Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXX

Coordinatore: prof. Salvatore Capozziello

Polymer physics models of the chromatin spatial organization in the cell nucleus

Settore Scientifico Disciplinare FIS/02

Dottorando

Simona Bianco

Tutore

Prof. Mario Nicodemi

Anni 2014/2017

Contents

Introduction

The spatial organization of the chromosomes in the nucleus of mammalian cells has a key role in cell vital functions like the regulation of gene transcription and expression (Misteli, 2007; Lieberman-Aiden *et al.*, 2009; Dekker *et al.*, 2013, Tanay & Cavalli, 2013; Bickmore & van Steensel, 2013). Understanding chromosome architecture, its folding mechanisms and its interplay with gene regulation is a current challenging problem. In the last decade, new experimental technologies have been developed, the Chromosome Conformation Capture (3C) techniques (Dekker *et al.*, 2013), that give information about inter and intra chromosome interactions, allowing to investigate, for the first time, the chromosomes three-dimensional spatial folding in a quantitative way. Briefly, these experiments measure the frequency of interactions between pairs of genomic regions, that are close in 3D space, but may have a large separation along the linear genomic sequence. An important variant of such technologies, Hi-C, has extended the possibility to map pairwise chromosome interactions genome-wide. Such novel technologies led to the discovery that chromosomes are characterized by a complex, non-random, 3D organization that occurs at different genomic length scales, through many local and long-range interactions (Lieberman-Aiden *et al.*, 2009). Chromosomes occupy distinct territories (Cremer&Cremer, 2001) and have preferred positions depending on cell type and transcription activity (Misteli, 2007; Tanay & Cavalli 2013; Bickmore & van Steensel, 2013). Within chromosomes, the genome is folded into a sequence of domains, called “topological associating domains” or briefly TADs (Dixon *et al.*, 2012; Nora *et al.*, 2012), in which segments of DNA interact frequently with each other. 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) and patterns are also seen within TADs (Sexton *et al.*, 2012; Philips-Cremins *et al.*, 2013). Chromatin 3D structure have crucial biological roles, as the control of the gene activity through the formation of loops between regulatory regions and genes. The disruption of such interaction network can alter the regular gene activity and produce effects on the fenotype (Spielmann & Mundlos, 2013; Lupianez *et al.*, 2015). Nevertheless, the mechanisms shaping chromatin spatial organization are still largely unknown. Polymer physics models have been proposed to interpret the rich amount of data from large-scale 3C experiments in terms of chromatin 3D structure and to identify key physical mechanisms underlying chromatin folding, in an innovative and fascinating research field at the border

between physics and biology. (Chiariello *et al.*, 2016; Fudenberg *et al.*, 2016; Tiana *et al.* 2016; Sanborn *et al.*, 2015; Nicodemi & Pombo, 2014; Giorgetti *et al.*, 2014; Jost *et al.*, 2014; Brackley *et al.*, 2013; Barbieri *et al.*, 2012; Rosa & Everaers, 2008; Marenduzzo *et al.*, 2006; Sachs *et al.*, 1995).

The research presented in this PhD thesis has been conceived in this general picture. It has been 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 work 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), the Development and Disease Group directed by Professor Stefan Mundlos, at Max Planck Institute for Molecular Genetics (Berlin), and the Genome Biology group directed by Professor Jim Hughes, at Oxford University.

This thesis is organized in four chapters. In Chapter 1, we try to highlight the importance of the chromatin spatial organization, and recall very briefly some concepts necessary to the comprehension of this research activity, as the Hi-C technique, the interpretation of the chromosome interaction data and the relationship between spatial organization and cell functionality. Then, we review the polymer models currently proposed to describe the genome three-dimensional architecture. In Chapter 2 we focus on the employment of polymer models to quantitatively explain the information contained in the Hi-C interaction data. In particular, we show how polymer thermodynamics concepts can be used to explain the long-range contact profile of chromosomes; then we try to schematically model the hierarchical structure of the DNA, and finally we present a theoretical study of the multiple co-localization contact landscape. In Chapter 3, we introduce more sophisticated polymer models by which we can reconstruct the 3D genome structure with very high accuracy, and we will study some important chromosome loci in detail. In Chapter 4, we will test the power of such polymer models to predict, for the first time, the effects on chromosome architecture of genomic mutations, like deletions, inversions and duplications. Genome rearrangements can result in a re-wiring of interactions between genes and regulatory sequences, leading to abnormal gene expression and disease (Lupiáñez *et al.* 2015). We will show how polymer modeling can be employed as a valid tool to predict interactions

and analyze the disease-causing potential of rearrangements *in-silico*, without performing extensive 3C-based experiments.

1. Chromatin spatial organization in the cell nucleus

1.1 Introduction

The spatial organization of the DNA in the cell nucleus of eukaryotic organisms is likely to be of great importance for fundamental biological functions like the regulation of gene expression. In recent years, many studies and experimental techniques have been developed to better understand how the genome is spatially organized and how such organization affects the cell functions. In this chapter, far from being exhaustive about this huge topic, we briefly review some recent, very important, results that are crucial in this research field and that will help the comprehension of our research activity described in the following chapters. Specifically, in Sections 1.2 and 1.3 we recall very elementary concepts of molecular biology together with some recent advances regarding gene regulation and epigenetics; in Section 1.4 we discuss the fundamental technologies that allow to quantitatively investigate the spatial architecture of the genome and in particular we focus on the Hi-C experimental technique; in Section 1.5 we report recent discoveries derived from the analysis of the data provided by these experimental methods, and we describe the emerging scenario about how DNA appears to be organized in the cell nucleus; finally, in Section 1.6 we review the most recent polymer models that aim to quantitatively explain and reconstruct the three-dimensional structure of the genome. Most of the results described in this chapter have been introduced and discussed in the papers: Lieberman-Aiden *et al.*, 2009, Dixon *et al.* 2012; Nora *et al.* 2012, Dekker *et al.*, 2013.

1.2 DNA packaging: chromosomes and chromatin

DNA (deoxyribonucleic acid) is a molecule that carries the genetic information used in the functioning of all known organisms. In eukaryotes nearly all the DNA is included in the nucleus, which occupies about 10% of the total cell volume. DNA is a double helix made of two paired long polymers of simple units called **nucleotides,** which are made of three components: a five-carbon sugar, a phosphate group, and a nitrogenous base. The base may be either adenine (A), cytosine (C), guanine (G), or thymine (T). The two strands of DNA

run in opposite directions to each other, one backbone is oriented 3' (three prime) and the other 5' (five prime): this notation refers to the direction along which the 3rd and 5th carbon of the sugar group of each nucleotide are encountered along the sequence. The two strands of DNA are linked together by hydrogen bonds between the nucleotides. Adenine only binds the opposite Thymine, while Cytosine only binds Guanine. It is the sequence of these four nucleobases along the backbone that encodes the **genetic information**.

Nuclear DNA is packaged into different linear filaments called **chromosomes,** occupying distinct spatial regions that are indicated as **chromosomal territories (CT)** and are clearly visible with microscopy techniques (**Fig. 1.1**, Cremer&Cremer, 2001). The genomic length, i.e. the number of **base pairs (bp)** composing the genome, and the number of chromosomes depends on the considered species. For instance, human genome consists of approximately 3.2×10^9 bp and it is distributed over 23 chromosomes. The majority of eukaryotic cells are diploid, i.e. they contain two copies of each chromosome, referred to as **homologous chromosomes**. The linear length of the total human genome is about 2m, included in a nucleus having a diameter of approximately 10÷15μm. This compaction level is achieved through an efficient interaction between DNA and proteins.

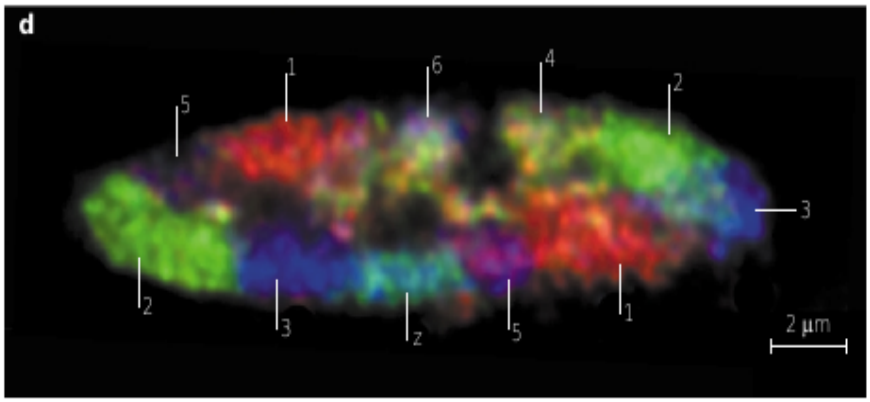

Figure 1.1: Chromosomes territories in the cell nucleus

In this microscopy image, chromosomes are marked by different colors. They tend to occupy regions in the nucleus that are distinct from each other, named chromosomal territories (CT). The image shows a mid-plane section of chicken fibroblast cells (figure adapted from Cremer&Cremer, 2001).

Chromatin is the complex of DNA and its bound proteins which is found in normal physiological conditions within the nucleus. This is the real physical structure containing the information to be processed. The organization of chromatin occurs at different genomic length scales and degree of compaction (**Fig. 1.2a**). **Histones** are responsible for the first and most basic level of chromosome packing, called **nucleosome**. Each individual nucleosome consists of a structure of eight histone proteins (two molecules each of histone H2A, H2B, H3 and H4) around which a double-strand of DNA is wrapped. The length of DNA associated with each nucleosome is 147 bp. This structure is called nucleosome core particle. Each nucleosome core particle (which is about 11nm) 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. On average, nucleosomes repeat at intervals of about 200 nucleotide pairs. 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. Linear arrays of nucleosomes fold into higher-order chromatin fibers, first through local chromatin interactions, but eventually giving rise to discrete 1 μm wide chromosome territories (CTs) [Cremer&Cremer, 2001]. Within a chromosome, it is possible to classify the chromosomal regions into two categories: **euchromatin** and **heterochromatin**. DNA in both types of chromatin is packaged into nucleosomes. Heterochromatic regions are composed by nucleosomal DNA that shows a high degree of compaction, while euchromatic nucleosomes are much less compacted (**Fig. 1.2b**). The high level of compaction reduces the accessibility of the DNA contained in these regions, which are therefore associated with a very low level of transcription.

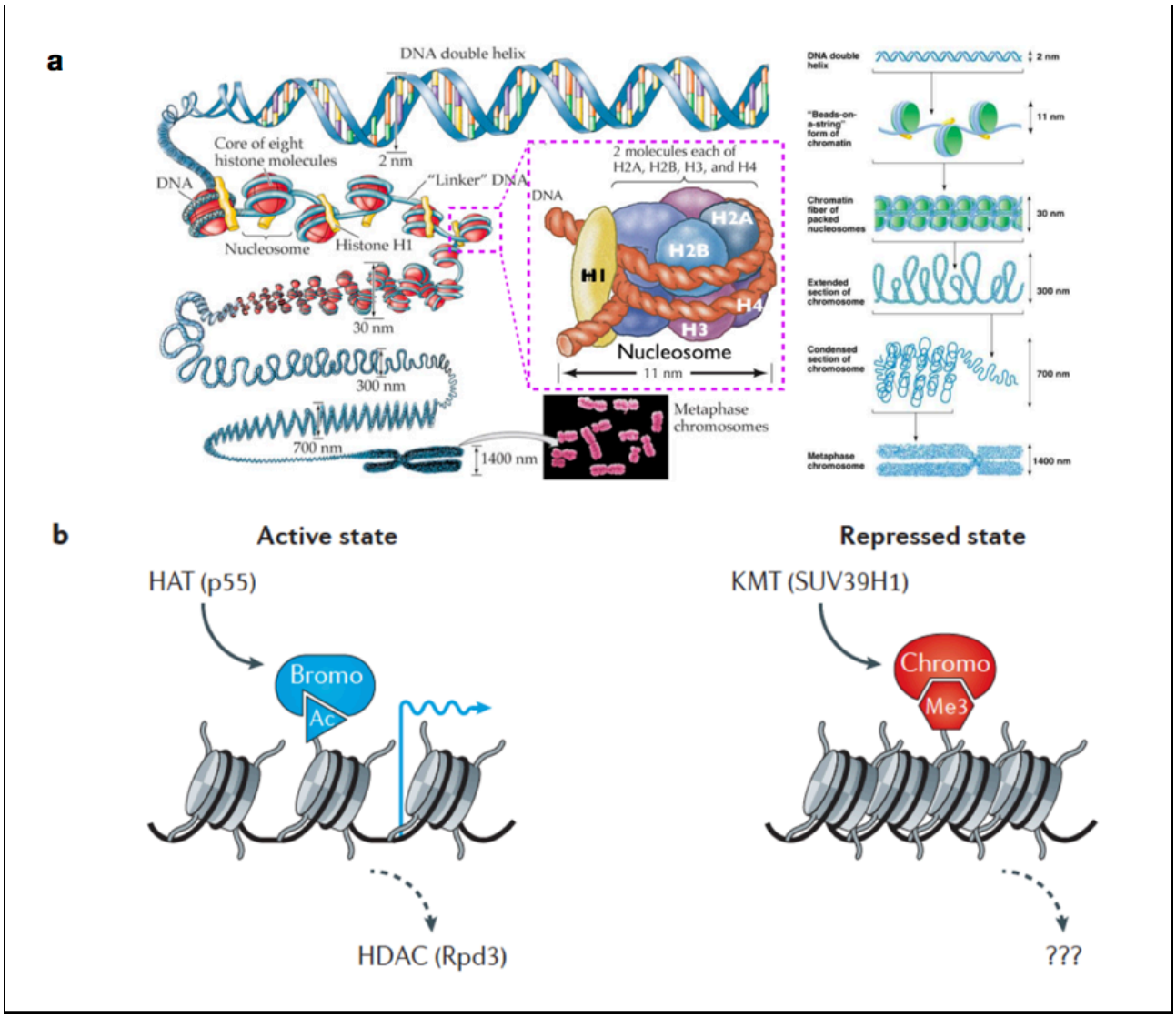

Figure 1.2: The organization of chromatin

a. Hierarchical folding of the chromatin and its basic unit, the nucleosome with its histone core octamer (Lodish *et.al.*, 2000). **b**. Left: Euchromatic nucleosomes are composed by nucleosomal DNA that shows a low degree of compaction. Acetylation (light blue triangle) of histone tails can be a transcription activating mark. Right: Heterochromatic regions are much more compacted and are associated with a very low level of transcription. Methylation (red hexagon) of histone tails can be a transcription repressive mark. (figure adapted from Allis&Jenuwein, 2016).

1.3 Genetic regulation and epigenetics

Every cell in a multicellular organism has the same DNA. Yet each cell type, such as muscle cells and skin cells in humans, has to generate a different set of proteins and even within a single cell type, its needs change throughout its life. **Transcriptional regulation of gene expression** plays a significant role in establishing the diversity of cell types and biological functions from a common set of genes. The components of regulatory control include **cis-regulatory elements**, such as *promoters*, *enhancers*, *silencers* and *insulators*, that act across immense genomic distances to influence the spatial and temporal distribution of gene expression (Noonan&McCallion, 2010). The transcriptional activity of a given gene is primarily regulated by the nearby DNA control sequence, called the gene **promoter**. The promoter, a sequence usually located in eukaryotes within 1kb upstream of the gene transcription starting site (TSS), is the fulcrum around which transcription starts, as it serves as an assembly point for the basal transcriptional machinery, including the polymerase holoenzyme, and defines the orientation and origin of transcription. The promoter activity is supplemented by other DNA regulatory regions. Among those, **enhancers** 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 rare cases even thousands, of kilobases (Calo&Wysocka, 2013; Bulger&Groudine, 2011; Ong&Corces, 2011, 2012). Regulatory proteins, known as **Transcription Factors (TFs),** bind there along with other regulatory factors (e.g., non coding RNAs), to activate the gene. Chromatin is, thus, folded into loops to bring distal regulatory elements into contact with their target genes.

Cis-acting elements can often be identified by their high evolutionary **sequence conservation** or by specific **epigenetic marks** associated to them. Epigenetics is the study of heritable changes in gene activity that are not caused by changes in the DNA sequence. Examples of mechanisms that produce such epigenetic changes are DNA methylation and **histone modification**, each of which alters how genes are expressed without altering the underlying nucleotide sequence. In fact, as discussed in the previous section, the packing of the eukaryotic genome into chromatin provides the means for compaction of the entire genome inside the nucleus, but, at the same time, this packing restricts the access to DNA of the many regulatory proteins essential for biological

processes like transcription, replication, DNA repair and recombination (Konberg&Lorch, 1999). Epigenetic mechanisms, including covalent modification of histone tails like acetylation, methylation, phosphorylation and ubiquitination, can counterbalance the repressive nature of chromatin, allowing access to nucleosomal DNA (de la Cruz *et al.* 2005) (**Fig. 1.2b**). In recent years, technological advances, such as **chromatin immunoprecipitation followed by next-generation sequencing (ChIP–seq)**, have enabled the genome-wide analysis of the epigenetic modifications at or near base-pair resolution. Briefly, ChIP-seq is a method to detect where a protein is bounded along the DNA, by use of specific antibodies that target the protein of interest. In particular, antibodies can recognise a histone protein with a specific modification. Such techniques lead to the identification of some consistent patterns of histone marks, used for example to better define DNA regulatory regions: for example H3K4me1 (histone H3 lysine 4 monomethilation) and H3K27ac (histone H3 lysine 27 acetylation) characterize enhancer regions, H3K4me3 (histone H3 lysine 4 trimethilation) promoter regions, H3K36me3 (histone H3 lysine 36 trimethilation) transcribed regions, H3K27me3 (histone H3 lysine 27 trimethilation) Polycomb-mediated repressed regions and H3K9me3 (histone H3 lysine 9 trimethilation) heterochromatin regions (Allis&Jenuwein, 2016).

1.4 The Chromosome Conformation Capture (3C) based techniques

As discussed above an important role in regulating gene expression is played by a complex network of short and long range contacts between genes and their regulatory sequences, so highlighting the central role of the three-dimensional organization of the chromatin in the functioning of the cell. During the last decade, a series of molecular and genomic approaches have been developed and can be used to study 3D chromosome folding with unprecedented accuracy. These methods are all based on the **chromosome conformation capture (3C)**. They allow the determination of the frequency with which any pair of loci in the genome is in close enough physical proximity (in the range of 10÷100nm) to become crosslinked (i.e. the pair can be bound by some molecule). Schematically, the steps common to all the 3C methods are the following: the cells in the population 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 and finally the DNA is purified and analyzed (Dekker *et al*. 2013). The difference among the specific methods is how the ligation product is detected. The most common methods are the 3C, 4C, 5C and HiC (see next subsections).

1.4.1. 3C, 4C and 5C methods

The biochemical experimental details can be found in the reference papers. Here, we will just discuss about what kind of data they produce. The **3C** (Dekker *et al.*, 2002) and **4C** (Simonis *et al.*, 2006) methods generate single interaction signals for individual loci. The 3C method yields a long-range interaction profile of a selected genomic region, like a gene promoter or other genomic element of interest, with chromatin in genomic proximity (**Fig. 1.3a**). The 4C method generates a genome-wide interaction profile for a selected locus (**anchor** or **viewpoint**, **Fig. 1.3b**). These data sets can be represented as single tracks that can be plotted along the genome. **5C** method (Dostie *et al.*, 2006) is not anchored on a single locus of interest but instead generate matrices of interaction frequencies that can be represented as two-dimensional heat maps (i.e. the intensity is indicated by the color scheme) with the genomic positions along the two axes (**Fig. 1.3c**). The Hi-C method will be discussed with some more detail in the next subsection.

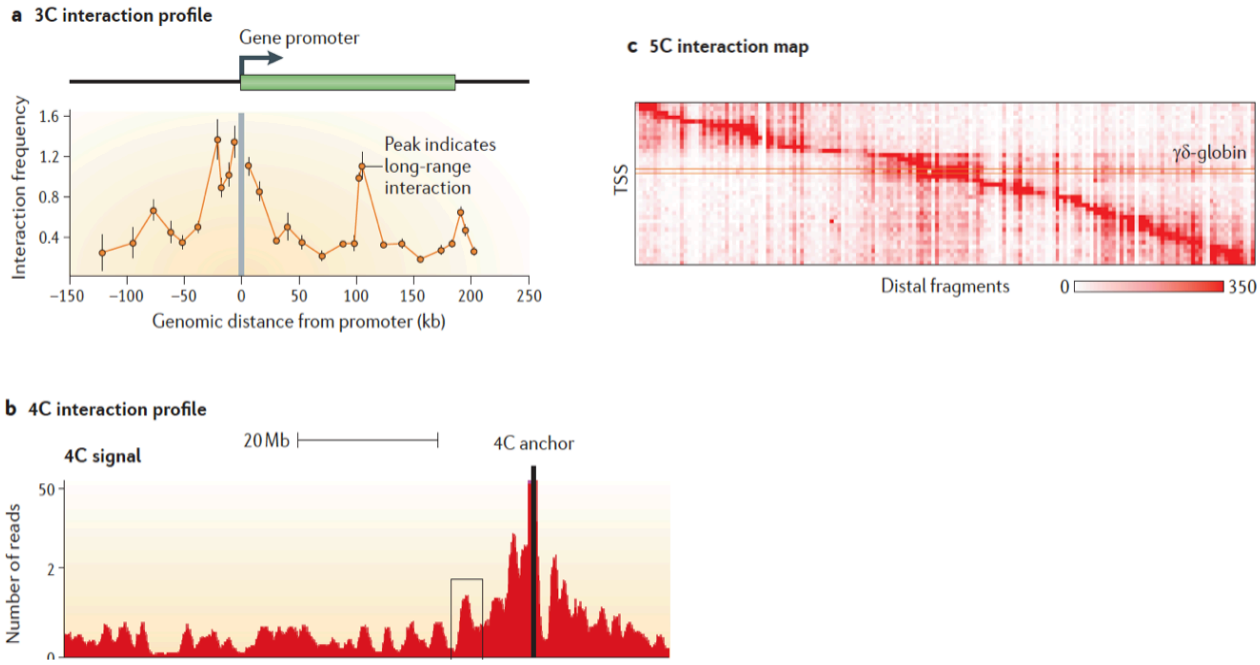

Figure 1.3.1: Examples of 3C, 4C and 5C data sets.

a. Example of chromosome conformation capture (3C) data. **b.** Example of 4C data from the mouse genome. In 3C and 4C, on the x-axis is reported the distance from the anchor point, or point of view. **c.** Example of a 5C interaction map for the ENCODE ENm009 region in K562 cells. The different rows contain an interaction profile of a transcription start site (TSS) in the 1 Mb region on human chromosome 11, that contains the β-globin locus. (Figure adapted from Dekker *et al.*, 2013.)

1.4.2. The Hi-C method

The **Hi-C** method (Lieberman-Aiden *et al.*, 2009) is the first genome-wide adaptation of 3C and include a further step in which, after restriction digestion, the staggered DNA ends are filled in with biotinylated nucleotides (**Fig. 1.4A**). 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. This facilitates selective purification of ligation junctions that are collected in a Hi-C library and then directly sequenced along the genome, producing a list of interacting fragments.

Then, data are organized in a genome-wide **contact matrix**, obtained by dividing the genome into windows (indicated as *loci*) of fixed length (in the first version, this length was 1Mb=1000000bp long). This important parameter defines the Hi-C data **resolution**. Each bin of the matrix x_{ij} contains the number of ligation products between the locus *i* and locus *j*. So, the extracted information is the contact frequency of any pair of loci *i* and *j* in the chromosome, that is obviously directly related to the chromosome spatial architecture.

In **Fig. 1.4B-D**, is reported an example of interaction matrix for an entire chromosome at 1Mb resolution. Since the Hi-C technique can detect interactions between any two loci in the genome, to each chromosome is associated its contact matrix (*Cis* data). Interactions between loci belonging to different chromosomes are also detected (with a much lower frequency), and are organized in *Trans* contact matrices. In all this work, we will focus only on *Cis* contact matrices.

Summarizing, Hi-C, as well as all the 3C-based methods gives information about the frequency in a cell population by which two loci *i* and *j* are in close spatial proximity. Anyway, the method does not give information about the nature of the contact, not distinguishing functional from non-functional associations and it does not reveal the mechanisms producing loci co-localization. Spatial proximity can be the result of contacts

mediated by protein complexes that bind them or of indirect co-localization to the same sub-nuclear structure (as the nuclear lamina). Loci co-localization can also be due to random collisions between distant regions of chromatin in the nucleus, caused by the chromosome flexibility. Also, the exact 3D structure of a specific region is highly variable from cell to cell, even if they are in the same differentiation stage. Each ligation product due to an interaction represents a contact involving a pair of loci in a single cell in the population. Thus, 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. The final value contained in the matrix bin x_{ij} represent the sum of interactions over a large cell population, and in each cell chromosomes conformation is determined by many different factors that act on the chromatin polymer and make the structure highly variable.

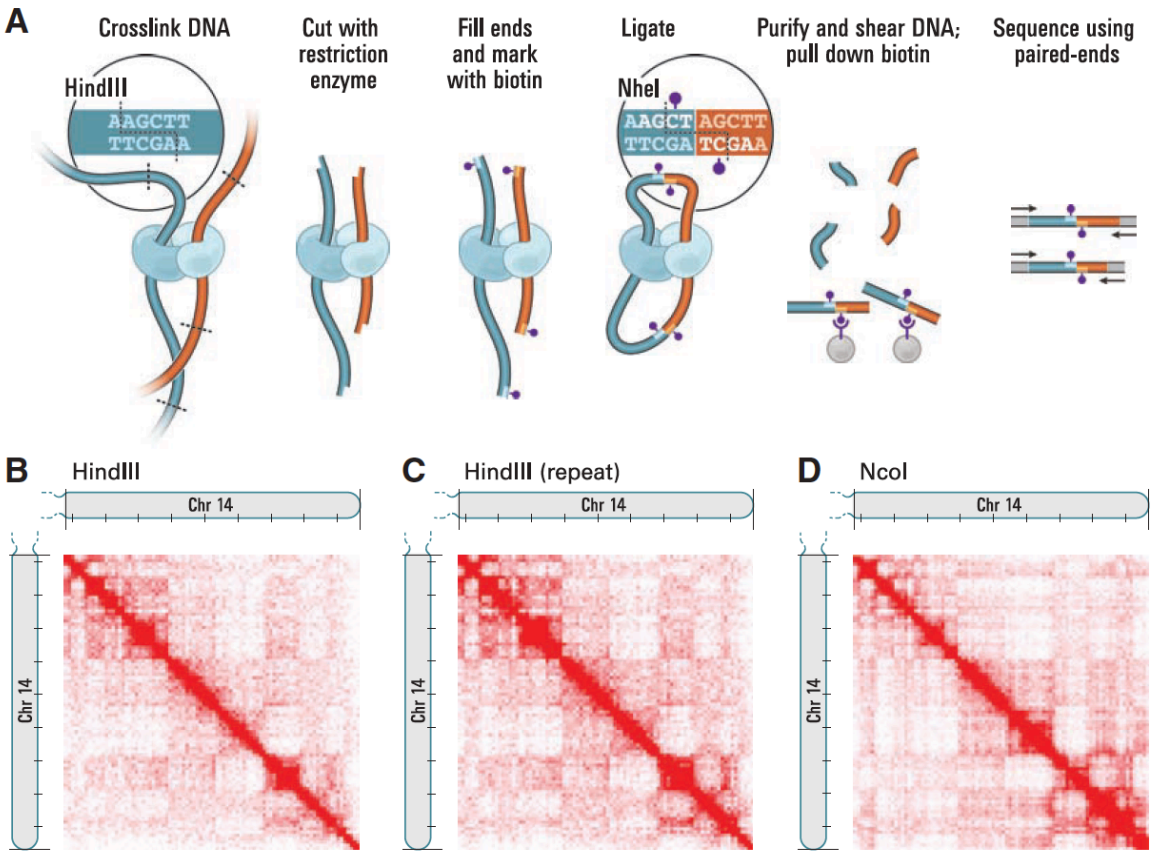

Figure 1.4.1: Hi-C technique

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 (for example HindIII or NcoI), and the resulting sticky ends are biotinylated; crosslinked fragments are then ligated to form unique hybrid DNA

molecules; DNA is purified and sheared. The labeling with biotin allow to efficiently detect the ligated fragments, which are finally identified by paired-end sequencing.
B. Example of Hi-C data output. Data are collected in a bidimensional heatmap (as in the 5C case). Hi-C data from the entire mouse chromosome 14 are shown, at 1Mb resolution, so that each pixel represents all interactions between a 1 Mb locus and another 1 Mb locus; intensity corresponds to the total number of reads (0 to 50).
C and **D:** Comparison between biological replicates using the same restriction enzyme [C] and using a different restriction enzyme [D, NcoI].
(Figure from Lieberman-Aiden *et al.*, 2009).

1.5 Chromatin 3D features from Hi-C data

1.5.1. A/B compartments

Since the 5C and Hi-C methods were introduced, interaction data have been analyzed to identify structural features of chromatin. A fundamental discovery, made through a principal component analysis (**Fig. 1.5a**) and also confirmed by microscopy experiments, is that each chromosome can be partitioned in two classes of regions, named **A and B compartments** (Lieberman-Aiden *et al.*, 2009, Rao *et al.*, 2014). Regions in the same compartment are enriched in mutual interaction while regions belonging to different compartments are depleted in mutual interaction. Compartments are considerably large regions of chromatin, having a characteristic size of 5÷10 Mb, and alternate along the chromosomes. Compartment A is typically associated with euchromatin, since it is less compact and correlates with gene enrichment, expression and accessibility, while compartment B has higher interaction values (Lieberman-Aiden *et al.*, 2009). This finding is compatible with the known presence in the nucleus of open and closed chromatin (**Fig. 1.5b**).

