Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN

FISICA

Ciclo XXXI

Coordinatore: prof. Salvatore Capozziello

Investigating the three-dimensional architecture of genomes by polymer physics

Settore Scientifico Disciplinare FIS/02

Dottorando
Carlo Annunziatella

Tutore
Prof. Mario Nicodemi

Anni 2015/2018

Contents

Acknowledgments

Appendices

A. Comparing different chromatin polymer models

Introduction

In mammalian cell nuclei, chromosomes have a spatial organization that is strictly related to cellular biological functions, such as regulation of gene transcription and expression (Bickmore and Van Steensel, 2013; Dekker et al., 2013; Lieberman-Aiden et al., 2009; Misteli, 2007; Tanay and Cavalli, 2013). However, still today, the three-dimensional organization of genome and the mechanisms driving its folding are not completely known and represent an open question in modern biology.

In recent years, new technologies have been developed to help to investigate, for the first time, the genome architecture in a quantitative way. These methods, such as Chromosome Conformation Capture (3C) techniques, measure the interaction frequencies between pairs of genomic regions across a cell population (Dekker et al., 2013). In particular, they have revealed that the genome is characterized by a complex non-random structure, that occurs at different genomic length scales through the formation of many local and long-range interactions (Beagrie et al., 2017; Lieberman-Aiden et al., 2009; Quinodoz et al., 2018). In the nucleus, chromosomes occupy distinct territories, whose preferred positions depend on cell type and transcription activity (Bickmore and Van Steensel, 2013; Misteli, 2007; Tanay and Cavalli, 2013). Within each chromosome, the genome is organized in self-interacting domains, called "topologically associated domains" (briefly, TADs) (Dixon et al., 2012; Nora et al., 2012), in which chromatin regions frequently interact with each other. Such domains are approximately 0.5-1 Mb long and result to be highly conserved across species, cell lines and tissue types (Dixon et al., 2012; Fraser et al., 2015; Nora et al., 2012; Phillips-Cremins et al., 2013; Sexton et al., 2012). At a higher order level TADs, in turn, interact with each other giving rise to a hierarchy of domains-within-domains, called meta-TADs, extending up to chromosomal scales (Fraser et al., 2015). This 3D architecture of chromatin has key functional roles, as for instance to control gene activity through the formation of physical loops between regulatory regions and target remote genes. The disruption of such an intricate network of interactions can alter the regular gene activity and produce effects directly on the phenotype (Lupiáñez et al., 2015; Spielmann and Mundlos, 2013).

To make sense of genome-wide contact data, and to explain the principles shaping chromosome 3D structure, models from polymer physics have been recently introduced (Barbieri et al., 2012; Brackley et al., 2016; Chiariello et al., 2016; Fudenberg et al., 2016; Giorgetti et al., 2014; Jost et al., 2014; Marenduzzo et al., 2006; Nicodemi and Prisco, 2009; R. K. Sachs, G. Van Den Engh, B. Trask, H.

Yokota, 1995; Rosa and Everaers, 2008; Sanborn et al., 2015; Tiana et al., 2016). This is an innovative and fascinating research field at the confluence of physics and biology.

In this framework, the research presented in my Ph.D. thesis has been developed. It has been conducted under the supervision of Professor Mario Nicodemi, in the group of Complex Systems at the Physics Department of the University of Naples "Federico II". Many results have been published or are currently under development in collaboration with the Epigenetic Regulation and Chromatin Architecture group directed by Prof. Ana Pombo, at Max Delbruck Centre For Molecular Medicine (Berlin), and the Development and Disease Group directed by Professor Stefan Mundlos, at the Max Planck Institute for Molecular Genetics (Berlin).

The thesis is organized in four principal chapters and one Chapter of Appendices. In Chapter 1, we highlight the importance of genome spatial organization and we briefly recall some basic concepts from biology, needed for the comprehension of this research activity, as the Chromosome Conformation Capture (3C) techniques, the interpretation of genome interaction data and the relationship between spatial organization and cell functionality. In Chapter 2, we describe a polymer physics model developed in our group to make sense of the complex pattern of genomic interactions and to explain, in a quantitative way the interaction network emerging from Hi-C contact data. In particular, we show that scaling concepts of classical polymer physics explain the large-scale behavior of contact data over three orders of magnitudes in genomic separation, across different cell types and chromosomes; we present a theoretical study of the multiple co-localization contact landscape and, finally, we schematically model the mechanisms underlying the self-assembly of topological domains. In Chapter 3, we introduce a more complex polymer physics model, by which we can reconstruct, with good accuracy, the 3D organization of real genomic regions. Finally, in Chapter 4, we test if such polymer model predicts the effect on chromatin architecture of structural variants (SVs), such as deletions, duplication or inversion. We show how polymer modeling emerges, in this scenario, as a valid approach for predicting pathogenic effects, facilitating the interpretation and diagnosis of this type of genomic rearrangements. In Appendix A, we show a direct comparison of different polymer physics models, which have shown to have an important role in explaining chromatin spatial organization.

References

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 and binders switch model. Proc. Natl. Acad. Sci. *109*, 16173–16178.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Bickmore, W.A., and Van Steensel, B. (2013). Genome architecture: Domain organization of interphase chromosomes. Cell *152*, 1270–1284.

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

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

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

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

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852–852.

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, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950–963.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Marenduzzo, D., Micheletti, C., and Cook, P.R. (2006). Entropy-driven genome organization.

Biophys. J. *90*, 3712–3721.

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

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

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 219683.

R. K. Sachs, G. Van Den Engh, B. Trask, H. Yokota, and J.E.H. (1995). A random-walk / giant-loop model for interphase chromosomes. Proc. Natl. Acad. Sci. USA *92*, 2710–2714.

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

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

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

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

Tanay, A., and Cavalli, G. (2013). Chromosomal domains: Epigenetic contexts and functional implications of genomic compartmentalization. Curr. Opin. Genet. Dev. *23*, 197–203.

Tiana, G., Amitai, A., Pollex, T., Piolot, T., Holcman, D., Heard, E., and Giorgetti, L. (2016). Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains. Biophys. J. *110*, 1234–1245.

Chapter 1: Three-Dimensional organization of the genome

The spatial architecture of the genome in the nucleus of eukaryotic cells is the principal way by which cells regulate their biological functions, such as transcription and regulation of gene expression. Although the link between 3D genome architecture and regulation of cellular functions is still not completely known, novel experimental protocols have been developed in the last years, and are still now, to investigate deeply and in a more quantitative way this open question.

In this first chapter, we briefly give an overall overview of the genome architecture problem, introducing the most important results achieved during the last decade in this research field, and the experimental methods that made possible to obtain them. This will help the comprehension of our research activity, described in more detail in the following chapters. In **Section 1.1** and **Section 1.2**, we summarize some fundamental concepts of molecular biology and recent advances about gene regulation and epigenetics. In **Section 1.3,** we introduce the fundamental technologies which have allowed to investigate the spatial organization of genomes. Finally, in **Section 1.4**, we summarize important findings obtained from the analysis of interactions data, provided by the described experimental technologies, and we briefly discuss the scenario that is now emerging, also thanks to the help of polymer physics models.

The results described in this chapter have been introduced and discussed in the papers from (Dekker et al., 2013; Dixon et al., 2012; Fraser et al., 2015; Lieberman-Aiden et al., 2009; Lupiáñez et al., 2015; Nora et al., 2012; Rao et al., 2014).

1.1 The genome in the cell nucleus

DNA (deoxyribonucleic acid) is a molecule carrying all genetic instructions needed for the cell development. DNA is a double helix of two chains, each made of monomer units, called nucleotides, bound to one another in the chain filament by covalent bonds. A nucleotide is composed of a sugar called deoxyribose, a phosphate group and one of four nitrogen-containing nucleobases, that are cytosine (C), guanine (G), adenine (A) or thymine (T). For making the double-stranded DNA, the nitrogenous bases of the two chains are bound together by hydrogen bonds, according to base pairing rules: A with T and C with G. The sequence of these four nucleobases encodes the genetic information. The DNA strand has a directionality that is defined by the orientation of the 3′ and 5′

carbons along the sugar-phosphate backbone. The two strands of double-helix structure run in opposite directions to each other and are thus antiparallel.

In the nucleus of the eukaryotic cells, the DNA is always associated with a variety of proteins, called histones, whose principal function is to package the DNA filament in a more compact way. Histones have also further functions, such as to control the gene expression, to prevent DNA damages and drive the DNA replication. The complexity of the chromatin packing allows, for instance, to include the entire mammalian genome, which would have a linear length of about 2 meters, into a nucleus of roughly 5÷15 μm diameter. The complex made of DNA and proteins is called chromatin. The basic units of the chromatin packing are the nucleosomes. Each nucleosome consists of a structure of eight histone proteins (consisting of two copies of each histone H2A, H2B, H3 and H4) and of 1.7 times wrapped DNA of approximately 146 base pairs (bps).

At a higher level, the chromatin is organized in chromosomes, each one restricted in a specific region, called chromosomal territory (CT), clearly visible using microscopy techniques, as shown in **Figure 1.1** (Cremer and Cremer, 2001). The total length of genome, which depends on the number of chromosomes and on the number of copies for each chromosome (named ploidy), varies across the different species. For instance, the human cells, as well as the most part of the eukaryotic cells, have two different copies per chromosome (diploid cells) and 23 different chromosomes, amounting to 6.4×10^9 base pairs. Each chromosome contains several hundreds of thousands of nucleosomes, and each nucleosome is separated from the next one by a filament of linker DNA, long up to about 80 bps. When viewed by microscopy, chromosomes assume the appearance of a string of beads where the beads are nucleosomes and string is the linker DNA.

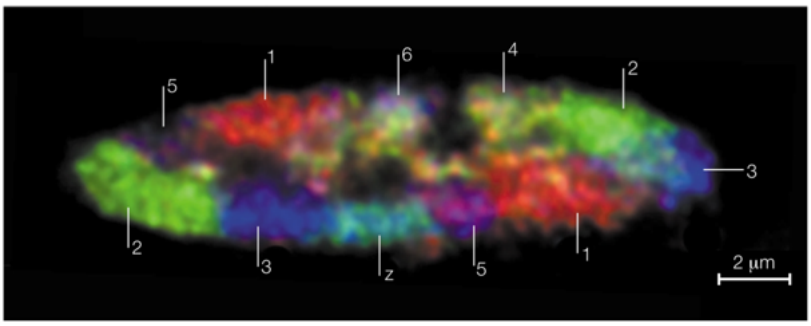

Figure 1.1: Chromosomes territory organization.

Microscopy image of the nucleus of chicken cell showing the organization of chromosomes in distinct regions, called chromosomal territories (CTs), each colored in a different way. Within the chromosomes, the chromatin at lower scale has a very complex three-dimensional organization, strictly related on gene expression of cells. However, this structure is still unknown and represent an open problem in modern biology. Figure adapted from (Cremer and Cremer, 2001).

Within the nucleus, the chromatin is found in two different structural forms, which play an important role in gene expression: heterochromatin and euchromatin. These forms were distinguished from the the different degree of compaction of "beads on a string" structure. Heterochromatin is a condensed form of chromatin, where nucleosomes are tightly packed, by making the DNA hardly accessible to the polymerase and therefore lowly transcribed. It is usually localized to the periphery, near the nuclear lamina. On the other hand, the euchromatin, that is the most part of the genome (up to 90%), is lightly packed DNA and is characterized by a high level of transcription.

