter for an accurate description of the epigenetic data analysis). The heatmap in **Figure 3.7f** shows the correlation coefficient between the genomic positions of the binding domains and a number of published chromatin features, such as histone modifications and transcription factors, from the ENCODE database (Dunham et al. 2012). We find that single colours do not correspond to single molecular factors, as each usually correlates with a combination of different marks. Many binding domains also correlate with CTCF, known to play an important role in chromatin architecture through the formation of chromatin loops (Fudenberg et al. 2016; Nora et al. 2017; Sanborn et al. 2015). However, they also correlate with other, different groups of ENCODE marks, returning the view that additional factors can aid, specify or constrain CTCF linked interactions. Few of the binding domains (e.g., type 8 and 9, **Figure 3.7a**) do not correlate with the considered epigenetic features, and result to be associated with the central, gene poor, region of the locus.

3.4 Predicting the effect of genomic mutations

In the previous two sections we have seen how the application of the SBS polymer model to real genomic loci is an important tool to study the chromosome spatial organization, both in normal and mutated genomes. In this section we will show how it is possible, with our model, to predict the effect on the spatial organization of a locus generated by a mutation along the genomic sequence. To this aim, we considered the *Xist* genomic region (chrX:100298000-101373000), since experimental data are available (Nora et al. 2012) for the Wild Type (WT) and for a deletion variant (indicated as *ΔXTX* deletion), so we can directly test the results of our simulated predictions with a completely independent dataset.

Precisely, we analysed a 5C data set from mouse ES cell line for the WT (**Figure 3.8b**, top triangle), and 5C data from XO mouse ES cell line for the deletion *ΔXTX* (**Figure 3.8c**, top triangle), both mapped at 20kb resolution. We studied a region 1.3Mb long around the *Xist* gene (**Figure 3.8a**). Starting from the WT, we used PRISMR to obtain the number of different colors (10 in this case) and their distribution along the polymer (**Figure 3.8a**) best describing the input data. Next, we implemented in silico the *ΔXTX* deletion on the WT polymer model and performed MD simulations starting from a set of completely independent initial configurations.

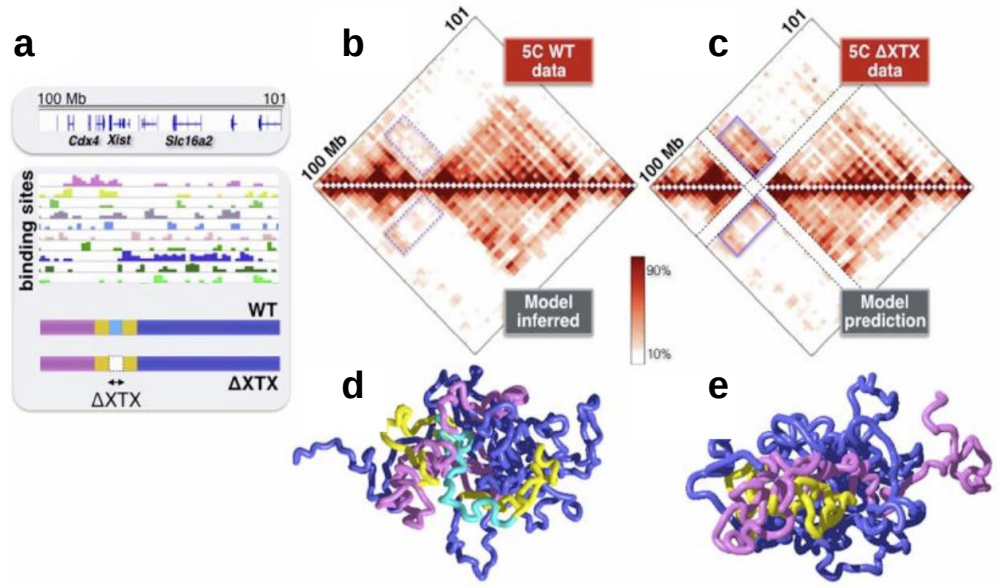


Figure 3.8: The SBS model predicts the effect of a deletion at the *Xist* mouse locus

a) Top panel: the *Xist* locus in mESC-E14 with some gene highlighted. Bottom panel: representation of the 10 binding domains inferred by the polymer modeling (one domain per histogram) and, below, the color scheme used in the polymer representation reflecting the abundance of the colors in the considered region. To help visualization, the *ΔXTX* deletion is colored in cyan and the flanking regions in yellow. **b)** The model inferred contact matrix (bottom) has a Pearson's correlation 0.96 with 5C experimental data (top). **c)** The contact matrix predicted by the WT model after the implementation of the *ΔXTX* deletion (bottom) reproduces with a high degree of similarity (correlation 91%) the ectopic interactions (magenta box). **d)** A snapshot of the Xist locus in its closed state. **e)** A snapshot of the predicted *ΔXTX* 3D structure where the yellow regions come closer in space after the deletion of the cyan segment of panel d. (Figure adapted from Chiariello et al. 2016)

As shown in**Figure 3.8c** (bottom triangle), the predicted matrix has a pattern of ectopic interactions compared to the WT case strikingly similar to the one detected in the experimental *ΔXTX* contact data (**Figure 3.8c**, top triangle), although the deletion was a very small change of the polymer model (N=540 beads in the WT and N=510 beads for *ΔXTX*). Interestingly, a good agreement between the experimental contact matrix and the map computed from the simulations was obtained in both WT and mutant cases; the values of the Pearson's correlation coefficient between model-derived and experimental matrices are r=0.96 (WT) and r=0.91 (*ΔXTX*). The inferred 3D structures of the locus allow us to visualize the effect of the deletion on its spatial organization. As we can see from the structure in **Figure 3.8d-e**, the yellow regions, close to the deletion (cyan part), are spatially repositioned with respect to each other, and contribute to form the ectopic contact between regions sharing the same binding sites.

3.5 Preformed topology at the *Shh* murine locus

In this last section, the mechanism of enhancer-promoter communication will be further investigated by a polymer physics SBS model of the ***Shh*** genomic locus in mouse limb buds, used as testbed. More precisely, the preformed interaction between the *Shh* gene and its limb-specific unique enhancer ***ZRS*** will be perturbed by using targeted genetic disruption of specific CTCF sites, known to be a crucial component of genome architecture (Fudenberg et al. 2016; Nora et al. 2017; Sanborn et al. 2015). The present study has been developed in collaboration with Prof. Stefan Mundlos' research group at Max Plank Institute, Berlin, who performed all the experimental part (Paliou et al. 2019).

3.5.1 *Shh* gene is regulated from a tissue-specific, unique enhancer

The *Shh* gene is expressed in the posterior part of the developing limb, within the zone of polarizing activity. This highly specific expression pattern is critical to ensure the development of limb extremities. In the limb bud, *Shh* is regulated by a single enhancer, the *ZRS*, the deletion of which results in a complete *Shh* loss of function in the limb, leading to digit aplasia (Sagai et al. 2005). The *ZRS* is located almost 1Mb away from the *Shh* promoter (**Figure 3.9a**), but despite this large genomic separation, FISH experiments have demonstrated complete colocalization of the *Shh* promoter and the *ZRS* in posterior limb buds, where *Shh* is expressed. Moreover, in contradiction to many enhancer–promoter interactions that are tissue- and time-specific, the two elements are found in close proximity even when inactive, suggesting a **preformed** mode of interaction (Amano et al. 2009; Williamson et al. 2016). In the limb buds, three major CTCF binding events occur on either side of the *ZRS*, which we termed **i4** (intron 4), **i5**, and **i9** (**Figure 3.9a**) and could account for the preformed interaction between the *ZRS* and *Shh*. However, how this preformed topology is established or how it relates to the expression of *Shh* in developing limb buds remains unclear.

3.5.2 Modeling of the Shh locus

To shed light on the mechanism regulating the *Shh-ZRS* interaction, we modeled the 3D architecture of the *Shh* locus by using CHi-C data produced in the E10.5 limb buds in two different cases: Wild Type (WT) and *ΔCTCF i4:i5*, where a homozygous deletions specifically targeting the CTCF binding sites i4 and i5 was performed. Precisely, we modeled the genomic region chr5:27,800,001-30,600,000 (mm9) encompassing the mouse *Shh* gene, in both WT and *ΔCTCF i4:i5*. Based on the CHi-C interaction data at 10kb resolution our PRISMR procedure returns a polymer model made of 12 different types of binding domains for each case, whose distribution along the genomic sequence is shown