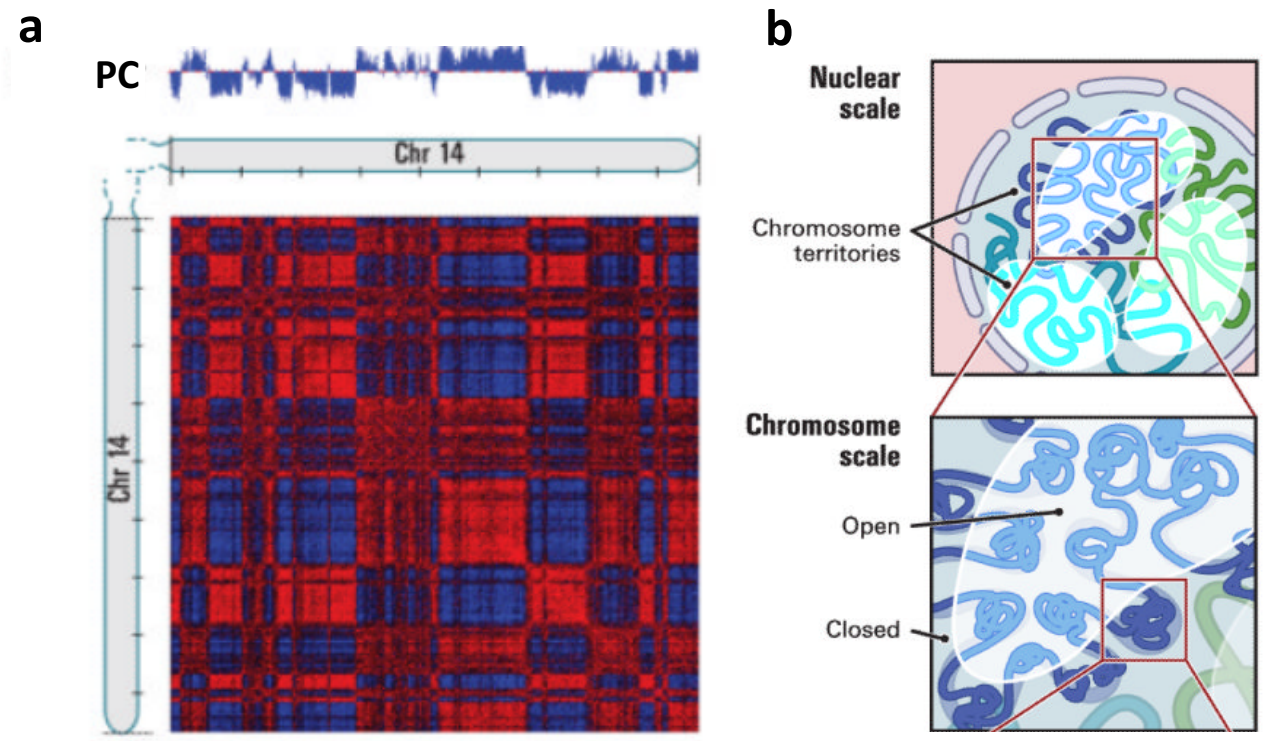

Figure 1.5: A and B compartments

a. Pearson correlation map of chromosome 14 and the principal component (PC) associated. 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.

1.5.2. Topologically associated domains (TADs)

Beyond the compartments A and B, the analysis of the patterns contained in the Hi-C contact data has brought to the discovery of other levels of organization and structural units. Importantly, a common feature among several organisms, from *drosophila melanogaster* to mouse and human, is the existence of discrete regions, much smaller than A and B compartments, where chromatin is marked by high levels of internal interactions. To indicate such domains, various names have been used in literature, as topological domains (Dixon *et al.*, 2012) and **topological associating domains** or, briefly, **TADs** (Nora *et al.*, 2012). As standardly used in literature now, we will use the latter in the following. In 5C or Hi-C matrices, TADs appear as squares of high intensity along the diagonal (**Fig.**

1.6a). From the structural point of view, this correspond to the fact that distinct loci located in the same TAD tend to interact with each other much more than two loci located in two different TADs. Microscopy experiments also confirm this scenario (Dixon *et al.*, 2012). TADs are found to be universal building blocks of chromosomes, as both mouse and human are composed by more than 2000 TAD domains, covering almost all the genome. Furthermore, they are largely conserved between different species (Dixon *et al.*, 2012). Their typical size (approximately 0.5÷1 Mb) is much smaller than the A and B compartment, and they can be active or inactive. To identify TADs several computational algorithms have been developed (Dixon *et al.*, 2012, Rao *et al.*, 2014, Fraser *et al.*, 2015). The mechanism that regulates the formation of TADs is still not clearly understood, and polymer models have been proposed to quantitatively describe it (Barbieri *et al.*, 2012, Brackley *et al.*, 2013, Sanborn *et al.*, 2015, Fudenberg *et al.*, 2016, Chiariello *et al.*, 2016).

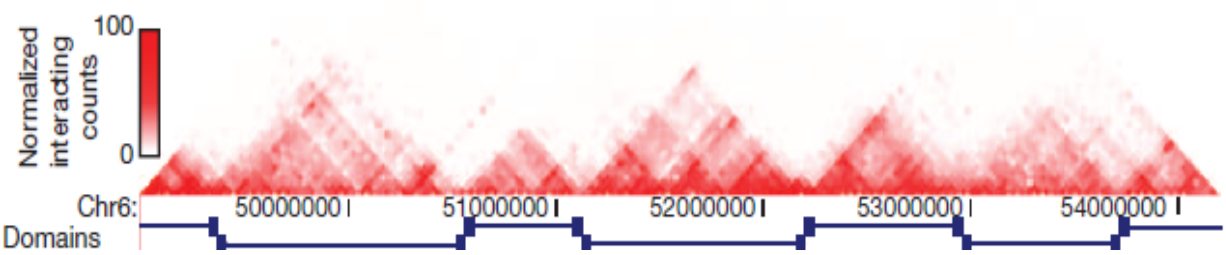

Figure 1.6: Topological Associating Domains (TADs)

Hi-C interaction 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 only the upper triangular matrix is represented, 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).

1.5.3. Further research developments

The results reviewed in this chapter represent only a limited part, yet fundamental, of the key points in the recent history of this research field. In the previous subsections, we have seen that chromosomes are folded into a sequence of mega-base sized domains of strong internal Hi-C interactions, named TADs (Dixon *et al.* 2012; Nora *et al.* 2012) and that larger structures have been also identified, such as the 10 Mb wide A/B compartments linked respectively to more transcribed and repressed chromatin sequences (Lieberman-Aiden *et al.* 2009). Yet, patterns are visible also within and across TADS at lower as much as at very large scales (Phillips-Cremins *et al.* 2013). In Fraser *et al* 2015, higher-order chromatin 3D structures were identified by analyzing the interactions between TADs, through mouse neuronal differentiation. Chromatin was found to be organized in a hierarchy of domains-within-domains, named "**metaTADs"**, up to chromosomal scales Importantly, MetaTADs, are correlated with a variety of epigenetic features, pointing towards a functional role of this organization. (Fraser *et al* 2015, reviewed in Bianco *et al*. 2017)

Another important feature that has recently been discovered is the formation of chromatin "**loops**" (Rao *et al* 2014), which specifically bring together pairs of sites that are distant along the linear sequence of the chromosome. Loops appear as point-like peaks of very high interaction in the Hi-C matrix and often occur at TAD boundaries (Rao *et al* 2014).

In general, as the technology quickly evolves, more sophisticated and refined experiments have been performed, producing better and higher quality data. In this way, very complex and more complete Hi-C datasets are available, with higher resolutions (up to 1Kb, Rao *et al.*, 2014, Dixon *et al.*, 2015) and for an increasing number of tissues and cell lines. In parallel, other experimental technologies have been developed to detect chromatin contacts (Khalor *et al.* 2011; Rao *et al.*, 2014; Beagrie *et al.*, 2017). Furthermore, experiments have been performed to evaluate the impact of chromatin structure alterations on health (as TADs disruption, described in Lupiáñez *et al.*, 2015, or neoTAD formation, described in Franke *et al.*, 2016). These notable works demonstrate the deep relationship between chromatin organization and individual fenotype, and confirm once more the importance of investigating the genome architecture in space. Overall, these more recent results allow to improve our knowledge (far anyway from being complete) about the chromatin organization, contributing to further enrich the scientific landscape about this interesting topic.

1.6. Polymer models of chromatin folding

Together with the improvements of the experimental techniques, also the theoretical technologies improve so to develop models that describe genome architecture. Many models have been proposed to explain quantitatively the behavior of chromatin in the nucleus, and in this subsection we will list very briefly some of them, for sake of completeness. We start considering the fundamental **String and Binders Switch (SBS)** model (Barbieri *et al.*, 2012), where a chromatin fiber is modeled 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. In the following chapters, we will use this model as starting point for our considerations about chromatin architecture. The idea of chromatin interacting with floating particles has been used also in other studies (Brackeley *et al.*, 2013, Chiariello *et al.*, 2016). After the developments of the Hi-C technology, the first proposed model as possible genome structure was the **Fractal Globule** (Lieberman-Aiden *et al.*, 2009), which emerges as result of polymer condensation during which topological constraints prevent knotting and slow down equilibration of the polymer (Dekker *et al.*, 2013). Another important model is the **Dynamic Loop** model (Bohn&Heerman, 2010), where chromatin moves under diffusional motion and when two sites co-localize, they form a loop with a certain probability for a certain lifetime. Another model consider chromatin as a sequence of regions characterized by an epigenetic state (Jost *et al.* 2014) with regions in the same state having specific interactions. Other models consider chromatin folding as the result of interaction of TAD boundary elements through dynamic mechanisms of **Loop Extrusion** (Sanborn *et al.*, 2015, Fudenberg *et al.*, 2016). In this process, cis-acting loop-extruding factors (as cohesin) form progressively larger loops but stop at TAD boundaries due to interactions with boundary proteins, like CTCF (Fudenberg *et al.*, 2016).

2. Polymer physics reproduces key features of the chromatin large-scale 3D architecture

2.1 Introduction

In this chapter, we will show how polymer physics can explain key features of the complex 3D organization of the chromatin. In Sections 2.2, we will introduce the Strings and Binders Switch Model (SBS) of chromatin, originally presented in the work by Barbieri *et al.* (2012), and, in Section 2.3, its Molecular Dynamics (MD) implementation. We will show the resulting phase diagram, with a novel thermodynamic stable state (Section 2.4). Next, in Section 2.5 we will show how, with few parameters, we are able to recapitulate the average behavior 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 the following Section 2.6, we will see how the SBS model can reproduce important features of chromatin organization like TADs and higher order structures. Furthermore, we will describe the theoretical multiple contact profile, which recently have been discovered to be very important (Olivares-Chauvet *et al.*, 2016, Beagrie *et al.*, 2017) for genome architecture and regulation (Section 2.7).

Most of the material presented in this Chapter, including figures, paragraphs and sentences, is adapted or taken literally from the published papers: Chiariello *et al.* 2016, Annunziatella *et al.*, 2016 and Annunziatella *et al., Submitted* (2017), which I coauthored.

2.2 The Strings and Binders Switch (SBS) model

In the **Stings and Binders Switch (SBS) model** (Nicodemi&Prisco, 2009, Barbieri *et al.*, 2012), a chromatin filament is represented as a classical **Self-Avoiding-Walk (SAW)** chain of beads. The beads can interact with molecular binders diffusing in the surrounding environment, though an attractive potential with interaction energy E_{int}. The binders have a concentration c, and can bridge the beads of the chain and fold spontaneously the polymer. Different equilibrium thermodynamics phases exist according to the value of control

parameters, E_{int} and c, giving rise to specific, corresponding conformational classes (see Section 2.4). A schematic representation of the SBS model is shown in **Fig. 2.1**, showing the case with only one type of binders and binding sites (red); yet, to describe more complex situations, different types of beads (and cognate binders) can be introduced, schematically represented by different "colors" (Bianco *et al.* Submitted (2017), Barbieri *et al.* 2017, Chiariello *et al.* 2016, Annunziatella *et al.* 2016).

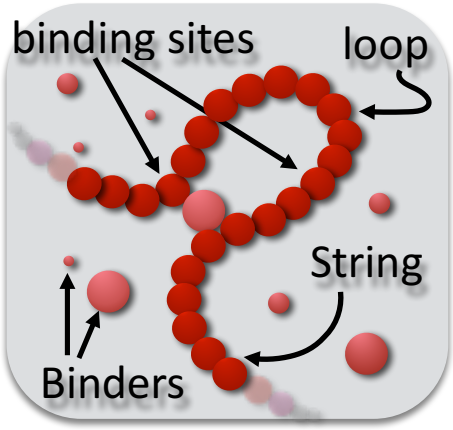

Figure 2.1: The Strings and Binders Switch (SBS) model

The SBS model is a self-avoiding chain of beads (the "string") interacting with diffusing molecules (the "binders") having a concentration, c, and a binding affinity E_{int}. The binders can bridge distant chain beads and loop the polymer. (Figure adapted from Chiariello *et al.* 2016)

2.3 Molecular Dynamics simulations of the SBS model

Molecular Dynamics (MD) techniques are very powerful and standard methods to investigate molecular structures, and are broadly used to study the folding processes and conformational properties of other important molecules like proteins (Di Carlo, Minicozzi *et al.*, 2015). For our simulations, we used the open source software **LAMMPS (Large Atomic Molecular Massive Parallel Simulator**, employing the Verlet algorithm (Plimpton, 1995)).

2.3.1. Langevin equation

In computer simulations, the chromatin filament is represented as a polymer composed by N beads, having each a diameter σ. The system composed by the polymer beads chain and its binders, is 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&Grest, 1990):

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

where *m* is the mass of the generic particle, *v*(*t*) the particle velocity, *V* is the potential acting on the particles (see next subsection) and *f*(*t*) a stochastic random force that 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 ζ=3πησ. As usual in MD simulations, we work in dimensionless units (Kremer&Grest, 1990). So, we set the diameter of the polymer bead σ equal to 1 (for simplicity, we do the same 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 is 1 and the temperature *T* is 1. For the dynamics, we set ζ=0.5 (Kremer&Grest, 1990; Rosa&Everaers, 2008; Brackley *et al.*, 2013). The Langevin equation is integrated using the Verlet algorithm (Plimpton, 1995). All these settings are standard choices and are well described in Kremer&Grest, 1990. 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 (see subsection 2.3.5).

2.3.2. Potentials

The potential energy *V*(x), of a particle having a position x, includes three components (**Fig. 2.2**). First, between two consecutive beads of the chain there is a potential that models a **finitely extensible non-linear elastic** (**FENE**, Kremer&Grest, 1990) spring:

(2) $$V_{FENE} = -0.5KR_0^2 \ln\left[1 - \left(\frac{r}{R_0}\right)^2\right]$$

where R_0 is the maximum extension of the spring and K is the strength of the spring. We set R_0=1.6σ and K=30k_BT/σ^2 (Kremer&Grest, 1990; Brackley *et al.*, 2013). Second, to consider excluded volume effects between any two particles, there is a hard-core repulsive force $V_{hard}(r)$, modeled by a shifted **Lennard-Jones (LJ)** potential:

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

The third contribution to the total potential in the system, is represented by the interaction between bead and binder. A bead of the polymer can interact with its cognate binder through an attractive, cut, Lennard-Jones $V_{int}(r)$:

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

where ϵ_{int}, in k_BT units, is the parameter that controls the strength of the interaction, r_{int} is the cut-off distance that regulates the interaction range and σ_{b-b} is the distance between bead and binder when they are close in space (i.e. the sum of their radii, therefore in our case is 1σ). In our simulations, we set $r_{int} = 1.3\sigma$, unless otherwise stated. The energy of the interaction between beads and binders is given by the minimum of the interaction potential $V_{int}(r)$:

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

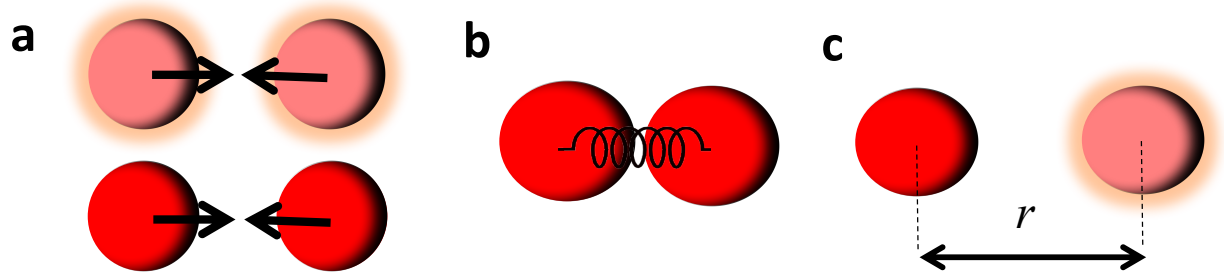

Figure 2.2: Schematic representation of the potentials

a. Between any two particles, there is a purely repulsive LJ potential, defined in equation (3), necessary to model the excluded volume effects. **b.** Between two consecutive polymer beads, the bond is a finite extensible non-linear elastic (FENE) potential, defined in equation (2). **c.** Between bead and binders the interaction is modeled with an attractive, shifted Lennard-Jones potential, defined in equation (4).

2.3.3. Preparation of the initial Self-Avoiding Walk states

In our simulations, the polymer is initially prepared in a random SAW configuration. To produce a SAW configuration, we use the following standard approach (Kremer&Grest, 1990): 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 (i.e., 0.97σ). Then, to remove overlaps between beads and binders, we let the system equilibrate, for some timesteps, where the hard-core LJ repulsion is replaced by a soft potential $V_{soft}(r)$ (Kremer&Grest, 1990; Brackley *et al.*, 2013):

$$V_{soft}(r) = \begin{cases} A\left[1 + \cos\left(\frac{\pi r}{2^{1/6}\sigma}\right)\right] & r < 2^{1/6}\sigma \\ 0 & otherwise \end{cases} \quad (6)$$

where the factor *A* increases linearly in time. To check that a SAW state has been approached it is convenient to monitor a set of physical quantities. An important one is the **gyration radius R_g** associated to the polymer, defined by the following relation:

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

where m_i and r_i are respectively the mass and the position of the i-th bead, M is the total mass of the polymer and r_{CM} is the position of the center of mass of the polymer. The gyration radius gives an estimation of the size of the average sphere enclosing the polymer.

In MD simulation, it is necessary to record R_g as a function of time t during the dynamics: when it reaches a plateau, the equilibrium SAW state should have been reached. An important additional check that the equilibrium state is attained is based on studying the scaling properties of the gyration radius R_g as function of the polymer length N. In fact, as known from polymer physics (de Gennes, 1979), for SAW states R_g exhibits a power-law behaviour, $R_g{}^2 \propto N^{2v}$ where the scaling exponent v is 0.588.

2.3.4. Polymer folding dynamics

Starting from a polymer in a SAW configuration, the binders are introduced, randomly located in the simulation box at the concentration, c, of interest. The interactions between bead and binders fold the system in its thermodynamics equilibrium state. That can be monitored by checking the time evolution of the polymer gyration radius R_g and its scaling properties at the steady state. For instance, in case the system starts from a SAW state and is folded into its globular state (see next section), R_g is initially high (SAW configuration) and then decreases until reaching a plateau value in the final equilibrium globular state (**Fig. 2.3**).

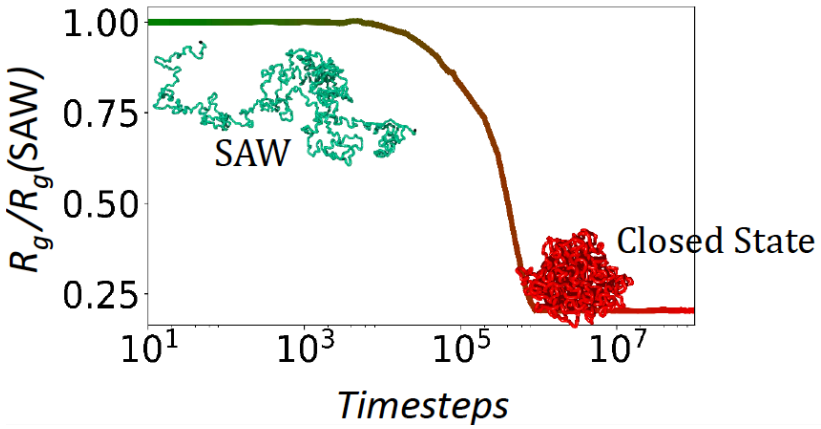

Figure 2.3: Polymer folding dynamics

The folding dynamics of the coil-globule transition is monitored by the gyration radius R_g (relative to its initial value) as a function of the MD time steps. As the dynamics proceeds, the polymer folds and and R_g decreases towards a plateau when equilibrium is reached. (Figure adapted from Annunziatella *et al.* Submitted (2017))

2.3.5. Mapping MD units into physical units

MD simulations use dimensionless units, called **Lennard-Jones or reduced units**. This means that σ, $\varepsilon=k_bT$ and m are taken as units of length, energy and mass respectively. The physical results can be easily obtained by a simple multiplication by a factor representing the specific physical unit, linked to the molecular details of the system or to experimental data (Allen&Tildesley, 1987). For instance, the physical unit of length can be obtained by imposing that the local density of chromatin equals the expected average density of DNA in the whole nucleus. Following this assumption, the expression for the physical length of the bead diameter is $\sigma = (s_0/G)^{1/3} D_0$, where G is the total genomic content of DNA in the cell, D_0 the average nuclear diameter of the considered cell type and s_0 the genomic content of each chain bead of the chromatin model (Barbieri *et al.*, 2012). The concentrations are estimated by 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. Analogously, the time scale τ is estimated by fixing the viscosity η and energy scale ε through the relation $\tau=\eta(6\pi\sigma^3/\varepsilon)$. So, by considering η=0.1P at room temperature T=300K, we obtain τ=0.03s. In the following polymer models, the physical units will be calculated in this way.

2.4 Phase diagram of a homo-polymer SBS chain

In this section, we will focus on the simplest SBS model system, where all the polymer beads are identical (homo-polymer model) and they all interact with the binders. We will use such a model to describe the average behavior of a generic region of the genome (i.e. a chromosome) and to extract quantitative information about its structure. Given the genomic length L of the region to be modeled, the corresponding genomic content (i.e. the number of bases) per bead is s_0=L/N, where N is the number of beads forming the chain. Here, we consider polymers made of N=1000 beads. Since a typical chromosome have a genomic length of approximately L=100Mb, each bead contains s_0=L/N=100Kb. We consider mouse embryonic stem cells, for which D_0=3.5μm and G=6.5Gb, so the resulting bead diameter of our polymers is σ=87nm (see subsection **2.3.5**).

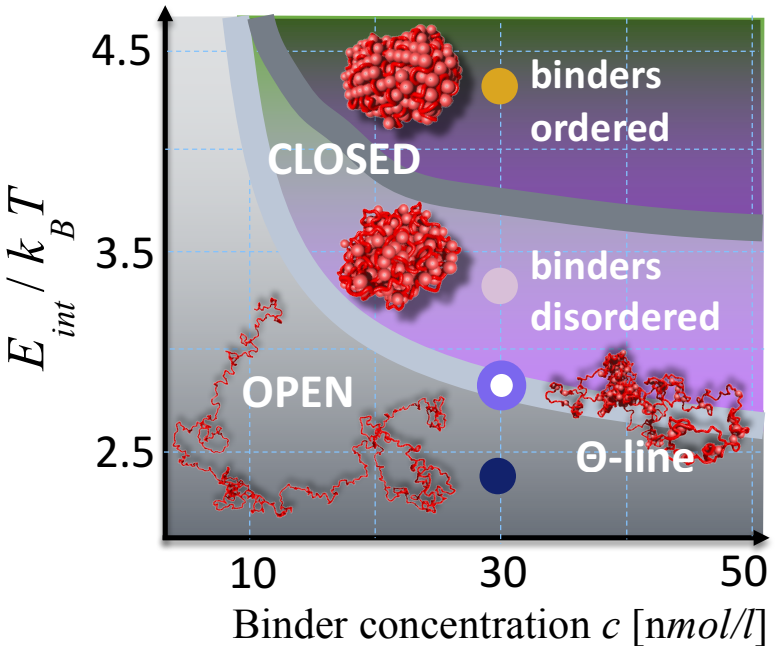

Figure 2.4: The Phase Diagram

As known in polymer physics the model stable architectural classes correspond to the different phases of its phase diagram: the polymer is open and randomly folded or, above its Θ-point transition, closed in more compact conformations; in the closed state, its binders can have a transition from a disordered to an ordered arrangement. Conformational changes can be sharply controlled (switch-like) by phase transitions driven by increasing *c* above threshold, e.g., by up up-regulation of the binder genes, or by chemical modifications of their binding sites, acting on E_{Int}. (Figure adapted from Chiariello *et al.* 2016).

2.4.1. Thermodynamic conformational classes

The equilibrium state and the conformational folding class of the system depend on two control parameters: the interaction energy E_{int} between bead and binders and the concentration c of binders. As known from polymer physics, there is a coil-globule folding transition, highlighted by a sharp drop of the gyration radius (that is the order parameter of this transition) when crossing the theta point in the phase diagram (in **Fig. 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, **Fig. 2.4b**, light blue box). On the contrary, in the globular state the polymer is in a **close** configuration, occupying a very small fraction of the open state volume (**Fig. 2.4b**, red box).

We identify also a new phase transition, occurring in the polymer globular phase, where the binders undergo an order-disorder transition, despite that they do not interact directly with each other. Two states exist: 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. The phase diagram is represented in **Fig. 2.4**. Such thermodynamic stable states are expected to play an important role in the chromatin organization. The different nature of these configurations is visually evident in **Fig. 2.5d**, and will be discussed in detail in the next subsection.

2.4.2. The order parameters of the transitions

To identify the values of E_{Int} and c which give the transitions discussed above, we proceed as follows. The coil-globule transition is identified (Barbieri *et al.*, 2012) by studying the **gyration radius** R_g associated to the polymer, defined above by equation (7), that is essentially a measure of the average linear size of the polymer (**Fig. 2.5**). The order-disorder transition is highlighted by two quantities associated to the **configuration of the binders: the pair distribution function** $g(r)$ and the **structure factor function** $S(k)$. They are defined as follows (Allen&Tildesley, 1987):

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

(9) $$S(k) = 1 + 4\pi\rho \int_0^\infty r^2 \sin(kr)/(kr)\, g(r) dr$$

where $\rho=N_b/V$ is the concentration of the binders attached to the polymer, δ is the Dirac delta function. The structure factor $S(k)$ is basically the Fourier transform of the pair distribution function. It is almost flat when the binders are in a disordered configuration, while it is characterized by sharp peaks when the binders are in an ordered configuration (**Fig. 2.6**). In our study, we consider as order parameter the ratio $S(k^*)/S_{MAX}$, where k^* is the position of the second peak in the $S(k)$ function and S_{MAX} is a normalization constant equal to the maximum value of $S(k^*)$ among the different studied cases, so to have a quantity normalized between 0 and 1. Such order parameter have a sharp jump at the order-disorder transition (**Fig. 2.5**). Analogous results are obtained if other peaks of $S(k)$ (for instance the first peak or the third peak) are taken.