1.2 Epigenetics and gene regulation

In a multicellular organism, all the cells share the same genome. Each specific cell line in the organism expresses a subset of all genes making up the genome of the species. During cellular differentiation, i.e., the process where a cell passes from one cell type to another, a change from one pattern of gene expression to another occurs. The transcriptional activity of a gene is regulated by a genomic region (long about 100/1000 bases), located near the transcription start site (TSS) of the gene, named promoter. Promoters provide an initial binding site for the transcriptional machine, including RNA polymerase and transcription factors. The gene activity can also be controlled by additional regulatory regions of DNA, named enhancers, which increase the probability of gene transcription. Enhancers have a key role in driving cell type-specific gene expression, and they can activate transcription of their target genes at great genomic distances, ranging from several hundreds, until to even thousands of bases (Bulger and Groudine, 2011; Calo and Wysocka, 2013; Ong and Corces, 2011; Phillips-Cremins et al., 2013).

By the term "epigenetics", we indicate all the features that affect the gene activity and its expression, without involving changes in genome sequence. Examples of mechanisms that produce such changes are DNA methylation and histone modification, each of which alters how genes are expressed without altering the underlying DNA sequence. Additionally, the gene expression can be controlled through the action of repressor proteins that attach to silencer regions of the DNA. In last years, new technologies have been developed to analyze genome-wide epigenetic modifications at base-pair resolution. Among these, chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq), a method that allows to identify the binding site along the DNA associated with a specific protein, by using specific antibodies that target the protein of interest. Such techniques allow to identify some consistent patterns of histone marks, used for example to better define DNA regulatory regions (Allis and Jenuwein, 2016).

1.3 Chromosomes Conformation Capture and further technologies

As discussed in the previous Sections, the three-dimensional organization of the chromatin has a key role in regulation of gene expression. It consists, indeed, of a complex network of contacts between genes and its corresponding regulatory elements, which allows the genes to be expressed or silenced. In the following Sections, we briefly describe two different molecular approaches that have been developed in last years to study the three-dimensional folding of chromosomes with increasing accuracy. On one side, the new genome-wide methods that allow to estimate the mean frequency of contact for any pairs of genomic regions. In particular, we focus on the methods based on the chromosome conformation capture (3C) technique. In this approach, two loci are considered in contact if their physical distance is close enough to become crosslinked (typically, around 100nm). On the other side, the FISH technique, an independent and conceptually different method, which enables to estimate the physical distances between a limited number of genomic regions at single cell level. Since that such information are not accessible with 3C-based methods, and a direct comparison is not trivial, they can be powerfully combined to bring comprehensive insights into genome folding.

1.3.1 3C-based techniques: 3C, 4C and 5C

In last decade, several experimental techniques have been developed to deeply investigate the three-dimensional organization of chromatin in mammalian cell nucleus. These approaches are based on chromosome conformation capture (3C) technique and allow to estimate the frequency of interaction, across a population of cells, between different genomic regions, which could be physically close even if separated by several nucleotides along the linear genome. Such interactions have a key role in gene expression during cellular differentiation, as, for instance, to drive the interaction between enhancer and promoter (see, **Section 1.2**).

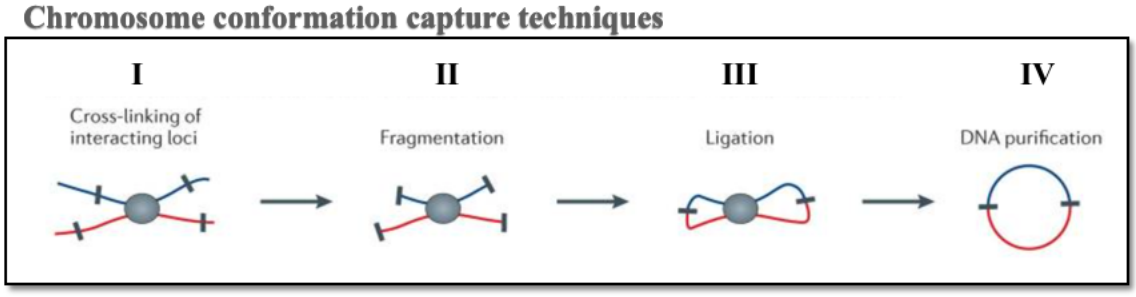

Figure 1.2: Chromosome conformation capture techniques.

Schematic representation of experimental steps that characterize the 3C methods. First (**Step I**), the chromatin is crosslinked with formaldehyde to create covalent bond between pairs of loci spatially close. Next, they are fragmented (**Step II**), ligated (**Step III**) and finally purified (**Step IV**). Figure adapted from (Dekker et al., 2013).

All 3C-based methods are characterized by following steps (schematically shown in **Figure 1.2**):

I) Chromatin in nucleus is cross-linked by formaldehyde, generating covalent bonds between different genomic regions which are physically close in the space;

II) By using restriction enzymes (e.g., HindIII, NcoI) during a digestion process, cross-linked chromatin is fragmented;

III) Cross-linked fragments are ligated to form a unique DNA molecule;

IV) DNA is purified and pairwise interactions are quantified,

The 3C (Dekker et al., 2002) and 4C (Simonis et al., 2006) techniques detect the interactions involving a specific genomic region. The 3C method identifies the interactions for a single pair of loci and can be used to test a candidate of interacting pair, e.g. enhancer-promoter (**Figure 1.3, Panel a**).

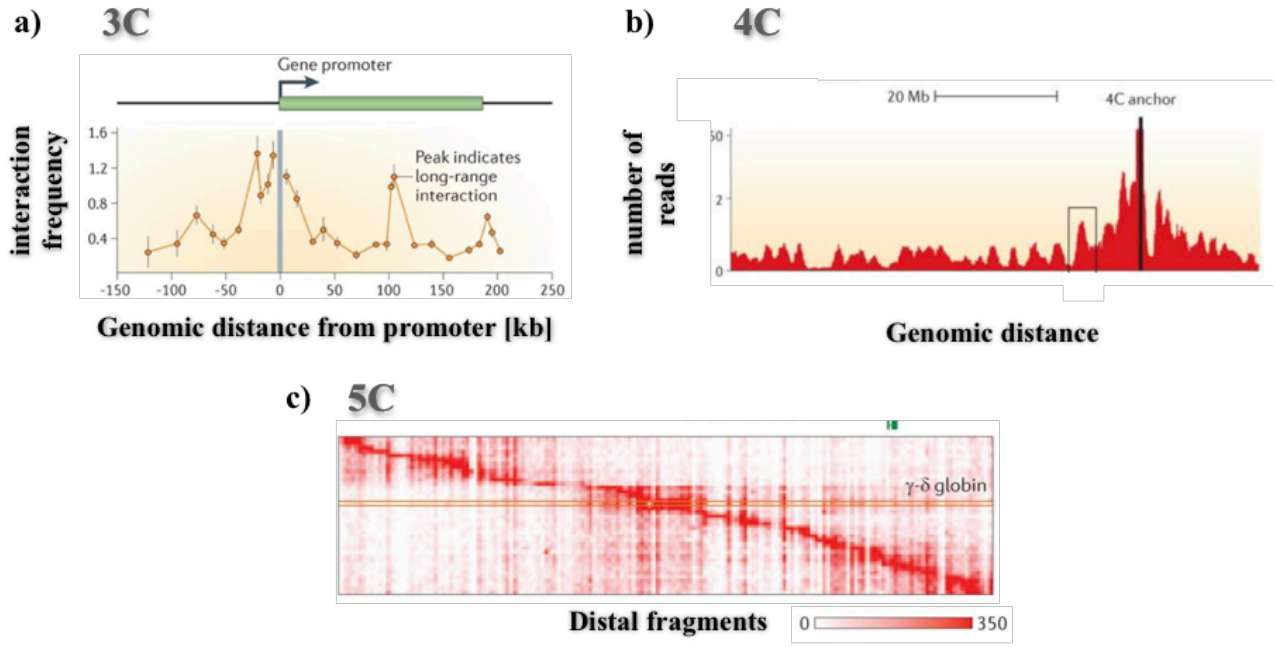

Figure 1.3: Comparison of output for different 3C-based techniques.

a) Example of Chromosome Conformation Capture (3C) data. The horizontal axis indicates the genomic distance from the anchor point, or point of view, here indicated by a grey line. **b)** Example of 4C data, where the anchor point is indicated, by a black line. **c)** Example of 5C interaction data for the ENCODE ENm009 region in K562 cells. Here, each row represents the interaction profile of a transcription start site (TSS) across the 1 Mb region on human chromosome 11 that contains the beta-globin locus. Figures adapted from (Dekker et al., 2013).

On the other hand, the 4C technique allows to quantify the interaction profile of a given locus with all the surrounding genomic regions (**Figure 1.3, Panel b**). This not requires the a-priori knowledge of both interacting loci. Finally, by 5C method (Dostie et al., 2006) instead, we can detect the interactions between all the pairs of loci within a given genomic region, which typically is no longer than a single mega-base (**Figure 1.3, Panel c**)

1.3.2 Hi-C technique

The Hi-C method was the first genome-wide extension of 3C techniques, which made possible to detect long-range interactions (Lieberman-Aiden et al., 2009). In this approach, once the cells are fixed by formaldehyde (binding the interacting regions by covalent cross-link), fragmented and ligated (as discussed in **Section 1.2.1**), the staggered DNA ends are filled in with biotinylated nucleotides. In this way, it results a genome-wide collection of ligation products, corresponding to pairs of chromatin fragments that were spatially close in the nucleus. Each of these ligation products is marked with biotin at ligation junction. The library is then sheared, and the junctions are pulled down from biotin. The purified junctions are then directly sequenced along the genome, generating a list of interacting fragments. Finally, the genome is divided into windows of fixed length, which defines the Hi-C data resolution (**Figure 1.4**). The resolution depends on depth of sequencing and on data quality: in the first experiments the resolution was 1 Mb but recent Hi-C or Hi-C-derived, e.g. in situ Hi-C (Rao et al., 2014) or cHi-C (Jäger et al., 2015), experiments can reach 1 kb of resolution.

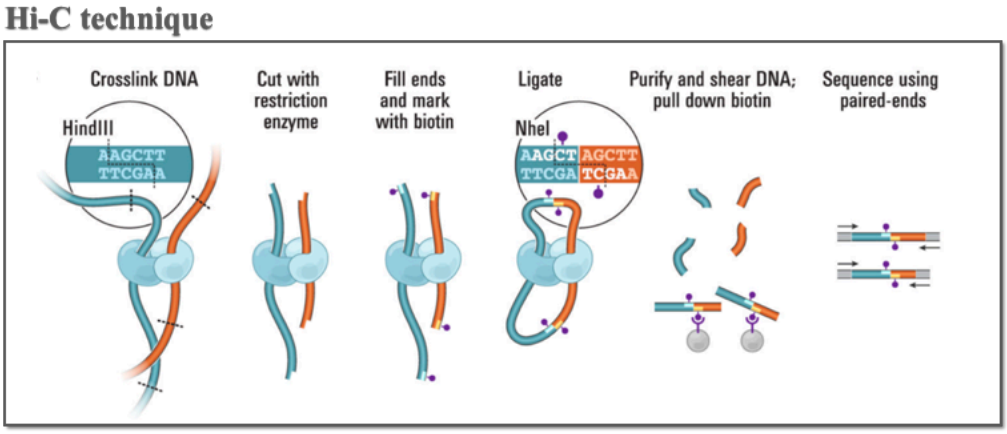

Figure 1.4: Hi-C method description.