in **Figure 3.9d-e**. In order to obtain an ensemble of 3D single-molecule conformations of the studied loci, we employed a polymer chain of N=3,360 beads and ran MD simulations starting from initial self- avoiding walk configurations (at least 10^2 independent simulations in each case). Then, we let the polymer evolve up to 10^8 timesteps to reach stationarity, using the interaction potentials described previously. A comparison between the experimental (**Figure 3.9b-c** top triangles) and the model obtained equilibrium contact matrices (**Figure 3.9b-c** bottom triangles) shows that the model well recapitulates the experimental contacts pattern, as also illustrated by the values of the Pearson's correlation coefficient, r, that equals to 0.97 in both cases, and by the value of the distance-corrected Pearson's correlation coefficient, r´ (see subsection 3.2.4 for a description of r´), that equals to 0.87 in the WT and 0.86 in *ΔCTCF i4:i5* model. The contact matrices of **Figure 3.9b** show a strong contact between *Shh* and the *ZRS* in WT limb. In our model this is explained by the presence of two peaks in the 8th binding domain (**Figure 3.9d**) in correspondence of the genomic location of the two elements (not present in the *ΔCTCF i4:i5* binding sites distribution). Although the genomic separation between *Shh* and *ZRS* is greater than the one between the *Mnx1* gene and *ZRS, Shh* and the *ZRS* are found in spatially close proximity and separated from *Mnx1*, as visualized in our 3D polymer model of **Figure 3.10a**, which illustrates a representative conformation of the locus. In contrast, in *ΔCTCF i4:i5* mutant limb buds, *Mnx1* is found closer to both *ZRS* and *Shh* (**Figure 3.10b**), whose mutual distance is increased as further confirmed by a shift in the distribution of distances across all of the polymer models derived from WT and mutant limb buds (**Figure 3.10c**). **Figure 3.10d** shows the relative distance changes among Shh and its regulatory regions, computed as $(d_{WT} - d_{i4i5})/d_{WT}$ (d_{WT} and d_{i4i5} being the average distances among the highlighted region in limb WT and *ΔCTCF i4:i5*, respectively). As a consequence of the increased *Shh–ZRS* distance, we detected, in *ΔCTCF i4:i5* limb buds, a 51% loss of *Shh* expression, which, however, is still active and indeed no limb malformation were detected in the animals bearing the *ΔCTCF i4:i5* mutation. This indicates that, in

the absence of the CTCF-driven chromatin interaction, *Shh* and the *ZRS* can still communicate in the 3D space of the nucleus, probably via an alternative mechanism such as molecular bridging of phase separation.

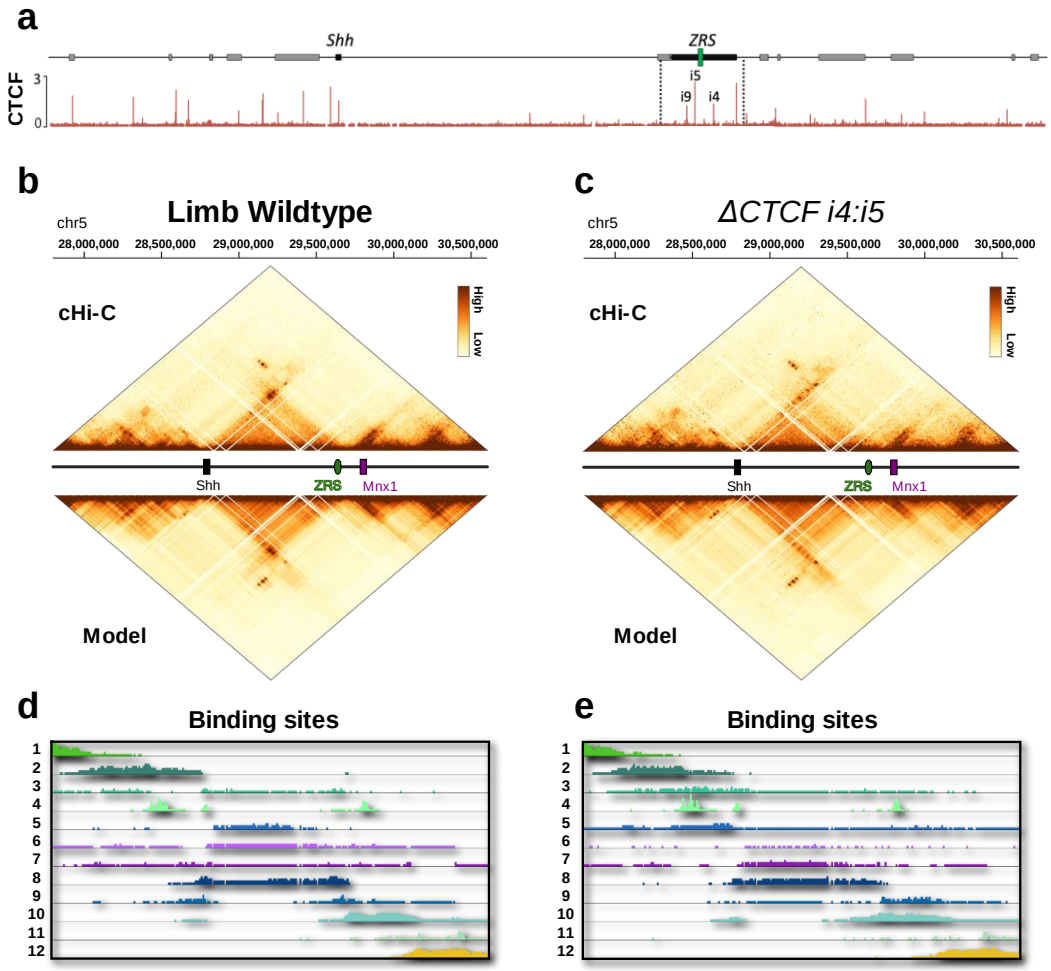


Figure 3.9: Shh locus 3D modeling

a) CTCF ChIP-seq enrichment in WT E10.5 limb buds at the *Shh* locus. Note the i4, i5, and i9 CTCF binding sites around the ZRS. **b)** Contact maps from CHi-C (above) and SBS model (below) in the limb WT have a Pearson correlation, r, and the distance-corrected Pearson correlation, r', respectively equal to r = 0.97, r´= 0.87. **c)** Contact maps from CHi-C (above) and SBS model (below) in the limb *ΔCTCF i4:i5* have a Pearson correlation, r, and the distance-corrected Pearson correlation, r´, respectively equal to r = 0.97, r´= 0.86. **d-e)** Distribution of the twelve binding domains (different colors) of the polymer model best explaining the experimental contact data in WT and *ΔCTCF i4:i5* (respectively); their abundance is represented as a histogram over the genomic sequence. (Figure adapted from Paliou et al. 2019)

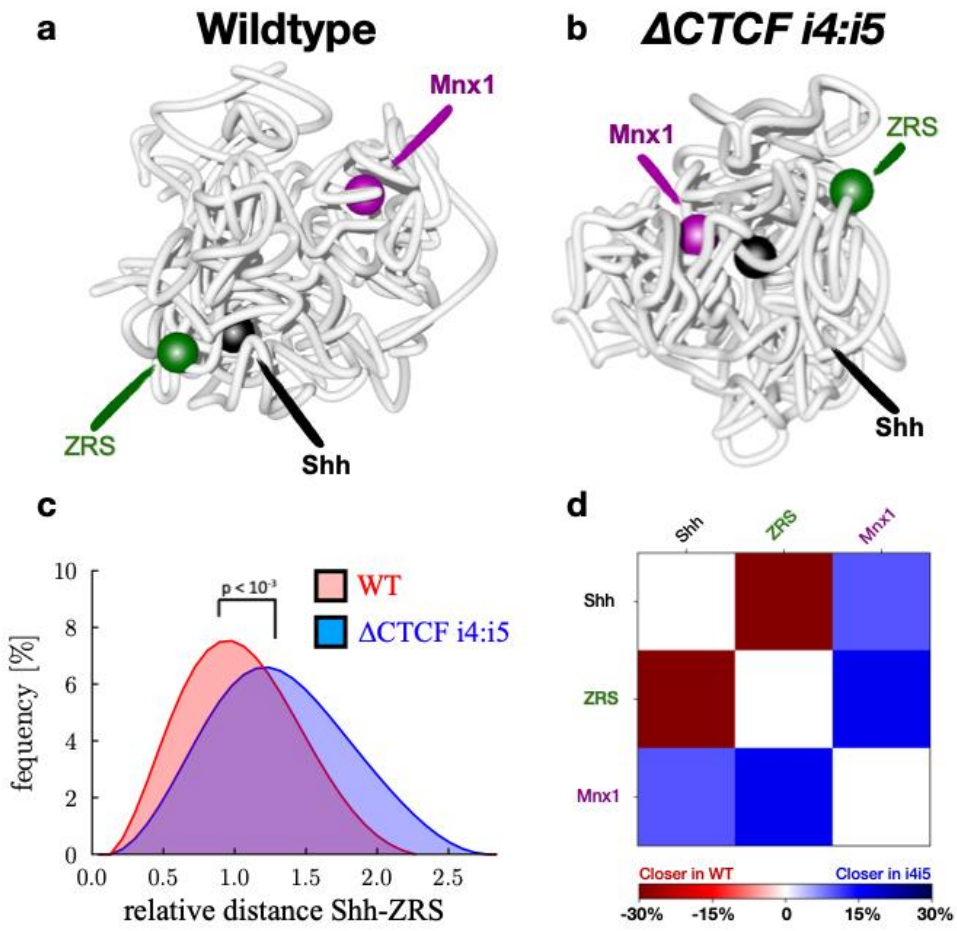


Figure 3.10: 3D structure and physical distance changes in the Shh locus

a) Representative 3D structure of the locus in WT limb buds. Note the proximity between Shh and ZRS and their distance from the Mnx1 gene. **b)** Representative 3D structure of the locus in ΔCTCF i4:i5 limb buds. Note the changes in proximity between Shh and ZRS and between Shh and Mnx1. **c)** Frequency plot of the distance distribution between Shh and the ZRS in WT and ΔCTCF i4:i5 limb buds. Note the increase in relative distance in the mutant limbs. P-value was calculated by Mann-Whitney test. **d)** Heatmap showing th relative distance changes among the three elements *Shh*, *ZRS* and *Mnx1* in the limb WT and *ΔCTCF i4:i5*, averaged over the single-molecule population from the polymer modeling. (Figure adapted from Paliou et al. 2019)