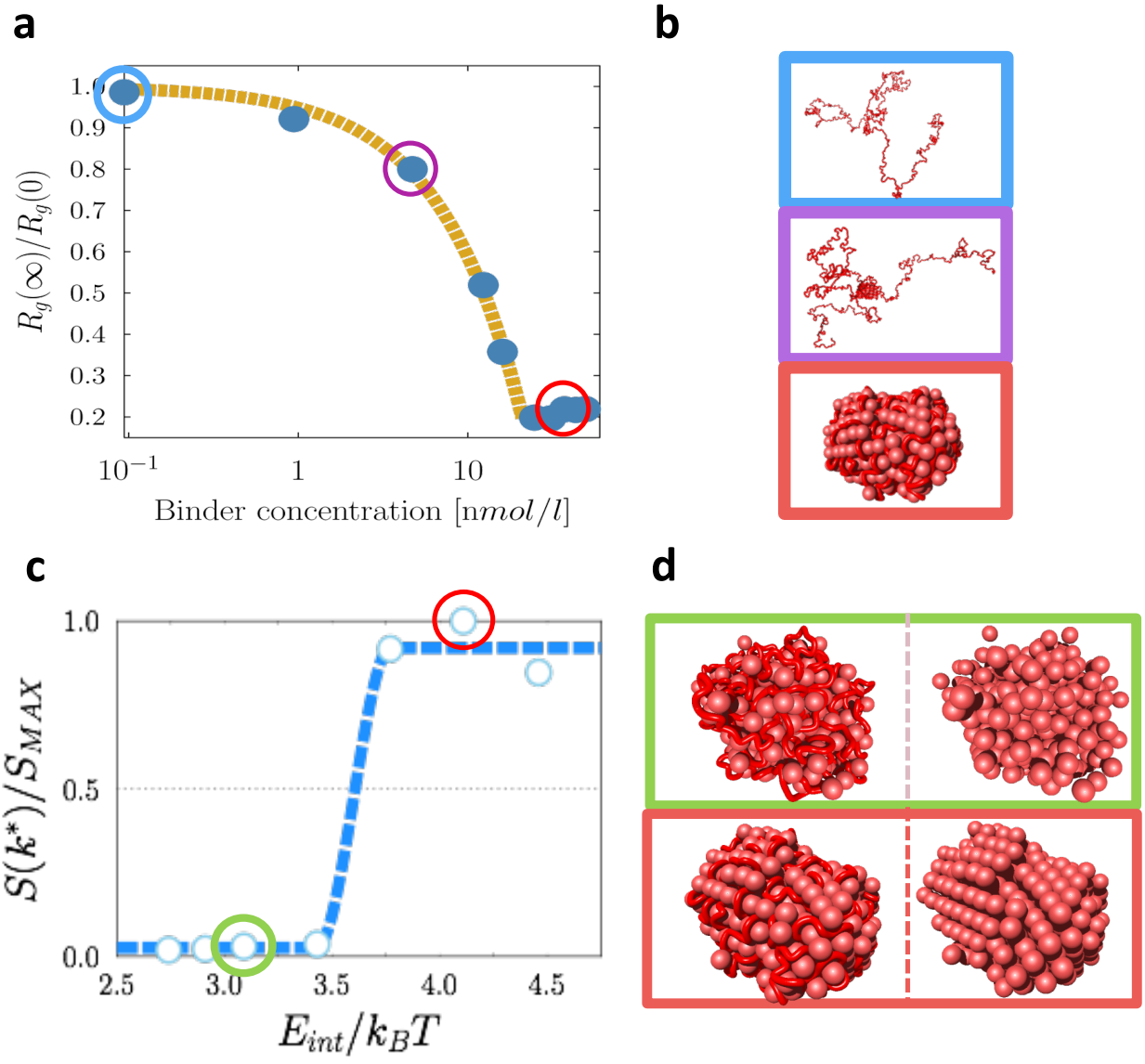

Figure 2.5: The order parameters of the transitions

a. The gyration radius of the SBS polymer, R_g, signals its coil-globule transition point as a function of the concentration of binders. **b**. Three different configurations at different concentration. **c.** The structure factor $S(k)$ peak marks the order-disorder transition in the arrangement of the binders around the folded polymer. **d.** The binders in disordered configuration (E_{int}=3.1k_BT, green box) and in an ordered configuration (E_{int}=4.1k_BT, red box). (Figure adapted from Chiariello *et al.* 2016)

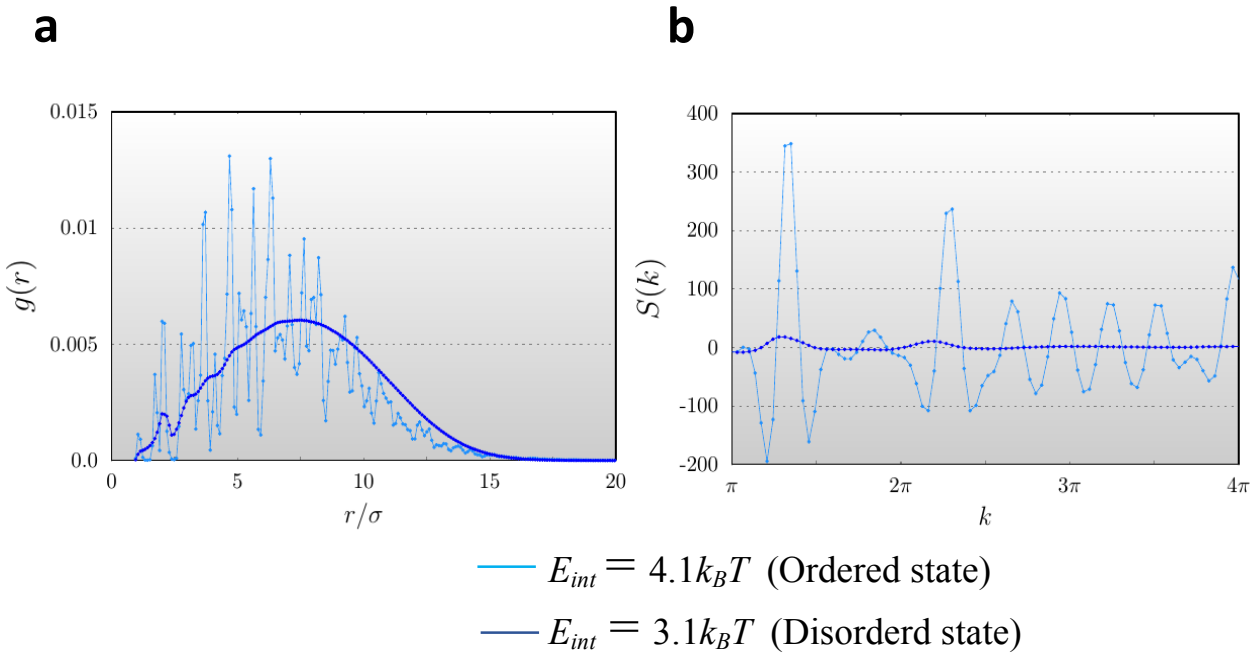

Figure 2.6: The pair function distribution *g*(*r*) and the structure function *S*(*k*)

a. An example of pair distribution function $g(r)$ defined by equation (8), in the closed state. In the disordered state (blue curve, interaction energy E_{int}=3.1k_BT) it is characterized by a smooth behavior, while in the ordered state (light blue curve, interaction energy E_{int}=4.1k_BT) it has several sharp peaks. **b.** The structure factor $S(k)$, i.e. the Fourier transform of the $g(r)$ function (eq. (9)), is practically flat for the disordered state, while it has sharp peaks in the ordered state.

2.5 Fit of the average contact probability of chromosomes

To characterize the folding state of our homo-polymer model, we calculated the average **contact probability** $P_c(s)$ of beads pairs separated by a given genomic distance s. The behavior of $P_c(s)$ depends on the state of the system (**Fig. 2.7a**). In the open state the probability decreases as a power law with s, i.e. $P_c(s) \sim s^{-\alpha}$, where the exponent α is about 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 the power law exponent equal to 0, in the latter it decreases with an observed exponent 1.0. The mean square distance between bead pairs $R^2(s)$ has a complementary behavior to the $P_c(s)$ function, as shown in **Fig. 2.7b**, so it depends on the thermodynamic state of the system. These properties are general features of this kind of systems. 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. 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. Furthermore, off-equilibrium, unstable conformations are also expected to be encountered in real chromosomal regions, for example during changes in the folding state.

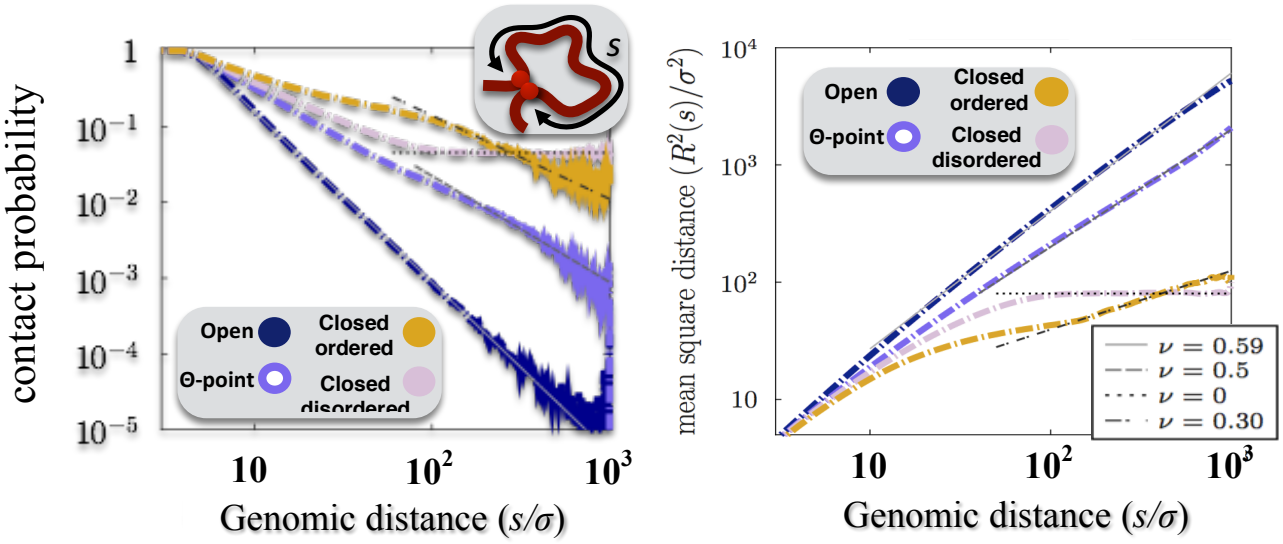

Figure 2.7: Contact probability $P_c(s)$ and mean square distance $R^2(s)$

a. The SBS homo-polymer contact probability as a function of the contour distance *s* (i.e. genomic distance), in its different thermodynamic phases. **b.** The mean square distance of two generic sites having a contour distance *s* along the polymer. (Figure adapted from Chiariello *et al.* 2016)

To compare our very simple model with the Hi-C experiments, we considered that a single chromosome is likely to be a mixture of differently folded regions, with some regions more compact than others, like euchromatin and heterochromatin (see **Chapter 1**). The regions can change their conformation from cell to cell according to functional purposes (Nagano *et al.*, 2013). At a first approximation, the conformation of each region must belong to the stable thermodynamic states (Nicodemi *et al.*, 2009; Barbieri *et al.*, 2012) previously identified (**Fig 2.4**), as schematically represented in **Fig. 2.8a**. So, in a simple coarse-grained approach where chromatin is approximately a homo-polymer, the model $P_c(s)$ is simply a linear combination of the different contact probability profiles represented in **Fig. 2.7a**. 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, by LSM we find the composition of the mixture of open and closed states that minimize the distance between the predicted $P_C(s)$ and the one derived from Hi-C data. Interestingly, 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. This results is valid for genome wide averaged data (**Fig. 2.8b**) and for single chromosomes data (**Fig. 2.8c**). 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 (hESC, from Dixon *et al.*, 2012), the open state is approximately 75%, while in the differentiated cells as IMR90 fibroblast (data

from Dixon *et al.*, 2012), this value is approximately 50%, in agreement with expectations. If we consider contact probabilities extracted from data obtained from different experimental techniques (Hi-C vs TCC, data from Dekker and from Kalhor *et al.* respectively), 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 **Fig. 2.8e**, for IMR90 cell type. For instance, chromosome X is very compact with a 75% of closed ordered state, while chromosome 1 has a 50% of open state. 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 (always less than 20%).

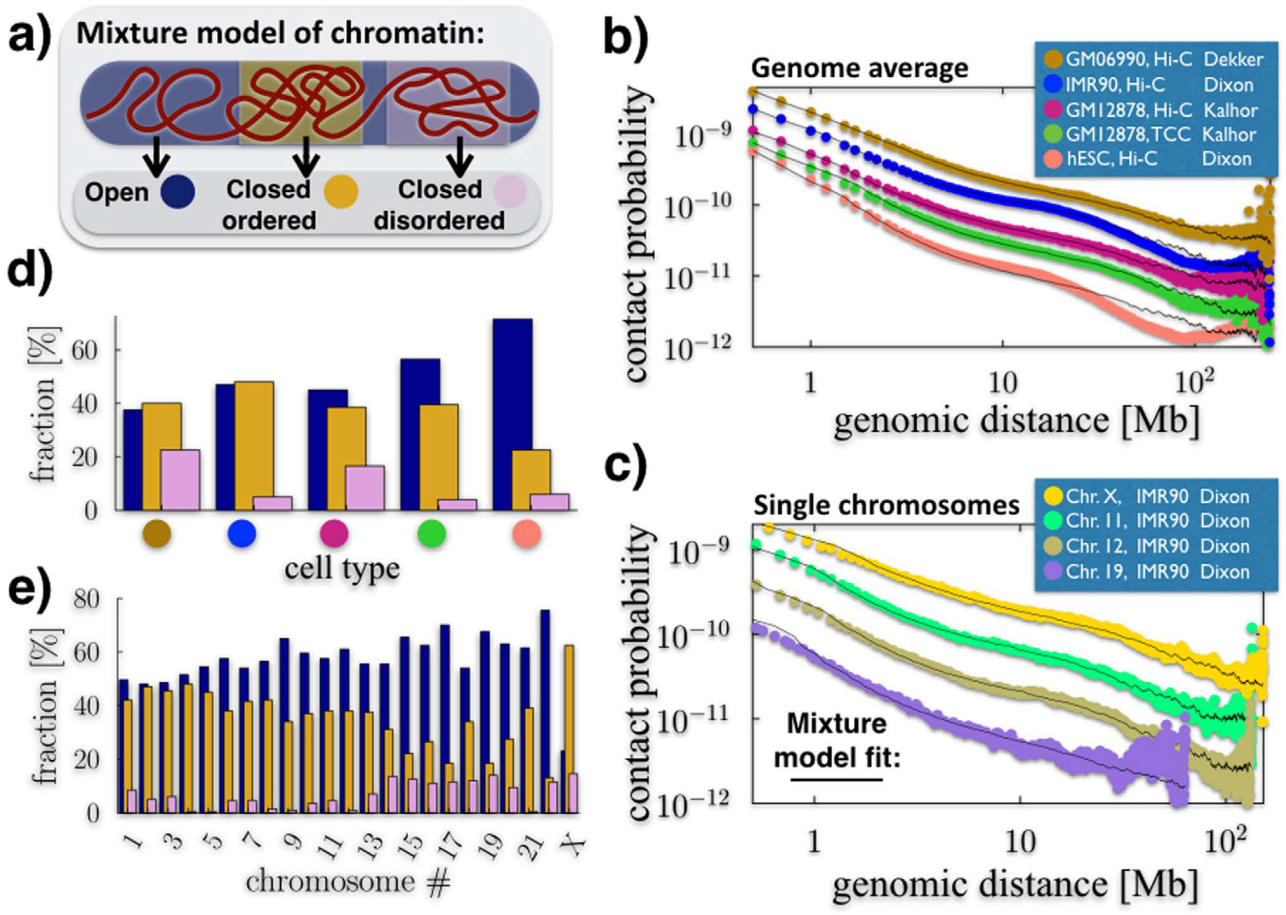

Figure 2.8: Chromatin is a mixture of regions folded in different thermodynamic states.

a. We model a chromatin filament as a mixture of differently folded regions, each belonging to one of the stable conformational classes. In this view, the average pairwise contact probability is only determined by the relative abundances of the states in the mixture, as

each state has a fixed, specific pairwise contact probability. **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 from IMR90 cells) can be similarly explained. **d.** Different cell types have a different chromatin composition, with hESC (orange circle) more open than differentiated cells, such as IMR90 (blue circle). **e.** Within a given cell type (here IMR90, as in panel c) 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)

2.6 The SBS model reproduces TADs and higher-order structures

2.6.1. Two colors polymer models

The model just discussed have one kind of bead that can interact with all the binders floating in the surrounding environment. Despite its simplicity, it recapitulates the average long-range contact properties of chromosomes. Nevertheless, the real Hi-C matrices have a very complex structure (see previous Chapters 1), and it is necessary to complicate the model to further investigate the patterns of the experimental data beyond the average long-range contact probability. To this aim, we now consider a block-copolymer, with two types of beads (visually represented by two colors, red and green), that can interact only with a specific kind of binder (red and green respectively). We consider as first case a 2-block co-polymer where each block is made of 500 beads, one red and one green, and the entire polymer is made of 1000 beads in total (**Fig. 2.9a**). To give a sense of length scales, we consider scales one order of magnitude lower than the chromosome modeling, which are typical genomic lengths where chromatin is known to be subjected to compartmentalization (Lieberman-Aiden *et al.*, 2009). Thus, 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 σ=64nm. The time step results to be 0.003s, assuming a viscosity of 2.5cP. 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, that can be interpreted as TADs. These objects correspond to enriched interaction squares along the diagonal of the contact matrix (**Fig. 2.9a**). The contact probability $P_c(s)$ and the average square distance $R^2(s)$ associated to this conformation are reported in **Fig. 2.9b**. It is apparent the crossover at the domain boundary, where the $P_c(s)$ has a sharp drop at s=N/2, and complementary, $R^2(s)$ reaches the maximum value as a plateau. In the second block co-polymer, the distribution of the colors along the polymer consists of four consecutive blocks (red-green-red-green, **Fig. 2.10a**), each block 250 beads long (as before, the total polymer is composed by 1000 beads). 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. (**Fig. 2.10a**, central matrix). When equilibrium is completely reached, the blocks of the same color interact, and the

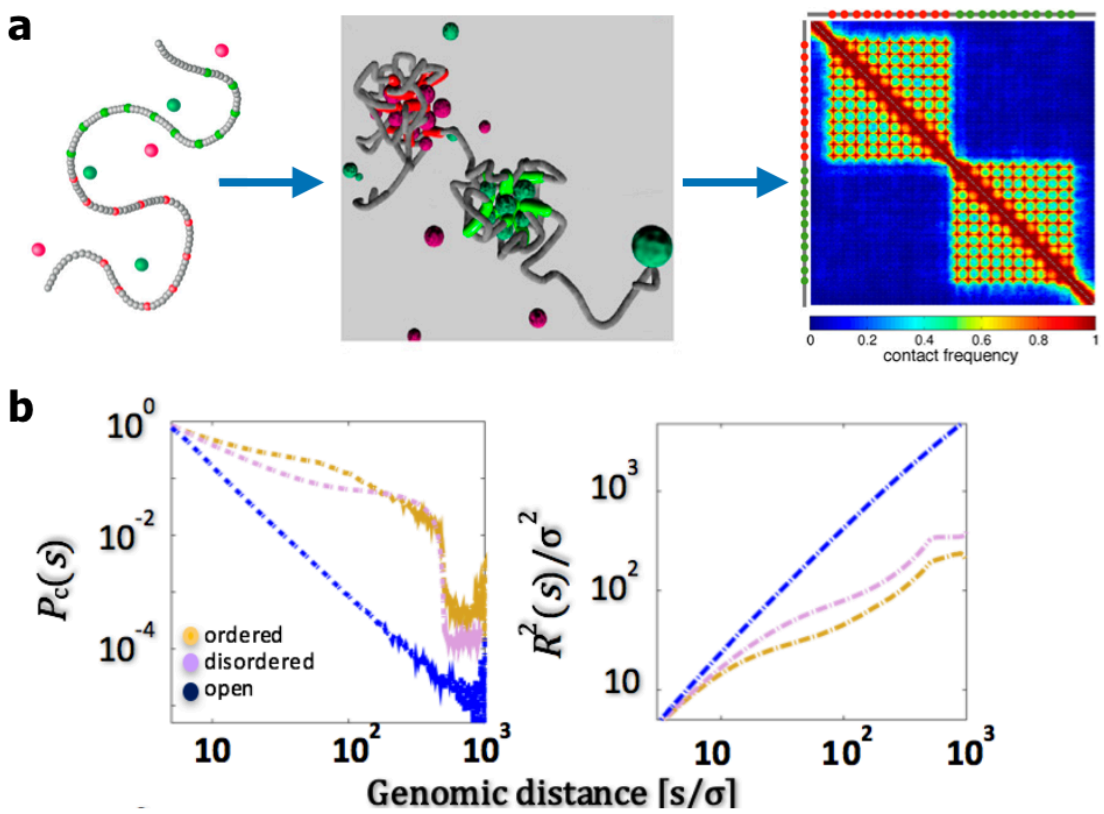

Figure 2.9: Formation of domains

a. The folding of a block co-polymer model made of two types of beads (red and green), interacting with two types of binders, bring to the formation of chromatin domains resembling TADs. (Figure adapted from Barbieri *et al.* 2012) **b.** Contact probability (left plot) and quadratic distance (right plot) in the 2-block co-polymer, for the three thermodynamic stable phases. (Figure adapted from Annunziatella *et al.* 2016).

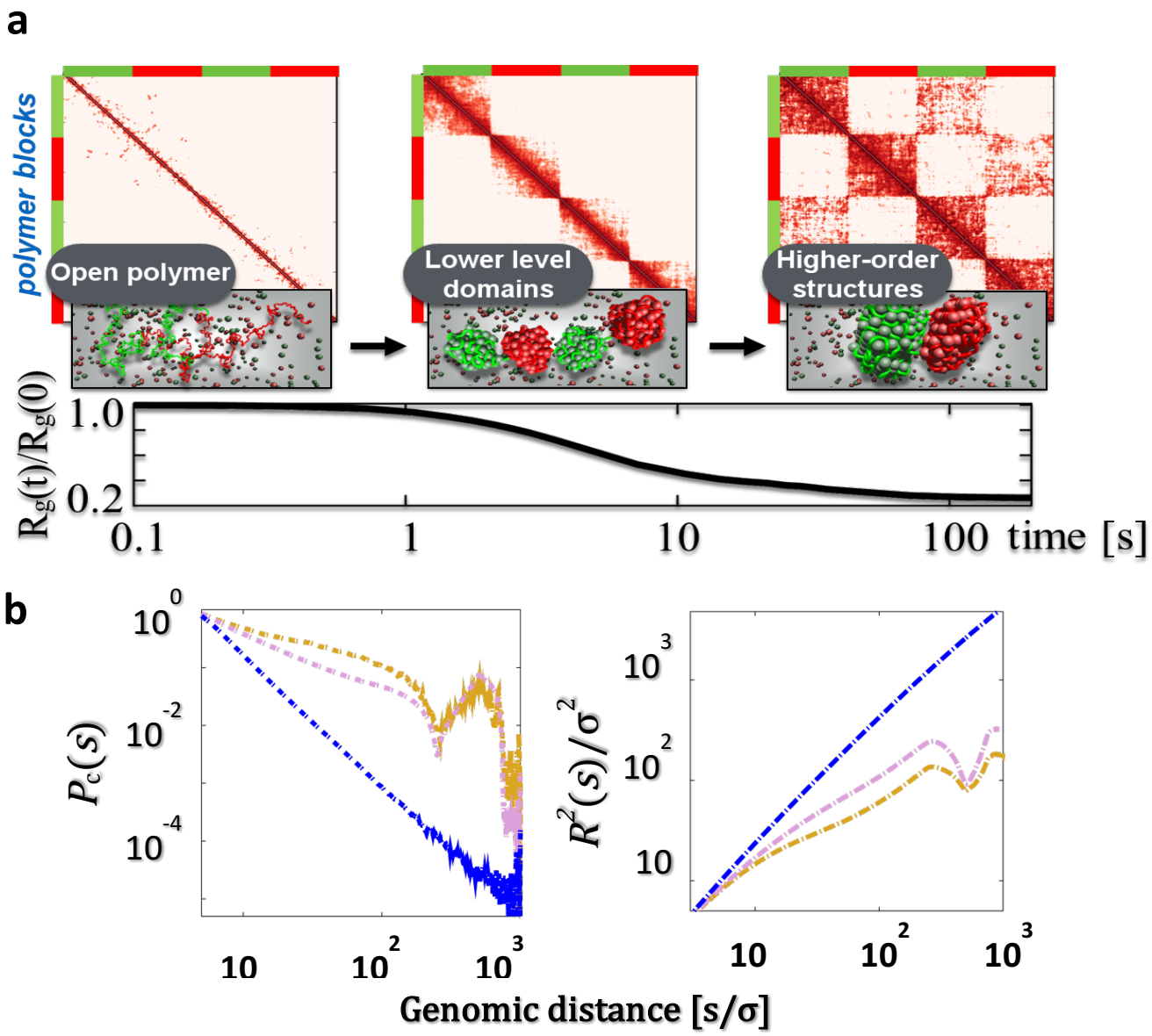

Figure 2.10: Hierarchical self-assembly of domains

a. Block co-polymer model made of four consecutive blocks alternating their color (green-red-green-red). The dynamics of the systems is marked by a decrease of the gyration radius, and a hierarchical self-assembly of domains spontaneously occur, as in the corresponding contact matrices (here E_{int}=4.1k_BT). (Figure adapted from Chiariello *et al.* 2016). **b.** Contact probability (left plot) and quadratic distance (right plot) in the 4-block co-polymer. (Figure adapted from Annunziatella *et al.* 2016).

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 (**Fig. 2.10a**, right matrix), such organization is represented by a chessboard-like pattern. $P_c(s)$ and $R^2(s)$

are reported in **Fig. 2.10b**, and they reflect the information contained in the matrix. In fact, we register a sharp drop at s=N/4, then it increases since there are higher order interactions, and then it sharply drops again at s=3N/4. In the framework of our model, such structural features naturally emerge by specialization of the involved molecular factors under the laws of polymer physics.

An interesting consequence of the self-assembly of domains, that probably can have functional roles, is 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 differently located with respect to the domain itself. Specifically, we consider two pairs of loci having the same genomic distance (i.e. the same contour distance along the polymer). We focus on two cases where the sites can be symmetrically or asymmetrically located with respect to the boundary of the domains. We find that the symmetrical pair have on average a larger spatial distance than the asymmetrical one, while in the open state (i.e. the SAW) no difference is observed. This is found for both closed states (ordered and disordered), as shown in **Fig. 2.11**. The details of the distances and contact matrices computation are given in the next subsection.

2.6.2. Computation of distance distributions

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=125σ. 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). To increase the statistics, we consider also the case where B is located at 5σ from the boundary and A at 120σ. The asymmetric distribution plotted in **Fig. 2.11** is an average of the two cases. The result is valid also if the distance from the boundary is higher (we checked the case 25σ from the boundary and similar results are found).

2.6.3. Computation of contact matrices

The polymer average pairwise contact frequency matrices for all polymer models discussed above are obtained in the following way. We fix a contact threshold distance Aσ, where σ is the length unit, and A is a dimensionless constant threshold, which we set to A=3.5. For a given 3D conformation of the polymer chain, we consider the distance r_{ij} between each bead pair i and j, (i≠j, where i and j are bead indices along the chain). If $r_{ij} < A\sigma$, 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.

The mean contact probability, $P_c(s)$, of a pair of polymer beads having a contour separation, *s* (genomic distance) is recorded in an analogous way by averaging also over all the bead pairs with the same given contour distance.

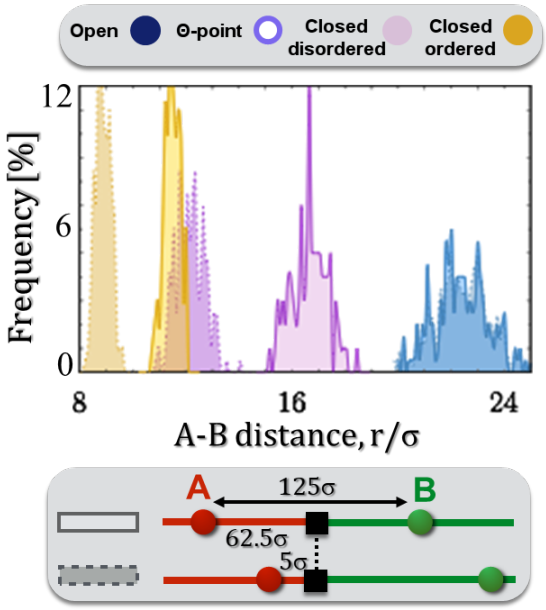

Figure 2.11: Symmetry-breaking mechanism in the physical distances

The physical distance distribution r (in dimensionless σ units) of a pair of sites having the same contour distance (here 125σ), differently located with respect to the TAD boundary position. In the open phase, as expected no difference is observed (blue distributions). Yet, in the closed phases (ordered, in yellow, and disordered, in purple) the symmetry is broken, and the loci with asymmetric positions (dashed line distributions) are closer in space than the symmetric pair (solid line distributions). (Figure adapted from Chiariello *et al.* 2016).

2.7. Multiple contacts interaction landscape

In this section, we discuss the many-body contacts, i.e. the probability of co-localization events of multiple sites. This is essentially a generalization of the pairwise contact interaction profile described in the previous section, where the dimension and the complexity of the interaction event is increased. To investigate this aspect of the polymer architecture, we first explore in details the probability of triple contact events $P_c(s_1,s_2)$ (i.e. triplets probability), where the three beads are separated by different genomic separations s_1 and s_2. Then we compute the frequency of observing n (n>3) sites in physical contact, and we do this in the three thermodynamic states identified previously. In particular, in the closed states many-body contacts are exponentially more frequent than in the open state.

2.7.1. Computational approach for the many-body contact

To estimate the average number of many-body contacts involving simultaneous interactions of *k* beads occurring in a given polymer conformation, we count the number of beads n_i that are in contact with the *i*-esim bead within the fixed threshold A (for this computation, we use as above A=3.5), and the number of possible combinations of *k* simultaneous contacts that contain the *i*-esim bead, $\binom{n_i}{k-1}$. 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 *k* particles with the *i*-esim bead, $\binom{\mathrm{N}}{k-1}$. In **Fig. 2.12a**, we show the value of this frequency as a function of the multiplet complexity n, computed in the homopolymer case.

2.7.2. The triplet surface

The calculation described in the previous subsection gives an estimate of the many-body contact average probability. A more accurate calculation is made for the computation of the multiple contact profile when the complexity n of the multiplet is 3 (i.e. the triplets). As in the pairwise contacts probability the mathematical object is a 1-dimensional curve (**Fig. 2.4**) and the parameter is the genomic distance *s*, here we need a 2-dimensional surface and the parameters are the two genomic separations s_1 and s_2 that separates the first bead and the second bead, and the second with the third bead, as schematically represented in **Fig. 2.12b**, bottom part. As expected, the surface is symmetric, and in the particular case where s_1 or s_2 equal to zero (i.e. two beads coincide) we recover the pairwise contact profile.