Schematic representation of the Hi-C method protocol. The biotinylated junctions allow to efficiently detect the ligated fragments genome-wide. Figure adapted from (Lieberman-Aiden et al., 2009).

The Hi-C data are organized in contact matrices, where each bin x_{ij} represents the frequency of contact (i.e., the number of pairs fragments) between the region i and the region j of genome. By definition,

x_{ij} is equal to x_{ji}, and the contact matrix is symmetrical. Since Hi-C is able to detect loci belonging to the same chromosome or to different chromosomes, in the first case we say *cis*- data and the associated matrix is squared by definition, while in the second one we say *trans*- data. In our work, we focus on *cis*- data (as for example in **Figure 1.5, Panel a**). Fixed a given genomic window (for instance, the entire chromosome), the size of contact matrix depends on data resolution: higher is the resolution, bigger is the size of contact matrix.

Importantly, in Hi-C contact matrix, a bin x_{ij} represents the interaction frequency averaged over a large population of cells. However, chromatin conformations, which are determined by several different factors, can have three-dimensional structures highly variable. Recently, single-cell Hi-C (scHi-C) technologies have been developed to investigate at single-cell level the 3D architecture of chromatin (Nagano et al., 2013; Stevens et al., 2017), revealing a high cell-to-cell variability. Such new methods provide a new approach to investigate these biological processes.

1.3.3 Ligation independent methods: GAM and SPRITE

As discussed in **Section 1.2.1**, the 3C methods developed to investigate genome-wide contacts are based on proximity ligation, which creates covalent bonds between regions spatially close. However, these technologies often fail to detect chromatin regions too far apart to directly ligate, although it has been proved that they have an important role in genome organization, as for example the nuclear bodies (Quinodoz et al., 2018). For this reason, two alternative ligation-free methods have been recently developed for more comprehensively understanding genome organization: Genome Mapping Architecture (GAM) and Split-Pool Recognition of Interactions by Tag Extension (SPRITE). Besides, these new approaches have made possible to investigate, besides pairwise interactions, also multi-way contacts, such as triplets, quadruplets, etc., which can help to shed light on the complexity of genome organization.

GAM (Beagrie et al., 2017) was the first genome-wide technology which allows to detect interactions between pairs of loci, without ligation process. Starting from a collection of slices obtained cryo-sectioning a population of nuclei in random directions, it is possible to estimate the frequencies of interaction between pairs of loci. The new idea is that two loci, which are frequently co-segregated in the same slice, will be also physically close in three-dimensional space (what would be not expected if the loci were independent and associated randomly). GAM technique summarizes the same results for interacting pairs found by Hi-C and, additionally, allows to investigate also multi-way contacts, helping to investigate the complex pattern of interactions characterizing genome organization. Furthermore, GAM enables the investigation of 3-d genome conformations at single-

cell level. To facilitate the comparison among these different techniques, an example of GAM contact matrix is shown in **Figure 1.5, Panel b,** for the same genomic region considered for Hi-C case.

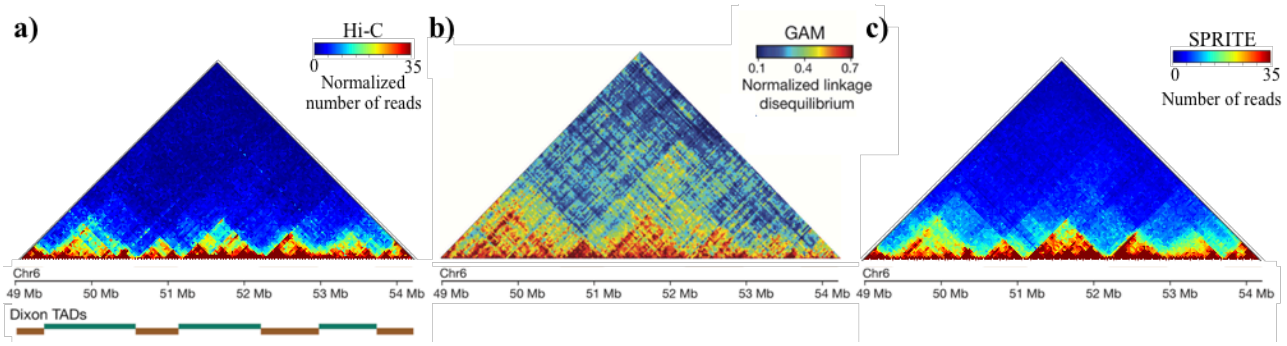

Figure 1.5: Comparison of average contact matrix from three different technologies.

Contact matrices from Hi-C, GAM and SPRITE technologies, for a genomic region on chromosome 6 (49 - 54.2 Mb) in mouse embryonic stem cells: **a)** Hi-C contact matrix from (Dixon et al., 2012). On bottom, we reported the TADs positions. **b)** Example of GAM matrix, where instead of co-segregation matrix, is reported the normalized linkage disequilibrium. The matrix shows a similar interaction pattern and an enrichment of long-range contact. Figure adapted (Beagrie et al., 2017). **c)** SPRITE matrix shows a pattern similar to those shown in Hi-C and GAM. (Quinodoz et al., 2018)

On the other hand, SPRITE (Quinodoz et al., 2018) is more similar approach to 3C-based methods, but it does not use the ligation process as well. After that chromatin is crosslinked and fragmented, the interacting molecules in a cluster are barcoded by using a split-pool strategy. Interactions are identified by sequencing and matching all the reads having the same barcode. The cluster obtained in this way are then converted in contact frequencies by counting all the contacts observed in a single cluster and weighting each contact by the total number of the molecules contained within the cluster. An example of contact matrix from SPRITE is shown in **Figure 1.5, Panel c**.

1.3.4 FISH technique

Fluorescent *in situ* hybridization (FISH) is a molecular technique that enables the measurement of physical distance between two target loci at single cell level, as it also allows to quantify the distribution of these distances across a cell population (Jefferson and Volpi, 2010). In FISH method, fluorescent probes bind target that can then be directly visualized using fluorescence microscopy, enabling its localization to be assessed in the context of the overall nuclear architecture and/or with respect to other genomic loci. Then, by indicating two targeted loci a and b, it is possible to estimate the associated probability distribution $P(r_{ab})$ measuring the variation of distance r_{ab} across the cell

population. Notably, this type of measure allows us to quantify the degree of variability of physical distances between pairs of genomic loci (**Figure 1.6**).

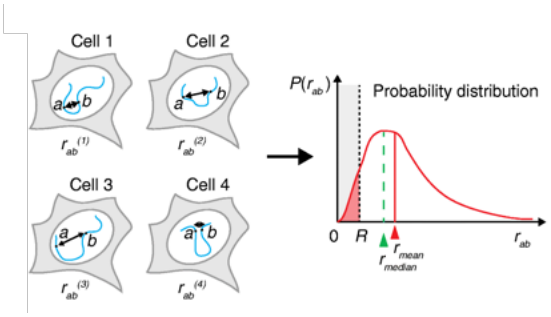

Figure 1.6: FISH experiments.

Starting from a population of cells, FISH method enables the measurement of cell-to-cell variability in distance r_{ab} between two genomic loci *a* and *b*. Indeed, the knowledge of probability distribution of distances $P(r_{ab})$ allows, for instance, to compute the mean (and median) distance, or the fraction of cells for which the distance r_{ab} is smaller than a certain threshold *R*. Figure adapted from (Giorgetti and Heard, 2016).

1.4 Nuclear organization of chromatin

By analyzing 5C and Hi-C data, some fundamental features of chromatin structure were discovered. In the following Sections we summary the most important findings achieved in this field during the last years.

1.4.1 Chromatin loops

A chromatin loop event occurs when two genomic regions on the same chromosome (in *cis*-), are brought close in physical space. This mechanism allows to bring together in 3-d space two regions that could be event apart along the chromosome. It is biologically driven by a number of architectural proteins, such as cohesin, transcription factor, etc. and represents the fundamental mechanism driving gene activation, since chromatin loops can be formed between gene promoter with one (or more than one) enhancer region, even if located up to 1Mb away from the gene (downstream or upstream from TSS position). As discussed in **Section 1.2**, the physical proximity between the gene and its enhancers increases the probability that transcription of the gene could occur. In human genome, about one-half of genes are involved in long-range chromatin loops.

Physical interactions have been also detected between regions falling on different chromosomes. Although they are not loops in a narrow sense of the term, these interactions show similar features. However, the exact mechanism of loops formation in not still understood.

1.4.2 A/B compartments

Through the principal component analysis (PCA) of Hi-C contact matrices (Lieberman-Aiden et al., 2009), and consequently confirmed by independent FISH experiments, it was discovered that the entire genome could be divided into two different classes of regions, named "A" and "B" compartments. Genomic regions in the same compartment tend to interact preferentially with regions belonging to the same compartment, rather than to regions associated with the other compartment (**Figure 1.7, Panel a**).

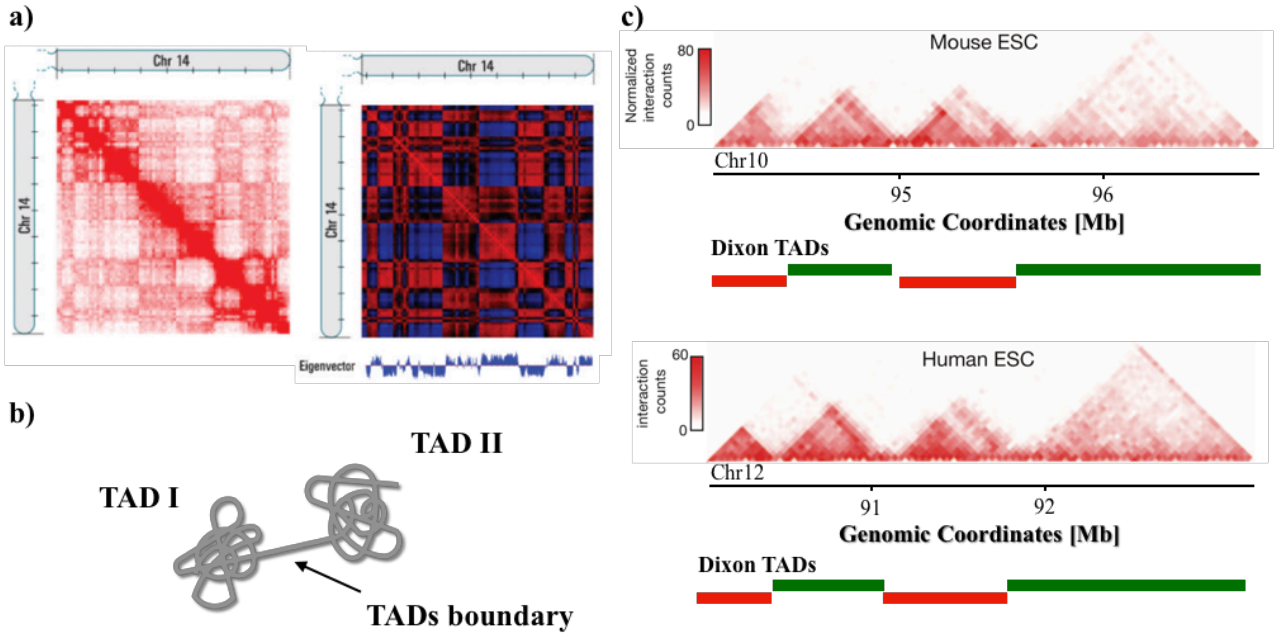

Figure 1.7: Compartments A/B and TADs organization of the genome.