Chapter 4 - Genome-wide analysis of pairwise chromatin contacts

In the previous Chapter 3, we have introduced and applied in some interesting cases the machine learning based PRISMR method, a powerful tool to describe the folding of chromosome loci and reconstruct their detailed 3D structure. That is achieved by use of the SBS polymer model (Chapter 2) together with information obtained from Hi-C experiments (Chapter 1). In this final chapter, we present the first genome-wide application of the algorithm. We will use PRISMR to infer, from just Hi-C data, the minimal polymer model best explaining the contact patterns across the nineteen somatic chromosomes of the mouse embryonic stem cell genome. Precisely, we will find the specific location and combination of the distinct binding sites whereby DNA contacts are spontaneously established, i.e. the molecular code underlying the 3D architecture of chromosomes. The inferred polymer model describes Hi-C data across mouse chromosomes with high accuracy, showing that it is sufficient to make sense of a large fraction of contact patterns (**Section 4.1**). In **Section 4.2** we will study the robustness and the structural feature of the inferred binding sites. Next, in **Section 4.3** we will investigate their molecular nature by showing that our domains can be used to bring together independently derived information on architecture and epigenetics, e.g., by crossing their genomic position with ENCODE databases. The different binding domains fall in similarity classes based on epigenetics, well matching functional chromatin states derived in linear epigenetic segmentation studies such as active, poised and repressed states. However, we discover that they have an overlapping, combinatorial genomic distribution at the current resolution of Hi-C experiments, lacking in linear segmentation studies, which is shown to be required to explain Hi-C contacts with high accuracy genome-wide (**Section 4.4**). The results presented in the last chapter have not been published yet and represent one of the current research projects of the group.

4.1 The inferred binding domains explain Hi-C data genome-wide

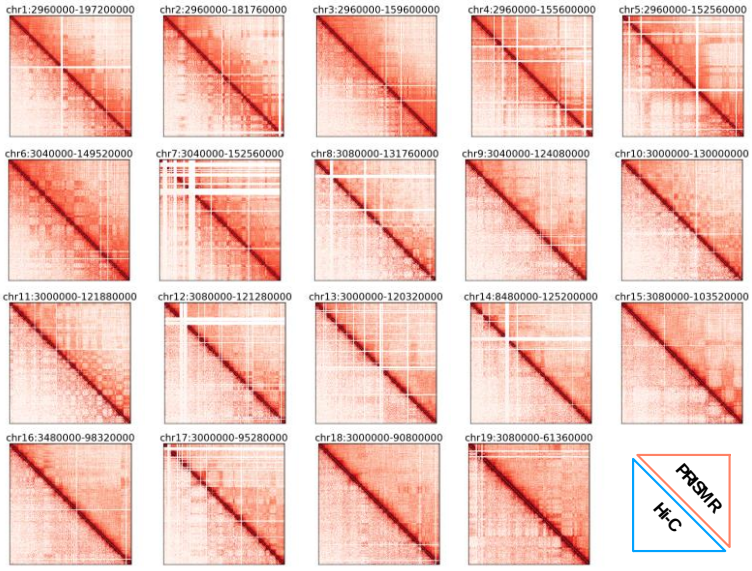
To dissect the molecular mechanisms that contribute to chromatin folding at large scales, we used our previously described (see Section 3.1) machine learning procedure PRISMR over the nineteen somatic chromosomes of the mouse genome, obtaining the SBS polymers from the experimental data set. Precisely, we computed, for each chromosome independently, the distribution of the binding sites along the polymer chain best explaining the contact matrix of each whole chromosome, given as input to PRISMR. We employed, as experimental dataset, published Hi-C data (Dixon et al. 2012) from mouse embryonic stem cell (mESC) at 40kb resolution and normalized according to the method described in (Yaffe and Tanay 2011). The optimal value of the parameters of the algorithm (i.e. the minimal number of colors *n*, the number of beads per bin *r*, and the Bayesian factor λ used to avoid the overfitting) has been estimated by repeating the simulated annealing procedure many times starting from different initial conditions and different values of *n*, *r*, and λ and selecting the values required to fit the data within a predefined accuracy. For the parameter's evaluation, we used as input the contact matrix of the chromosome eleven (chr11), a medium-sized chromosome, and obtained $n=150$, $r=30$, and $\lambda=1$. The same values of the parameters have been used to run PRISMR for all the other chromosomes.

To check whether the model can explain Hi-C data genome-wide, we compared the PRISMR-derived SBS contact matrices to the original Hi-C data (**Figure 4.1a-b**). The global pattern obtained by our model is highly correlated with the experimental one as quantified by the comparatively high values of the Pearson's (r), distance-corrected Pearson's (r') and stratum-adjusted (SCC) correlation coefficients averaged over the different chromosomes, respectively equal to $r = 0.95$, $r' = 0.60$, and $SCC = 0.80$ (see Subsection 4.1.1 for the details on the calculation of the similarity measures).

The PRISMR method is highly generalizable across different experiment and data resolution. To test that, we also applied our method to a recent high resolution (5kb), mESC Hi-C data set (Bonev et al. 2017). As the task is computationally demanding, we only considered the chromosome 19 and used the same parameters discussed above, obtaining correlations values comparable to those reported above for the 40 kb data ($r = 0.95$, $r' = 0.51$, and $SCC = 0.74$, see **Figure 4.2a**). Additionally, we modeled at 5kb resolution a specific genomic region around the *Sox9* gene (chr11:109140000-115140000, mm10) and, in this case, we further evaluated the similarity between experiment and model by calling loops in both matrices and obtaining that most of the experimental loops were correctly captured in the model (**Figure 4.2b**).

Taken together, the high correlations found between the SBS model and Hi-C contact data support the view that transcription factor (TF) mediated interactions between the inferred, different sets of binding sites can explain an important component of the molecular mechanisms shaping chromosome architecture. The binding domains inferred genome-wide by PRISMR identify the system architectural code as they contain key information required for TFs to spontaneously fold chromatin in its 3D structure through just the basic laws of physics.

a



b

Chr.	r	r'	SCC
1	0.94	0.55	0.85
2	0.94	0.59	0.86
3	0.94	0.59	0.84
4	0.95	0.64	0.89
5	0.94	0.62	0.87
6	0.98	0.73	0.89
7	0.96	0.68	0.88
8	0.94	0.55	0.81
9	0.94	0.55	0.78
10	0.94	0.52	0.77
11	0.95	0.60	0.83
12	0.94	0.58	0.81
13	0.94	0.54	0.77
14	0.95	0.56	0.79
15	0.95	0.58	0.77
16	0.94	0.55	0.76
17	0.95	0.61	0.82
18	0.95	0.58	0.77
19	0.96	0.69	0.81

Figure 4.1: Evaluation of the similarity between Hi-C and model-inferred matrices

a) Contact maps across chromosomes from the PRISMR inferred SBS model (upper triangle) and from Dixon et al. 2012 Hi-C data (lower triangle). Note the high similarity between the model and experimental contact patterns. **b)** Pearson's (r), distance-corrected Pearson's (r'), and stratum adjusted correlation coefficients (SCC) between SBS model and Hi-C data. SCC values were computed using HiCRep (Yang et al. 2017).

4.1.1 Details on the calculation of correlations and loops

The agreement between experiment and model has been quantified using the Pearson's correlation coefficient, r, between Hi-C contact matrices and the ones inferred by PRISMR. We also used two additional measures: 1) the distance corrected Pearson correlation coefficient, denoted by r', that is the Pearson's correlation coefficient between the two matrices where we subtracted from each diagonal (corresponding to a given genomic distance) their average contact frequency; 2) the stratum-adjusted correlation coefficient, denoted by SCC, from the HiCRep method (Yang et al. 2017) with a smoothing parameter h=1 and an upper bound of interaction distance equal to 5Mb. These two measures have been used to put aside the obvious decreasing trend of the pairwise contact frequency with genomic distance, that tend to dominate in the simple Pearson's correlation

values. Furthermore, as a control case, we computed HiCRep correlations between the Hi-C contact map of a chromosome from a different cell type, i.e. mouse Cortex cells (Dixon et al. 2012), with the corresponding chromosome contact matrix from our model in mESC. For instance, for chr11 we found $SCC_{Cortex\_VS\_Model}$ = 0.49 (std ≈ 0.01) and $r'_{Cortex\_VS\_Model}$ = 0.16. These correlations are comparable with analogous values computed between mESC and Cortex HiC data: $SCC_{Cortex\_VS\_mESC}$ = 0.56 (std ≈ 0.01) and $r'_{Cortex\_VS\_mESC}$ = 0.17. Moreover, they are significantly lower than the corresponding values between our model and Hi- C in mESC: $SCC_{mESC\_VS\_Model}$ = 0.83 (std ≈ 0.01), $r'_{mESC\_VS\_Model}$ = 0.60.