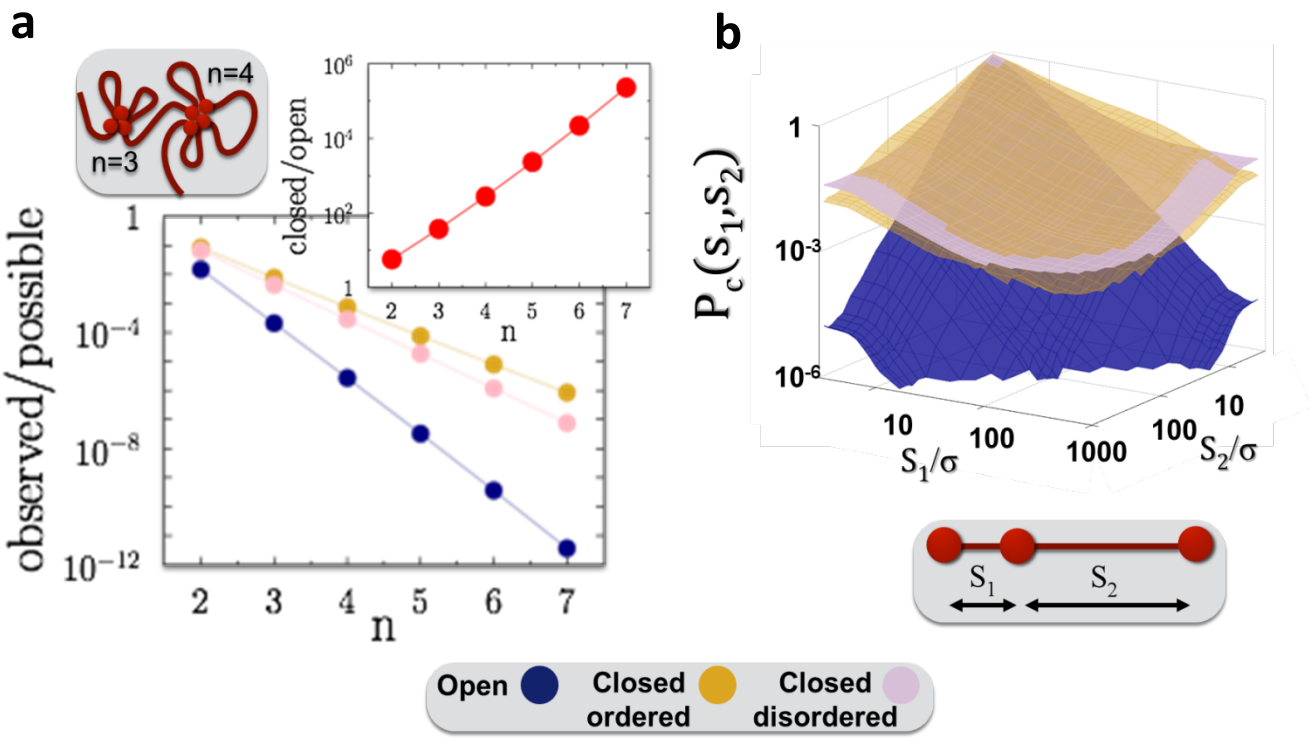

Figure 2.12: The theoretical multiple contact interaction profile

a. 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 top-left inset shows the ratio in the compact-disordered and open states. **b.** The plot shows the contact probability of bead triplets at different contour separations, $P_c(s_1, s_2)$, along the SBS homo-polymer in its different thermodynamics phases. (Figure from Chiariello *et al.* 2016)

2.7.3. Importance of multiple contacts

Multiple interactions cannot be detected by Hi-C, yet our model highlights that they are likely to be an abundant structural component of chromatin, as is emerging from new researches in the field (Olivares-Chauvet *et al.*, 2016, Beagrie *et al.*, 2017). 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. Taken together our results support a view whereby basic mechanism of polymer folding could play key functional roles in the regulation of the genome by controlling the spatial organization of chromatin.

3. Polymer models of specific DNA loci

3.1. Introduction

A **locus** (plural **loci**) in genetics is any fixed portion of DNA along a chromosome. In this Chapter, we will show how the SBS model can be used to describe and reconstruct the 3D architecture of real loci in the DNA. In the first part of the previous chapter we showed how with a very simple model (homo-polymer model), polymer physics is able to recapitulate with a good degree of accuracy the average behavior of the chromosome structure in a wide range of genomic lengths (from the sub-Mb scale up to the whole chromosome scale). Next, we introduced an extended model, with the introduction of a second type of bead (the red-green polymer models), in order to explain other aspects of the chromatin architecture and to highlight mechanisms that could have important functional roles: the existence of TAD domains, the symmetry-breaking in the distance distribution, and the hierarchical structure contained in the experimental Hi-C contact matrices, occurring in a spontaneous self-assembly process. In this Chapter, we will see how the SBS model can capture the finer spatial structure of specific regions of the genome. To this aim, we generalize the SBS model by introducing a multicolor polymer, where each color can interact only with its cognate type of binder (**Fig. 3.1a**). Based on this generalized SBS polymer model, we developed a Simulated Annealing Monte Carlo optimisation procedure, named **PRISMR (Polymer-based Recursive Statistical Inference Method**), 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 and no additional or tunable parameters. The PRISMR algorithm takes as input an experimental Hi-C matrix, and gives as output the optimal SBS polymer model of the corresponding genomic locus, with the minimum number and types (“colors”) of required binding sites (**Fig. 3.1b**).

The details of the PRISMR procedure will be described in Section 3.2. Although PRISMR polymer models are derived from Hi-C pairwise contacts, they can be used to derive any further aspect of folding, that cannot be directly obtained from Hi-C, such as the ensemble of 3D conformations that the given locus assumes, its higher-order contacts, or the physical distances of genes and regulatory regions. As a first application, in Section 3.3 we

will present the results about the modeling of the *Epha4* locus, which is a key developmental region associated with different types of limb malformations (Lupiáñez et al. Cell 2015). In Section 3.4 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). In Section 3.5 we will model the *7q* locus, which is important for Neurogenetics applications. Finally, in Section 3.6 we will discuss the general validity of the method. In the next Chapter, we will use the results of Section 3.2 on *Epha4*, to discuss another, fundamental, application of PRISMR polymer models, that is the prediction of the effect of genomic rearrangements on loci folding.

Most of the material presented in this Chapter, including figures, paragraphs and sentences, is adapted or taken literally from the following papers, which I coauthored: Bianco *et al., Submitted* 2017 (Sections 3.2, 3.3); Chiariello *et al.*, 2016 (Section 3.4); Chiariello *et al.*, 2017 (Section 3.5).

3.2. The PRISMR method

For sake of definiteness, we focus on the case where a chromatin locus filament is modeled with the SBS polymer model, but the method can be straightforwardly generalized to other cases. In the generalized SBS (**Fig. 3.1a**), as mentioned above, 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 "excluded volume" effects. We name *n* the number of different colors allowed in the model.

Our **PRISMR (Polymer-based Recursive Statistical Inference Method)** algorithm 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 an input contact matrix of a given chromosomal locus (**Fig. 3.1b**). PRISMR is based on a standard **Simulated Annealing Monte Carlo (SA)** optimization procedure that minimizes the distance between the predicted polymer model and the Hi-C contact matrix, under a Bayesian weighting factor to avoid overfitting.

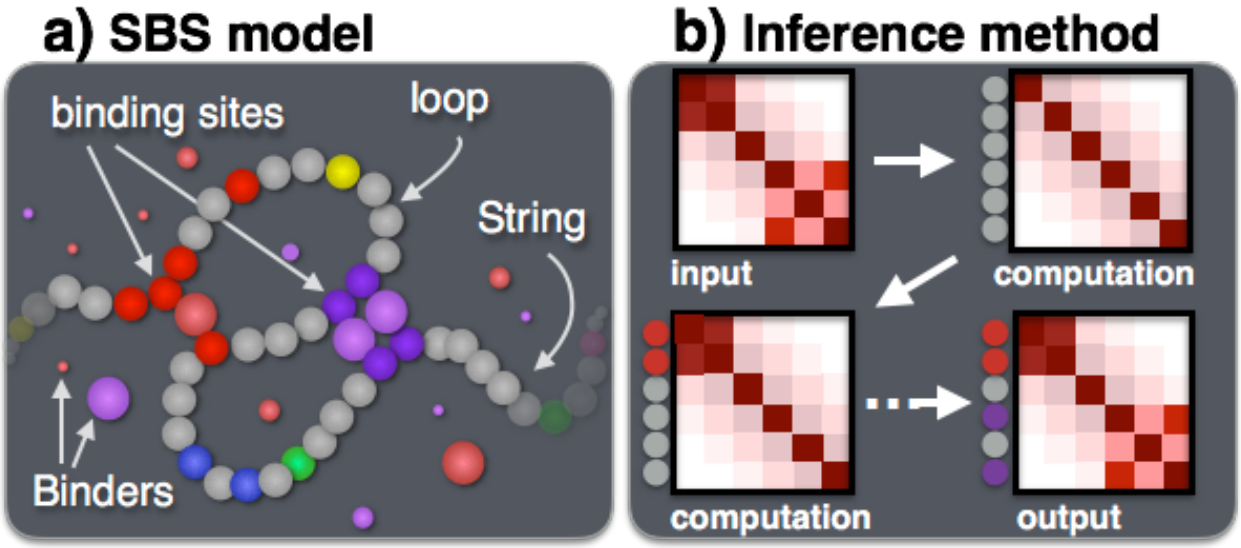

Figure 3.1: The PRISMR method.

a. In the generalized SBS model, different types of binding sites (and their cognate bridging molecules) are visualized by different colors. **b.** PRISMR samples the thermodynamics ensemble of the conformations of a given SBS model to derive its contact matrix from polymer physics. By Simulated Annealing, it iteratively finds the model that best describes the input contact matrix. (Figure adapted from Bianco *et al.* Submitted (2017))

For a given experimental genomic resolution of the contact matrix of the DNA region of interest, such as a Hi-C pairwise matrix, the locus is divided in bins. Below, we name L their total number. For instance, a 6Mb locus at a 10kb resolution is partitioned in L=600 bins. As a single DNA bin along the locus could include many binding sites, we consider a polymer chain that has a resolution r times larger than the bins of the corresponding Hi-C data. 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 (see below). In the practical cases discussed here, the number of binding sites within a bead is typically smaller than the total number of different types of binders, hence as we show below r=n is a safe assumption. For simplicity, in the description of the algorithm we consider the case with r=n.

A given SBS polymer model is fully assigned by the arrangement of binding sites along the chain and by the sequence of their types (colors). A given arrangement of beads along the polymer chain, i.e., a given SBS model, is characterized by the set of its “color” variables

c_i = 0, 1, ..., n (0 corresponding to gray, inert sites), where i = 1,...,N is an index identifying the *i-th* bead.

PRISMR aims to find the color arrangement $\{c_i\}$ of the chain, i.e., the polymer model that at equilibrium best describes the input contact matrix under a given **cost function**, H. In the approach considered here, H includes two terms accounting for two main requirements, the need to fit well the data and the attempt to avoid overfitting:

a) one term, H_0, considers how far is the input Hi-C matrix, $C_{exp}(i,j)$, from the one, $C(i,j)$, predicted by polymer physics thermodynamics given the arrangement of binding sites on the polymer, $\{c_i\}$. H_0 is normalized to the average contact frequency and to the total number of sites. A constant scale factor F is used to map the total counts in Hi-C matrix data, $C_{exp}(i,j)$, onto the derived physical contact frequencies of loci in our 3D models, $C(i,j)$; the value of F is returned by the optimization algorithm itself.

b) another component is a Bayesian term (a chemical potential in Statistical Mechanics), $H_{\lambda,}$ that penalizes the addition of new interacting beads to avoid over-fitting. H_λ is proportional, through a parameter λ, to the total number of binding sites in the polymer model and it is normalized to the total number of beads of the polymer chain, N.

In our approach the weight of the Bayesian term is given by a factor λ: if λ=0 there is no Bayesian term, the larger λ the stronger its impact, because as λ grows, the value of the cost function deteriorates. In the λ=0 case, PRISMR would find the best arrangements of bead colors to minimize the difference between the experimental and the SBS predicted contact frequencies, irrespective of noise and experimental error. A non zero λ is used to avoid overfitting, i.e., to select the minimum number of interacting beads required to explain the input matrix within a given statistical accuracy, as described below. For a given color arrangement $\{c_i\}$ of the chain, the model average contact frequency, $C(i,j)$, between any bead pair i and j, is derived within PRISMR from polymer physics by sampling the thermodynamics ensemble of allowed 3D conformations of the SBS polymer, in its different thermodynamics conformational states (see below).

For a given value of the two parameters n and λ, PRISMR samples the space of all allowed color arrangements, $\{c_i\}$, of the chain in order to find the one, named $\{c_i\}_m$, that minimizes the above cost function:

$\{c_i\}_m = \text{argmin}_{\{ci\}} [H(\{c_i\})]$

To find $\{c_i\}_m$ in the huge space of all possible color arrangements on the chain (which has $(n+1)^N$ elements), our approach employs a standard Simulated Annealing (SA) iterative procedure (MacCallum *et al.* 2015, Parisi G. 1998) whereby the color of a randomly chosen bead is changed at random, the average contact matrix of the new polymer computed and the cost function weighted until convergence is approached (see Section 3.2.3 on the SA procedure).

Finding $\{c_i\}_m$ is only an intermediate output of the method. The procedure is repeated to search for the minimal allowed value of n, n^*, and then for the maximum of λ, λ^*, required to fit the data within a predefined accuracy (see Section 3.2.2. on the SA procedure). The best color arrangement $\{c_i\}_m$, here named $\{c_i\}^*$, corresponding to n^* and λ^* is the final output of the algorithm, returning the minimal required number of binding domains (colors) and their best positioning along the SBS polymer to explain the input data within the predefined accuracy (**Fig. 3.4a**).

A computationally demanding step of PRISMR is the calculation of the equilibrium thermodynamics average contact frequency, *C(i,j)*, for the sites of a given polymer model, i.e., a polymer with a given color arrangement, $\{c_i\}$. That can be achieved, for instance, by Molecular Dynamics (MD) computer simulations, which may require huge computational efforts, or by enhanced folding algorithms (see, e.g., MELD, MacCallum *et al.* 2015), which albeit approximate can be much faster.

3.2.1. The Mean-Field approximation

To speed up the computation of *C(i,j)* and to make our procedure feasible over genomic scales, we considered an approximation typical of *mean-field* methods of Statistical Mechanics (Parisi G. 1998). In our approach, the average contact frequency, *C(i,j)*, 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. More specifically, under the above mean-field approximation, the contact matrix, *C(i,j)*, is approximated as follows. The contact frequency, *C(i,j)*, of two gray sites *i* and *j* (i.e., not interacting) is approximated to the average one between two non-interacting beads in the corresponding SBS homo-polymer made of N gray sites. Analogously, in case *i* and *j* are binding sites of the same color, *C(i,j)* takes the

value of the average contact frequency of the two sites in the corresponding SBS homo-polymer of interacting beads (Barbieri *et al.* 2012). In particular, as envisaged by polymer thermodynamics (Barbieri *et al.* 2012, de Gennes P. G. 1979), the interacting SBS homopolymer can fold in two conformational classes, corresponding to the coil and to the globule thermodynamics state. PRISMR also finds the optimal value of the fraction, f, of loci folded in the coil state across the sampled cell population (Barbieri *et al.* 2012). For clarity of presentation, below we illustrate the case f=0. Finally, in case *i* and *j* belong to different color types, *C(i,j)* is set to the value of the interaction frequencies between beads belonging to different types in different domains in the SBS model (Barbieri *et al.* 2012), which is typically found to be negligible with respect to the other cases. A limitation of the above approximation is that it neglects, in particular, correlation effects induced by the other beads on the chain. The impact of such an approximation is partially moderated by the existence of TADs along chromosomes, i.e., chromatin stretches having strong local interactions, and thus a predominant color.

We tested by extensive **Molecular Dynamics (MD)** simulations that such approximation performs comparatively well. Specifically, to test the above approximation, we simulated the derived optimal model by full-scale Molecular Dynamics (MD) simulations and we find that the contact matrices obtained by MD have a correlation with the mean-field contact matrices ranging from r=0.91 to r=0.95 across the studied cases (Bianco *et. al* Submitted (2017)). The major advantage of the mean-field approximation is that the contact matrices can be calculated in a comparatively easy way, and used throughout the SA Monte Carlo procedure to make computation times feasible. Next, we can run MD simulations (i.e., full Langevin dynamics, see Chapter 2) of the optimal model found by the SA procedure to derive its contact matrix without any approximation.

3.2.2. The Simulated Annealing procedure and its parameters

To optimize the cost function H, we customized a standard Simulated Annealing (SA) procedure based on a repeated sequence of Monte Carlo sweeps at decreasing temperatures in sawtooth-like discrete steps (Kirkpatrick *et al.* 1983, Salamon& Frost 2002). Here we report the details of our SA procedure referring to the in-situ Hi-C data of the mouse ***Epha4*** **locus**, that will be discussed in the next Section 3.3, but similar results are found in the other cases. Specifically, we refer to the case where the input matrix is from in-situ Hi-C data with KR normalization in CH12-LX cells (Rao *et al.* 2014), concerning a

6Mb long region (chr1:73000000-79000000, mm9) around the *Epha4* gene at a 10kb resolution. For sake of definiteness, we illustrate the case with r=n=20 and λ=0, but we applied the same criteria discussed below to the other cases.

The algorithm that we employ is based on standard procedure used in Simulated Annealing (Kirkpatrick *et al.* 1983, Salamon& Frost 2002). The inverse temperature values, β, used in the sweeps of single SA saw-tooth steps were selected in order to have a range of acceptance rates in the Monte Carlo moves spanning from 1 (at the highest temperature) to 10^{-3} (at the lowest temperature), equally spaced on a log scale. In the SA procedure, the temperature in each saw-tooth step is hold fixed for a time $\tau(\beta)$, depending on β, long enough to ensure convergence of the cost function to its equilibrium value at that temperature. For a given β, we measure the number of sweeps, $\tau_0(\beta)$, required to decrease the cost function 95% of its asymptotic spanned range. In the cases studied here, $\tau_0(\beta)$ ranges up to 10^2 Monte Carlo sweeps. The scale of $\tau(\beta)$ is chosen 10 times larger than $\tau_0(\beta)$, except at the highest temperature where $\tau = \tau_0$, as the acceptance rate is 1, and at the lowest temperature where $\tau = 50\tau_0$ to most accurately locate the local minimum of H. The values of β and $\tau(\beta)$ for r=n=20 and λ=0 are identified by very general criteria and, in practical terms, we find that they can be safely employed in the other cases we dealt with, as no major changes are found. The plot in **Fig. 3.2** shows how the convergence of the cost function is achieved during a saw-tooth run of our SA procedure. Finally, the SA saw-teeth steps are iteratively repeated until the changes in the estimated optimal value of the cost function are within 0.1%. Additionally, we considered the impact of the initial values on the SA algorithm outcome. To this aim, we employed the standard strategy in Simulated Annealing that is to consider a number of runs from independent initial states. Thus, the entire optimization procedure is repeated over up to $5x10^2$ independent runs from different initial random states, and their absolute minimum used as output. Importantly, we find that the final values of the cost function, obtained for the different initial conditions, differ from each other of less than 0.5%, with the best 10 minima differing less than 0.1%, so that the algorithm outcome is largely independent from the initial conditions, i.e. different runs, considered.

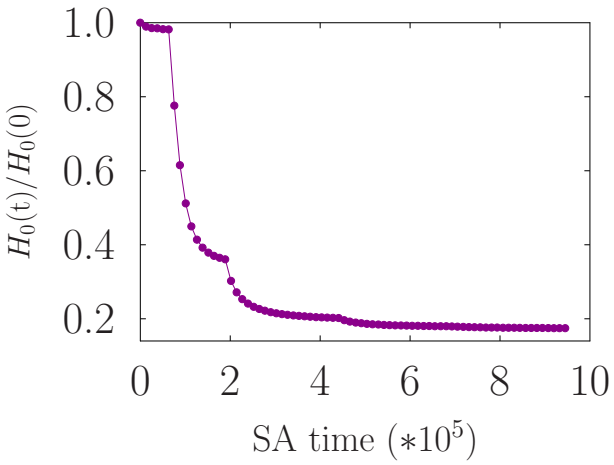

Figure 3.2: Convergence of our SA algorithm during a saw-tooth run.

The convergence of the cost function during a single saw-tooth run of our SA algorithm to its asymptotic value is shown. The different visible steps in the plot correspond to the different SA temperatures sampled by the algorithm (from very high, initial region, to almost zero, final step). The plot shows that a stable minimum is approached well within the time scales of our simulations. (Figure from Bianco *et al.* Submitted (2017))

To find the minimal number of colors, n^*, required to describe the input contact matrix, we evaluate the cost function, $H_0(n)$, as a function of n with our SA procedure. As expected, $H_0(n)$ decreases with n, towards an asymptotic plateau (**Fig. 3.3a**). We choose n^* as the value where $H_0(n)$ has decreased below 90% of its spanned range. Analogously, next we find λ^* by computing H_0 v.s. λ. H_0 increases with λ, and λ^* is the value where it has increased above 10% of its initial plateau (**Fig. 3.3b**). Finally, the minimal required value of r, r^*, is determined in an analogous way by finding where H_0 v.s. r decreases below 90% of its spanned range (**Fig. 3.3c**). As r^* is smaller than n^* in all the cases considered here, for simplicity we set $r^*=n^*$. In the case of the in-situ Hi-C data on the *Epha4* locus in CH12-LX cells (Rao *et al.* 2014), the above procedure returns $n^*=21$ and $\lambda^*=1.0$. In the following, unless otherwise stated, we simplify the notation and n^* and r^* are renamed n and r.

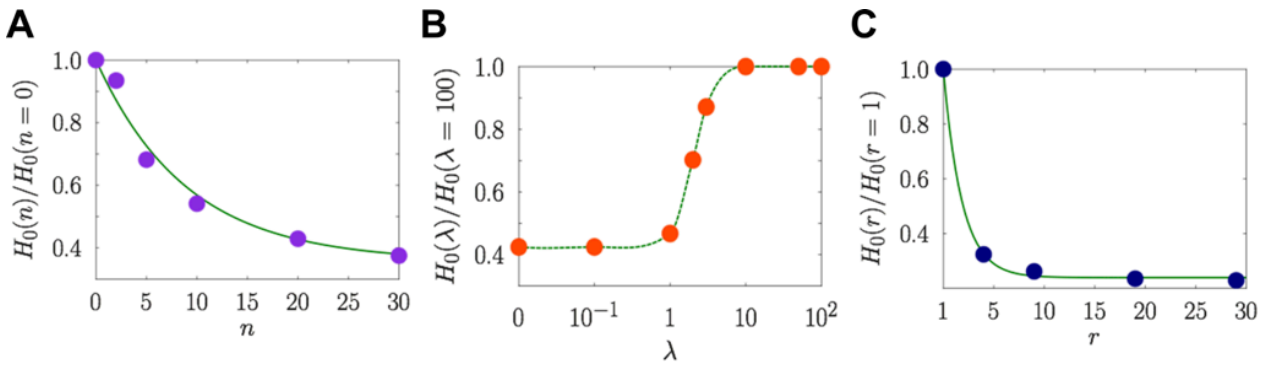

Figure 3.3: Determination of PRISMR parameters.

a. The decay of the minimum of the cost function, H_0, in our SA algorithm as a function of the total allowed number of different types of binding sites, *n*, in the studied murine *Epha4* locus (case shown λ=0). **b.** The increase of the minimum of the cost function, H_0, in our SA algorithm as a function of the cost to add an additional binding site, λ, in the studied locus (case shown *n*=21). **c.** The decay of the minimum of the cost function, H_0, in our SA algorithm as a function of the resolution of a single polymer bead, *r*, in the studied locus (case shown λ=1, *n*=21). (Figure from Bianco *et al.* Submitted (2017))

3.3. Modeling of the *Epha4* locus in mouse CH12-LX cells

As a first application of the model we consider the *Epha4* locus employed above to illustare the details of the SA procedure. Specifically, we consider a 6Mb long DNA sequence around the *Epha4* gene (chr1:73000000-79000000, mm9) in mouse CH12-LX cells. *Epha4* is a key developmental region associated with different types of limb malformations. A variety of genomic rearrangements, like deletions, inversions and duplications cause distinct phenotypes (brachydactyly, syndactyly, polydactyly) by altering the chromatin organization of the locus, thereby causing rewiring of enhancer-promoter contacts and gene misexpression (Lupiáñez *et al.* Cell 2015). The results obtained here about the *Epha4* modeling, will be used in the next Chapter to test the capability of our polymer models to predict the effect of genomic rearrangements.

The Hi-C datsets used to infer the polymer model are published from Rao *et al.*, 2014, in mouse CH12-LX cells, at 10kb resolution, and are shown in **Fig. 3.4b,** upper part . The experimental data used are normalized with the KR (Knights and Ruiz) normalization, that is a standard precedure (Knight& Ruiz, 2013). A scheme of the locus is shown in **Fig. 3.4b**,

middle part. Applying the PRISMR algorithm decribed in the previous section, we identify the different "**binding domains"** of the locus, i.e., the sets of binding sites of the same type (i.e., color) determining of the folding patterns (**Fig. 3.4a**). In this case, as showed above, the model predicts n=21 different binding domains. To check the robustness of our SA procedure we showed that its best minima return comparable fits to Hi-C data and similar corresponding polymer models (p-value<1e-250, subsection 3.3.4); additionally, the optimal model is robust to changes to the number of the colors (see subsection 3.3.4). The model binding domains can be broadly divided into 2 categories (**Fig. 3.4a**). One category includes domains largely overlapping with annotated TADs (**type-I**), and another with domains extending over several TADs or even the whole analyzed region (**type-II**) (see subsection 3.3.3).

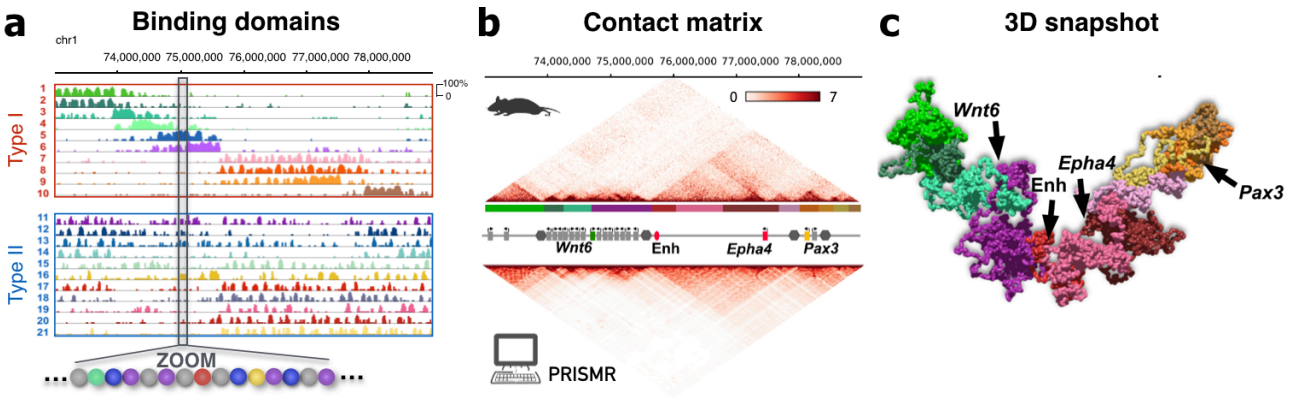

Figure 3.4: Polymer physics recapitulates 3D conformation at the *Epha4* locus

a. PRISMR identifies the different binding sites along the locus that shape its 3D structure. The binding sites of the same type (color) form the shown different *binding domains*. In case of *Epha4* locus in CH12-LX cells, there are 21 statistically significant (Wilcoxon's rank sum test; p-value <1.1e-7, see subsection 3.3.4) different binding domains. The bar plots represent their abundance (in percentage) along the genomic sequence. Type-I domains are spatially restricted whereas Type-II are ubiquitous, covering the whole region. Their overlapping genomic positions produce the observed complex interaction patterns. **b.** Published in-situ Hi-C data (Rao *et al*. 2014, top matrix) compare well with the contact matrix derived by PRISMR (bottom matrix). Their Pearson correlation, r is 0.91. In the middle, the *Epha4* locus is represented, with some important genes (*Wnt6*, *Epha4*, *Pax3*) and *Epha4* enhancer (Enh) highlighted. **c.** The shown 3D conformation is a snapshot of the model of the *Epha4* locus with the relative positions of genes and regulator highlighted. The color scheme used is reported in panel b along the linear sequence of the locus.
(Figure adapted from Bianco *et al.* Submitted (2017))

Asstated in the previous Section, once we have obtained the optimum arrangement of the binding sites along the polymer, we performed MD simulations to recostruct the 3D structure of *Epha4*. To test the accuracy of our structures, we compute, from the ensemble of configurations formed in the dinamics process, the average contact maps and compare it with the experimental data. As shown in **Fig. 3.4b**, bottom matrix, the simulated contact matix obtained is very similar to experimental data. The Pearson correlation coefficient between in-situ Hi-C data and simulated contact frequency matrix is r=0.91, proving that our model captures relevant features of the mechanisms determining the folding of *Epha4*. In **Fig. 3.2**, a snapshot of a single typical 3D configuration, obtained from the dinamics, when the polymer is fully equilibrated, i.e. when it is in the closed state. Here, we represent the relative positioning of *Epha4*, *Wnt6* and *Pax3* and the Enh regulator.