a) On left, an example of genome-wide Hi-C contact matrix for chromosome 14 from a karyotypically normal human lymphoblastoid cell line. On right, the Pearson correlation matrix of the same chromosome, and the principal component associated analysis. This last panel shows that PC correlates the checked pattern in matrix, which respectively defines A (positive values) and B compartment (negative values). **b)** Schematic cartoon showing TAD organization of chromosomes in the cell nucleus. Genomic regions within the same TAD interact each other much more frequently than regions belonging to a different TAD. TADs are separated by genomic region, called "boundary". **c)** Comparison of Hi-C matrix over a systemic region for mouse (top) and human (bottom) stem cell, and corresponding TADs positions. The TADs are highly conserved across the different species. Figures adapted from (Dixon et al., 2012; Lieberman-Aiden et al., 2009).

A/B compartment-associated regions have typical size of some Mb (5÷10) and correlate with eu- and hetero-chromatin respectively (**Section 1.1**). While the A compartment tends to be less compact and

enriched of genes, correlating with higher expression and accessible chromatin, the B compartment, instead, tends to be more compact and gene-poor, with higher interaction values. The presence of A/B compartments is in full agreement with the known presence of open and closed chromatin in the cell nucleus.

1.4.3 Topologically Associated Domains

Besides the A/B compartments, chromatin shows a lower level of structural organization. Recent findings have shown that genome is organized in self-interacting domains (Dixon et al., 2012), called in literature Topologically Associating Domains (TADs) (Nora et al., 2012). The principal feature of TADs is that regions within a domain interact most frequently with regions within the same domain, rather than regions outside it (schematic cartoon in **Figure 1.7, Panel b**). Typically, TADs have size of about 0.5÷1 Mb (then smaller than the A/B compartments) and they are formed through the interaction of architectural proteins with DNA, which gives rise to several chromatin loops within it. TADs strictly correlate with regulation of gene expression, since they can be associated with active or inactive transcription and are almost conserved between different species (**Figure 1.7, Panel c**). As both mouse and human are composed by more than 2000 domains, covering almost all the genome, TADs are found to be universal building blocks of chromosomes. In Hi-C contact matrix, a TAD appears as a square along the principal diagonal, characterized by high interaction level of interaction (**Figure 1.6, Panel a bottom**). By using this observation, different computational algorithms have been developed to identify TADs from experimental data (Dixon et al., 2012; Fraser et al., 2015; Oluwadare and Cheng, 2017; Rao et al., 2014).

TADs represent physically isolated units along the genome, characterized by two distinct functional features: the regulation of genes within them, that allows chromatin interaction among loci within the same domain, and the separation of gene activity of two neighbouring TADs. Recent studies have shown that deletion of TAD boundary can lead to ectopic expression of several developmental regulator genes during limb formation, and to several congenital diseases. However, the mechanism that regulates the formation of TADs is still not clear, and several polymer models have been developed to help to quantitatively describe them (Barbieri et al., 2012; Bianco et al., 2018; Brackley et al., 2017; Chiariello et al., 2016; Fudenberg et al., 2016; Sanborn et al., 2015). Some of them (Brackley et al., 2017; Fudenberg et al., 2016; Sanborn et al., 2015) are based on observed enrichment of CTCF binding proteins at TAD boundaries (Rao et al., 2014), proving that CTCF is an important insulating factor in mammalian cell. However, such models do not take into account other possible factors, that have an important role in the formation of these domains (Barbieri et al., 2017; Dixon et al., 2016; Kundu et al., 2018; Yan et al., 2017).

1.4.4 Besides TADs: the meta-TADs structures

In the previous sections, we showed that chromosomes are organized in megabase-sized self-interacting domains, named TADs, which are arranged at a higher order level in A/B compartments, i.e., in nuclear domains enriched of active or repressed chromatin states. However, this scenario is too simplistic to efficiently explain the complex pattern of interactions found in the Hi-C data. As visible from data (**Figure 1.8, Panel a**), TADs (indicated by Arabic numbers) in turn, interact with each other at higher-order level of interaction, giving rise to a hierarchical structure of domains-within-domains, called meta-TADs (Latin number), extending across genomic scale up to the entire chromosome length (Fraser et al., 2015). This structure can be well investigated, whatever the cell type (human or mouse), by a tree-like structure (**Figure 1.8, Panel b**). The meta-TADs organization has been proved to correlate with several epigenomic features, and its changes during cell differentiation correlate with transcriptional state of the cell. Therefore, these hierarchical structures seem to have an important role in chromatin compaction and help the chromatin to re-organize itself and to activate or silence a specific genomic region, according to transcriptional state of the cell.

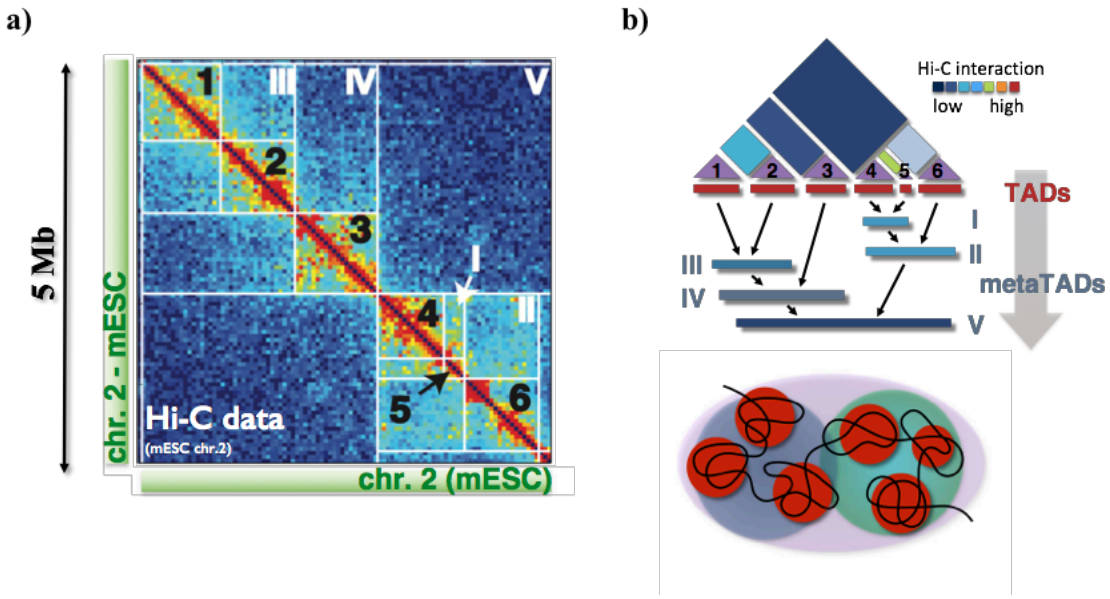

Figure 1.8: Meta-TAD structure.

a) Contact matrix Hi-C of chromosome 2 (53-58 Mb) from mouse embryonic stem cell, where we indicate the TADs by Arabic number. How appears clearly in the matrix, TADs, in turn, can interact with each other giving rise to higher order structures, the meta-TADs, that are here indicated by Latin numbers. **b)** Starting from matrix Hi-C, meta-TAD can be identified by single-linkage clustering. Figures adapted from (Fraser et al., 2015).

Within TADs, in turn, there are smaller interacting domains, generally called "sub-TADs" (Phillips-Cremins et al., 2013; Rao et al., 2014). They seem to have similar features of TADs, but, on the contrary, they consistently differ across different cell lines. Even in this case, cell-type specific organization of sub-TADs appears to be related to cell-type specific regulatory events, as for instance, driving the activation of a gene (**Figure 1.9**).

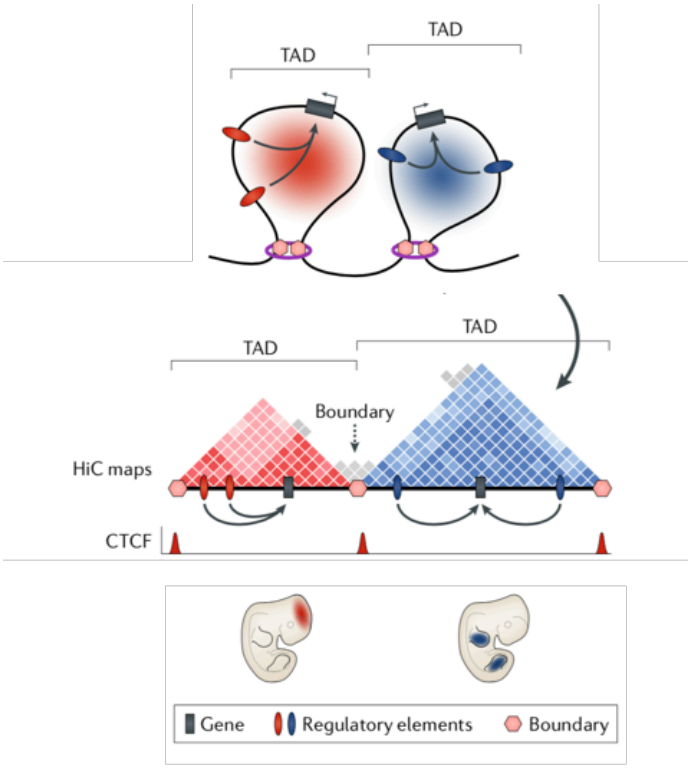

Figure 1.9: TADs and sub-TADs organization.

Boundary regions (in brief, "boundary") separating two TADs, are generally enriched in architectural proteins, such as transcriptional repressor CTCF. Hi-C data at high resolution have revealed the existence of lower level of chromatin organization, such as sub-TADs, which are smaller spatial domains (around 100 kb) that display a more dynamic nature and tissue specificity. Both TADs and sub-TADs can be schematized as loops, sometimes associated with enhancer-promoter interaction; they are bound by mediators and the cohesin complexes, displaying a dynamic and tissue-specific nature. The CTCF–cohesin complex proteins play a key part in the looping process, which explains some of the features observed in Hi-C interaction map. Figures adapted from (Spielmann et al., 2018).