Chromatin loops are thought to be a basic unit of interphase nuclear organization, since they can provide contacts between regulatory regions and gene promoters. Loops are often formed between TAD borders (S. S. P. P. Rao et al. 2014) and are recognizable on a heat map as points of strong increased contact frequencies at the top corner of the TAD (**Figure 4.2b**). As loop-calling method, we used the HiCCUPS algorithm from Juicer Tools (Durand et al. 2016) with 5kb, 10kb and 25kb resolution and a False Discovery Rate (FDR) equal to 0.001 for each case. In particular, assuming as a reference the loops detected from Hi-C, we computed for the model data the fraction of actual positives that are correctly identified (sensitivity) and the fraction of actual negatives that are correctly identified (specificity), within a 50kb window. We obtained a sensitivity equal to 0.64 and a specificity equal to 0.998. We also checked the effect of using different FDR cut-off values, finding that both sensitivity and specificity are only marginally affected.

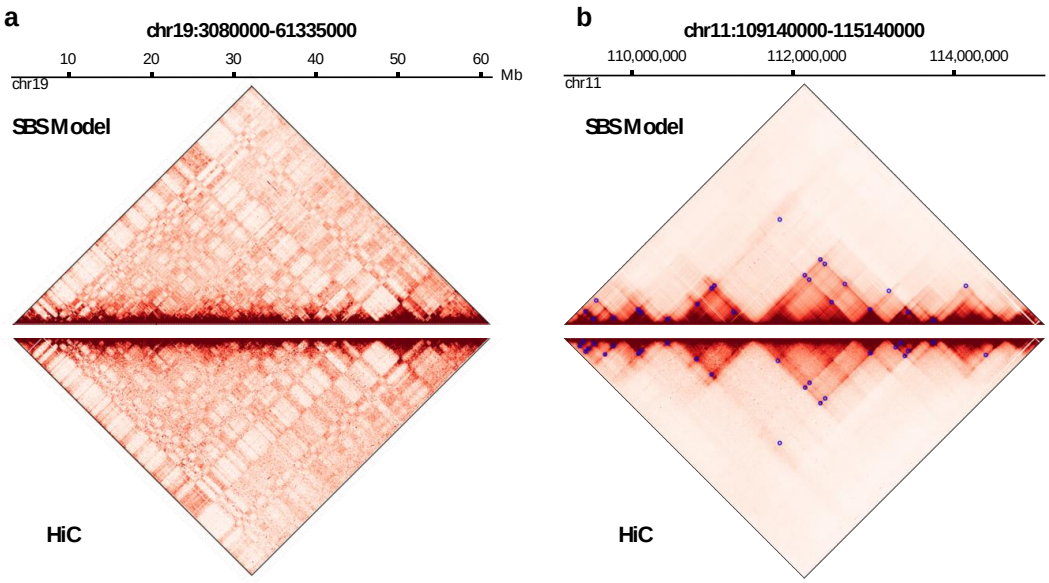


Figure 4.2: The SBS model works well across different data resolution and detects contacts at the scale of genomic loops

a) Heat-maps from the PRISMR inferred SBS model (upper triangle) and from Bonev et al. 2017 Hi-C data at 5kb resolution (lower triangle) relative to the entire chromosome 19. Pearson, distance-corrected Pearson and stratum adjusted (SCC) correlations are r=0.95, r'=0.51, SCC=0.74, respectively. **b)** Heat-maps from the PRISMR inferred SBS model (upper triangle) and from Bonev et al. 2017 Hi-C data at 5kb resolution (lower triangle) around the genomic region spanning the *Sox9* gene. Pearson, distance-corrected Pearson and stratum adjusted (SCC) correlations are r=0.95 and r'=0.59, SCC = 0.85 respectively. The blue circles indicate the loops found using the HiCCUPS algorithm from Juicer Tools (Durand et al. 2016).

4.2 Binding domains structural features

The model binding domains, i.e., the sets of homologous binding sites along the chromosomes, are the output of PRISMR. To study how the different binding domains (colors) span along the genome we employed two measures. The first one, that measures the domain size, is the genomic coverage, i.e., the fraction of beads of a given color multiplied by the length of the chromosome it belongs to. Averaging over all the domains identified by PRIMSR across chromosomes, we find that the genomic length covered by each domain is on average 0.75 Mb, with

a standard deviation of 0.2 Mb, a value close to the mean-size of a TAD. To measure, instead, how long is the extension of the interaction due to a single domain we defined r_{int} as two times the standard deviation of the center of mass of a domain (the cartoon in **Figure 4.3a** gives a visual impression of what r_{int} is measuring). Precisely, as the location of the different colors obtained by PRISMR can be specified indicating the coordinates of their binding sites along the genome, we can compute the center of mass of each domain as the average of the coordinates of a given domain weighted over the occurrence number of the binding sites in each genomic window. The distribution of r_{int} extends far beyond the size of the single domain, ranging from a few mega-bases to more than 100 Mb (**Figure 4.3a**). To check the statistical significance of the domains identified by PRISMR, we compared $P(r_{int})$ from our binding domains (red curve in **Figure 4.3a**) with a random control model obtained by randomly bootstrapping the location of our binding sites along the genome (blue curve in **Figure 4.3a**) and we found that the two distributions are significantly different (p-value < 10^{-3}, Wilcoxon's rank sum test). We also found that $P(r_{int})$ is asymptotically consistent with a power-low scaling $P(r_{int}) \sim {r_{int}}^{-1}$, as shown in **Figure 4.3a** where the right-hand side of the distribution is well described by a power-low fit (dotted red curve in the graph).

To further test the level of randomness of the binding domains identified by PRISMR, we measured their genomic overlap q and compared it to the level of overlap expected in the random model of bootstrapped domains mentioned before. For a generic pair of binding domains $k_1$and k_2, their overlap, which measures their similarity, is defined as:

$$q(k_1, k_2) = \frac{\sum_{i=1}^{L} f_i(k_1) * f_i(k_2)}{\sqrt{\sum_{i=1}^{L} f_i^2(k_1) * \sum_{i=1}^{L} f_i^2(k_2)}} \qquad (6)$$

where $f_i(k_j)$ is the number of beads of the domain k_j in the i-th bin of chromosome of length L (the cartoon in **Figure 4.3b** gives a visual impression of

what q is measuring). We found that the distribution $p(q)$ of the overlaps of the binding domains predicted by PRISMR is significantly different (p-value<0.001, Wilcoxon's rank sum test) from the one expected in the random control model (red and blue curves in **Figure 4.3b**, respectively).

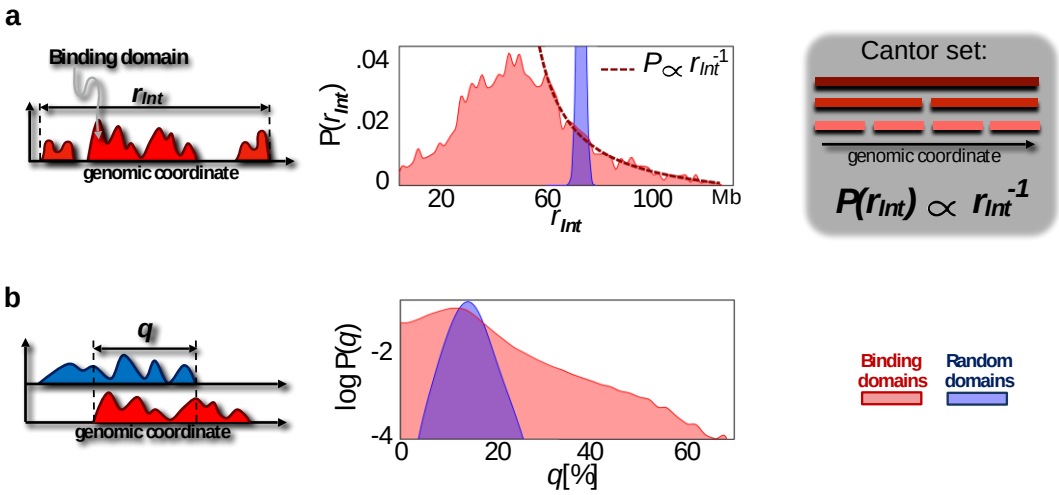


Figure 4.3: Characterization of the identified binding domains

a) Distribution of the range of interaction r_{int} of the PRISMR inferred binding domains genome-wide. The blue curve corresponds to a random model where the binding sites are bootstrapped. A Cantor set has hierarchically nested domains: the distribution of their ranges scales as an inverse power law. **b)** The distribution of overlaps between all the possible pairs of binding domains on the same chromosome (red) compared to the one expected in a random model (obtained by bootstrapping, blue).

4.3 Molecular nature of the binding domains

4.3.1 Epigenetic profile of the binding domains

To shed light on the nature of the model inferred binding sites we compared their genomic locations with a set of **epigenetic marks** available in mES cells from the ENCODE database (Dunham et al. 2012). We considered 8 histone modification signals available (H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K9me3, H3K27me3 and H2AUb), which mark the different

transcriptional state of chromatin (see section 1.2). In our analysis, we retained only statistically significant correlation values, i.e., those above a random control model with sites having bootstrapped genomic positions (see the next subsection 4.3.3 for the details on the epigenetics analysis). As the different binding domains tend to fall in groups with similar epigenetic profiles, we clustered them to identify genome-wide significantly distinct epigenetic classes. The Akaike Information Criterion, (Akaike 1974) returns a set of 10 statistically different groups (**Figure 4.4a-b**), a result supported by the structure of the branching tree obtained via a simple hierarchical clustering procedure (**Figure 4.4c**).