3.3.1. Simulation details

To model the 3D structure of the *Epha4* locus in CH12-LX cells, we use a chain made of N=12600 beads. Since the region to model is L=6Mb long, the corresponding genomic content per chain bead is s_0=L/N= 476bp. Furthermore, to speed up the folding of the polymer, we start the simulation with a shorter polymer made of N/3 beads, then we add the remaining beads by reducing the original bead diameter of a factor 1/3, and the other MD parameters change accordingly in order to keep the same interaction energy. The total binder concentration, *c*, was sampled in the range from zero to 250n*mol/l* and the interaction energy E_{int}=0k_BT or 8.1k_BT, corresponding to the polymer coil and globule state respectively. The dimensionless friction coefficient is set to ζ=0.5 (Kremer, K. & Grest, 1990, Chiariello et al. 2016). The MD integration time step is Δt=0.012 (Rosa & Everaers, 2008) and we let the system evolve up to 10^9 time steps, to reach stationarity. Our ensemble averages span up to $4x10^2$ independent runs for each set of system parameters. The 3D structure presented in **Fig. 3.2**, is obtained from the polymer dynamics, in the equilibrium closed phase. Mathematically, it is a smooth curve described by a third order polynomial spline passing through the centers of each polymer bead.

3.3.2. Computation of contact maps

The contact maps presented in this Chapter, are computed following the approach described in the previous Chapter, with a variant where only contacts between beads of

the same type are considered. In the case of *Epha4*, the parameter for the interaction threshold is set to A=9 (in the same notation of Chapter 2, see section 2.6.3). The same approach is used for the models presented in the following sections and chapter, where we computed the matrices with the parameter A ranging from 2 to 10, and we find similar results in all cases. In the framework of the SBS model, we consider 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). Then, we consider a mixture of the contact matrices in the coil and in the globule states, as discussed in Barbieri *et al*. (2012); Chiariello *et al*. (2016), and we seek the open-closed mixture that maximizes the Pearson correlation coefficient between model inferred and Hi-C data. For example, we find that the *Epha4* locus discussed here is made of 89% of open and 11% of closed state. As Hi-C matrices are computed from sequencing reads, a scale factor must be used in the comparison. The corresponding matrix for the *Epha4* locus is shown in Fig.3.4b.

3.3.3. Characterization of the identified binding domains

The different n^* binding domains identified by PRISMR in the best polymer $\{c_i\}^*$ are specified by the coordinates (in bases) of their binding sites along the locus. To quantify the similarity between pairs of binding domains we measured their **genomic overlap**, q. For a generic pair of binding domains (colors) k_1 and k_2, q is defined as:

$$q(k_1,k_2) = \sum_{i=1}^{L} f_i(k_1)*f_i(k_2)/\mathrm{sqrt}\left[\sum_{i=1}^{L} f_i^2(k_1)* \sum_{i=1}^{L} f_i^2(k_2)\right] \quad (10)$$

where $f_i(k_j)$ is the occurrence number of the binding sites of domain k_j in the i-th bin of the genomic sequence of length L.

We also measured the overlaps of binding domains with the locus TADs annotated in Dixon *et al.* 2012. For a given TAD, we define as above a signal f_i that is equal to 1 if the i-th bin of the polymer chain is inside the TAD and equal to 0 if not. For the TADs at the edges, which extend beyond the boundaries of the locus, we cut their coordinates at the border so to consider just the part inside the given locus.

To assign the binding domains to the **type I-II classes** described above, we used their overlaps with TADs (Dixon *et al.* 2012): specifically, a binding domain is of type I if it strongly

overlaps only one TAD or two consecutive TADs, else it is of type II. We considered as 'strong overlaps' values that exceed the median of the overlaps between all pairs of TADs and binding domains.

3.3.4. Statistical significance and robustness of the binding domains

To check the robustness of our approach and, more specifically, of the different types of binding sites and their locations along the polymer chain identified by PRISMR, we compared the minimum $\{c_i\}^*$ (i.e., the best polymer model) found for n^*=21 (and λ^*=1.0) (**Fig. 3.1a**) with the minima found when the allowed number of colors, n, is increased or decreased by 30% (i.e., with the minima for n=27 and n=15), and against a random control model. We find that the in-situ Hi-C contact matrix has a Pearson correlation coefficient equal to r=0.95 with the PRISMR predicted contact matrix in the optimal n^*=21 case. A r=0.95 correlation is also found in the case where n=27, decreasing to r=0.93 for n=15. Such a comparison supports the view that n^*=21 is a good estimate of the required number of different types of binding domains (colors) in the model of *Epha4*, as it strikes a good balance between overfitting and returning a good description of the data because comparatively higher values of n would not return significantly better correlations.

To check the level of randomness inherent to the binding domains identified by PRISMR in $\{c_i\}^*$, we compared their overlap (see overlap definition subsection 3.3.3) with each other against the expected overlap in a control random model. More specifically, we first measured the overlap, q, between different colors in the optimal output $\{c_i\}^*$, i.e., the overlap of the positions of the beads belonging to two different types of sites in the n^*=21 case. Then, we measured the positional overlaps between pairs of binding domains (different colors) in a random model obtained from the optimal configuration $\{c_i\}^*$ by bootstrapping. We found that the average overlap between domains within $\{c_i\}^*$ is q=15%, which is significantly smaller (p-value = 1.9e-130, Wilcoxon's rank sum test) than the average overlap found in the random control, q_{rand}=40% (the distribution of random overlaps has a standard deviation σ_{rand}=3%). Furthermore, the body of the distribution of the values of q extends from zero up to 35%, remaining thus below the average value of the random control case. Those results show that the binding domains identified by PRISMR are far from randomly positioned in the *Epha4* locus.

Next, to test the robustness of our results to changes in the algorithm procedure, we compared the similarity of the binding domains found for n=27 with those for n^*=21, i.e., the overlap of the colors in the two cases. Specifically, we measured the positional overlap between the beads of all the possible pairs of colors in the n=27 and n^*=21 cases. We then linked, in an exclusive way, a given color type in the n^*=21 case with the most overlapping color in the n=27 case and found that for 90% of domains the overlap is larger than $q_{rand}+2\sigma_{rand}$, spanning a range from 98% down to 41%. Similarly, the comparison of the domains identified for n=15 and n^*=21 shows that 87% of domains have an overlap larger than $q_{rand}+2\sigma_{rand}$. Hence, the color domains found in the case n=15 and n=27 are similar, in a statistically significant way (p-values = 1.1e-7 and 2.4e-13 respectively, Wilcoxon's rank sum test), to those of the optimal case n^*=21.

Finally, to check the robustness of the absolute minimum found by PRISMR, we compared the best 10 discovered minima (whose related cost functions differ less than 0.1% (see **subsection 3.2.2.**) in the optimal case n^*=21 and found that the overlap between the top 30% overlapping binding domains is above 90%, while overall the average overlap between corresponding binding domains is 65%, significantly higher than the average overlap found in the random control model (p-value<1e-250, Wilcoxon's rank sum test). Such a result shows that if a comparable fit to real Hi-C data is obtained by two different runs, then the two runs produce polymer models that are very similar to each other.

Taken together our results support the view that the optimal polymer $\{c_i\}^*$ identified by PRISMR and its binding domains are far from random and robust to changes in the parameters of the algorithm. Similar results are found in the other datasets considered here and in the next Chapter.

3.3.5. Comparison of the binding domains with epigenetic features

As an initial step to investigate the molecular nature of the factors contributing to define the different types of binding sites ('colours') envisaged by PRISMR, we derived their epigenomic barcode. In murine erythroleukemia CH12-LX cells, chromatin data are available from the ENCODE project database (ENCODE Project Consortium, 2004; Wu *et al.* 2011) that we use to characterize the binding domains identified by PRISMR. We crossed the information about their genomic positions with a number of published chromatin features, such as histone modifications and transcription factors (see Section

1.3 for some basic information about epigenetics). Specifically, for each binding domain ('colour') and for each chromatin feature we calculated the Pearson correlation coefficient between the number of binding sites of that domain at a 10kb resolution and the number of called peaks present in those 10kb wide bins (by at least a base pair) as identified by the widespread *bedtools coverage* tool (Quinlan A. R, 2014, Quinlan&Hall, 2010) (**Fig. 3.4**). Afterwards, in order to find only statistical significant correlations, we employed a random control model where Pearson correlations are computed between chromatin marks and random binding domains, obtained from ours by bootstrapping. Correlations with a specific chromatin mark are considered significant if above the 95th percentile or below the 5th percentile of the corresponding random correlations distribution. We find that single colours do not correspond to single molecular factors, as each usually correlates with a combination of different marks.

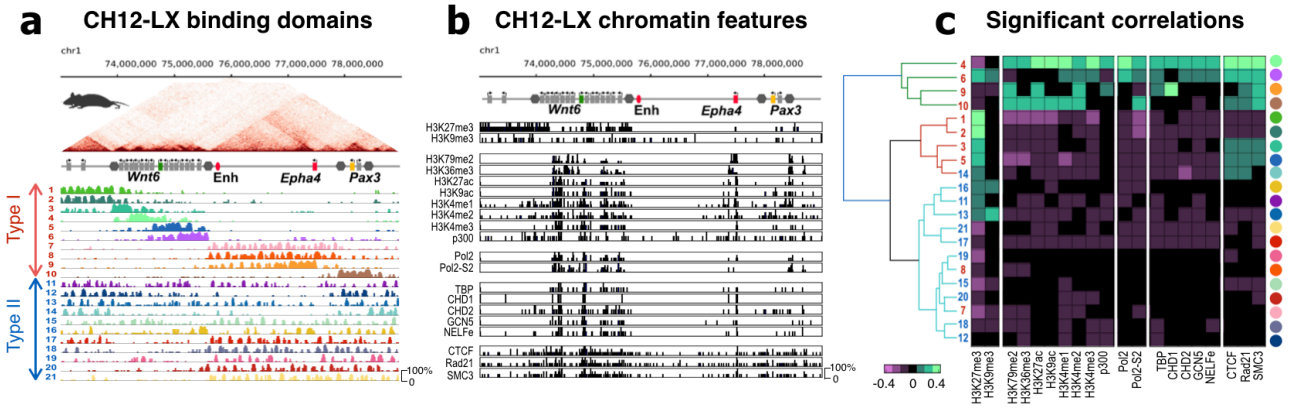

Figure 3.4: Epigenetic barcoding of the binding domains envisaged by PRISMR in the *Epha4* locus of CH12-LX cells

a. In the *Epha4* locus of CH12-LX murine cells, PRISMR envisages *n*=21 different binding domains (colors). Their genomic position and abundance are here recapitulated. **b.** ENCODE chromatin features (ENCODE Project Consortium, 2004; Wu *et al.* 2011) for the same DNA region are here listed. **c.** Matrix with the statistically significant Pearson correlation coefficients of the different binding domains of panel (a) with the ENCODE signals of panel (b). The domains have been clustered according to the similarity of their epigenetic barcode. (Figure from Bianco *et al.* Submitted (2017))

Finally, a hierarchical clustering was performed on the significant correlations matrix by using the *Python SciPy* clustering package. From the clustering analysis a nontrivial relationship emerges between binding domains and epigenetic features. For instance, we find that Type-I binding domains can be broadly subdivided in two categories linked respectively to repressive epigenetic marks (e.g., H3K27me3) and active marks (e.g., H3K4me1/2/3 and Pol-II). Many Type-I binding domains also correlate with the CTCF/Cohesin (CTCF/Rad21/Smc3) system, known to play an important role in chromatin architecture through the formation of chromatin loops (Nora *et al.* 2017, Schwarzer *et al.* 2016, Sanborn *et al*. 2015, Fudenberg *et al*. 2016). 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. This is consistent with recent experiments showing that targeted depletion of CTCF can have a minor effect on chromatin organization (Kubo *et al.* 2017). Our finding that other factors, beyond CTCF, may play a role in chromatin organization is also consistent with recent exciting developments in the literature where additional players are being identified, such as PRC1 (Kundu *et al.* 2017), MLL3/4 (Yan *et al.* 2017), Active/Poised Pol-II (Barbieri *et al.* 2017), etc. Additionally, many other colours have no significant correlation with CTCF/Cohesin. Type-II colors can be subdivided, as well, in a group correlated to H3K27me3 and in a group anti-correlated with H3K27me3. However, they are mainly characterized by lack of significant correlations with most of the other available ENCODE marks, which could point towards the existence of other, yet unidentified structurally relevant chromatin factors.

Taken together, our epigenetic analysis shows that the different types of binding sites, and their cognate binders, envisaged by PRISMR do not simply correspond to a single molecular factor associated to chromatin, but rather to combinations of different factors. It supports the view that several, structurally relevant chromatin organizers exist beyond CTCF/Cohesin, including factors yet unmapped in ENCODE, which act in combinations to induce, specify or constrain folding. This is consistent with recent developments in the literature where novel factors are being discovered linked to chromosome folding (Kubo *et al.* 2017, Kundu *et al.* 2017, Yan *et al.* 2017, Barbieri *et al.* 2017, Hug *et al.* 2017).

3.4. Modeling of the *Sox9* locus in mouse ES cells (mESCs)

In this Section, we apply the PRISMR method just described above to another locus, to test the general validity of our approach and consider it as a powerful tool to reconstruct and visualize the 3D architecture of real loci in the genome. We focus on the *Sox9* locus (chr11:109000000-115000000, mm9), a 6Mb long region around the *Sox9* gene, which is a very important locus linked to congenital diseases (Franke *et al.*, 2016), including gene rich regions and gene desert regions, as shown in **Fig. 3.5a**, upper part. The experimental Hi-C data used to infer the polymer are published from Dixon *et al.*, 2012, in mouse ESC-J1 cell line, at 40kb resolution, and are normalized as described in Yaffe&Tanay, 2011 (**Fig. 3.5b**). The PRISMR inference procedure gives in this case a total of n=15 different binding domains (colors), as represented in **Fig 3.5a**. As seen for the *Epha4* locus before, the binding domains tend to overlap with the different TADs existing in the locus, but they also overlap with each other and produce interactions between TADs, giving the hierarchical structure (metaTADs) visible in the original experimental matrix (**Fig. 3.5b**, top matrix). Our method returns a contact matrix very similar to the experimental data, with a Pearson correlation coefficient r=0.95 between model and data, as shown in **Fig. 3.5**. A snapshot of a single typical configuration, in the closed state, is shown in **Fig. 3.5c**, where the relative positioning of *Sox9* and other genes in the locus, across its different higher order domains, can be visualized. For instance, the transcription starting sites (TSSs) of the *Sox9* and *Kcnj2* genes have a genomic separation almost four times larger than the TSSs of *Sox9* and *Slc39a11* (1.72 Mb v.s. 0.46 Mb), but the average physical distances of the two pairs are proportionally closer (1.19 μ m v.s. 0.59 μ m) as the three genes belong to consecutive regional areas. The *Sox9* locus is marked by many-body contacts (see Chapter 2 Section 2.7) that are exponentially more abundant than expected in a randomly folded conformation (**Fig. 3.5d**, error bars within symbol size). As shown **Fig. 3.5e**, the self-assembly of the locus spatial structure starts from a totally random SAW initial state (open confromational class) and proceeds hierarchically, passing through early local domains folding into larger and larger domains that cover the whole locus. The same snapshot of the full 3D structure of the *Sox9* locus in **Fig. 3.5c**, along with a comparison with its average contact matrix and TADs, is shown in **Fig. 3.6** with an alternative color scheme that follows the coordinates of the original TADs identified in Dixon *et al.* 2012. This way, for example, it is visible that TAD D (red) has a complex internal 3D structure, with a large part of it mostly associated to TAD

E. Additionally, the dynamics of the interactions between the genes within the locus and their regulators can be derived. Summarizing, the variety of information on *Sox9* and its folding mechanisms that can be inferred from polymer physics extends well beyond the Hi-C pair-wise contact data used to infer the model.

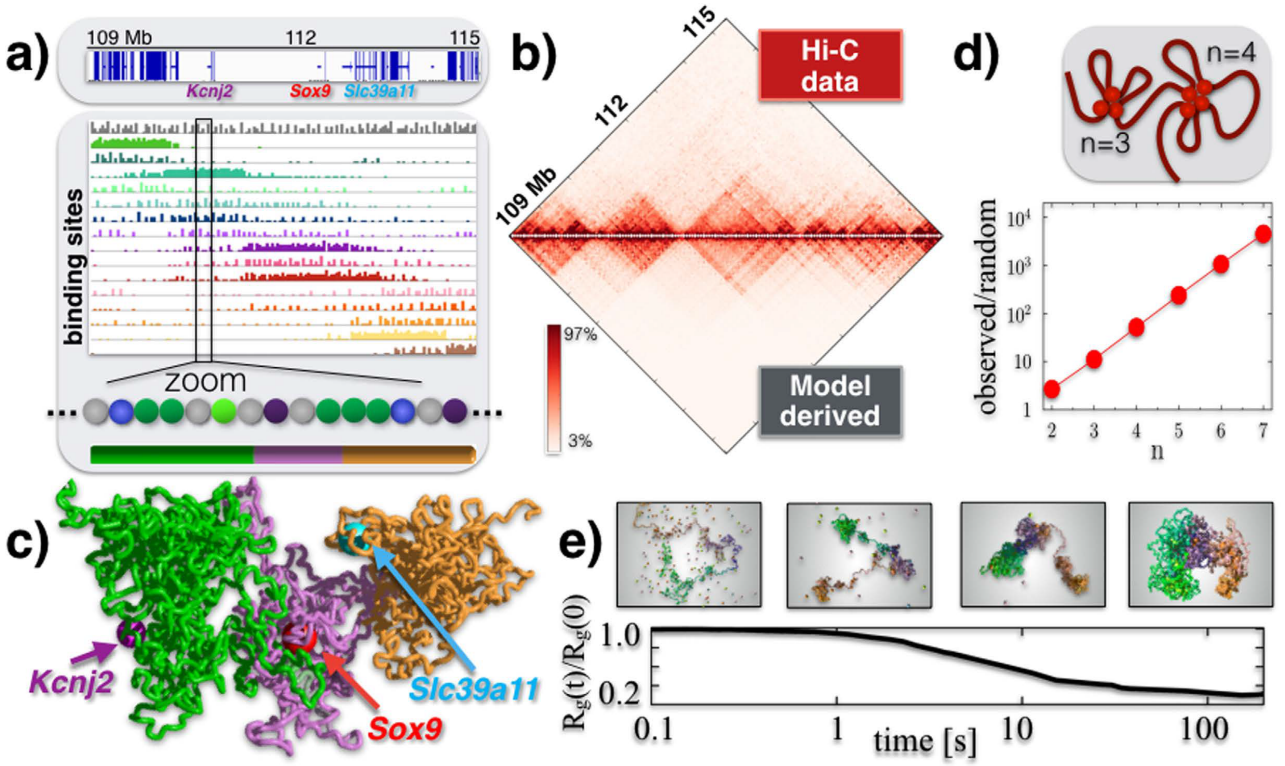

Figure 3.5: Polymer physics captures the folding of the Sox9 locus.

a. Top: the considered Sox9 locus in mESC-J1 cells, with a few marker genes. Bottom: the SBS polymer model that best explain the Hi-C contact map of the Sox9 region has the shown different types of binding sites, as seen in the zoom (different colors); 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.** The model derived pairwise contact frequency matrix (bottom) has a 95% Pearson correlation with Hi-C experimental data (top). **c.** A snapshot of the Sox9 locus in its closed disordered state as derived by the polymer model, with the position of TSSs of some key genes highlighted. **d.** In the locus, many-body contacts of n sites are exponentially more abundant than in random SAW conformations (the ratio of the average number is plotted v.s. n), which could help the simultaneous colocalization of multiple functional regulatory regions. **e.** The Sox9 locus folding dynamics from an initially open conformation towards the closed disordered state is represented by the gyration radius, Rg(t). Chromatin domains self-assemble hierarchically in higher-order structures, in approx. 20s.
(Figure from Chiariello *et al.* 2016)

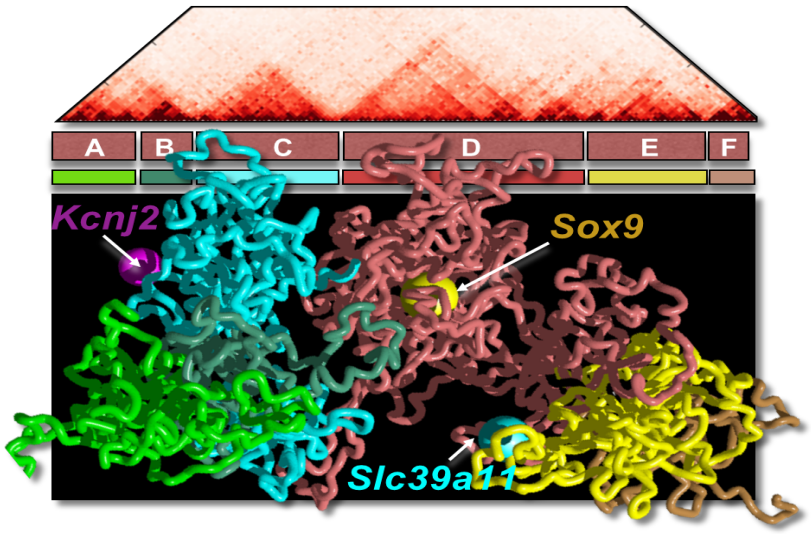

Figure 3.6: 3D reconstruction of the *Sox9* locus in mESC-J1 cells, alternative color scheme

The same configuration of the full 3D structure of the *Sox9* locus in mESC-J1 represented in **Fig. 3.5c**, with a color scheme reflecting the TADs position (from Dixon *et al.* 2012). It allows to visually interpret the patterns seen in its Hi-C map and the relative organization and interactions of TADs.
(Figure adapted from Bianco *et al.* 2017)

3.4.1. Simulation details of the *Sox9* polymer model

For the considered 6Mb region around the *Sox9* gene in mESC-J1, the PRISMR procedure returns a polymer composed by n=15 colors and made of N=2250 beads. So, the elementary bead of the polymer has a genomic content of L/N=2.67Kb. The size of the bead is thus 26nm, as follows from the calculation described in Section 2.3.5. The parameters used in the MD simulations (potentials, interaction energies and concentrations) are the same used for the modeling of the *Epha4* locus. The 3D structures are produced as previously described.

3.5. Modeling of the murine orthologue of the *7q11.23* human locus

In this Section, as a further example of application of our method, we show the results of the modeling of the 7q11.23 locus, a genomic region of great interest in Neurogenetics. In this region, structural variants (deletions and duplications) are associated with a variety of neurological disorders. For example, the 7q11.23 duplication syndrome is associated with speech problems and behavioral issues such as increased anxiety levels and autism (see Berg et al., 2007; Merla et al., 2010; Ramocki et al., 2010; Ebert et al., 2014). Deletions at the 7q11.23 locus, encompassing a couple of dozens genes, are instead associated to the Williams-Beuren syndrome (WSB), a complex developmental disorder (see Nature Research Highlights, 2011; Sanders et al., 2011; Chailangkarn et al., 2016). The attribution of the various features of WBS to specific genes is a complex, on-going effort. Besides the role of the genes in the deleted/duplicated region or epigenetic mechanisms, a factor that may be implicated in determining the genotype-phenotype relationship is the effect of such structural variants on the 3D folding of the locus. With such motivation, we report the first, albeit initial polymer physics exploration of the 3D conformation of the 7q11.23 locus.

Using the same procedure described above for the *Epha4* and *Sox9* loci, we modelled a 8Mb region (chr5:129500000-137500000, **Fig. 3.7a**) in mouse ESC-46C cell line, syntenic with the 7q11.23 locus in human genome. The dataset used is from Fraser et al. (2015), binned at 50kb, with ICE normalization (Imakaev *et al.*, 2012). The PRISMR inferred contact matrix is highly correlated with the experimental matrix (**Fig. 3.7a**), with a Pearson correlation coefficient r = 0.97. The polymer model of the locus involves n=15 different types of binding sites, whose position and abundance along the DNA sequence is represented by the different color histograms in **Fig. 3.7c**. The associated inferred conformations of the polymer model help explaining the 3D features of this locus. In **Fig. 3.7b**, two possible configurations, obtained from independent simulations, are showed. For sake of clarity, we color in green, orange and cyan respectively the three major domains visible in the matrix (labeled again as A, B and C, **Fig. 3.7a**), so we can easily compare the contact pattern with the spatial reconstruction. At a first visual analysis, we can recognize the A, B and C domains as distinguishable and individual blocks, in agreement with the experimental data.

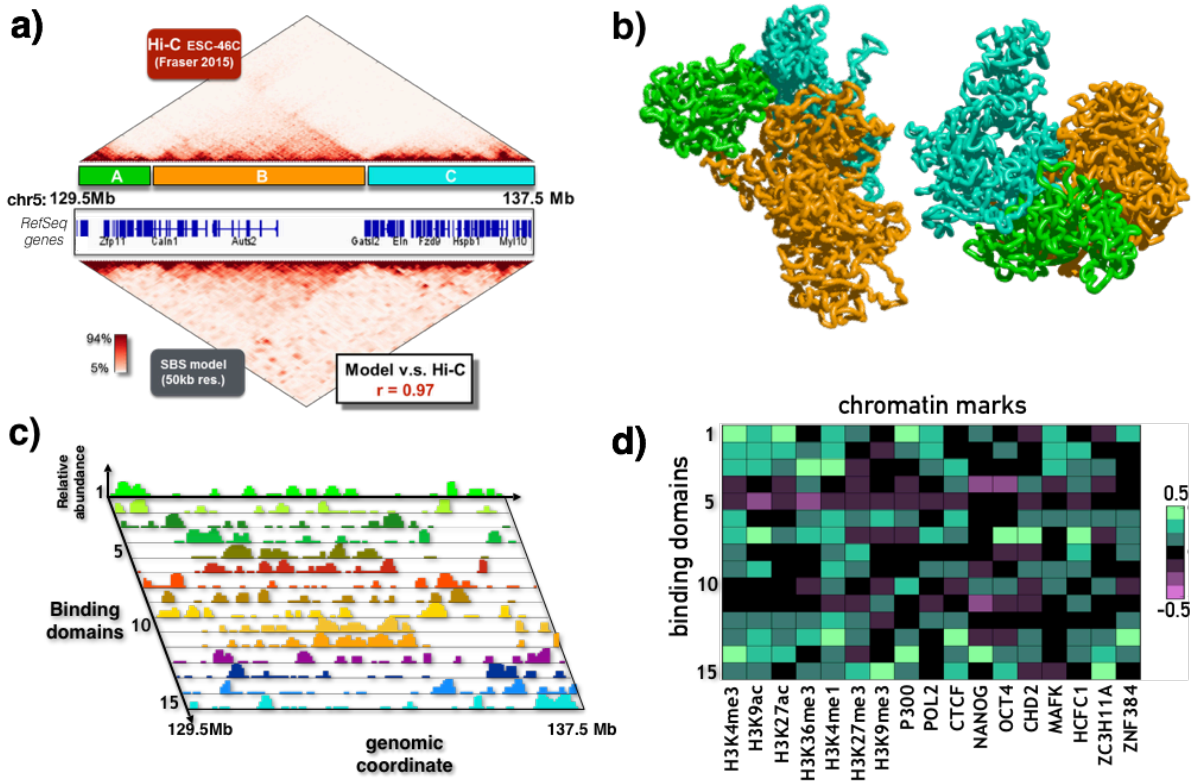

Figure 3.7: 3D reconstruction of the 7q11.23 locus in ESC-46C cells.
a. The modeled 8Mb chromosome region in mouse ESC-46C cells is shown. This region is syntenic with the 7q11.23 locus in human genome, which is linked to neurological disorders such as autism. The Hi-C experimental data (top matrix) are reproduced by the model derived contact data (bottom matrix) with good accuracy (Pearson r = 0.97). **b.** Two examples of independent 3D structures derived by Molecular Dynamics simulations. The three major contact domains (A, B, C) have complex, long range interactions with each other (e.g., the contacts between the green and cyan domains), in agreement with the Hi-C data. **c.** The SBS inferred n=15 binding domains of the locus, which drive its folding, are here shown. **d.** Pearson correlation coefficients between the relative abundance of the SBS binding domains with epigenetic chromatin features. (Figure from Chiariello *et al.* 2017)

A more deeper inspection reveal a non-random contact between domain A and domain C, which is again in agreement with the long-range interaction contained in the experimental matrix, even though it is a higher-order detail. Importantly, a collection of much smaller domains (the strong red triangles in the contact matrix) close to the diagonal is evident, and a complex pattern of higher order interactions among them is present, so to give the typical hierarchical internal substructure to the major domains. The model also captures such

lower level organization. For sake of clarity, we do not color all these domains in the polymer representation of **Fig. 3.7b**. As the model does capture not only general aspects of the locus organization, but also its finer features, it can be used to derive relevant biological implications of the 3D structure. To this aim, as made for the *Epha4* locus in Section 3.3 (see **Fig 3.4**), we analyze databases of epigenetic features and cross them with the relative abundance of the binding sites (Bianco *et al.*, 2017, Bianco *et al.*, *Submitted* (2017)). In **Fig. 3.7d** are reported the results, which reveal again a complex, not trivial, pattern of correlations. Interestingly, several binding domains exhibit high correlations with many of the considered features, reflecting the biological complexity of the locus, highly enriched in genes (**Fig. 3.7a**). Conversely, few of the binding domains (e.g., type 5 and 11, **Fig. 3.7c,d**) do not correlate with the considered epigenetic features, and result to be associated with the central, gene poor, region of the locus.

In summary, our polymer model of the murine genomic region syntenic with the 7q11.23 locus in human provides a first reconstruction of the ensemble of 3D conformation of the locus. A complex network of higher-order interactions of the locus emerges from our analysis, whose rewiring could be important to understand the effects of disease associated structural variants.