1.4.5 Interpretation of the structural data and further information

Spatial proximity between different genomic regions can be the result of specific contacts mediated by protein complexes bridging them, or co-proximity near the same nuclear structure (i.e., nucleolus, nuclear lamina, etc.). All the experimental methods we described in previous **Section 1.3**, give information about the relative frequency of contact across a population of cells between pairs of loci. However, these do not give information about specificity of contacts; indeed, they do not distinguish the functional associations from non-functional ones, that could be caused by random collisions between different genomic regions, and made possible by chromatin flexibility. Similarly, they cannot even individuate what are the mechanisms driving the chromatin folding, that are still completely unknown.

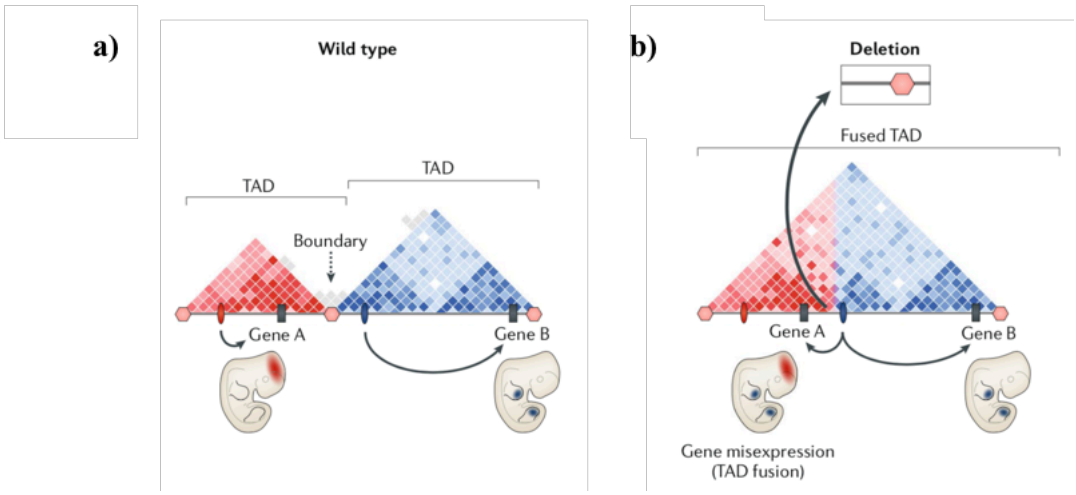

Figure 1.10: Boundary deletion causes a TAD disruption in Epha4.

a) In the schematic representation of the wild-type genomic locus, gene A is expressed in the developing brain and gene B in the developing limbs. Both genes are regulated by their own tissue-specific *cis*-regulatory elements (red and blue, respectively) located in different TADs separated by boundary elements. **b)** An inter-TAD deletion of a boundary element can cause TAD fusion and enhancer adoption; the relocation of enhancer elements into a neighbouring TAD causes mis-expression and disease. Through the deletion of the boundary, the enhancer of gene B (blue) is free to act on gene A, driving ectopic expression in the developing limbs. Figures adapted from (Spielmann et al., 2018).

In order to shed light on these mechanisms of chromatin organization and to investigate the impact on health of its structural alteration, more and more experiments have been recently performing (Franke et al., 2016; Lupiáñez et al., 2015; Spielmann et al., 2018). These research activities have found that structural variations, such as duplications or deletions, involving even little genomic regions can be pathogenic and they often cause of congenital diseases (as shown, for instance, in **Figure 1.10**). Additionally, they have proved that high levels of structural variations are linked to human cancer genome. These results prove how chromatin organization in space and phenotype are very closely related, and the knowledge of mechanisms by which that structure is regulated is fundamental to prevent and recover congenital diseases. In the following chapter, we show how, by polymer physics approach, is possible to predict the effect of 3D spatial organization, due to structural variations, using as input information the data available for healthy subjects.

References

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. *17*, 487–500.

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 and binders switch model. Proc. Natl. Acad. Sci. *109*, 16173–16178.

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., De Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *24*, 515–524.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R., and Marenduzzo, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. Phys. Rev. Lett. *119*.

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

Calo, E., and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell *49*, 825–837.

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

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science (80-.). *295*, 1306–1311.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

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

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell *62*, 668–680.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D.,

Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schï¿½pflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852–852.

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., and Heard, E. (2016). Closing the loop: 3C versus DNA FISH. Genome Biol. *17*.

Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H.E., Heindl, A., Whiffin, N., Carnicer, M.J., Broome, L., Dryden, N., et al. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat. Commun. *6*.

Jefferson, A., and Volpi, E. V. (2010). Fluorescence in situ Hybridization (FISH) for Genomic Investigations in Rat (Humana Press, Totowa, NJ).

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2018). Erratum: Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation (Molecular Cell (2017) 65(3) (432–446.e5) (S1097276517300357) (10.1016/j.molcel.2017.01.009)). Mol. Cell *71*, 191.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Oluwadare, O., and Cheng, J. (2017). ClusterTAD: An unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC Bioinformatics *18*.

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.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 219683.

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

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Yan, J., Chen, S.-A.A., Local, A., Liu, T., Qiu, Y., Lee, A.-Y., Jung, I., Preissl, S., Rivera, C.M., Wang, C., et al. (2017). Histone H3 Lysine 4 methyltransferases MLL3 and MLL4 Modulate Long-range Chromatin Interactions at Enhancers. BioRxiv 110239.

Chapter 2: 3D chromatin investigation by polymer physics models

In the previous Chapter, we described the complex architecture of the chromatin in the nucleus of cells. To make sense of genome-wide contact data and to expose the principles shaping three-dimensional structure of chromosomes, several theoretical models have been developed from polymer physics. For the sake of completeness, we briefly describe some of these models, recently proposed, which have had a key role in quantitatively explaining the spatial organization of chromosomes.

Initially, as a possible structure of chromatin in the nucleus, the Fractal Globule model was proposed (van Berkum et al., 2010; Lieberman-Aiden et al., 2009; Mirny, 2011). Here, a compact polymer not-equilibrium state emerges during polymer condensation due to topological constrains and prevents one genomic region to pass across another one. This model was independently introduced in (Grosberg et al., 1988), but experimental evidences in biology were not found until the Hi-C paper (Lieberman-Aiden et al., 2009). Shorty later, another model was introduced, named Dynamic Loop model (Bohn and Heermann, 2010), where chromatin moves under diffusional motion and functional loops can be formed when two specific sites co-localize, thanks to the presence of mediating proteins, such as CTCF or transcription factors (TFs). These loops can be formed with a certain probability, and dissolve after a certain lifetime. Another important model, introduced in (Jost et al., 2014), tries to link structural and epigenetic information: starting from 1D epigenetic data, it is possible to associate to each chromatin region a specific epigenetic state. This model can explain TADs formation by introducing a specific interaction between regions characterized by the same epigenetic state. A similar approach has been later used to explain chromatin folding at chromosomal scales (Di Pierro et al., 2016). At the moment, however, two chromatin models are mainly considered: the String&Binders Switch (SBS) model (Nicodemi and Prisco 2009; Barbieri et al. 2012), that was also used in other independent studies (Brackley et al., 2013), and the Loop Extrusion (Fudenberg et al., 2016; Sanborn et al., 2015), together with the Slip-Link model (Brackley et al., 2017).

In this Chapter, we will focus on the SBS model, which is having an important role in genome 3-dimensional reconstruction and that we will use in more detail in this and the following Chapters for our considerations about chromatin architecture. In **Section 2.1**, we describe the SBS model, as introduced in (Barbieri et al., 2012; Nicodemi and Prisco, 2009), and how we implemented it for the

first time by using a Molecular Dynamics (MD) approach. In **Section 2.2**, we will discuss the resulting phase diagram for the polymer model, which shows novel thermodynamic stable states. In **Section 2.3** and **Section 2.4**, we show how, just using few parameters, besides recapitulating the average behaviour of chromatin folding at chromosomal scales, we are able to explain by the SBS model the formation of interacting domains and the hierarchical organization of higher-order structures of chromatin.

Most of the results shown in this chapter, including figures, paragraphs and sentences, is adapted or lifted verbatim from the following papers, which I co-authored: (Annunziatella et al., 2016, 2018, Chiariello et al., 2016, 2017).

2.1 String & Binders Switch (SBS) Model and its implementation by MD simulations

In the following Section, we describe in detail the SBS model (Barbieri et al., 2012; Nicodemi and Prisco, 2009) and how we implement it by a Molecular Dynamics (MD) approach, which is widely used in the computational community to investigate such models of chromatin. In the MD approach, the trajectory of each particle in the system is determined by numerically solving its equations of motion (e.g., by Verlet algorithm); the interaction with the other particles are taken into account by introducing appropriate potentials. Unlike the Monte-Carlo method, used, e.g., in (Barbieri et al., 2012), the MD approach allows to investigate not only the equilibrium properties of the system but also its dynamics. Our simulations are run via LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) (Plimpton, 1995), a MD program that is optimized for parallel computing, allowing to drop significantly the time of simulations.

2.1.1 The SBS model

In the *String&Binders* Switch (SBS) model, a chromatin filament (we call the "string") is represented as a self-avoiding walk (SAW) polymer chain made of consecutive beads. The beads interact with diffusing molecules (the "binders"), in solution at a given concentration c, which can bring two beads in physical proximity and loop the polymer. The scale of such interaction is indicated by E_{int}. The interaction between binders and polymer beads drives the folding of the chain. Different equilibrium thermodynamics phases exist according to the value of the control parameters, E_{int} and c, giving rise to specific, corresponding conformational classes. A schematic cartoon of the SBS model is represented in **Figure 2.1, Panel a**, in the simplest 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 "colours". (Annunziatella et al., 2016; Barbieri et al., 2012, 2017; Bianco et al., 2018; Chiariello et al., 2016)

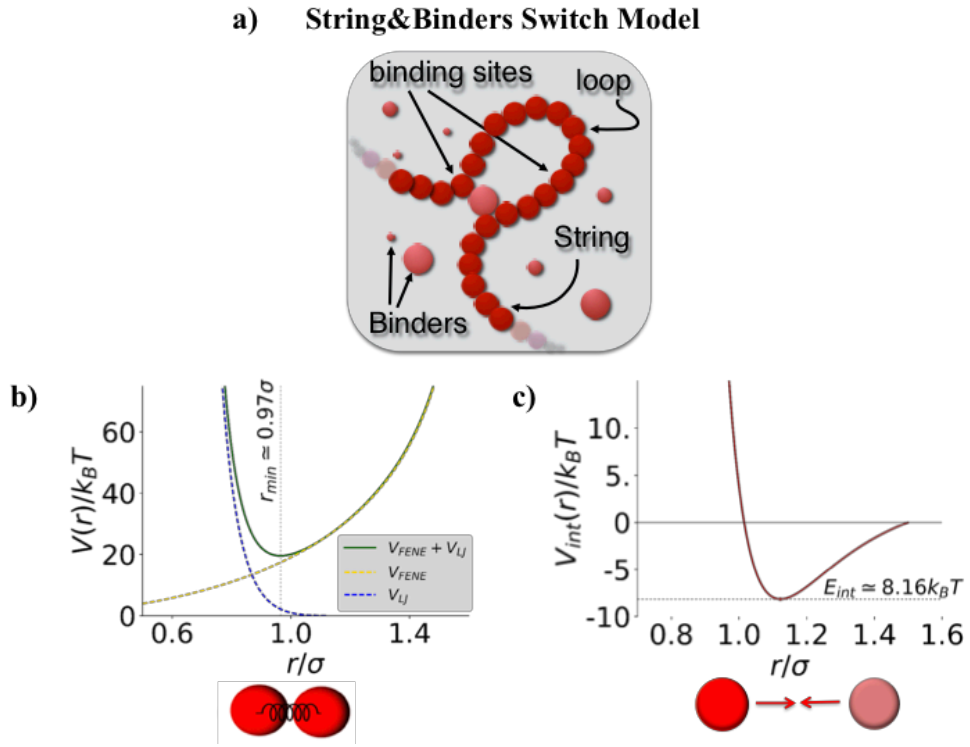

Figure 2.1: String&Binders Switch (SBS) polymer model describing chromatin folding.

a) In the String&Binders Switch (SBS) model chromatin folding is driven by the interactions between the polymer chain of beads (the 'string') and the binding molecules (called 'binders'). **b)** Interaction potential between consecutive beads making up the polymer chain. This is a combination of repulsive Lennard-Jones potential (V_{IJ}, blue dashed line) and FENE potential (V_{FENE}, yellow dashed line). Here, we set ε=1, σ=1, R_0=1.6σ and k_{FENE} = 30k_BT/σ^2. **c)** Attractive potential between chain bead and cognate binders $V_{int}(r)$, modeled by a truncated-shifted LJ potential. The absolute value of the minimum of $V_{int}(r)$ defines the scale of interaction energy, E_{int} (horizontal dashed line in figure), Here, we set ε_{int}=12k_BT, σ_{bb}= 1σ, and r_{int}= 1.5σ. Figures adapted from (Annunziatella et al., 2018; Chiariello et al., 2016).