Two main classes of binding domains strongly correlate with active chromatin marks (**Figure 4.4a**), but they are clearly distinct from an epigenetic point of view. Class 1 is broadly enriched across all available active histone marks, whereas class 2 is only enriched for H3K4me1 and H3K36me3, associated especially to active enhancer regions (Boettiger et al. 2016; Gifford et al. 2013; Ho et al. 2014; Javierre et al. 2016). Interestingly, the genomic positions of the sites of the two classes are partially correlated (correlation coefficient = 0.3, **Figure 4.5c**). Their histone signatures are also consistent with gene annotation and transcription data (**Figure 4.4a**). That supports the view that the binding sites in class 1 and 2 are responsible, genome-wide, especially for specific contacts between transcribed and regulatory regions, mediated by factors such as active Pol-II, as experimentally demonstrated at a number of specific loci (see e.g., Mariano Barbieri et al. 2017). Class 3 has the typical signature of bivalent chromatin, with H3K27me3 combined with active marks. Its binding sites could be responsible for interactions between regions including, for instance, poised genes and their regulators, as seen in FISH co-localization experiments. Class 4 is significantly correlated with only H3K27me3 and could be responsible of the experimentally observed self-interacting domains of PRC repressed chromatin (Kundu et al. 2017). Interestingly, classes 1, 2, 3, and 4 are the only to include active and inactive promoters and the only ones to correlate with CTCF binding

sites. That confirms the significance of CTCF in the regulation of chromatin architecture and gene activity (see, e.g. Tang et al. 2015), also highlighting that its role can be modulated by different sets of histone marks and molecular factors.

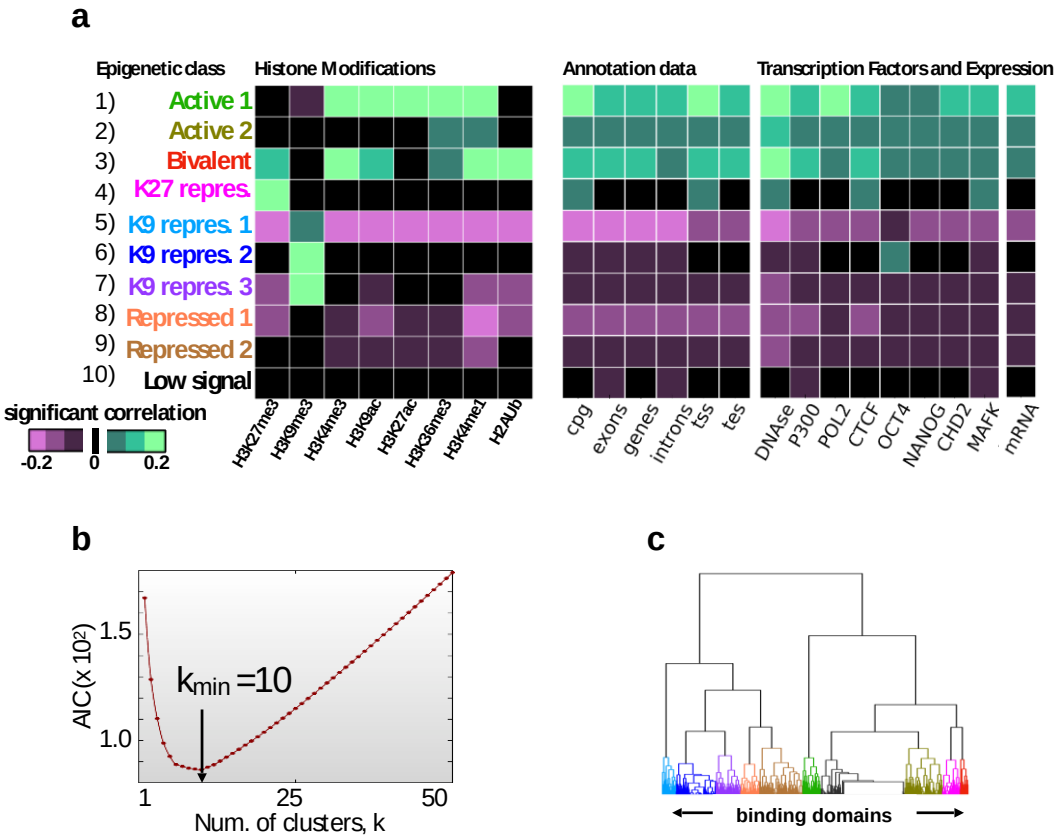


Figure 4.4: Epigenetic profile of the 10 classes of the model binding domains

a) The model binding domains, albeit inferred from Hi-C data only, correlate each with a specific set of ENCODE tracks. They cluster in 10 main classes genome-wide according to their Pearson's correlations with the shown ENCODE histone marks. The epigenetic profile of the centroid of each class is shown in the heat-map, together with their correlation with a set of annotation data. The 10 classes match well chromatin states derived in epigenetic segmentation studies (e.g., active, poised, repressed states). **b)** The AIC statistical criterion has a minimum at k=10 clusters of binding domains based on their epigenetic profile shown in panel a). **c)** Hierarchical clustering of the model inferred binding domains with the 10 identified classes highlighted with different colors.

Classes 5, 6 and 7 are distinct in their epigenetic profile, but all significantly correlated with H3K9me3, a mark usually associated to constitutive

heterochromatin and lack of transcription factor binding. Their genomic positioning is also weakly, yet significantly correlated (correlation coefficient around 0.15, **Figure 4.5c**). Classes 8 and 9 are gnomically partially overlapping with class 5, 6 and 7 and anti-correlated with active marks too. Finally, class 10 (named "low signal") has no significant correlation with available histone marks. However, consistently with previous studies (Ho et al. 2014), it covers almost 25% of the genome, while the other classes range from around 5% to 10% in genomic coverage (**Figure 4.5a**). Interestingly, the different classes are significantly differently enriched over the different chromosomes and not consistent with a uniform random genomic distribution (**Figure 4.5b**, p-values < 0.05).

4.3.2 Classes removal

To understand the relative importance of the different types of binding domains in shaping chromatin architecture, we conducted a set of in-silico experiments with mutant models where each class, one at the time, is erased. Specifically, given the list of the binding domains of each class, we removed all these domains along the different chromosomes by replacing their PRISMR-inferred binding sites with gray, non-interacting, elements. Then, we computed the contact matrices of the modified SBS polymers and compared them with the corresponding Hi-C matrices by measuring the variation of the Pearson's correlation coefficient with respect to the 'wild-type' value, that is the mean value over the different chromosomes of the correlations obtained without the epigenetic class removal (r =0.95). The variation is found to be proportional to the genomic coverage of the different classes, with the exception of the "low signal" class, whose removal has an impact much lower than expected by coverage (**Figure 4.5d**). That implicates that no binding class has a special role in holding the architecture of the genome in place. The proportionality relation whereby the removal of, say, 10% of binding sites genome-wide roughly results in a 10% reduction of r, highlights the structural stability of the system: the

removal of a small fraction of binding sites proportionally alter the structure, but does not produce a sudden collapse of the architecture, as reported by recent experiments (Barutcu et al. 2018; Kubo et al. 2017; Nora et al. 2017; S. S. P. Rao et al. 2017; Rodríguez-Carballo et al. 2017).

Importantly, the 10 classes of binding domains here identified match well the classes found by previous epigenetic genome segmentation studies (Ernst et al. 2011; Gifford et al. 2013; Ho et al. 2014). However, our binding domains are inferred from only Hi-C data without previous knowledge of epigenetics. Hence, they bring together independent information on architecture and epigenetics. In particular, a crucial feature of the model binding domains to explain contact data is that the different types do overlap with each other along the genome at the resolution of the considered Hi-C data. Therefore, they naturally provide each DNA window with a complex barcode made of the list of the different included binding site types. This is an important difference with 1D epigenetic segmentation classes: by definition, those have no genomic overlap thus each DNA window is associated to only one of such classes.

4.3.3 Computational details of the epigenetic study of the binding domains

We downloaded, from the ENCODE database (Dunham et al. 2012), a set of 8 histone modifications peak-called tracks files (H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K9me3, H3K27me3 and H2AUb) relative to the mES cells. We then used the bedtools coverage tool (Quinlan and Hall 2010) to identify the number of called peaks contained within each genome-wide window for each chromatin feature; the signal thus obtained represent our histone mark profile. To measure the similarity of binding domains with histone marks we computed the Pearson's correlation coefficient between the number of binding sites of each domain and each histone mark profile. Furthermore, we employed a control model to retain only the statistical significant correlations: first, we

computed the Pearson correlation between the chromatin mark signals and randomized binding domains signals obtained by bootstrapping the actual ones along the genomic locations; then, we considered significant only correlations above the 95th or below the 5th percentile of the distribution of the random correlations. Data are then collected in a rectangular matrix X, in which the element $X(i,j)$ is the significant correlation between the i-th binding domains and the j-th histone mark or zero if the correlation does not result significant. Since each row of X represents the correlation profile of a binding domain with respect to the used histone modifications, we refer to them as the **epigenomic signature** of a binding domain. In order to find binding domains with similar epigenomic signatures, we performed a hierarchical clustering analysis on X using the *Python SciPy* clustering package with "Euclidean" distance metric and "Ward" linkage method. To assess the number of clusters in the hierarchical clustering output, we cut the dendrogram at different values (ranging from one to the number of binding domains) and evaluated the Akaike Information Criterion (Akaike 1974), or briefly AIC, as the number of clusters k is varied. As shown in **Figure 4.4b**, while no sharp transitions are present, the curve has a global minimum at k =10. We therefore grouped all the different rows of X in ten different classes according to their affinity to each cluster. The centroid of each cluster, that is the average of the epigenomic signature of the domains belonging to it, was considered as the epigenetic signature of the entire class. To assign biologically meaningful labels to the obtained partition, we looked at the enrichment of several types of functional annotations in the different classes. More precisely, we first binned each annotation track using the same windows of the binding domains (taking the sum in each bin), and then, for each pair of annotation mark and epigenetic class, we computed the average of the Pearson correlation values between that mark and each binding domains of that class (**Figure 4.4a**). The set of functional annotations considered in this study is the follow (where all the coordinates are relative to the mm9 version of the mouse

genome): (1) CpG islands, exons, introns, genes, transcription start sites (TSS) and transcription end sites (TES) downloaded from the UCSC Table Browser. (2) Transcription factors binding sites and DNase peaks obtained from the ENCODE database in the mESC cell line. We also correlated the binding domains signal with ENCODE expression data in mESC, where the transcription level has been obtained based on GENCODE annotation and normalized to FPKM (Fragments Per Kilobase Million) values using the Cufflinks software (Trapnell et al. 2010).