3.6. General applicability of the PRISMR method

We have discussed in this Chapter how polymer physics can reproduce with high accuracy (Pearson correlation coefficients higher than 0.9) the details of the structure of specific DNA regions, as emerge from Hi-C experiments. And we have seen how the polymer models inferred from Hi-C, can be used to derive further information about the 3D organization of the corresponding loci, beyond pairwise Hi-C contacts data. In particular, we focused on three important loci of DNA (*Epha4, Sox9, 7911.23*), linked to development and diseases. Using the same approach, we have modelled a dozen of other genomic loci, each 2-10Mb long, in different cell lines and tissues, obtaining similar levels of accuracy and various insights into the 3D chromatin architecture. Some of these studies are made in collaboration with the group of Professor Stefan Mundlos, at Max Planck Institute for Molecular Genetics, Berlin (Bianco *et al.*, *Submitted* (2017), Kragesteen *et al., Submitted* (2017)), and the group directed by Professor Jim Hughes, at Oxford University (Oudelaar *et*

al, in preparation) and are not yet published. Others can be found in the papers Chiariello *et al.*, 2016, Annunziatella *et al.*, 2016, Bianco *et al.*, 2017, Chiariello et al. 2017. These results will not be reported here for brevity. Here, we just want to highlight that in the different studied loci we used experimental contact data produced with different techniques (as 5C, Hi-C, in-situ Hi-C, etc.), at different resolutions and normalized in completely different ways (e.g., KR or ICE normalization). In this way, we enforce the validity of the method, that results to be unaffected by the underlying experimental technique and data treatments. Furthermore, in a collaboration with the group of Professor Ana Pombo at Max Delbruck Center for Molecular Medicine in Berlin, we adapted the PRISMR algorithm to work with a completely new type of experimental data, produced by GAM method (Beagrie *et al.* 2017), which, unlike all the others, is not a 3C-based technique (see Section 1.4). Moreover, in case of mouse ESC cells, we have extended the modeling genome-wide, i.e. to the entire set of 19 chromosomes of the mouse genome. The results of these studies have not been published yet and represent current research projects of the group.

4. Prediction of the effects of Structural Variants (SVs) on chromatin organization

4.1. Introduction

In the previous Chapter, we have introduced the PRISMR method as a powerful tool to describe the folding of DNA 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 will present a stringent test of the model. We will try to predict how the folding of a DNA locus changes after genomic rearrangements along its genomic sequence. In fact, the chromosome 3D folding, and especially the organization of TADs (Section 1.5.2), can be disrupted by genomic rearrangements, such as deletions, duplications or inversions, collectively called **structural variants (SVs)** (Lupiáñez et al. 2015, Hnisz et al. 2016, Lupiáñez et al. 2016, Franke et al. 2016). SVs can result in a re-wiring of enhancer-promoter contacts (Section 1.3), gene misexpression and disease. However, it is currently difficult to predict such ectopic interactions without performing extensive 3C-based experiments (Section 1.4) in cells or tissues carrying the rearranged chromosomes. In this scenario, polymer modelling by PRISMR emerges as a valid approach to predict interactions *in silico,* thereby providing a tool for analyzing the disease-causing potential of SVs.

In this Chapter, we will analyze a set of SVs, along the *EPHA4* locus, associated with different limb malformations (Lupiáñez *et al.* 2015), in four different mouse and human cell types (Section 4.2). In Section 4.3 we will model with PRISMR the wild-type, i.e. non-rearranged, *EPHA4* locus using all the four datasets. In Sections 4.4 and 4.5 we will discuss the model predictions, in mouse and human respectively, of the effects of SVs and will compare them to Capture Hi-C data generated from mouse limb buds and patient-derived fibroblasts.

Most of the material presented in this Chapter, including figures, paragraphs and sentences, is adapted or taken literally from the paper Bianco *et al., Submitted* (2017), which I coauthored.

4.2. Capture Hi-C (cHi-C) experiments and studied datasets

To test the predictive value of PRISMR, we chose the *EPHA4* locus, a key developmental region associated with different types of limb malformations (Lupiáñez *et al.* 2015). This study has been developed in collaboration with the group of Professor Stefan Mundlos at Max Planck for Molecular Genetics in Berlin, that performed the **Capture Hi-C (cHi-C)** experiment (Franke et al. 2016), one variant of Hi-C methods, and produced the datasets that we used for modeling the *EPHA4* locus. In a previous study by Mundlos group, the authors showed, using 4C experiments, that **structural variants (SVs)** at the *EPHA4* locus, like deletions, inversions and duplications, cause distinct phenotypes (brachydactyly, syndactyly, polydactyly) by altering the chromatin organization of the locus, thereby causing rewiring of enhancer-promoter contacts and gene misexpression (Lupiáñez *et al.* 2015). As seen in Section 1.4.1, 4C experiments only give information about the contacts of a single genomic position (the viewpoint) with the rest of the locus. To analyze the 3D configuration of the entire locus, as part of this new study, we performed cHi-C experiments from E11.5 mouse limb buds and human skin fibroblasts. We also analyzed the same locus from previously published Hi-C datasets from CH12-LX murine and IMR90 human cells (Rao *et al.* 2014).

To test the model predictions, new experimental cHi-C dataset was next produced from mouse limb buds carrying previously reported **homozygous** structural variants (Lupiáñez et al. 2015): a 1.6 Mb deletion, *DelB* (coordinates mm9 chr1: 76388978-78060839), a 1.5 Mb deletion, *DelB*^{s} (coordinates mm9 chr1: 76388978-77858974), and a 1.1 Mb inversion, *InvF (*coordinates mm9 chr1: 74832836-75898707). The term homozygous means that both alleles at the *Epha4* locus on the homologous chromosomes (Section 1.2) carry the mutation. On the contrary, a structural variant is **heterozygous** if present only on one allele, the other being of wild-type kind. To test the potential of PRISMR to predict the effects of heterozygous SVs on chromatin organization as they are commonly observed in human patient samples, fibroblasts obtained from human patients were used to perform cHi-C. In particular, we analyzed a 1.6 Mb deletion associated with brachydactyly (coordinates hg19 chr2: 221276849-223021152, similar to mouse *DelB*), a 900 Kb duplication (coordinates hg19 chr2: 219875536-220789199, *DupP*) associated with polydactyly and *IHH* activation, and a 1.4 Mb duplication (coordinates hg19 chr2: 219713606-221090946, *DupF*) associated with syndactyly and *WNT6 gene* activation (Lupiáñez *et al.* 2015).

The murine *Epha4* locus discussed here is the 6Mb long region around the *Epha4* gene introduced in Section 3.3 (coordinates mm9 chr1: 73,000,000-79,000,000). We employ in-situ Hi-C data from wild type (wt) CH12-LX cells (Rao *et al.* 2014) and cHi-C data in *wt* and mutants E11.5 mouse limb buds, at a 10kb resolution. The studied human *EPHA4* locus in skin fibroblasts is 5.77Mb long (coordinates hg19 chr2: 218,320,000-224,090,000); in that system cHi-C data for human healthy control [or wild-type (WT)] and patients are produced at 10kb resolution. We also studied the *EPHA4* locus in WT human IMR90 cells, where we used previously published in-situ Hi-C at 10kb resolution ((Rao *et al.* 2014); the considered locus is 8Mb long (coordinates hg19 chr2: 217,000,000-225,000,000). All the datasets have been normalized applying the KR (Knights and Ruiz) normalization (Knights&Ruiz, 2013), a matrix balancing algorithm that ensures equal sums for all rows and columns of the map. The underlying assumption for this type of normalization is that all loci should have an equal representation in the map. Since we did not work directly to the experimental stage, all the biological and chemical details of the cHi-C experiment, biological samples, preparation of the libraries and sequencing, will not be discussed here.

Figure 4.1: PRISMR recapitulates 3D conformation at the *EPHA4* locus

Hi-C data from Rao *et al.* 2014 (left) and our capture Hi-C data (right) compare well with the contact matrices derived by PRISMR. Their Pearson correlation, r, and distance-corrected Pearson correlation coefficient, r', are: r=0.91, r'=0.56 in CH12-LX (see also Fig. 3.4), r=0.94, r'=0.60 in limb tissue E11.5, r=0.92, r'=0.64 in IMR90 and r=0.93, r'=0.69 in human fibroblasts.
(Figure adapted from Bianco *et al.* Submitted (2017))

4.3. Polymer models of the wild type *EPHA4* locus in mouse and human cells

4.3.1. PRISMR models of the *EPHA4* locus

In **Fig. 4.1** (top matrices) the different wild-type Hi-C datasets mentioned in the previous section are shown. Regardless of the cell or tissue type or the species, we observe in Hi-C maps a subdivision of the wild-type *EPHA4* locus in one large TAD-containing only *EPHA4*, a smaller TAD-containing the genes *PAX3* and *SGPP2*, and a gene dense region on the centromeric side that shows no clear TAD structure. Differences within the *EPHA4* TAD were apparent, that are likely to reflect cell and tissue-specific patterns of interaction and gene regulation.

Towards developing predictive models of the architecture of the *EPHA4* locus across the different cell types, we applied PRISMR to all four datasets. As shown in the previous Chapter, PRISMR identifies the different *binding domains* of the locus, i.e., the sets of binding sites of the same type (i.e., color) determining the folding patterns. In case of the 10kb resolution in-situ Hi-C data on the *Epha4* locus in CH12-LX cells (Rao *et al.* 2014), we have seen that the model predicts n=21 different binding domains and λ=1.0. (Chapter 3, **Figs. 3.3**, **3.4**). In the case of 10kb resolution, KR normalized, cHi-C data in mouse limb tissue similar values are found and we also use n=21 and λ=1.0. Applied to 10kb resolution, KR normalized, in-situ Hi-C data in human IMR90 cells (Rao *et al.* 2014) on the considered human *Epha4* locus, PRISMR gives n=16 and λ=1.0. For 10kb resolution, KR normalized, cHi-C data in human fibroblast produced in this study we find n=24 and λ=1.0. The model contact matrices, derived from Molecular Dynamics (MD) simulations of the identified polymer models as discussed in the previous Chapter, are similar to the original Hi-C, not only recapitulating the global TAD conformation of the locus, but also cell-specific intra-TAD organization (**Fig. 4.1**): the Pearson correlation coefficient, r, range up to r=0.95 (**Fig. 4.1**).

As the simple Pearson correlation is dominated by distance-dependence of Hi-C data, we also considered a finer measure to compare experimental and model matrices, the **distance-corrected correlation**. Thus, we have measured the Pearson correlation, r', *after* the effect of genomic distance is subtracted from a contact map. (**Fig. 4.2**). Specifically, we

subtracted from each diagonal of the contact matrices (experimental and predicted) their average contact frequency (corresponding to a fixed genomic distance), and then calculated the Pearson correlation coefficient, excluding the effect of strong outliers (>99.9th percentile) in the murine datasets. As control case, we considered the correlation between distance-corrected experimental matrix and distance-corrected predicted matrix with bootstrapped diagonals where we obtained values of r‘ lower than 0.05, well below the values for the real data that range up to r’=0.69 (**Figs. 4.1, 4.2**). Furthermore, the correlation values found between our model matrices and experiments are comparable to the estimation of the correlation levels between biological replicates of the KR normalized cHi-C maps; for example, in human we find that r’ is around 0.65 across replicates (Bianco *et al.*, Submitted (2017)). That is an indication that our models accurately describe experimental data.

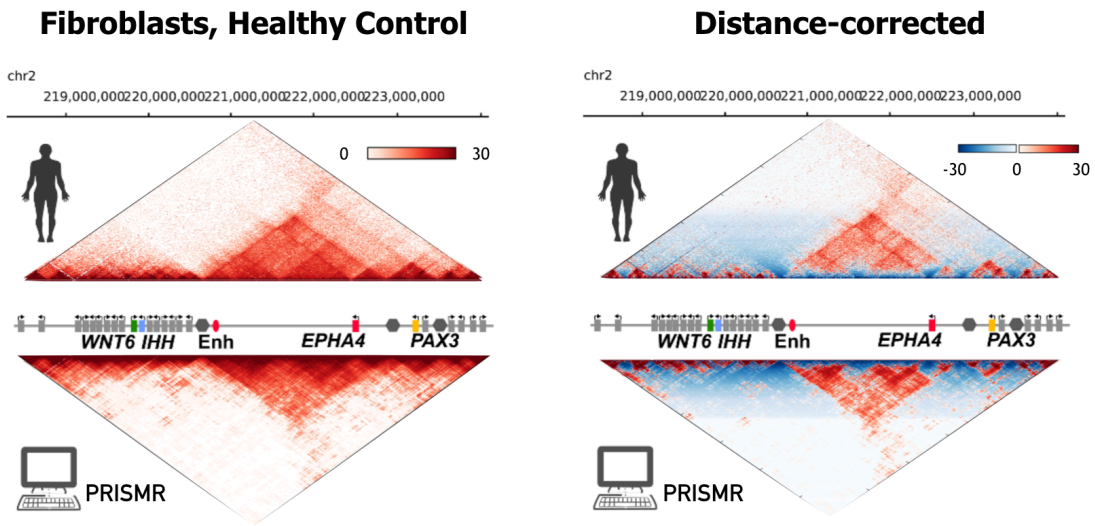

Figure 4.2: Comparison of cHi-C and model data in human fibroblast cells after correcting for genomic proximity effects.

The model derived contact matrix (bottom left) of the *EPHA4* locus in human fibroblast cells has a high Pearson correlation with our cHi-C data (top left, r=0.93). The matrices in the right panel are obtained by subtraction of the average interaction at a given genomic distance. The patterns in the data are still captured by the model after the effects of genomic distance are subtracted: the Pearson correlation coefficients remains high (r’=0.69). (Figure from Bianco *et al.* Submitted (2017))

4.3.2. PRISMR+CTCF models of the *Epha4* locus

One major factor known to organise chromatin folding is the architectural protein CTCF, a DNA-binding transcription factor thought to facilitate the formation of chromatin loops (Sanborn *et al.* 2015, Guo *et al.* 2015, Fudenberg *et al.* 2016, Nora *et al.* 2017). Interestingly, as discussed in the previous Chapter (subsection 3.3.5), some binding site types envisaged by PRISMR correlate with CTCF. As our method does not exploit prior information on binding sites and factors, to test its reach we considered a variant of the model where we include previous knowledge on the location of CTCF binding sites in the locus, which are added to interact with an additional type of binders bridging opposed (forward/reverse) CTCF sites. In limb tissue E11.5 cells, for instance, this variant (named 'PRISMR+CTCF') has correlations with Hi-C data similar to the initial model: it improved the visualisation of the large *Epha4* TAD, mainly by strengthening the loop anchors characteristic for CTCF-associated loops; however, it also results in additional contacts in the neighbouring gene dense region that are not present in the original cHi-C data (**Fig. 4.3**, top right). Conversely, a model with only CTCF (named 'CTCF only') can describe some of the loops seen in the data (Sanborn *et al.* 2015, Fudenberg *et al.* 2016), but poorly captures the global contact patterns of the *Epha4* locus (**Fig. 4.3**, bottom right), resulting in a lower correlation coefficient (r'=0.05). These results indicate that other factors besides CTCF are important in chromatin folding and TAD configuration and that our approach can recapitulate most of the interactions of Hi-C data without *a-priori* information on binding factors. Nevertheless, such information can be added to adapt and improve model predictions.

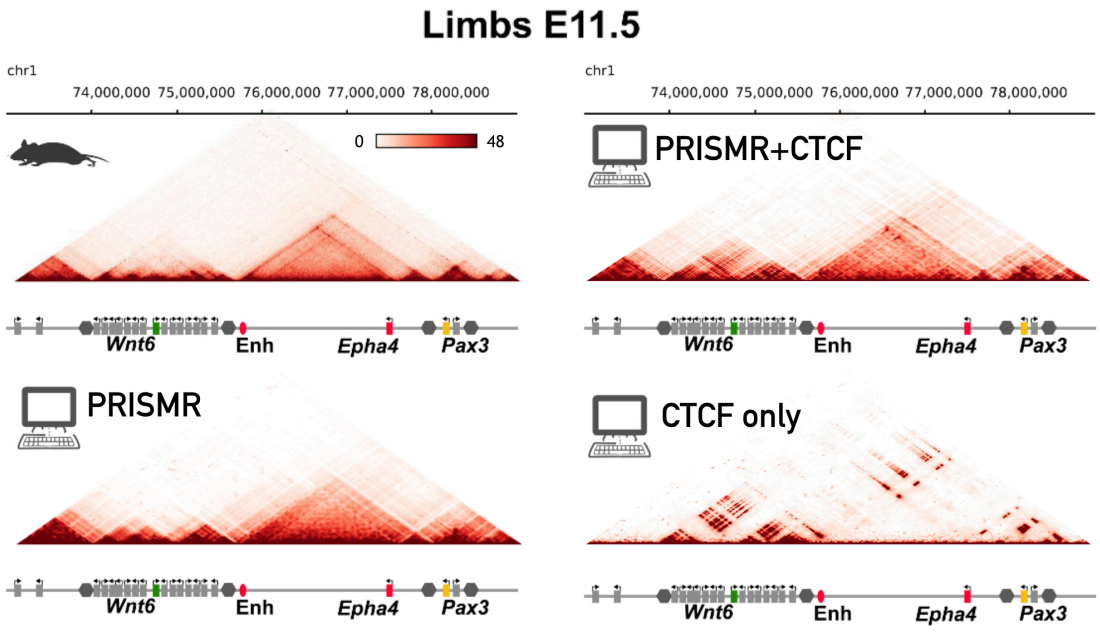

Figure 4.3: comparison of original model predictions with the model also including prior knowledge on CTCF and with a model with only CTCF

The figure shows the contact matrices from our cHi-C data in mouse E11.5 limb bud tissue (top-left) and from three different models. The bottom left panel reports the results derived by our PRISMR method: they have a Pearson, r, and distance-corrected Pearson correlation, r', with cHi-C data equal to, respectively, r=0.94 and r'=0.60. The top right panel shows the data from a variant of PRISMR (the 'PRISMR+CTCF' model) that takes into account prior knowledge of the CTCF binding sites of the locus; its correlations with cHi-C data are r=0.95 and r'=0.52, comparable to the initial PRISMR model. Conversely, a model that only includes CTCF sites (bottom right) has a lower correlation with cHi-C data (r=0.89, r'=0.05). (Figure from Bianco *et al.* Submitted (2017))

4.3.3. Simulations details and calculation of contact matrices

To simulate the model derived from the wild type *Epha4* locus in mouse E11.5 cell, we used the same MD parameters used for CH12-LX cells, given in Section 3.3.1. Similarly, to model the *InvF* inversion, and the $DelB^s$ and *DelB* deletions in mouse, that will be discussed in the next Section, we used polymers with N=12600, N=9093 and N=9513 beads respectively.

Analogously, to model the human wild-type *EPHA4* locus in human skin fibroblast cells we used a polymer made of N=13848 beads, so the genomic content results s_0=417bp. As in the mouse case, to speed up simulations, we start with a shorter polymer, made of N/4 beads in this case, and then we add the remaining beads by reducing the original bead diameter of a factor 1/4. The polymer models of the *DelB* deletion and the *DupP* and *DupF* duplications, that will be discussed in Section 4.5, were made of N=9672, N=16032 and N=17160 beads respectively. The range explored of the total binder concentration, *c*, was from zero to *c*=300n*mol*/*l*, and the interaction energy used is the same than in the mouse cases. Finally, our polymer models of the *EPHA4* locus in IMR90 human cells, were made of N=12800, while interaction energy and concentrations were in the same range of the other simulations.

To simulate the PRISMR+CTCT model, to the binding sites of our PRISMR polymer we added new specific binding sites corresponding to the genomic locations of CTCF peaks. We used peak-called CTCF ChIP-seq data (ENCODE Project Consortium, 2012) and to avoid background effects we only considered the peaks having a score higher than a stringent threshold. We performed a standard motif finding analysis (using the FIMO tool in

the MEME Suite online software) (Grant *et al.,* 2011) to identify the best matching peak, within the considered 10Kb bin, with the CTCF binding motif (Barski *et al.*, 2007). Analogously, an orientation was attributed to the motif according to its location on the forward or reverse strand (Grant *et al.,* 2011). In the PRISMR model of the locus, we add two new types of binding sites, one for forward and one for reverse CTCF binding sites. We also add a new type of binder that can only bind and bridge opposed oriented CTCF sites. The binding energy is taken to be the same used for all other binders in our model. To speed up MD simulations of the novel model (i.e., PRIMR with CTCF sites), the system starts initially from the already folded configurations of the original PRISMR model. To speed up the preparation of the initial conformation, elastic springs are used to bring in close physical proximity nearest neighbor forward-reverse CTCF site pairs.

Finally, to try to dissect the specific effects of CTCF alone, we considered a simpler polymer model where only the above described CTCF binding sites are included (i.e., the different 'colours' of the PRISMR model are not considered). As above, opposite CTCF site pairs can interact with their specific binders and the system is prepared and equilibrated as before.

To extract the average pairwise contact frequency matrices of the polymer models, we proceed as discussed in Chapter 2 (section 2.6.3, see also section 3.3.2 of Chapter 3). As mentioned in section 3.3.2, we used a threshold parameter A that ranges from 2 to 10, with similar results in all cases. For example, as already stated there, in the model of the mouse wild-type *Epha4* locus in CH12-LX cells, we find a 89%-11% open-closed mixture and a correlation coefficient r=0.91 with in-situ Hi-C data (**Fig. 3.4** and **Fig. 4.1**). Analogously, in the case of our human fibroblast cell data (**Fig. 4.1**, log color scale), the correlation coefficient between model and cHi-C is r=0.93, with a 70%-30% mixture. In all cases, we find correlation coefficients between model and experimental contact maps from r=0.88 to r=0.94 (**Tables 4.1** and **4.3**).

4.4. Prediction of the effects of homozygous SVs on *Epha4* locus folding in mouse

To test if PRISMR was able to predict the effects of homozygous SVs on chromatin folding, we investigated three previously reported variants (Lupiáñez *et al.* 2015) at the *Epha4* mouse locus: a deletion (*DelB*) encompassing a large part of the *Epha4* TAD and the telomeric TAD boundary (associated with brachydactyly, due to misexpression of the gene *Pax3*), a slightly smaller deletion (*DelBs*) that leaves the TAD boundary intact (no misexpression, no phenotype), and a balanced 1.1 Mb inversion (*InvF*) (causing misexpression of *Wnt6*) (see Section 4.2). We implemented the mutations in the polymer models of the wild-type (wt) CH12-LX (in Fig. 3.4a) and E11.5 limbs cells inferred by PRISMR (see previous Section) and re-run Molecular Dynamics simulations to derive an average contact matrix for the mutated locus. For E11.5 limb tissue, we tested both the PRISMR polymer model (subsection 3.4.1) and the PRISMR+CTCF version (subsection 3.4.2) with the addition of CTCF sites (**Figs 4.4A**, **4.5A**, **4.6A** top matrices).

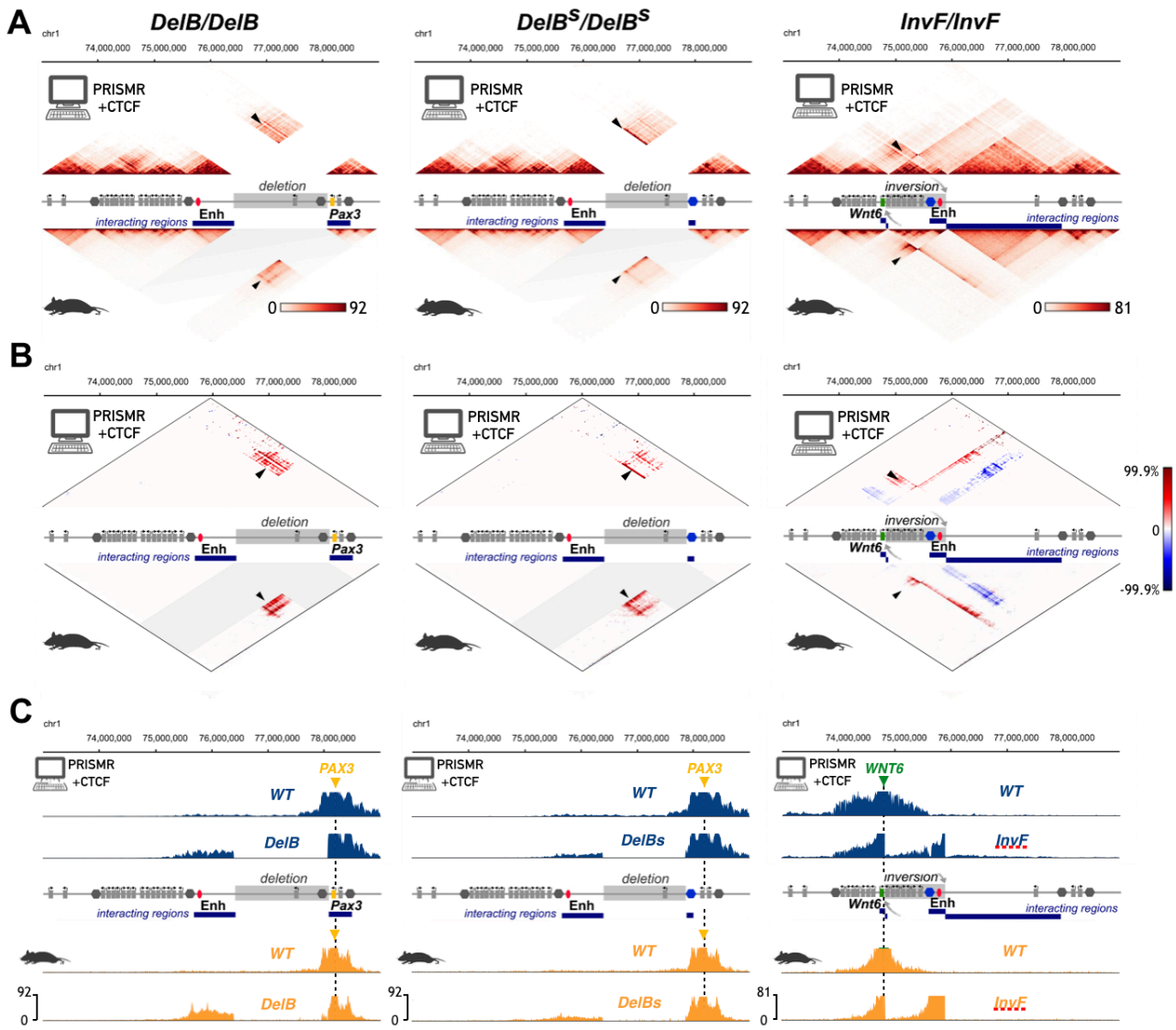

Figure 4.4: PRISMR+CTCF model based on wild-type mouse limbs E11.5 cHiC data predicts the effects of homozygous structural variants on chromatin architecture

A. Contact matrices are shown from model predictions based on our cHi-C obtained from wt E11.5 limb bud tissue (top). Data from cHi-C in mutant E11.5 limb bud tissue are shown below. The genomic region with its genes is shown schematically. Deleted/inverted regions are in grey.

DelB/*DelB*: PRISMR predicts the 3D chromatin effects of a 1.6 Mb deletion encompassing parts of the *Epha4* TAD and the *Epha4*/*Pax3* boundary (r=0.95, r'=0.41). Arrowhead and blue bars indicate the regions of interaction between the remaining *Epha4* and *Pax3* TADs. *DelBs*/*DelBs*: PRISMR predicts the chromatin effects of a 1.5 Mb deletion encompassing parts of the *Epha4* TAD but not the *Epha4*/*Pax3* boundary (r=0.95, r'=0.50). Arrowhead and blue bars indicate the interaction between the remaining *Epha4* TAD and the *Epha4*/*Pax3* boundary (blue hexagon). *InvF*/*InvF*: PRISMR predicts the chromatin effects of a 1.1 Mb homozygous inversion (r=0.95, r'=0.60). Arrowheads and blue bars mark the region of interaction between the enhancer region and the genomic location of *Wnt6*. Note that

contacts are interrupted at the position of the inverted boundary element. A small genomic region at the centromeric inverted region (containing the *Wnt10a* gene) also interacts de-novo with the *Epha4* TAD. The centromeric *Epha4* boundary retains its functionality despite its inversion (blue hexagon).

B. Subtraction maps produced between wt and mutants from predictions and cHi-C data. Above threshold gain of interaction is displayed in red and loss in blue. Arrowheads and blue bars indicate regions of ectopic interaction between the *Epha4* TAD and other regions of the genome.

C. Virtual 4C plots derived from predictions and cHi-C data from the viewpoint on the respective phenotype causing gene.
DelB/*DelB*: note increased interaction of *Pax3* promoter with the remaining *Epha4* TAD, including the enhancer cluster in both, the prediction and experimental data. The immediate downstream gene *Spp2* shows no increased interaction with the enhancer cluster. *DelBs*/*DelBs*: note that the promoter of *Pax3* interacts less frequently with the *Epha4* TAD compared to *DelB*/*DelB* mutants. *InVF*/*InvF*: note increased interaction between *Wnt6* gene promoter and the *Epha4* enhancer cluster in both predictions and experimental data.