2.1.2 The MD potentials

In our MD simulations, the SAW polymer chain is composed by N consecutive beads, having each a diameter σ. To model hard-core repulsion and prevent physical overlap among particles, between any two beads i and j we introduce a truncated Lennard-Jones (LJ) potential V_{LJ}, described by the following expression:

$$V_{LJ}=\begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12}-\left(\frac{\sigma}{r_{ij}}\right)^{6}\right]+\varepsilon & r_{ij}<2\sigma^{1/6} \\ 0 & \text{otherwise} \end{cases} \qquad (2.1)$$

where σ is the diameter of a bead, $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is the center-to-center distance and $\varepsilon = k_B T$ the strength of the potential (T temperature of the system and k_B Boltzmann constant). This is a continuous decreasing positive function of r_{ij} that becomes zero for $r_{ij} = 2^{1/6}\sigma$, as shown in **Figure 2.1, Panel b** (blue curve). Excluded volume effects between beads are taken into account by such term, which drastically hampers physical overlaps between beads.

To model the bond between two consecutive beads in the chain, an established approach (Kremer and Grest, 1990) considers that between any pair of consecutive beads there is the *finitely extensible nonlinear elastic* (FENE) potential V_{FENE}:

$$V_{FENE} = -\frac{k_{FENE} R_0^2}{2} \ln\left[1 - \left(\frac{|\mathbf{r}_{i+1} - \mathbf{r}_i|}{R_0}\right)^2\right] \qquad (2.2)$$

where $\boldsymbol{r}_i$ and $\boldsymbol{r}_{i+1}$ are the position of neighboring bead on polymer, k_{FENE} is the strength of the FENE spring and R_0 is its maximal extension. The FENE potential is close to a harmonic potential for values of the distance $r = |\boldsymbol{r}_{i+1} - \boldsymbol{r}_i|$ near to zero ($r \longrightarrow 0$) and diverges for $r \longrightarrow R_0$, which represent the maximal length of the bond (**Figure 2.1, Panel b**, yellow curve).

The resulting total potential, $V(r) = V_{LJ} + V_{FENE}$ (shown in **Figure 2.1, Panel b**, green curve), is a function whose minimum corresponds to the mean distance between consecutive beads on the chain. The value of the minimum depends on the potential parameters and is in general taken to be approximately equal to σ. Typical values for parameters used in the FENE potential are $k_{FENE} = 30 k_B T/\sigma^2$ and $R_0 = 1.5\sigma$, which have been also typically employed in other chromatin models (Brackley et al., 2013; Kremer and Grest, 1990; Rosa and Everaers, 2008).

The binding molecules (binders) are also modeled as hard-core particles, so they interact with any other bead or binder through the above LJ potential of equation (**Eq. 2.1**). Moreover, to model the attractive interaction between a binder and its cognate beads on the polymer, we use using the truncated LJ potential described above, where a higher cut-off value is used in order to include an attractive part in the potential. Hence, the attractive potential V_{int} between a diffusing binder and its cognate binding site on the polymer chain is:

$$V_{int} = \begin{cases} 4\varepsilon_{int}\left[\left(\frac{\sigma_{bb}}{r}\right)^{12} - \left(\frac{\sigma_{bb}}{r}\right)^{6} - \left(\frac{\sigma_{bb}}{r_{int}}\right)^{12} + \left(\frac{\sigma_{bb}}{r_{int}}\right)^{6}\right] & r < r_{int} \\ 0 & \text{otherwise} \end{cases} \qquad (2.3)$$

where σ_{bb} is the sum of bead and binder radii (for example, to model binders and beads having the same radius, $\sigma_{bb} = 1\sigma$), ε_{int} sets the attractive interaction intensity scale, r is the center-to-center distance between the binder and the polymer bead and r_{int} is the cut-off distance that sets the interaction range. As V_{int} goes to zero when $r = r_{int}$, in this framework beads and binders do interact only if their distance is shorter than the range r_{int}. The interaction energy scale E_{int} is set to be the minimum (absolute value) of the interaction potential V_{int}:

$$E_{int} = |\min(V_{int})| = \left| 4\varepsilon_{int} \left[\left(\frac{\sigma_{bb}}{r_{int}} \right)^{6} - \left(\frac{\sigma_{bb}}{r_{int}} \right)^{12} - \frac{1}{4} \right] \right|$$

In **Figure 2.1, Panel c**, V_{int} is shown for $r_{int} = 1.3\sigma$, $\varepsilon_{int} = 12k_BT$ and $\sigma_{bb} = 1\sigma$.

2.1.3 Langevin equation

The above described system, composed by the polymer 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 the Langevin equation (Allen and Tildesley, 1989; De Gennes, 1979; Kremer and Grest, 1990):

$$m \frac{d^2\mathbf{x(t)}}{dt^2} = - \zeta \frac{d\mathbf{x(t)}}{dt} - \nabla V + \boldsymbol{\xi}(t) \qquad (2.4)$$

where m and $\boldsymbol{x(t)}$ are respectively the mass and the position (in vectorial notation) of the particle, ζ is the friction coefficient, V the total potential on the particle, and $\boldsymbol{\xi}(t)$ is the random noise term representing the collisions with the molecules in the fluid. The components of the noise term have a Gaussian probability distribution with zero mean and a time correlation given by:

$$\langle \xi_i(t)\, \xi_j(t') \rangle = 2k_B T \zeta\, \delta_{ij} \delta(t-t') \qquad (2.5)$$

where, again, T *is* the temperature of the system and $\xi_i(t)$ is the *i-th* component of the noise vector. In MD simulations the dimensionless friction coefficient needs to be set; as discussed in a classical study of polymer simulations, a typical value is ζ=0.5 (Kremer and Grest, 1990), which has been also used in a number of investigations on chromatin modelling (Annunziatella et al., 2016; Barbieri et al., 2017; Bianco et al., 2018; Brackley et al., 2013; Chiariello et al., 2016; Rosa and Everaers, 2008).

Typically, in MD simulations the energy scale is set by $\varepsilon = k_B T = 1$, the length scale by $\sigma = 1$, and the mass is set to $m = 1$. Change in the ratio of the binder and bead masses leads to a shift in the time constant, but importantly does not change the equilibrium state of system (Kremer and Grest, 1990). In our simulations, the system is confined within a cubic simulation box with edge size D. Usually, periodic boundary conditions are employed: one particle can cross a box boundary and re-enter from the opposite side. A rule of thumb is to take the size, D, of the box edge at least as large as the gyration radius of the polymer in its open SAW conformation (see below), in order to minimize finite size effects. Once all the parameters are set, the system can be simulated. In general, the optimum integration time-step dt, necessary to the numerical integration of the Langevin equation, depends on the simulation parameters. For instance, for the *Verlet* algorithm an integration timestep $dt = 0.012\ \tau$ has been used in (Annunziatella et al., 2016; Bianco et al., 2018; Chiariello et al., 2016), where τ is the time scale (see following Sections).

2.1.4 Lennard-Jones dimensionless units

Usually, MD simulations use dimensionless units, called Lennard-Jones or reduced units. This means that σ, $\varepsilon = k_B T$ 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 and Tildesley, 1989). To estimate physical unit of length σ for simulation of chromatin organization within the cell nucleus, typically two different approaches are used: the first approach consists in comparing distances between particles derived by simulations against experimental data (FISH data, **Section 1.4**) (Brackley et al., 2016; Giorgetti et al., 2014); the second one, that is a less accurate but more straightforward strategy, consists in imposing that the local density of chromatin equals the expected average density of DNA in the whole nucleus; this assumption gives the expression for the physical length of the bead diameter:

$$\sigma = (s_0/G)^{1/3} D_0 \qquad (2.6)$$

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; Chiariello et al., 2016). Once estimated σ, the molar concentration of binders can be obtained by the relation $c = P/N_A V$, where N_A is the Avogadro's number, P the number of binders in the simulation box, and V its volume (in physical units). Analogously, the energy scale is set by choosing the temperature value T (e.g., $T = 300$K at usual lab room conditions). Finally, the time-

scale τ of the MD simulation is dimensionally linked to the scales σ, ε and m, via the relation: $\tau = \sigma\sqrt{m/\epsilon}$. Additionally, τ can be also related to the viscosity of the embedding fluid. More precisely, the friction coefficient ζ for a spherical particle can be expressed in terms of its size σ and of the solvent viscosity η by the Stokes law $\zeta = 3\pi\eta\sigma$. Since $\zeta=0.5m/\tau$ in physical units, by use of the Stokes law, τ can be given as function of η:

$$\tau = 6\pi\sigma^3\eta/\varepsilon \tag{2.7}$$

Such a relation permits to derive a rough estimation of the MD time scale τ from η, based on typical values of the order of magnitude of the nucleo-plasmic fluid viscosity, $\eta \sim$ 1-10cP (Brackley et al., 2013; Chiariello et al., 2016).

2.1.5 Preparation of the initial configurations

In a computer simulation the typical initial state of the polymer is a Self-Avoiding Walk (SAW) conformation. In order to obtain a SAW state, a nice method has been described, e.g., in (Kremer and Grest, 1990): first, a Random-Walk (RW) chain configuration with fixed steps is easily generated. The RW average bond length is taken to be equal to the minimum of the bonding potential (e.g., 0.97σ, see **Figure 1, Panel b**, green curve). Then, to softly remove any excess overlap between the beads of the chain, the hard-core repulsive Lennard-Jones is replaced with a soft potential:

$$V_{soft} = A\left(1+\cos\frac{\pi r}{2^{1/6}\sigma}\right) \tag{2.8}$$

where A is a normalization factor that is linearly increased in time during the simulation. As the soft potential does not diverge at small distances, the Langevin equations can be easily integrated for enough time-steps to remove the overlap and to reach the equilibrium SAW state.