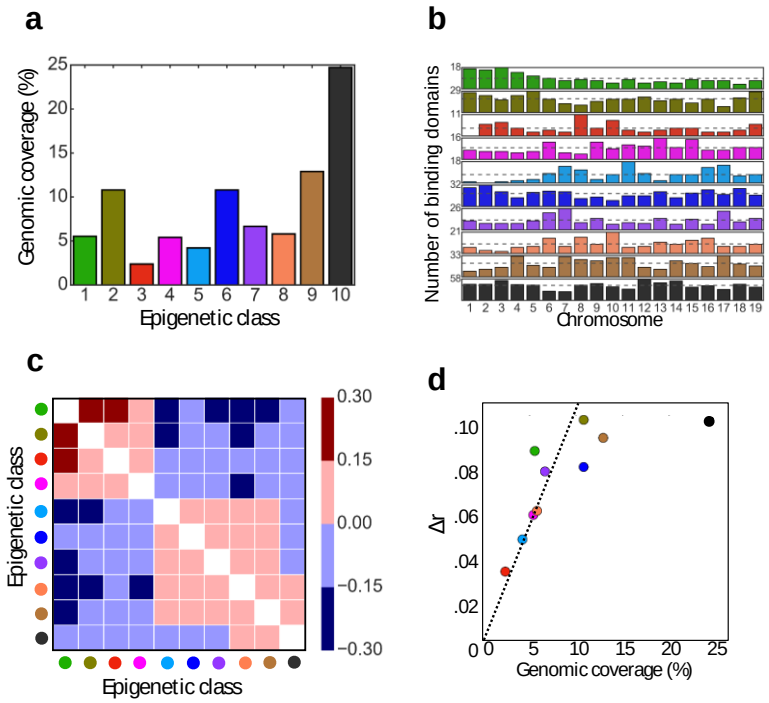


Figure 4.5: Characterization of the identified epigenetic classes

a) Genomic coverage of the 10 classes of the model binding domains. **b)** Number of the binding domains of the different classes across chromosomes. The distribution is not uniform (p-value<0.05). **c)** Pearson's correlation coefficient of the genomic location of the different classes over the chromosomes. **d)** Effect of the withdraw of a class of binding sites on the architecture measured by the variation of the Pearson's correlation with respect to the wild-type model. Δr is the difference between r in the wild-type model (r=0.60) and in a model where the domains of a given class are removed, averaged over chromosomes.

4.4 Epigenetic linear segmentation only partially captures chromatin folding

To deepen our comprehension of the interplay of chromosome epigenetics and folding, we investigated the architectural information content retained in 1D epigenetic segmentations of the genome and compared it with the more complex DNA barcoding given by the classes of our binding domains. As done in previous studies (Boettiger et al. 2016; Gifford et al. 2013; Ho et al. 2014; Javierre et al. 2016), we segmented chromosomes in 10 epigenetic classes based only on ENCODE histone marks. For simplicity, we opted for 10 classes to match the number of different types of binding domains found above. Such a number of classes is comparable to those in previous segmentations studies, but our results are not affected by more complex choices of segmentation (until the scale of the single binding domain is reached). Next, we derived in-silico the contact maps predicted by a polymer model based only on such a 1D epigenetic segmentation. Specifically, we considered a polymer where chromatin physical interactions only occur between homologous 1D-segmented epigenetic regions (Jost et al. 2014). Interestingly, while the overall contact patterns from such a model visually resemble Hi-C patterns (r=0.80), their distance-corrected Pearson correlation, r', with Hi-C data is low, r'= 0.05 (**Figure 4.6b**). Hence, the patterns derived from a polymer model constructed from 1D epigenetic segmentation is only marginally better than one where Hi-C pair-wise interactions are replaced by the average value corresponding to that genomic separation. Conversely, a 10 color SBS model based on its epigenetics classes (see subsection 4.4.2), with overlapping binding domains, has r= 0.88 and r'= 0.20 and, as discussed, the model with the full set of inferred binding domains has r= 0.95 and r'= 0.60.

To understand the partial failure of 1D epigenetic segmentation in explaining contact data, we identified, for each pair of genomic sites, the binding domain that mostly contributes to their pair-wise interaction within the full SBS model

(**Figure 4.6d-e**). For clarity, we focus on a case-study 34 Mb wide region on chromosome 11. Plaid-patterns are visible in its Hi-C contact map, as expected from A/B compartments; they are also visible in the matrix of the most contributing binding domains (see subsection 4.4.1), where rich and fine substructures appear as well. Consider, for instance, the TAD associated to region C in **Figure 4.6c**. The interactions within that TAD are mainly related to binding domains in class 5 (cyan, **Figure 4.4a**), which is indeed the most abundant within the genomic region where C is located (**Figure 4.6d**). However, the interactions between region C and B cannot be traced back to class 5, but they stem from binding domains in class 4 (magenta), which is the 2nd and 3rd most abundant for B and C, respectively. Such an example illustrates that a linear epigenetic segmentation model with homotypic interactions fails to account for the complexity of the observed contact pattern because an interaction between B and C would only occur if the two regions belong to the same class. Analogously, the contacts between region A and C (and between A and B) originate from different binding domains included in those regions. A similar reasoning can be extended to the plaid-pattern of A/B compartments (which is a specific example of a two-classes genome 1D segmentation) capturing the overall interactions between homologous active and repressed regions respectively. Yet, a much more complex and finer structure of contacts exists (including interactions across A and B compartments). Indeed, it has been shown that polymer models based on a linear epigenetic classification of domains are forced to include combinatorial heterotypic interactions to accurately explain Hi-C data (Di Pierro et al. 2016).

Summarizing, homotypic interactions between the domains of a coarse-grained linear epigenetic segmentation of the genome, such as compartment A/B, are not enough to explain Hi-C patterns with high accuracy, since a complexity of relevant heterotypic contacts exist between those regions. That is captured by the binding domains of our model. They associate a barcode to DNA segments, containing the information required to produce, through physics, the system 3D

conformations. On the other hand, by bringing together independent architectural and epigenetics data, our binding domains form epigenetic classes well matching those found in segmentation studies.

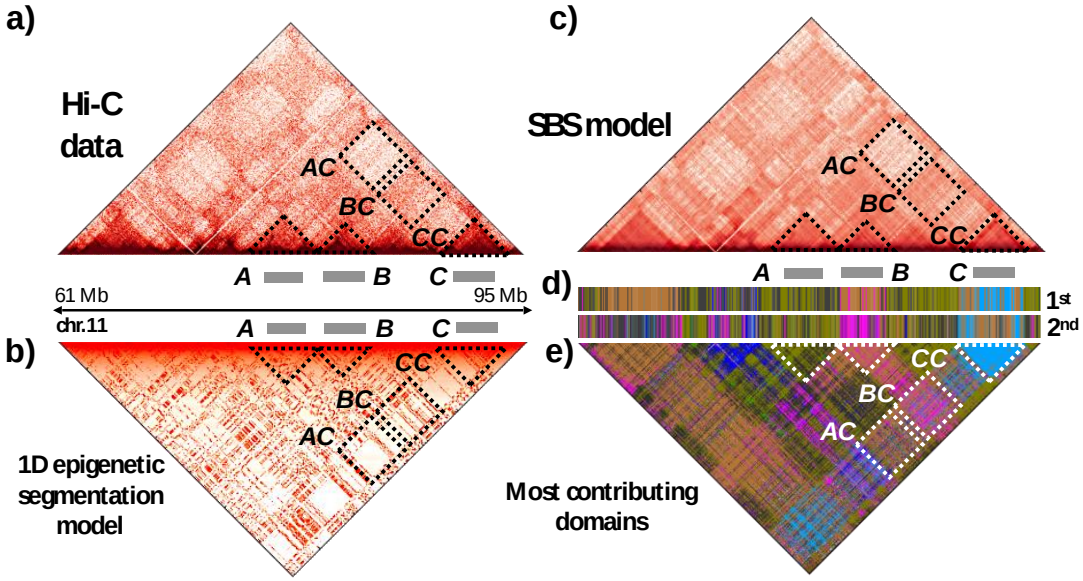


Figure 4.6: Contact patterns are only partially captured by linear epigenetic segmentation.

a) Hi-C data (Dixon et al. 2012) of a 34 Mb wide region on chromosome 11 in mESC with highlighted TAD patterns around region A, B and C, and some pair-wise contacts (AC, BC, CC). **b)** The contact map of a model based only on homotypic interactions between linear segmented epigenetic domains has a Pearson correlation r=0.84 with Hi-C data, but a distance corrected correlation r'=0.06, showing only a marginal improvement over a control model where each interaction is replaced by the average at the corresponding genomic separation. **c)** The contact map of the SBS model in the shown region has r=0.96 and r'=0.71. **d)** The PRISMR inferred 1st and 2nd most abundant binding site types of the SBS model along the zoomed region. **e)** The SBS most contributing binding domain to each pairwise contact highlights that the complexity of interaction patterns is captured by the combinatorial overlap of different binding site types along the sequence. For example, contacts within C (CC) are mainly mediated by the 1st most abundant binding site type in C (cyan), but the interactions of C with A and B (AC and BC) are mediated by different binding domains (respectively the 2nd (brown) and the 3rd (magenta) most abundant in C).