(Figure from Bianco *et al.* Submitted (2017))

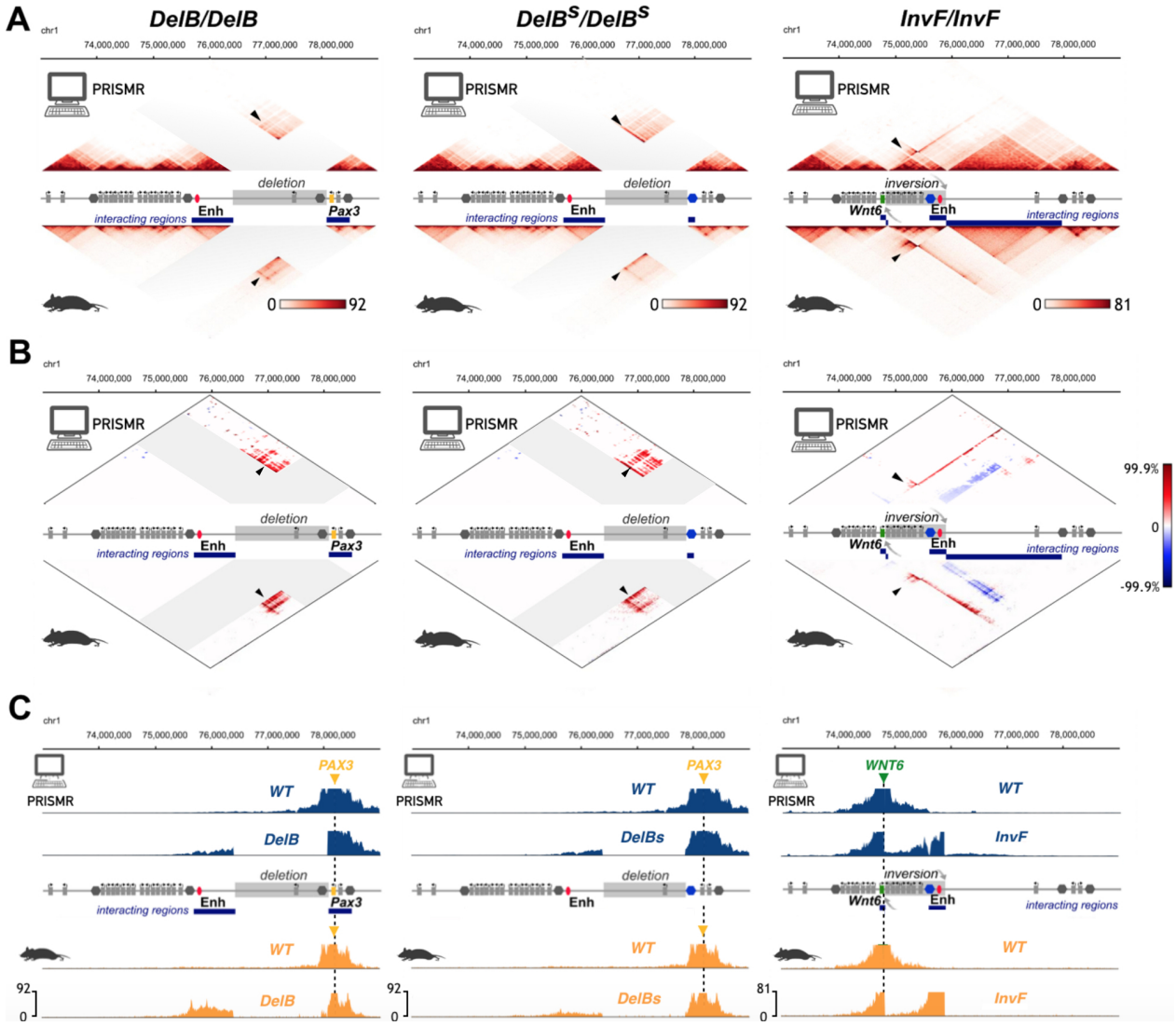

Figure 4.5: PRISMR model based on wild-type mouse limbs E11.5 cHiC data predicts the effects of homozygous structural variants on chromatin architecture

As in Fig. 4.4, for PRISMR model (without ad-hoc addition of CTCF) predictions based on our capture Hi-C obtained from E11.5 limb.

(Figure from Bianco *et al.* Submitted (2017))

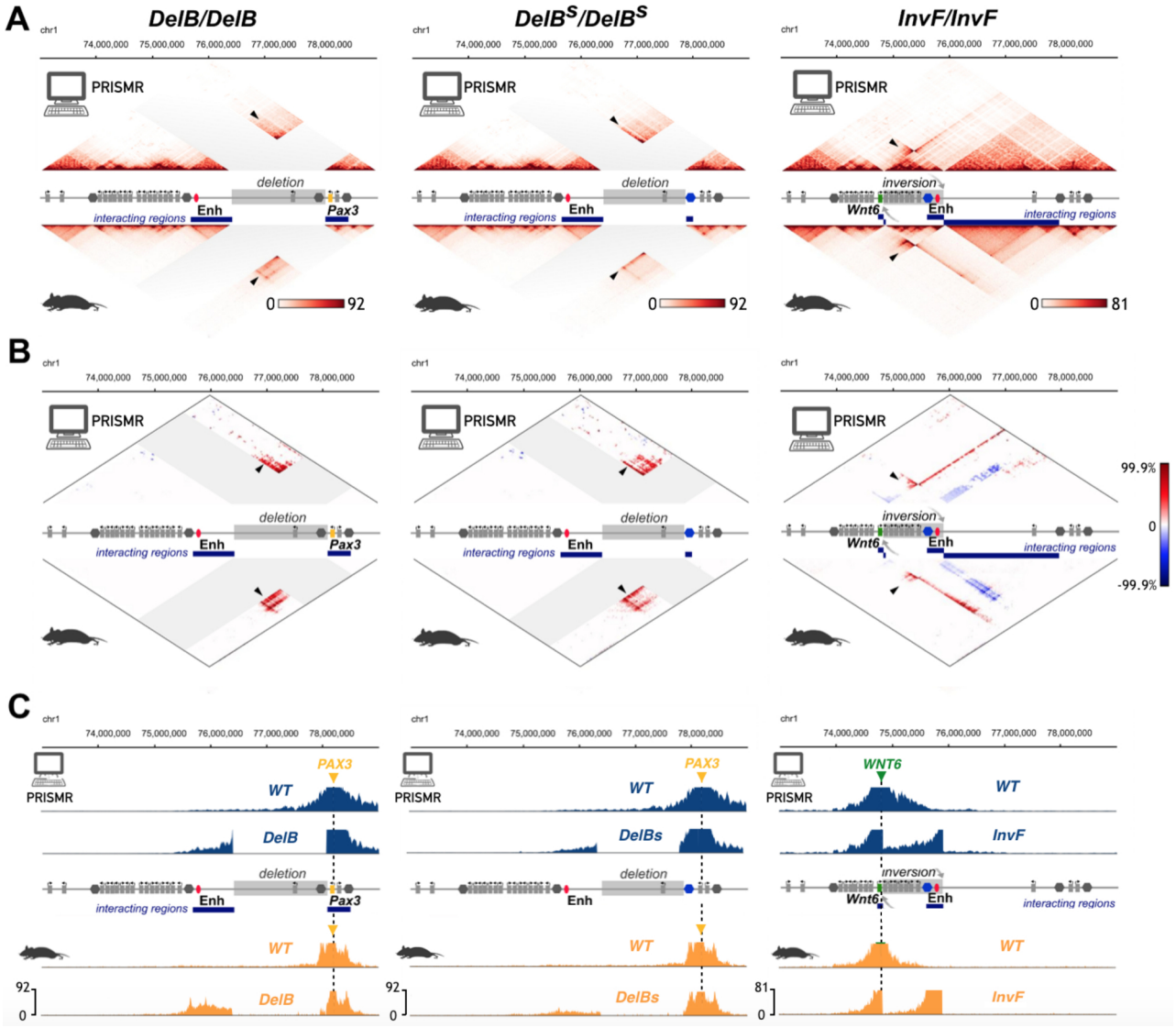

Figure 4.6: PRISMR model based on mouse wild-type CH12-LX Hi-C data predicts the effects of homozygous structural variants on chromatin architecture

As in Fig. 4.4, for PRISMR model predictions based on in-situ Hi-C data from Rao *et al.* 2014 in CH12-LX cells.

(Figure from Bianco *et al.* Submitted (2017))

To identify the regions of statistically significant ectopic interactions in each predicted rearrangement, we subtracted each mutant matrix from the wt (**Figs. 4.4B**, **4.5B**, **4.6B**, see subsection 4.4.2 for details). Although the studied locus is populated by more than 40 genes, our matrices predicted that only certain regions, which contain a limited number of genes, displayed changes in the interaction profiles**.** For example, in the larger deletion (*DelB*) including the *Epha4* TAD boundary we identified new contacts predicting a fusion between the remaining *Epha4* and *Pax3* TADs, thus facilitating the association between *Epha4* enhancers and *Pax3* that results in ectopic gene activation and a pathogenic phenotype (Lupiáñez *et al.* 2015). Ectopic contacts between the same regions were also predicted in the smaller deletion (*DelBs*), which leaves the *Epha4*/*Pax3* boundary intact.

However, virtual 4C analysis derived from our predictions showed that the enhancers/*Pax3* ectopic interaction is diminished, consistent with the absence of *Pax3* activation in these mutants (**Figs. 4.4C**, **4.5B**, **4.6C**, see subsection 4.4.3 for details). The inversion (*InvF*) was predicted to result in the rearrangement of the genomic content of the two adjacent TADs with interaction hotspots between *Epha4* enhancers and a gene dense region (3 genes affected) that is consistent with the ectopic *Wnt6* activation reported previously. In addition, we also observed ectopic interactions between a region near the centromeric breakpoint containing the *Wnt10a* gene and the remaining *Epha4* TAD. Therefore, PRISMR identified specific and localized regions of ectopic interactions across the entire locus, highlighting a very limited number of genes whose regulation might be directly influenced by large genomic rearrangements.

As a next step, we tested the accuracy of our predictions by comparison against a new experimental cHi-C dataset from mouse limb buds carrying the homozygous mutations. Our dataset revealed the same regions of ectopic interaction and displayed a remarkable high agreement with PRISMR predictions, not only across the entire locus but also when the regions of ectopic interaction were compared (**Figs. 4.4, 4.5, 4.6, 4.7 and Tables 4.1, 4.2**). Again, our results confirmed that the larger deletion in *DelB* mutant led to a fusion of the *Epha4* and *Pax3* TADs, not ocurring in the smaller *DelBs* mutation, where the TAD boundary remains intact. In the inversion, ectopic contacts are observed between *Wnt6* and the *Epha4* enhancer region, which facilitate *Wnt6* activation as previously observed *in vivo* (Lupiáñez *et al.* 2015), and between a region at the centromeric breakpoint with the entire *Epha4* TAD. Interestingly, the observed ectopic interaction is interrupted by the

Epha4 centromeric boundary that, although inverted, appears to retain its functionality. Hence, deletions and inversions that include boundary elements can result in TAD fusions or TAD reorganisations, respectively.

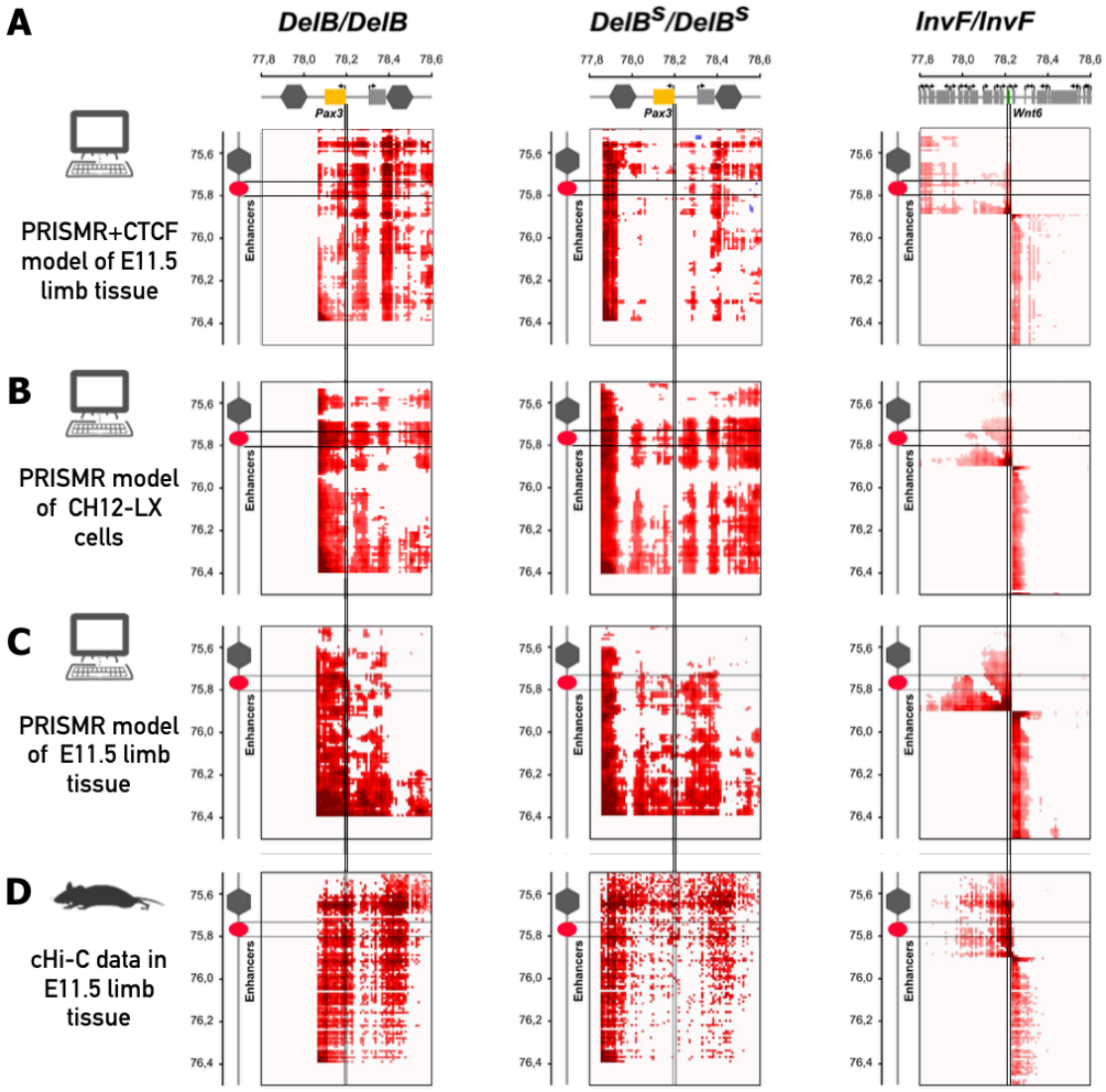

Figure 4.7: Regions of ectopic interaction in murine cell mutants.

Zoom of the regions exhibiting significant ectopic interactions within the subtraction matrices from:
(A) PRISMR+CTCF model of E11.5 limb tissue (**Fig. 4.4b**).
(B) PRISMR model of E11.5 limb tissue (**Fig. 4.5b**).
(C) PRISMR model of CH12-LX cells (**Fig. 4.6b**).
(D) Experimental cHi-C data in E11.5 limb tissue (**Fig. 4.4b**).
The distance-corrected correlation coefficient between model and experiment is reported in **Table 4.2** for all the shown cases.
(Figure from Bianco *et al.* Submitted (2017))

Dataset	Genotype	Coefficient r	Coefficient r'
E11.5 limbs cHi-C v.s. PRISMR+CTCF	wild type	r=0.95	r´=0.52
	DelB	r=0.95	r´=0.41
	DelBs	r=0.95	r´=0.50
	InvF	r=0.95	r´=0.60
E11.5 limbs cHi-C v.s. PRISMR	wild type	r=0.94	r´=0.60
	DelB	r=0.94	r´=0.50
	DelBs	r=0.95	r´=0.55
	InvF	r=0.93	r´=0.52
E11.5 limbs cHi-C v.s. PRISMR derived from mouse CH12-LX (Rao et al, 2014)	wild type	r=0.91	r´=0.56
	DelB	r=0.93	r´=0.45
	DelBs	r=0.93	r´=0.46
	InvF	r=0.92	r´=0.49

Table 4.1: **Pearson correlations between models and experimental data in mouse.** Summary of Pearson correlations (r) and distance corrected Pearson correlations (r') for all the considered datasets and variants in mouse. (Table adapted from Bianco *et al.* Submitted (2017))

Dataset	Mutation	Coefficient r'
E11.5 limbs cHi-C v.s. PRISMR+CTCF	*DelB*	r'=0.52
	DelBs	r'=0.75
	InvF	r'=0.55
E11.5 limbs cHi-C v.s. PRISMR	*DelB*	r´=0.57
	DelBs	r´=0.69
	InvF	r´=0.64
E11.5 limbs cHi-C v.s. PRISMR derived from mouse CH12-LX (Rao et al, 2014)	*DelB*	r´=0.59
	DelBs	r´=0.65
	InvF	r´=0.61

Table 4.2: distance-corrected Pearson correlations in regions of ectopic interactions.

We report here the distance-corrected Pearson correlation coefficients between model predicted and Hi-C contact change matrices (i.e., the correlation between the 'subtraction maps', wt – mutants, from model and Hi-C) restricted to the regions with significant ectopic contacts, for all the mutations considered. (Table adapted from Bianco *et al.* Submitted (2017))

4.4.1. 3D conformations of the *Epha4* locus and its mutations

By use of our polymer models we derived not just the pairwise contact matrix for each given locus/mutation, but also the ensemble of the corresponding 3D conformations. In our MD simulations such 3D conformations are breathing in time, even at stationarity. The examples shown in **Fig. 4.8** are time-snapshots from such an ensemble of conformations at equilibrium. The shown polymer is obtained by a geometric interpolation with a smooth spline curve mathematically described by a third-order polynomial, passing through the coordinates of the beads of the polymer chain. The snapshots of the predicted 3D structures help clarifying, e.g., the relative positions of regulatory regions and promoters, and the nature of the changes in folding captured by the pairwise contact matrix. The 3D snapshots in **Fig. 4.8** refer to the *Epha4* locus in mouse CH12-LX cells, where the wt case is inferred by PRISMR from published in-situ Hi-C data (Rao *et al.* 2014) and the mutations are predicted as described above and in the Main Text. The 3D snapshots illustrate that the deletions, beyond producing such specific interactions, bring in closer proximity regions that in wt are genomically distant, thus increasing their generic overall contacts.

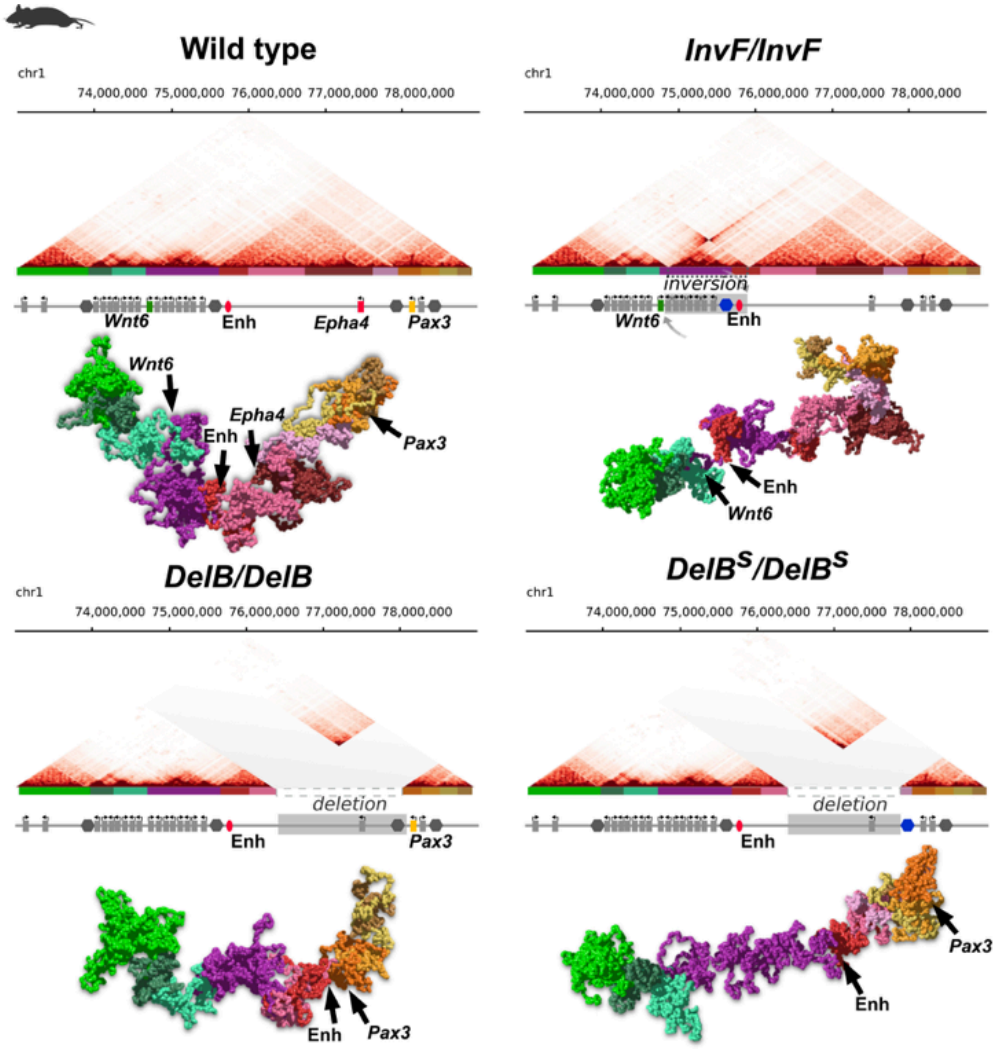

Figure 4.8: PRISMR predicted 3D conformations of the *Epha4* locus in murine CH12-LX cells.

Top-left: the PRISMR model based on published Hi-C data in murine CH12-LX cells recapitulates (r=0.91, r'=0.56) the experimental pairwise contact matrix (see also Fig. 4.1). The shown 3D conformation is a snapshot of the model of the locus with the relative positions of genes and regulator highlighted. Bottom-left: the PRISMR model inferred from the above wt data is informed with the *DelB*/*DelB* deletion and the effects on chromatin folding predicted (see also Fig. 4.6). The shown 3D conformation is a snapshot of the model bearing the *DelB* deletion. Top-right: Analogous results for the *DelBs*/*DelBs* shorter deletion. Bottom-right: Analogous results for the *InvF*/*InvF* inversion. (Figure from Bianco *et al.* Submitted (2017))

4.4.2. Determination of significant ectopic interactions

In order to identify the statistically significant ectopic interactions in the contact matrices after SVs from experimental data and from PRISMR model predictions, we consider the

normalized difference matrices (**Figs 4.4B**, **4.5B**, **4.6B** and **4.7**) between the mutated contact matrix and the wt contact matrix. Specifically, we multiply the matrix corresponding to the mutation (experimental and simulated) by a factor that equalizes the reads count equivalent of the regions that are not involved in the mutation, then we subtract from the mutated matrix the wt matrix. To take into account the genomic distance bias, we normalized the difference matrix by dividing each sub-diagonal by the average wt reads count at its corresponding pairwise genomic distance.

Next, in order to identify the statistically significant differences, we only retain the values of the normalized difference matrix falling above two standard deviations of the distributions of values in each sub-diagonal (that corresponds to an average one tail p-value less than 0.1 across genomic distances, over the different samples). In the calculation of the standard deviations, we filter out the points above 96th percentile in the cases where the data are marked by strong outliers, as in the human deletion data discussed in the next Section. Finally, to correct for finite size effect, we used a higher threshold (four standard deviations) near the edge of the matrix (within the 5% of the matrix size). The same higher threshold is used when the data sample gets smaller, as in the case of genomic distances larger than half of the matrix size. To check our results, we also tested a procedure where the threshold is increased linearly with the genomic distance along the contact matrix, without finding major differences; this is shown in the case of human mutations (next Section). Since the mouse cHi-C matrices are homozygous mutants, the data corresponding to the deleted genomic segments are not represented (**Figs. 4.4**, **4.5**, **4.6**). The experimental subtraction matrices were computed on the raw count cHi-C maps.

4.4.3. Virtual 4C analysis

In order to better highlight ectopic interactions, we produced virtual-4C plots (see Section 1.4.1 for a brief description of real 4C-seq experiments) from the viewpoint of the phenotype causative genes in each mutant in mouse (**Figs. 4.4C**, **4.5C**, **4.6C**) and in human (next Section). Virtual 4C are obtained by plotting the column in the contact matrix corresponding to the considered viewpoint. To have a fair comparison between wt and mutation, we first normalised the wt matrix by equalizing the number of its reads to the total reads in the mutation, as described in the previous subsection.

4.5. Prediction of the effects of heterozygous SVs on *EPHA4* locus folding in human

In the previous Section, we have seen how PRISMR was able to predict, to a large extent, the effects of three different homozygous SVs on the *Epha4* locus folding, in mouse cell and tissue. Here, we want to test its potential to predict the effects of heterozygous SVs on chromatin folding as they are usually observed in human patients (see Section 4.2). We considered a 1.6 Mb deletion associated with brachydactyly (analogous to mouse *DelB*), a 900 Kb duplication (*DupP*) associated with polydactyly and *IHH* gene activation, and a 1.4 Mb duplication (*DupF*) associated with syndactyly and activation of *WNT6* gene (Lupiáñez *et al.* 2015). The starting point is the PRISMR model of the wild-type *EPHA4* locus, that was inferred from our cHi-C data of healthy human fibroblast and discussed in Section 4.3 (**Fig. 4.9**).

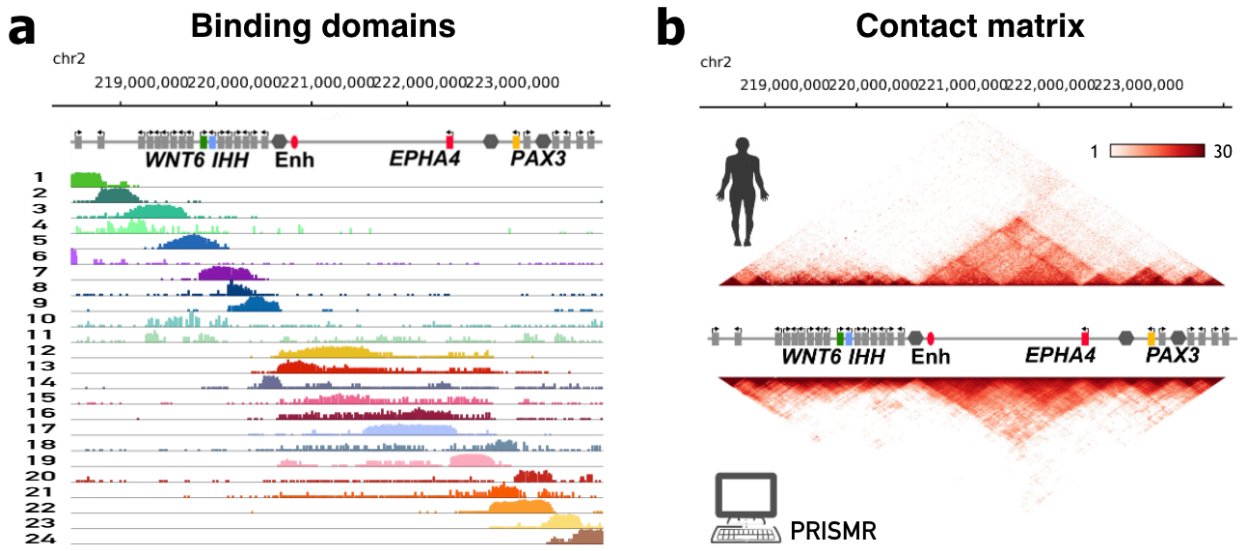

Figure 4.9: PRISMR model of the wild-type (WT) human fibroblasts cHiC data.

a. PRISMR identified *n*=24 different binding domains (colors) in the model of the *EPHA4* locus in healthy human fibroblasts (see Section 4.3). **b.** Experimental cHiC (top) and model derived (bottom) contact matrix of the *EHPA4* locus in human fibroblasts (already shown in Fig. 4.1, bottom right). Pearson correlation is r=0.93 and distance-corrected correlation r'=0.69. (Figure adapted from Bianco *et al.* Submitted (2017))

The considered SVs were implemented in such polymer model, in order to predict their effects on the locus contact matrices (**Figs. 4.10A**, top matrices). To obtain models for heterozygous mutants, the corresponding simulated contact matrices are equally averaged with the simulated wild-type matrix (for details about MD simulations of the human fibroblast polymer models, see Section 4.3.3.). To test the model predictions, we performed cHi-C on fibroblasts obtained from human patients carrying the different heterozygous Svs (**Fig. 4.10A**, bottom matrices).