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 (De Gennes, 1979), indicated by R_g:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{CM})^2} \tag{2.9}$$

where N is the number of beads, $\boldsymbol{r}_{CM}$ is the position of the center of mass of the chain and $\boldsymbol{r}_i$ is the position of its i-th bead. The gyration radius gives an estimation of the size of the average sphere enclosing the polymer. In a real 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 (**Figure 2.2, Panel a**). 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 **Figure 2.2, Panel b**, the values of the gyration radius for RW and SAW equilibrium states, obtained from real MD simulations, are shown. As expected from polymer physics (De Gennes, 1979), they both exhibit a power-law behaviour, $R_g^2 \propto N^{2\nu}$ where the scaling exponent ν is 0.5 and 0.588 for RW and SAW polymer states respectively. Note that also the equilibration time of the chain grows as a power law of the number of its beads, N.

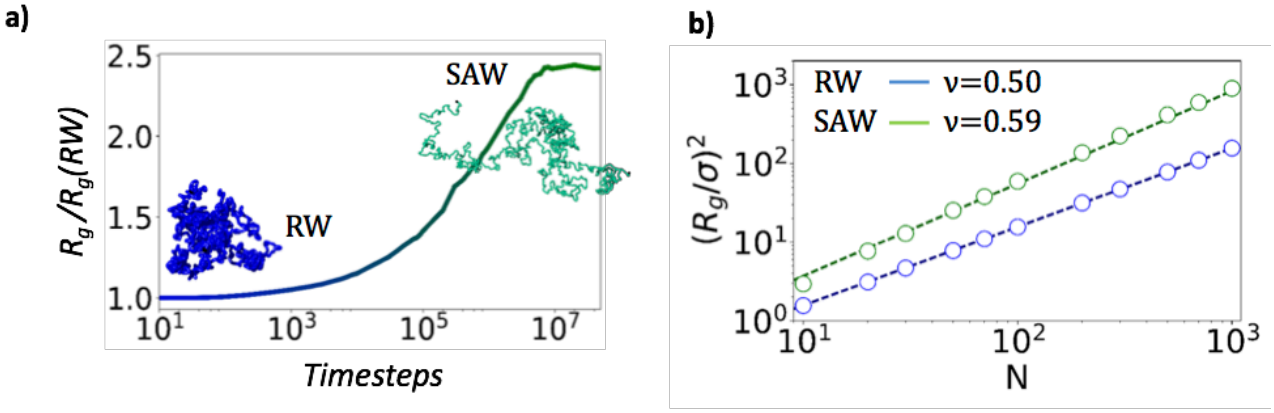

Figure 2.2: Scaling proprieties of Random and Self-Avoiding Walk.

a) Preparation of the SAW state. The gyration radius R_g is shown as function of the MD time steps, for a polymer that is initially prepared in a Random Walk (RW) state (blue) and evolves at equilibrium into a SAW (green) configuration, by the action of a soft-potential as described in **Eq. 2.8**. **b)** A log-log plot of R_g^2 as function of the number of beads of the polymer, N, highlights its scaling properties. The circles are the average values for the RW (blue) and the SAW (green) state from MD simulations of an ensemble of 10^4 different conformations at equilibrium. The dashed lines are the theoretical power-law behavior predicted by polymer physics. Figures adapted from (Annunziatella et al., 2018).

Once the polymer has been prepared in its initial SAW configuration, the binders are introduced at the concentration, c, of interest. Typically, they are randomly distributed in the simulation box.

2.2 Phase Diagram and structural characterization

In this Section, we investigate the SBS model in simplest case where all the beads making up the polymer are equal, i.e., all beads can interact with the same type of binders (homo-polymer case). We investigated the thermodynamic stable conformations for this type of system, varying the model parameters: concentration of binders in the environment and the energy of interaction between bead-

binders. Once we identified stable states, we characterized each one by studying their structural proprieties, such as their shape or the power-law trend of average contact probability between two sites that are at fixed genomic distance along the polymer.

2.2.1 Phase Diagram and order parameter

Initially, to characterize the thermodynamic features of the SBS model, we focus on the simplest case of homo-polymer where all beads can interact with the same type of binders in suspension (both colored in red, as in **Figure 1, Panel a**).

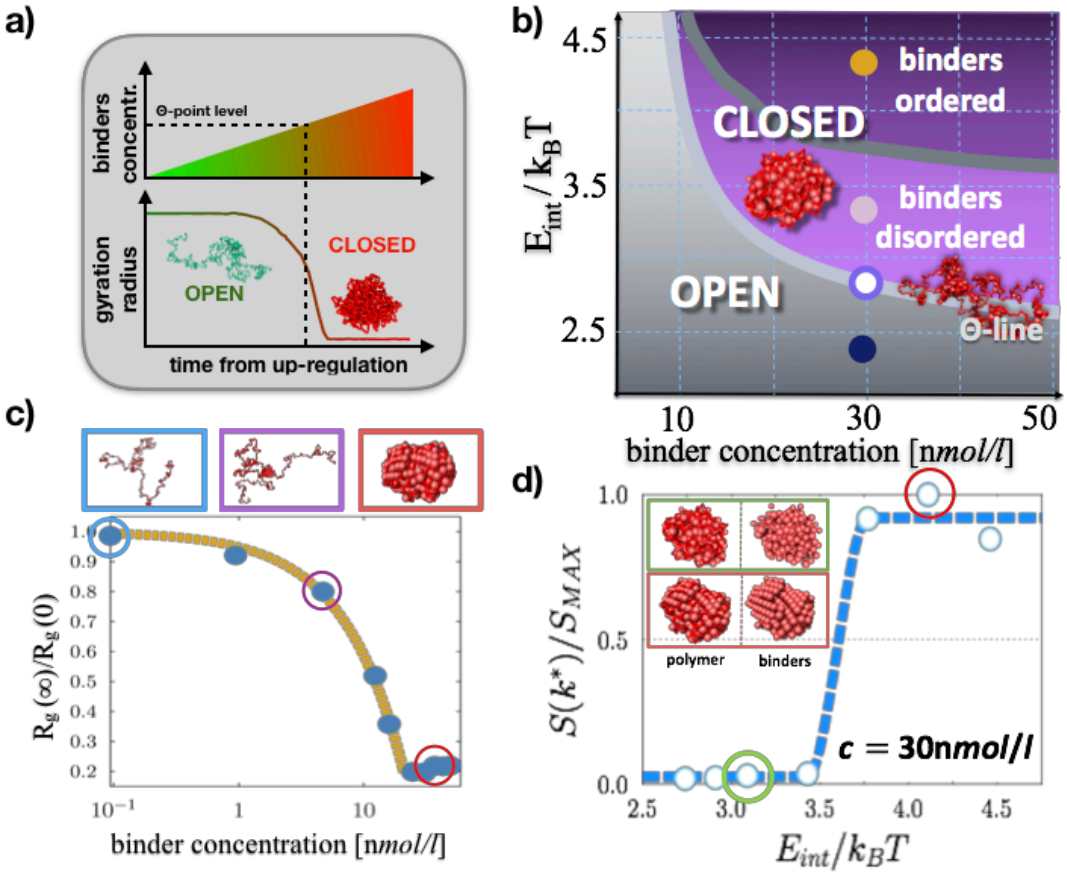

Figure 2.3: Stable thermodynamical states predicted by polymer physics and its characterization.

a) The binders in solution drive the folding of polymer, from open to closed state, as quantified by decreasing of gyration radius. **b)** Phase diagram, as a function of binder concentration c and energy of interaction E_{int}, of thermodynamical pure states predicted by polymer physics: at low E_{int} or c, the polymer is open and randomly folded in its *coil* phase; above its Θ-point transition, the polymer collapses in the *globule* phase assuming a compact conformation. In the globule state, at higher values of E_{int} or c, binders have a transition from a disordered to an ordered arrangement. Here, the unit length scale is calibrated on chromosomal scale. **c)** The gyration radius of the SBS polymer, R_g, signals its coil-globule transition point as a function of the concentration of binders. **d)** The Structure Factor peak marks the order-disorder transition in the arrangement of the binders around the folded polymer. Figures adapted from (Chiariello et al., 2016).

We study the equilibrium states of the system and the corresponding conformational folding classes, as function of model parameters, i.e., concentration *c* and energy of interaction E_{int} (**Figure 2.3, Panel b**). At small values of E_{int} and *c*, we find that the polymer has an open SAW conformation (we call it "*coil*" state), because of binders can establish only few and unstable loops between the beads of polymer. By increasing the values of E_{int} and *c* above a threshold (indicated by Θ-point), the system undergoes its coil-globule transition; the polymer collapses in a more compact conformation thanks to the formation of several and stable loops bridging pairs of beads (**Figure 2.3, Panel a-c**). The transition to globule state can be monitored by checking the time evolution of 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 below), R_g is initially high (SAW configuration) and then decreases until reaching a plateau value in the final equilibrium globular state (**Figure 2.3, Panel a**). Interestingly, we find at globule state two different regimes are possible: for lower values of E_{int} and *c*, the binders form a disordered lump around the chain; however, increasing E_{int} and *c*, they form instead of an ordered aggregate, although they do not have any direct interaction between each other. To quantify the disorder-order transition, we computed the structure factor S(k), defined as follows (Allen and Tildesley, 1989):

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

where, g(r) is the pair distribution function:

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

where we indicated by $\rho = N_b/V$ the concentration of the binders bound to the polymer and δ the Dirac delta function. As *S(k)* is the Fourier transform of the pair distribution function *g(r)*, it is almost constant when the binders are in a disordered configuration, while it shows sharp peaks when the binders form an ordered structure. In our analysis, we use as transition order parameter the ratio $S(k^*)/S_{MAX}$, where k^* is the position of the second peak in *S(k)* and S_{MAX} is a normalization coefficient taken to be equal to the maximum value of *S(k*)* across the different considered cases. As expected, the ratio has a jump at the order-disorder transition (**Figure 2.4, Panel d**). Analogous results are found in case other peaks of *S(k)* are considered, but the signal to noise ratio can be higher. The phase

diagram and the other results here discussed refer to simulations of a polymer made of $N = 1000$ beads, but they are independent of the system size (De Gennes, 1979).

To set the physical scales of our model, we have to consider the molecular details of the considered system. Since we are interested on average behavior at chromosomal scale, we can impose that the genomic length for the polymer is, e.g., 100 Mb (an average number of bases for mouse chromosomes): in this way we obtain a genomic content per bead equal to $s_0 = 100$ kb. Imposing a liquid viscosity of the order of estimates of the nucleoplasm environment, $\eta = 10$ cP, and a nuclear diameter equal to 3.5 μm as in mouse embryonic stem cells (mESc), by using **Eq. 2.6** and **Eq. 2.7**, we get as length and time scale $\sigma = 87$ nm and $\tau = 0.03$s respectively.