4.4.1 Most abundant and most contributing domains to pairwise contacts

As the different binding domains can overlap with each other, to better visualize their locations along the genome, we show in **Figure 4.6d** the 1st and 2nd most abundant binding domain, i.e. the one and the second with the largest number of binding sites type per bin. In both cases, to help the visualization, the domains are colored with the color of their epigenetic class. The contribution of the different binding domains in forming the interactions between bins pairs is then highlighted in **Figure 4.6e** where the colors of the most contributing binding domains are shown. Specifically, for a given pair-wise contact we defined the contribution of a binding domain to that contact as the number of pairs of its binding site type between the two considered bins. The binding domain having the highest number of binding site pairs is then considered as the most contributing one and is shown in **Figure 4.6e** with the color corresponding to the epigenetic class it belongs to.

4.4.2 Epigenetic linear segmentation model

To build a model based only on linear epigenetics, we considered the same dataset of eight histone modifications previously discussed and assigned to each genomic window the sequence of the number of peaks of each histone mark. Then, we performed a hierarchical clustering analysis to gather the genomic windows with similar histone profile in 10 different classes, in order to match them with the 10 different types of binding domains found above. The obtained linear segmentation has been employed to define a polymer model for the chromosome 11 with 10 different colors corresponding to the different linear epigenetic classes, in such a way that the interacting elements belong to the same 1D epigenetic segmented region. Finally, we derived in-silico the contact map of such a model and compared it with the experimental matrix (**Figure 4.6a-b**). We found that the Pearson's correlation and distance-corrected Pearson's correlation

between the matrices are r = 0.80 and r'= 0.05, respectively. We have also built a model by assigning at each of the different binding sites of chr.11 the colour of the epigenetic class it belongs to. We found that, this 10 color SBS model with overlapping binding domains has r = 0.88 and r'= 0.20.

Conclusions and perspectives

Recent advancements in Molecular Biology have revealed that chromatin has a complex spatial organization, intimately linked to its biological functions. Thanks to the development of new experimental techniques, such as Hi-C, a large amount of data is now available, making it possible to address the problem of chromatin folding in a more quantitative way. Yet, a unified theoretical framework describing the molecular mechanisms of DNA folding is still lacking and polymer modeling can help to tackle this problem. This work wants to be a contribution towards this challenging goal. In particular, here we focused on the Strings and Binders Switch (SBS) model, in which chromatin 3D conformations form through the interaction of diffusing molecular binders with binding sites along the polymer chain. Firstly, we showed that a very simple polymer can recapitulate emerging aspects of chromatin structure such as its spontaneous hierarchical folding. A generalized SBS model in combination with the machine learning inference method PRISMR, can be used to accurately reconstruct the landscape of real genomic loci in 3D. For instance, we discovered that two different tissue-specific shapes regulate the correct limb development at the *Pixt1* genomic locus and that the perturbation of a sequence including the *Pitx1* enhancer can spatially revert the state of the locus and, consequently, lead to gene malfunctions. This highlights that the dynamic 3D chromatin architecture can play a determinant role in modulating the transcriptional activities. Next, we extended our method to the whole genome and inferred, from Hi-C data, the specific genomic location of the distinct sets of putative binding sites required to establish chromatin architecture. The increase of statistics obtained from the genome-wide study, allowed us to cross the architectural information with a set of available epigenetics data to investigate the molecular nature of the complexes that mediate chromatin interactions. Our results show that the molecular architectural code enclosed in the inferred binding domains of the SBS model

and its folding mechanisms must be capturing some general and important principle of regulation of chromatin structure. They provide an interpretation of the link between epigenetics, architecture and function, which can also help understanding the impact of genomic structural variations and epigenetic changes in diseases such as cancer and congenital disorders. The diseases potential of genetic and epigenetics modifications is most frequently hard to predict with current screens. Our results progress the research of new quantitative, in-silico methods for the medical interpretation of the phenotypic impact of such variations. Some results, which for brevity are not presented here, have already been obtained in this sense. New research lines we are following concern the employment of our polymer models to study the cell-to-cell variability of chromatin architecture and to capture the structural differences of a specific genomic region during differentiation or in any two different cell types. Our aim is to realize a reliable tool, able to elucidate the many, still unknown, mechanisms involved in the genome organization, and to predict the effects of genomic mutations on the genome architecture and, ultimately, on the cell functionality.

References

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control* 19(6): 716–23.

Allis, C. David, and Thomas Jenuwein. 2016. “The Molecular Hallmarks of Epigenetic Control.” *Nature Reviews Genetics*.

Amano, Takanori et al. 2009. “Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription.” *Developmental Cell*.

Annunziatella, Carlo et al. 2018. “Molecular Dynamics Simulations of the Strings and Binders Switch Model of Chromatin.” *Methods* 142: 81–88.

Barbieri, M. et al. 2012. “Complexity of Chromatin Folding Is Captured by the Strings and Binders Switch Model.” *Proceedings of the National Academy of Sciences* 109(40): 16173–78. http://www.pnas.org/cgi/doi/10.1073/pnas.1204799109.

Barbieri, Mariano et al. 2017. “Active and Poised Promoter States Drive Folding of the Extended HoxB Locus in Mouse Embryonic Stem Cells.” *Nature Structural and Molecular Biology* 24(6): 515–24.

Barutcu, A. Rasim et al. 2018. “A TAD Boundary Is Preserved upon Deletion of the CTCF-Rich Firre Locus.” *Nature Communications* 9(1).

Beagrie, Robert A. et al. 2017. “Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping.” *Nature* 543(7646): 519–24.

Bianco, Simona et al. 2018. “Polymer Physics Predicts the Effects of Structural Variants on Chromatin Architecture.” *Nature Genetics* 50(5): 662–67.

Bickmore, Wendy A., and Bas Van Steensel. 2013. “Genome Architecture: Domain Organization of Interphase Chromosomes.” *Cell* 152(6): 1270–84. http://dx.doi.org/10.1016/j.cell.2013.02.001.

Boettiger, Alistair N. et al. 2016. “Super-Resolution Imaging Reveals Distinct Chromatin Folding for Different Epigenetic States.” *Nature* 529(7586):

418–22. http://www.ncbi.nlm.nih.gov/pubmed/26760202%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4905822.

Bohn, Manfred, and Dieter W. Heermann. 2010. "Diffusion-Driven Looping Provides a Consistent Framework for Chromatin Organization." *PLoS ONE* 5(8).

Bonev, Boyan et al. 2017. "Multiscale 3D Genome Rewiring during Mouse Neural Development." *Cell* 171(3): 557-572.e24. https://www.sciencedirect.com/science/article/pii/S0092867417311376?via%3Dihub (March 5, 2019).

Boyle, S. 2001. "The Spatial Organization of Human Chromosomes within the Nuclei of Normal and Emerin-Mutant Cells." *Human Molecular Genetics*.

Brackley, C. A. et al. 2013. "Nonspecific Bridging-Induced Attraction Drives Clustering of DNA-Binding Proteins and Genome Organization." *Proceedings of the National Academy of Sciences* 110(38): E3605–11. http://www.pnas.org/cgi/doi/10.1073/pnas.1302950110.

Bulger, Michael, and Mark Groudine. 2011. "Functional and Mechanistic Diversity of Distal Transcription Enhancers." *Cell*.

Calo, Eliezer, and Joanna Wysocka. 2013. "Modification of Enhancer Chromatin: What, How, and Why?" *Molecular Cell*.

Chiariello, A.M. Andrea M. et al. 2016. "Polymer Physics of Chromosome Large-Scale 3D Organisation." *Scientific Reports* 6.

Cremer, T., and C. Cremer. 2001. "Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells." *Nature Reviews Genetics*.

Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. 2013. "Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data." *Nature Reviews Genetics*.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromatin Conformation." *Science*.

DeLaurier, April, Ronen Schweitzer, and Malcolm Logan. 2006. "Pitx1 Determines the Morphology of Muscle, Tendon, and Bones of the Hindlimb." *Developmental Biology* 299(1): 22–34.

Dixon, Jesse R. et al. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485(7398): 376–80. http://www.ncbi.nlm.nih.gov/pubmed/22495300%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3356448.

Dostie, Josée et al. 2006. "Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements." *Genome Research*.