Subtraction maps were computed to identify the precise regions and intensity of significant ectopic interactions (**Figs. 4.10B**, **4.11**, see Section 4.4.2). In the brachydactyly associated deletion, both PRISMR and the fibroblast derived cHi-C data detected the same region of ectopic interaction as seen for the equivalent mouse mutant *DelB*, displaying increased interaction between *PAX3* and the *EPHA4* enhancer cluster. In the duplication *DupF*, we observe ectopic interactions not only between the enhancer cluster and *WNT6*, but also with the neighboring gene *WNT10a*. In *DupP* the disease-causing gene *IHH* displayed increased interaction with the enhancers, as well as the two neighboring genes, genes *CCDC108* and *NHEJ1*. Comparison of the PRISMR predictions with cHi-C data from patient fibroblasts revealed a high correlation (**Tables 4.3 and 4.4**), demonstrating that PRISMR can also predict the effects of SVs on misfolded chromatin contacts in heterozygous samples, thus facilitating the identification of disease-causing genes.

human fibroblasts cHi-C v.s. PRISMR	WT	r=0.93	r´=0.69
	DelB	r=0.93	r´=0.61
	DupF	r=0.88	r´=0.52
	DupP	r=0.90	r´=0.56

Table 4.3: **Pearson correlations between models and experimental data.** Summary of Pearson correlations (r) and distance corrected Pearson correlations (r') for all the considered datasets and variants. (Table adapted from Bianco *et al.* Submitted (2017))

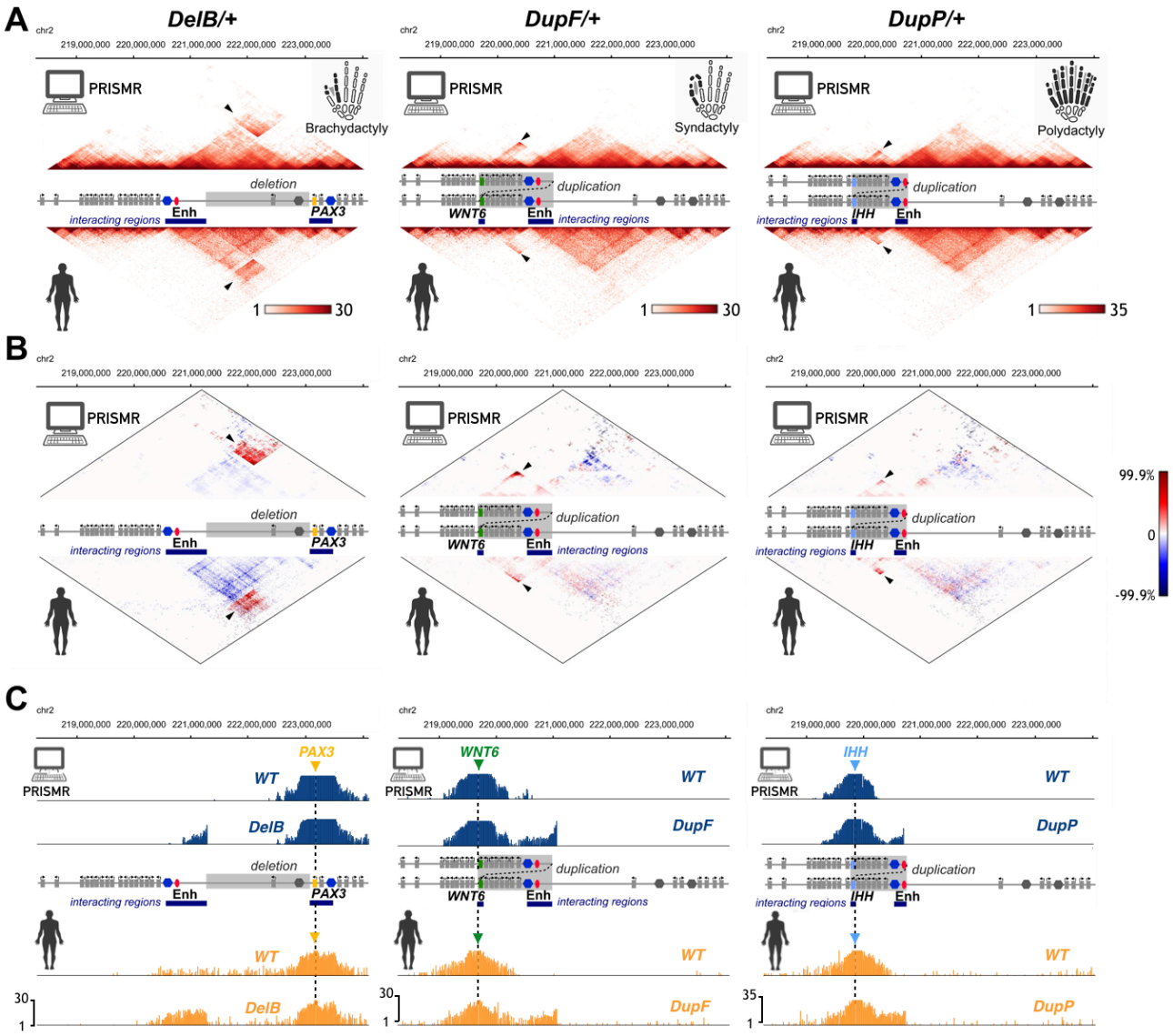

Figure 4.10: PRISMR predicts the effects of human heterozygous structural variants on chromatin architecture

(**A**) Contact matrices are shown from model predictions derived from WT data (top) and cHi-C experiments in mutation carrying cultured human skin fibroblasts (bottom). The genomic region and its genes are indicated schematically. The human phenotype associated with the rearrangement is indicated on right. See Methods for sample collection.

DelB/+: PRISMR predicts the chromatin effects of a 1.6 Mb heterozygous deletion (r=0.93, r'=0.61). Arrowhead and blue bars mark the region of interaction between the remaining *EPHA4* and *PAX3* TADs resulting in misexpression of *PAX3* and brachydactyly. *DupF*/+: PRISMR predicts the chromatin effects of a heterozygous 1.4 Mb duplication (r=0.88, r'=0.52). Arrowhead and blue bars mark the regions of interaction between the *EPHA4* enhancer cluster and a genomic region containing *WNT6*. *DupP*/+: PRISMR predicts the chromatin effects of a heterozygous 900 bp duplication (r=0.90, r'=0.56). Arrowhead and blue bars mark the regions of interaction between the *EPHA4* enhancer cluster and a genomic region containing *IHH*.

(**B**) Subtraction maps produced between wt and mutants from predictions and cHi-C data. Above threshold gain of interaction is displayed in red and loss in blue. Arrowheads and blue bars mark regions of interaction involving *EPHA4* enhancers and disease causing genes. See Methods for description of statistics.

(**C**) Virtual 4C plots derived from predictions and cHi-C data from the viewpoint on the respective phenotype causing gene.

DelB/+: note increased interaction of *Pax3* promoter with the remaining *EPHA4* TAD, including the *EPHA4* enhancer cluster in both, the prediction and experimental data. *DupF*/+: note increased interaction of *WNT6* promoter with the *EPHA4* enhancer cluster in both, the prediction and experimental data. *DupP*/+: note increased interaction of *IHH* promoter with the *EPHA4* enhancer cluster in both, the prediction and experimental data. (Figure from Bianco *et al.* Submitted (2017))

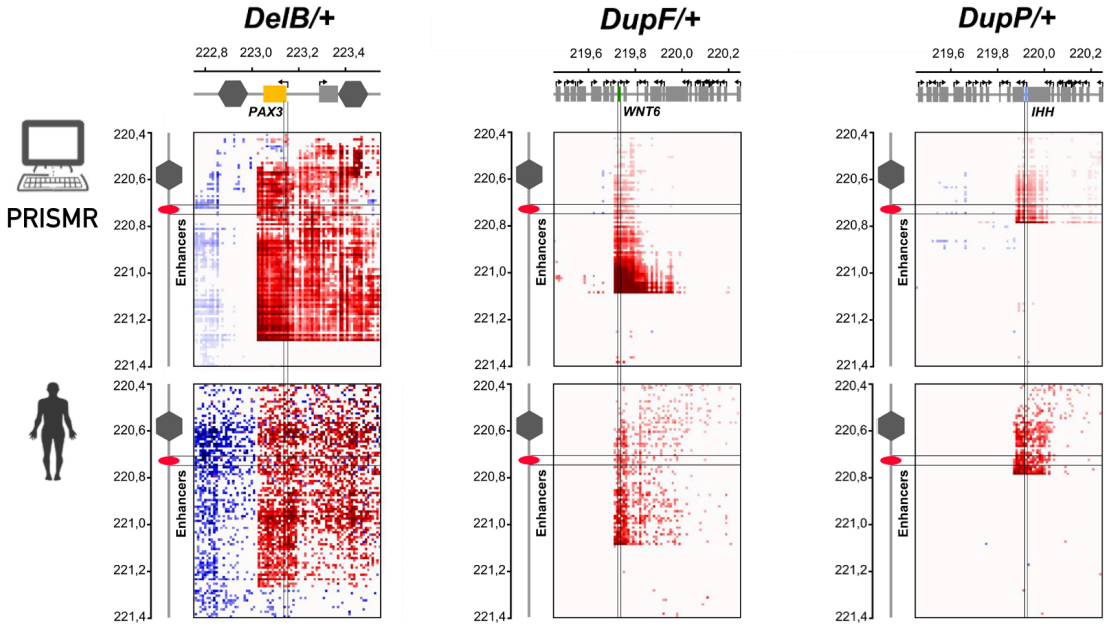

Figure 4.11: Regions of ectopic interaction in human fibroblast cells.

Zoom of the regions exhibiting significant ectopic interactions within the subtraction matrices from the PRISMR model (top) and cHiC data (bottom) in human fibroblast cells. (Figure from Bianco *et al.* Submitted (2017))

human fibroblasts cHi-C v.s. PRISMR	*DelB*	r´=0.41
	DupF	r´=0.54
	DupP	r´=0.49

Table 4.4: distance-corrected Pearson correlations in regions of ectopic interactions. (Table adapted from Bianco *et al.* Submitted (2017))

In **Fig. 4.12** snapshots of 3D conformations are shown, derived by MD simulations for the model of the WT and mutant *EPHA4* locus in human fibroblasts. The 3D snapshots of the duplications illustrate, for instance, that part of the ectopic contacts discussed in **Fig. 4.10** are produced because the duplicated segments tend to twist back in a loop onto each other. The color code in the mouse case (**Fig. 4.8**) is derived from the one of the human case based on their synteny (as determined by the liftOver tool in the UCSC Genome Browser).

Collectively, our results demonstrate that PRISMR is an efficient tool to predict alterations in chromatin contacts induced by disease-associated SVs, in both homozygous and heterozygous samples and even in complex genomic regions with high gene density. PRISMR predictions can be used to identify regions of ectopic interaction that can then be scanned for their content, i.e. the presence of genes and enhancers that could interact. Furthermore, our results indicate that PRISMR can be used in cases where affected tissues or equivalent cell types are not available. The recent advance in Next Generation Sequencing (NGS) has boosted the identification of SVs (Gilissen *et al.*, 2014, Hehir-Kwa *et al.*, 2016, Newman *et al.*, 2015). In this scenario, polymer modelling by PRISMR emerges as a valid method to predict pathogenic effects, facilitating the interpretation and diagnosis of this type of genomic rearrangements.

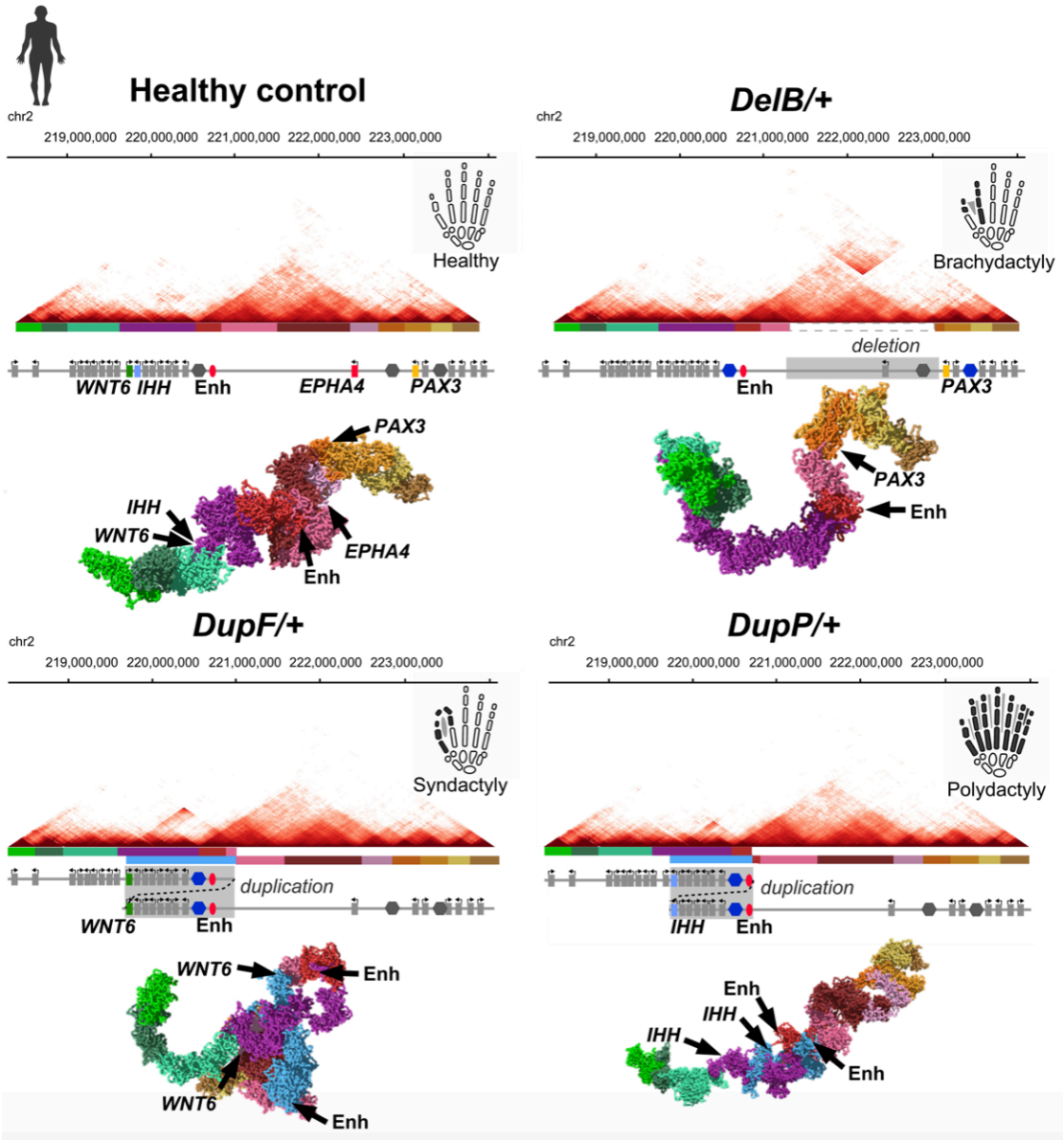

Figure 4.12: PRISMR predicted 3D conformations of the EPHA4 locus in human fibroblast cells.

Top-left: the PRISMR model based on our cHi-C data in healthy human fibroblast cells recapitulates (r=0.93, r'=0.69) the experimental pairwise contact matrix (see also **Fig. 4.1**). The shown 3D conformation is a snapshot of the model of the locus with the relative positions of genes and regulator highlighted.
Top-right: The PRISMR model inferred from the above wild-type control data is informed with the *DelB* heterozygous deletion (*DelB*/+) and the effects on chromatin folding predicted (see also Fig. 4). The shown 3D conformation is a model derived snapshot of the mutated locus.
Bottom-left: Analogous results for the *DupF*/+ duplication.
Bottom-right: Analogous results for the *DupP*/+ duplication.
(Figure from Bianco *et al.* Submitted (2017))

Conclusion and perspectives

In this work, we have discussed the application of polymer physics models to investigate quantitatively the chromatin three-dimensional structure in mammalian genomes. Such work is motivated by the observation that chromatin architecture has a crucial role in vital biological functions and that its abnormal folding is often linked to severe human diseases, like congenital diseases and cancers. New experimental techniques, such as Hi-C, revealed that chromatin have a complex spatial structure, with interesting features like the formation, along its sequence, of TAD domains of strong internal interactions, that weakly interact with each other. However, a unified quantitative framework describing spatial chromatin organization is still lacking and polymer modeling can help to face such challenging problem. In this work, we focused on the Strings and Binders Switch (SBS) polymer model, in which chromatin 3D conformations form through the interaction of diffusing molecular binders with binding sites along the polymer chain describing chromatin. As a first step, we recapitulated with a very simple and essential model some important aspects of chromatin folding as, for instance, the formation of TADs and the spontaneous hierarchical folding mechanism. Next, we generalized the SBS model and developed an innovative method, PRISMR, to obtain highly accurate 3D reconstructions of specific genomic regions. Such method is interesting also because it gives information on the position of binding sites for bridging binders along the chromatin filament, that can be crossed with several available epigenomic datasets to try to infer the molecular nature of the complexes that mediate chromatin interaction. Such epigenetics analysis revealed that the different types of binding sites, and their cognate binders, do not simply correspond to a single molecular factor associated to chromatin, but rather to combinations of different factors, including known factors shaping chromatin structure like CTCF/Cohesin complex. However, such analysis is still at an early stage and the extension of the modeling genome-wide and at higher resolutions, together with further experimental work seems a promising means to unravel the molecular determinants of chromatin folding. Finally, we showed that our polymer models can be employed to predict with a high degree of accuracy the effect of structural variants in the genomic sequence on the 3D architecture, thus providing a valuable tool for analyzing their disease-causing potential. We analyzed a set of deletions, inversions and duplications. However, our method can be straightforwardly extended to model translocations or insertions of regions deriving from the same locus. The case of

translocations/insertions deriving from a distinct genomic region would require modeling also the other DNA region or additional hypothesis on the structure of the polymer segment to be inserted, for example by exploiting the epigenetic barcode of that region. Another aspect we are currently working on is the employment of our polymer models to capture the structural differences of a fixed genomic region during differentiation or in any two different cell types. In summary, we are following new researches lines, not described in this thesis, in order to improve the predictive power of our model and to investigate at a deeper level the numerous, still unknown, mechanisms involved in the genome organization.

References

Allen MP & Tildesley, DJ, *Computer simulation of liquids* (1987), Oxford University Press

Artus, J. & Hadjantonakis, A. K. Generation of chimeras by aggregation of embryonic stem cells with diploid or tetraploid mouse embryos. *Methods Mol Biol* **693**, 37-56 (2011).

C. David Allis and Thomas Jenuwein (2016),
"The molecular hallmarks of epigenetic control" *Nat. Rev. Genet. **17** 487-500*

Annunziatella C., Chiariello A. M., Esposito A., Bianco S., Fiorillo L. and Nicodemi M. Molecular Dynamics simulations of the Strings and Binders Switch Model of chromatin. *Submitted to Methods (under review 2017)*

Annunziatella C, Chiariello AM, Bianco S and Nicodemi M. (2016) Polymer models of the hierarchical folding of the HoxB chromosomal locus, *Phys Rev E* **94**: 042402

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.M., Dostie, J., Pombo, A. and Nicodemi, M. (2012) Complexity of chromatin folding is captured by the Strings & Binders Switch model. *Proc. Natl. Acad. Sci. U S A* **109**: 16173-1678.

Barbieri, M., Xie S. Q., Torlai Triglia E., Chiariello A. M., Bianco S. *et al.* Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. *Nat Struct Mol Biol* **24**, 515-524 (2017)

Barski, A. *et al.* (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837.

Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ., Barbieri M, de Santiago I, Lavitas LM, Branco MR, Fraser J, Dostie J, Game L, Dillon N, Edwards PAW, Nicodemi M & Pombo A (2017) Complex multi-enhancer contacts captured by genome architecture mapping, *Nature* **543**: 519-524

Berg JS, Brunetti-Pierri N, Peters SU, Kang SH, Fong CT, Salamone J, Freedenberg D, Hannig VL, Prock LA, Miller DT, Raffalli P, Harris DJ, Erickson RP, Cunniff C, Clark GD, Blazo MA, Peiffer DA, Gunderson KL, Sahoo T, Patel A, Lupski JR, Beaudet AL, Cheung SW (2007) Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. Genet Med 9(7):427-41.

Bianco S, Chiariello AM, Annunziatella C *et al*. (2017) Predicting chromatin architecture from models of polymer physics, *Chrom Res* **1**: 25-34.

Bianco S, Lupiáñez DG, Chiariello AM., Annunziatella C., Kraft K, Schöpflin R, Wittler L, Andrey G, Vingron M, Pombo A, Mundlos S, Nicodemi M. Polymer Physics Predicts the Effects of Structural Variants on Chromatin Architecture. *Submitted to Nat. Gen. (under review 2017)*

Bickmore WA, van Steensel B. Genome architecture: domain organisation of interphase

chromosomes. (2013) *Cell* **152**:1270-84.

Bohn M, Heermann DW (2010) Diffusion-driven looping provides a consistent framework for chromatin organisation. *PLoS ONE*, **5**: e12218.

Brackley CA, Taylor S, Papantonis A, Cook PR, and Marenduzzo D, (2013) Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and and genome organisation. *Proc Natl Acad Sci U.S.A.* **110**: E3605-11.

Branco MR, Pombo A (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4:** e138

Brookes E, de Santiago I, Hebenstreit D, Morris KJ, Carroll T, Xie SQ, Stock JK, Heidemann M, Eick D, Nozaki N, Kimura H, Ragoussis J, Teichmann SA, Pombo A (2012) Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10:** 157-170

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339.

Chailangkarn T., Trujillo CA, Alysson et al. (2016) A human neurodevelopmental model for Williams syndrome, Nature 536, 338–343.

Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A. & Nicodemi, M. (2016) Polymer physics of chromosome large-scale 3D organisation. *Sci Rep* **6**: 29775.

Chiariello, A. M, Esposito A., Annunziatella C., Bianco S., Fiorillo L., Prisco A. and Nicodemi M. (2017). A polymer physics investigation of the architecture of the murine orthologue of the 7q11.23 human locus, *Frontiers in Neuroscience* Vol. 11 N. 559.

Cremer T. and Cremer C (2001) Chromosome territories, nuclear architecture and agene regulation in mammalian cells. *Nat. Rev. Gen.***2**: 292

de Gennes PG (1979) Scaling Concepts in Polymer Physics (Cornell Univ Press, Ithaca, NY).

Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organisation of genomes: interpreting chromatin interaction data. *Nat. Rev. Gen*. **14**(6): 390-403.

Di Carlo MG, Minicozzi V, Foderà V, Militello V, Vetri V, Morante S and Leone M (2015) Thioflavin T templates Amyloid b(1-40) Conformation and Aggregation pathway *Biophysical Chemistry* 10.1016/j.bpc.2015.06.006.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485:** 376-380

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature* **518:** 331-336

Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98 (2016).

Eliezer Calo and Joanna Wysocka (2013). Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell* **49**, 825-837

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Encode Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640.

Ebert G, Steininger A, Weißmann R, Boldt V, Lind-Thomsen A, Grune J, Badelt S, Heßler M, Peiser M, Hitzler M, Jensen LR, Muller I, Hu1 H, Arndt PF, Kuss AW, Tebel K and Ullmann R (2014) Distribution of segmental duplications in the context of higher order chromatin organisation of human chromosome 7, BMC Genomics 15:537

Franke M, Ibrahim D M, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K,Kempfer R, Jerković I, Chan W L, Spielmann M, Timmermann B, Wittler L, Kurth I, Cambiaso P, Zuffardi O, Houge G, Lambie L, Brancati F, Pombo A, Vingron M, Spitz F & Mundlos S, (2016) Formation of new chromatin domains determines pathogenicity of genomic duplications, *Nature* **538**: 265-269

Fraser, J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, Xie SQ, Morris KJ, Itoh M, Kawaji H, Jaeger I, Hayashizaki Y, Carninci P, Forrest ARR, FANTOM, Semple CA, Dostie J, Pombo A, and Nicodemi M. (2015) Hierarchical folding and reorganisation of chromosomes are linked to transcriptional changes during cellular differentiation. *Mol. Sys. Bio.* **11**: 852.

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

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**: 950-63.

Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-347 (2014).

Grant, C. E., Bailey, T. L. & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018.

Guo, Y. *et al.* (2015) CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910

Hagege, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* **2**, 1722-1733 (2007).

Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).

Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458 (2016).

Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228 e219 (2017).

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9:** 999-1003

Jost D, Carrivain P. Cavalli G, Vaillant C (2014) Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.* **42**: 9553-61.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. (2011) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotech.* **30**, 90–98

Kirkpatrick, S., Gelatt, C. D., Jr. & Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* **220**, 671-680.

Knight, P. A. & Ruiz, D. (2013) A fast algorithm for matrix balancing. **33**, 1029–1047.

Konberg RD, Lorch Y. (1999). Chromatin-modifying and remodelling complexes. Curr Opin Genet Dev **9**:148–151

Kragesteen B.K., Spielmann M., Paliou C., Heinrich V., Schoepflin R., Esposito A, Annunziatella C, Bianco S, et al., "Dynamic 3D Chromatin Architecture Determines Enhancer Specificity and Morphogenetic Identity During Limb Development", *Submitted* (2017)

Kremer, K. & Grest, G. S. (1990). Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chern. Phys* **92**, 5057 (1990).

Kreth G, Finsterle J, von Hase J, Cremer M, Cremer C (2004): Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophys J*, **86**:2803- 2812.

Kubo, N. *et al.* Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. *BioRxiv* (2017)

Kundu, S. *et al.* Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell* **65**, 432-446 e435 (2017)

Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8:** 104-115

Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289-293.

Lupiáñez, DG *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012-1025.

Lupianez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet* **32**, 225-237 (2016).

MacCallum, J. L., Perez, A. & Dill, K. A. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci U S A* **112**, 6985-6990.

Marenduzzo D, Micheletti C, Cook PR (2006) Entropy-driven genomeorganisation. *Biophys J* **90**: 3712-3721.

Merla G, Brunetti-Pierri N, Micale L, Fusco C. (2010) Copy number variants at Williams-Beuren syndrome 7q11.23 region. Hum Genet. 128(1):3-26.

Misteli T (2007) Beyond the sequence: cellular organisation of genome function. *Cell* **128:** 787-800.

Nagano T, Lubling Y, Stevens T, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A and Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **302**: 59-64.

Nagy, K. N., J. in *Advanced Protocols for Animal Transgenesis* 431-455 (Springer Protocols Handbooks, Springer, Berlin, 2011).

Newman, S., Hermetz, K. E., Weckselblatt, B. & Rudd, M. K. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am J Hum Genet* **96**, 208-220 (2015).

Nature Research Highlights (2011) Neurogenetics: Extended hunt for autism genes Nature 474, 254–255.

Nicodemi M, and Pombo A (2014) Models of chromosome structure. *Current Opinion in Cell Biology* **28**:90–95.

Nicodemi, M. and Prisco, A. (2009) Thermodynamic pathways to genome spatial organisation in the cell nucleus. *Biophys. J.* **96**: 2168-2177.

James P. Noonan and Andrew S. McCallion (2010) Genomics of Long-Range Regulatory Elements. Ann. Rev. Gen. & Hum. Genet. **11**, 1-23

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Bluthgen N, Dekker J, Heard E (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485:** 381-385

Nora, E. P. *et al.* (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922.

Olivares-Chauvet P, Mukamel Z, Lifshitz A, Schwartzman O, Elkayam NO, Lubling Y, Deikus G, Sebra RP, Tanay A (2016) Capturing pairwise and multi-way chromosomal conformations using chromosomal walks*. Nature* **540**: 296-300.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. **12**, 283–293.

Ong, C.T., and Corces, V.G. (2012). Enhancers: emerging roles in cell fate specification. EMBO Rep. **13**, 423–430.

Oudelaar A. M., Chiariello A. M., Bianco S., Higgs D. R., Nicodemi M., and Hughes J. R., Higher-order architecture of cis-regulatory elements associate with complex regulation of the globin loci**,** *In preparation*

Parisi, G. *Statistical field theory*. (Westview Press, New York, 1998).

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG (2013) Architectural protein subclasses shape 3D organisation of genomes during lineage commitment. *Cell* **153:** 1281-1295.

Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. *J Comput Phys* **117**: 1–19.

Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34 (2014).

Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

Ramocki MB, Bartnik M, Szafranski P, *et al*. (2010) Recurrent distal 7q11.23 deletion including HIP1 and YWHAG identified in patients with intellectual disabilities, epilepsy, and neurobehavioral problems. Am J Hum Genet. 87:857–865.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665-168

Rosa A, Everaers R (2008) Structure and dynamics of interphase chromosomes. *PLoS Comput Biol* **4**: e1000153.

Sachs RK, Van den Engh G, Trask B, Yokota H, Hearst JE (1995) A random-walk/giant-loop model for interphase chromosomes *Proc. Natl. Acad. Sci. U S A* **92**: 2710-14.

Salamon, P. S., P.; Frost, R. *Facts, conjectures, and improvements for simulated annealing*. (SIAM, Philadelphia 2002).

Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, Geeting KP, Gnirke A, Melnikov A, McKenna D, Stamenova EK, Lander ES, Aiden EL. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U.S. A.* **112**: E6456-65.

Sanders, S. J. et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863–885.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G (2012) Three-dimensional folding and functional organisation principles of the Drosophila genome. *Cell* **148:** 458-472.

Simonis, M. *et al.* (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genet.* **38**, 1348–1354.

Spielmann M, Mundlos S (2013) Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays* **35:** 533-543.

Schwarzer, W. *et al.* Two independent modes of chromosome organization are revealed by cohesin removal. *BioRxiv* (2016).

Tanay A, Cavalli G (2013) Chromosomal domains: epigenetic contexts and functional implications of genomic compartmentalization. *Current Opinion in Genetics & Development* **23:** 197-203.

Tiana G, Amitai A, Pollex T, Piolot T, Holcman D, Heard E, Giorgetti L, (2016) Structural fluctuations of the chromatin fiber within topological associating domains, *Biophys. J.* **110**: 1234

Xavier de la Cruz, et al. (2005) Do protein motifs read the histone code? *BioEssays* 27:164–175

Yan, J. *et al.* Histone H3 Lysine 4 methyltransferases MLL3 and MLL4 Modulate Long-range Chromatin Interactions at Enhancers. *BioRxiv* (2017)

Watson JD, Baker TA, Bell SP, Gann A, Levine M and Losick R (2008) *Molecular biology of the gene*. Pearson Benjamin Cummings, San Francisco.

Wingett, S. *et al.* (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310

Wu, W. *et al.* Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* **21**, 1659-1671 (2011).