Dunham, Ian et al. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489(7414): 57–74.

Durand, Neva C. et al. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3(1): 95–98.

Ernst, Jason et al. 2011. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types." *Nature* 473(7345): 43–49.

Esposito, Andrea et al. 2018. "Models of Polymer Physics for the Architecture of the Cell Nucleus." *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*.

———. 2019. "Higher-Order Chromosome Structures Investigated by Polymer Physics in Cellular Morphogenesis and Differentiation." *Journal of Molecular Biology*. http://www.sciencedirect.com/science/article/pii/S002228361930717X.

Franke, Martin et al. 2016. "Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications." *Nature* 538(7624): 265–69.

Fraser, James et al. 2015. "Hierarchical Folding and Reorganization of Chromosomes Are Linked to Transcriptional Changes in Cellular Differentiation." *Molecular Systems Biology* 11(12): 852–852. http://msb.embopress.org/cgi/doi/10.15252/msb.20156492.

Fudenberg, Geoffrey et al. 2016. “Formation of Chromosomal Domains by Loop Extrusion.” *Cell Reports* 15(9): 2038–49. http://dx.doi.org/10.1016/j.celrep.2016.04.085.

Fudenberg, Geoffrey, and Leonid A. Mirny. 2012. “Higher-Order Chromatin Structure: Bridging Physics and Biology.” *Current Opinion in Genetics and Development*.

De Gennes, P. G. 1979. “Scaling Concepts in Polymer Physics. Cornell University Press.” *Ithaca N.Y.,*.

Gifford, Casey A. et al. 2013. “Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells.” *Cell* 153(5): 1149–63.

Giorgetti, Luca et al. 2014. “Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription.” *Cell*.

Ho, Joshua W.K. et al. 2014. “Comparative Analysis of Metazoan Chromatin Organization.” *Nature* 512(7515): 449–52. http://dx.doi.org/10.1038/nature13415.

Hsieh, Tsung-Han S et al. 2019. “Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding.” *bioRxiv*.

Javierre, Biola M. et al. 2016. “Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters.” *Cell* 167(5): 1369-1384.e19.

Jost, Daniel, Pascal Carrivain, Giacomo Cavalli, and Cï¿½dric Vaillant. 2014. “Modeling Epigenome Folding: Formation and Dynamics of Topologically Associated Chromatin Domains.” *Nucleic Acids Research* 42(15): 9553–61.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. “Optimization by Simulated Annealing.” *Science* 220(4598): 671–80.

Kragesteen, Bjørt K. et al. 2018. “Dynamic 3D Chromatin Architecture Contributes to Enhancer Specificity and Limb Morphogenesis.” *Nature Genetics* 50(10): 1463–73.

Kremer, Kurt, and Gary S. Grest. 1990. “Dynamics of Entangled Linear Polymer Melts: A Molecular-Dynamics Simulation.” *The Journal of Chemical*

Physics 92(8): 5057–86.

Kubo, Naoki et al. 2017. "Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells." *bioRxiv*: 118737. http://biorxiv.org/content/early/2017/03/20/118737.

Kundu, Sharmistha et al. 2017. "Polycomb Repressive Complex 1 Generates Discrete Compacted Domains That Change during Differentiation." *Molecular Cell* 65(3): 432-446.e5. http://dx.doi.org/10.1016/j.molcel.2017.01.009.

Lieberman-Aiden, Erez et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326(5950): 289–93.

Lupiáñez, Darío G. et al. 2015. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." *Cell* 161(5): 1012–25.

MacCallum, Justin L., Alberto Perez, and Ken A. Dill. 2015. "Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference." *Proceedings of the National Academy of Sciences of the United States of America*.

Misteli, Tom. 2007. "Beyond the Sequence: Cellular Organization of Genome Function." *Cell* 128(4): 787–800.

Nicodemi, Mario, and Ana Pombo. 2014. "Models of Chromosome Structure." *Current Opinion in Cell Biology* 28(1): 90–95.

Nicodemi, Mario, and Antonella Prisco. 2009. "Thermodynamic Pathways to Genome Spatial Organization in the Cell Nucleus." *Biophysical Journal* 96(6): 2168–77.

Nir, Guy et al. 2018. "Walking along Chromosomes with Super-Resolution Imaging, Contact Maps, and Integrative Modeling." *PLoS Genetics*.

Nora, Elphège P. et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre." *Nature* 485(7398): 381–85.

———. 2017. "Targeted Degradation of CTCF Decouples Local Insulation of

Chromosome Domains from Genomic Compartmentalization." *Cell* 169(5): 930-944.e22.

Olivares-Chauvet, Pedro et al. 2016. "Capturing Pairwise and Multi-Way Chromosomal Conformations Using Chromosomal Walks." *Nature* 540(7632): 296–300. http://dx.doi.org/10.1038/nature20158.

Ong, Chin Tong, and Victor G. Corces. 2011. "Enhancer Function: New Insights into the Regulation of Tissue-Specific Gene Expression." *Nature Reviews Genetics*.

Paliou, Christina et al. 2019. "Preformed Chromatin Topology Assists Transcriptional Robustness of Shh during Limb Development." *Proceedings of the National Academy of Sciences of the United States of America*: 201900672. http://www.pnas.org/lookup/doi/10.1073/pnas.1900672116 (May 31, 2019).

Phillips-Cremins, Jennifer E. et al. 2013. "Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment." *Cell*.

Di Pierro, Michele et al. 2016. "Transferable Model for Chromosome Architecture." *Proceedings of the National Academy of Sciences* 113(43): 12168–73. http://www.pnas.org/lookup/doi/10.1073/pnas.1613607113.

Plimpton, Steve. 1995. "Fast Parallel Algorithms for Short-Range Molecular Dynamics." *Journal of Computational Physics* 117(1): 1–19.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26(6): 841–42.

Quinodoz, Sofia A. et al. 2018. "Higher-Order Inter-Chromosomal Hubs Shape 3D Genome Organization in the Nucleus." *Cell* 174(3): 744-757.e24. https://doi.org/10.1016/j.cell.2018.05.024.

Rao, Suhas S.P. et al. 2017. "Cohesin Loss Eliminates All Loop Domains." *Cell* 171(2): 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026.

Rao, Suhas S.P. P et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159(7): 1665–

80. http://www.ncbi.nlm.nih.gov/pubmed/25497547%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5635824.

Rodríguez-Carballo, Eddie et al. 2017. "The HoxD Cluster Is a Dynamic and Resilient TAD Boundary Controlling the Segregation of Antagonistic Regulatory Landscapes." *Genes and Development* 31(22): 2264–81.

Rosa, Angelo, and Ralf Everaers. 2008. "Structure and Dynamics of Interphase Chromosomes." *PLoS Computational Biology*.

Sachs, R. K. et al. 1995. "A Random-Walk/Giant-Loop Model for Interphase Chromosomes." *Proceedings of the National Academy of Sciences of the United States of America*.

Sagai, Tomoko et al. 2005. "Elimination of a Long-Range Cis-Regulatory Module Causes Complete Loss of Limb-Specific Shh Expression and Truncation of the Mouse Limb." *Development*.

Salamon, Peter, Paolo Sibani, and Richard Frost. 2002. SIAM monographs on mathematical modeling and computation *Facts, Conjectures, and Improvements for Simulated Annealing*. http://www.loc.gov/catdir/enhancements/fy0708/2002029215-d.html%5Cnhttp://www.loc.gov/catdir/enhancements/fy0708/2002029215-t.html.

Sanborn, Adrian L. et al. 2015. "Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes." *Proceedings of the National Academy of Sciences* 112(47): E6456–65. http://www.pnas.org/lookup/doi/10.1073/pnas.1518552112.

Sexton, Tom et al. 2012. "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome." *Cell*.

Shen, Yin et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature*.

Simonis, Marieke et al. 2006. "Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture-

on-Chip (4C).” *Nature Genetics*.

Spielmann, Malte, and Stefan Mundlos. 2013. “Structural Variations, the Regulatory Landscape of the Genome and Their Alteration in Human Disease.” *BioEssays* 35(6): 533–43.

Tanabe, Hideyuki et al. 2002. “Non-Random Radial Arrangements of Interphase Chromosome Territories: Evolutionary Considerations and Functional Implications.” *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*.

Tanay, Amos, and Giacomo Cavalli. 2013. “Chromosomal Domains: Epigenetic Contexts and Functional Implications of Genomic Compartmentalization.” *Current Opinion in Genetics and Development*.

Tang, Zhonghui et al. 2015. “CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.” *Cell* 163(7): 1611–27.

Tiana, Guido et al. 2016. “Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains.” *Biophysical Journal*.

Trapnell, Cole et al. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature biotechnology* 28(5): 511–15. http://www.ncbi.nlm.nih.gov/pubmed/20436464%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3146043.

Williamson, Iain, Laura A. Lettic, Robert E. Hill, and Wendy A. Bickmore. 2016. “Shh and ZRS Enhancer Colocalisation Is Specific to the Zone of Polarising Activity.” *Development (Cambridge)*.

Yaffe, Eitan, and Amos Tanay. 2011. “Probabilistic Modeling of Hi-C Contact Maps Eliminates Systematic Biases to Characterize Global Chromosomal Architecture.” *Nature Genetics* 43(11): 1059–65.

Yang, Tao et al. 2017. “HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient.” *Genome Research* 27(11): 1939–49.

List of Figures