2.2.2 Pairwise and multi-way contacts

To further characterize the folding state of our polymer model, we computed the average pairwise contact probability, $P_C(s)$, of bead pairs at a given contour distance, s. The contact probability $P_C(s)$ is obtained by computing the number of pairs at given genomic distance s along the polymer, whose physical distance is less than (or equal to) a fixed threshold λσ (λ is a dimensionless constant threshold, here we set to λ = 3.5), and then averaging over the total number of pairs with the same, given contour distance. $P_C(s)$ trend only depends on the thermodynamic state of the system (**Figure 2.4, Panel a**). In the coil state, $P_C(s)$ decreases asymptotically as a power law with s, $P_C(s) \sim s^{-\alpha}$, with an exponent $\alpha \sim 2.1$ in the SAW universality class, whilst at the Θ-point, the exponent becomes $\alpha \sim 1.5$, as known in polymer physics (De Gennes, 1979). In the globule state, $P_C(s)$ depends on whether the system is in the disordered state, where after an initial decrease, a long plateau is found ($\alpha = 0$), or whether it is in the ordered state, where an exponent close to $\alpha \sim 1.0$ is observed. Analogously, the mean square distance of site pairs, $R^2(s)$, which can be accessed by FISH measurements, depends on the system thermodynamics phase. In that case, $R^2(s) \sim s^{-2\nu}$, where ν is equal to 0.59 and 0.5, for coil and Θ-point state respectively, while $\nu = 0.3$ for ordered globule state and $\nu = 0$ for the disordered case (**Figure 2.4, Panel b**).

Beyond pairwise interactions, we can investigate the occurrence of "many-body" contacts, i.e., co-localization events where multiple sites come simultaneously in physical proximity. 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-th bead within the above fixed threshold, and the number of possible combinations of k simultaneous contacts that contain the i-th 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*-th bead, $\binom{N}{k-1}$. First, we computed the contact probability of bead triplets on the same polymer at different genomic separations, $P_C(s_1, s_2)$, shown in **Figure 2.4, Panel c**. Next, we measured the frequency of observing *n* sites in physical contact (**Figure 2.4, Panel d**). As expected, in the closed states many-body contacts are exponentially more frequent than in the open state as *n* grows.

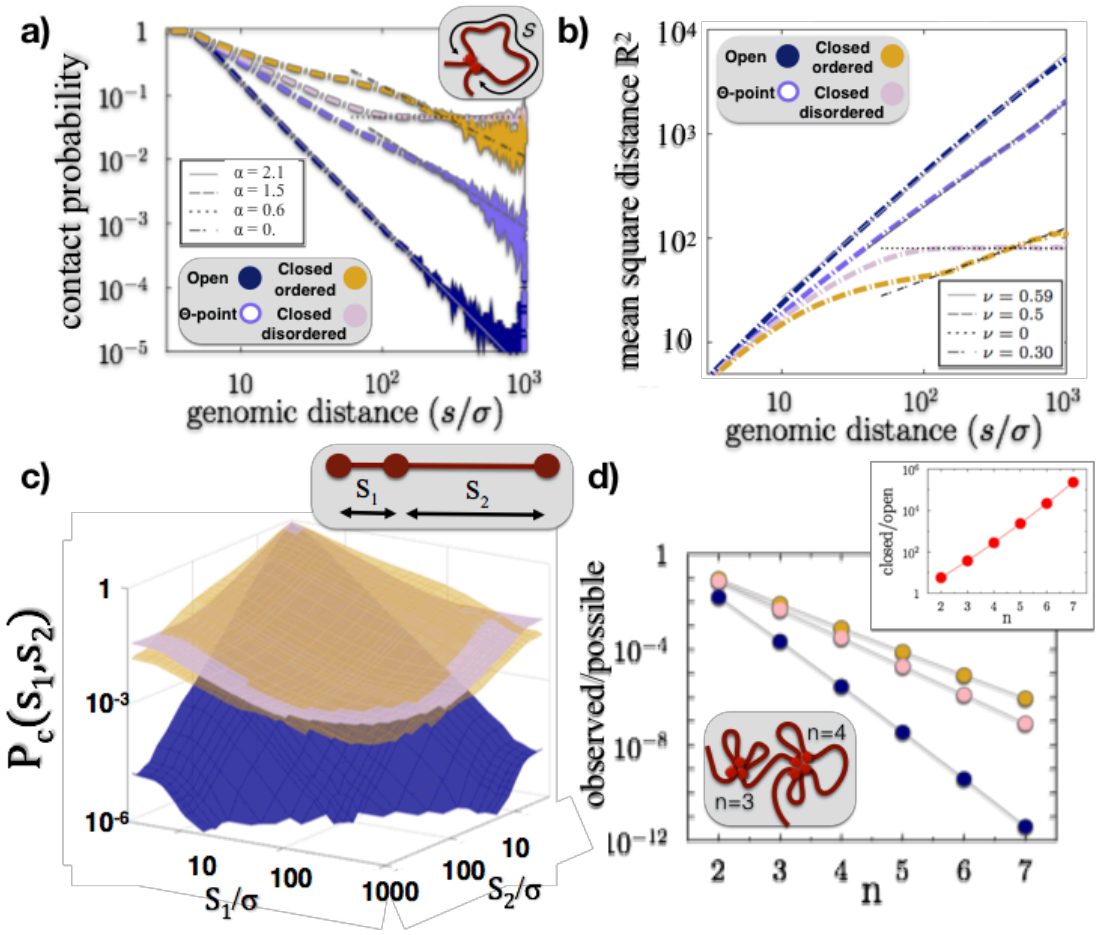

Figure 2.4. Study of contact frequencies for pair and triplets of beads for different stable states.

a) Average contact probability and **b)** mean square distance R^2 as a function of the contour distance s (i.e. genomic distance), in the different thermodynamic phases predicted by the SBS model. **c)** Average contact probability for triplets of beads, as function of their genomic distances s_1 and s_2 along the polymer for the stable conformation. **d)** Plot showing 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. On top-left panel, the ratio of the same quantity in the compact-disordered and open states. Figures adapted from (Chiariello et al., 2016).

Although multiple interactions are not detected by 3C-based methods, such as Hi-C, new GAM (Beagrie et al., 2017) and SPRITE (Quinodoz et al., 2018) technologies (**Section 1.3**) have highlighted that multiple interactions are an abundant structural component of chromatin. They show,

for instance, an abundance of three-way contacts between region highly transcribed and super-enhancer regions. That hints towards an until recently underestimated functional role of closed chromatin domains whereby multiple regulatory regions (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.

2.2.3 Characterization of phases by their shapes

To investigate the shape of the polymer in the different phases predicted by the SBS model, we also calculated its inertia tensor T, defined as:

$$T_{jk} = \sum_{i=1}^{N} m_i(r_i^2 \delta_{jk} - x_{ij}x_{ik})$$

where j and k are the indices of the space axes, j, k {x, y, z}, 'i' is a bead index, m_i is the mass of the i-th bead and x_{ij} its j-th coordinate. By diagonalizing *T*, we derive its three eigenvalues, which are the system principal momenta of inertia, I_1, I_2, I_3. The ratio $e_I = 2I_1/(I_2 + I_3)$, where $I_3 \geq I_2 \geq I_1$, returns a measure of the degree of ellipticity of the polymer shape: in a perfectly spherical conformation $e_I = 1$, while the higher the level of ellipticity the lower is e_I.

We find that in the coil SAW state $e_I \simeq 0.5$, in the ordered globular state $e_I \simeq 0.7$ and in the disordered globular state $e_I \simeq 0.9$ (**Figure 2.5, Panel a**). Hence, even in the SAW state, the polymer is more elongated along one axis, which in this case we found to be statistically aligned with the end-to-end direction of the polymer (**Figure 2.5, Panel b**). Our results on asphericity of SAWs are in full agreement with previous findings from polymer physics (Bishop and Michels, 1985). Interestingly, experimental measures suggest that many chromosomal territories have regular ellipsoid-like shapes with an ellipticity falling within the range 0.7−0.9 (Sehgal et al., 2014).

2.2.4 Folding dynamics

As previously discussed, the polymer folding process from a SAW configuration to a globule state is driven by the formation of loops produced by the binders. However, the details of the process depend on the specific choice of the system parameters, i.e., its interaction energy E_{int} and binder concentration *c*. In particular, by looking at total potential energy, E_{pot}, i.e., the bead-binders LJ and bead-bead FENE interactions, as function of MD simulation time, the folding process becomes more complex and two different dynamical regimes appear: the first related to folding process leading to a

compact conformation, and the second one, instead, related to binders reorganization in ordered structures (**Figure 2.5, Panel c**).

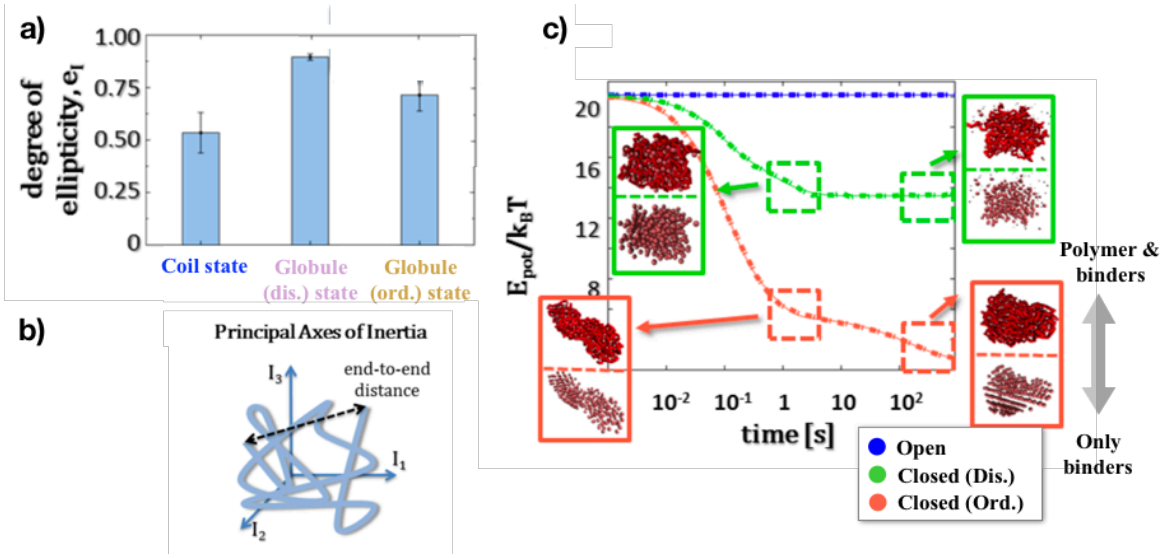

Figure 2.5. Structural properties of stable state at equilibrium and during dynamics.

a) The polymer ellipticity ratio, defined as $e_I = 2I_1/(I_2 + I_3)$, where $I_1 < I_2 < I_3$ are the polymer principal momenta of inertia, for different polymer state predicted by the SBS model. By definition, $e_I \simeq 1$ in a spherical conformation. All the stable conformations e_I is smaller than 1, and globule ordered state shows an ellipticity greater than the disordered state ($e_I \simeq 0.7$ v.s. $e_I \simeq 0.9$, respectively). **b)** Schematic representation of the principal axes of inertia of the polymer and their reference system. **c)** Trend of the total potential energy, E_{pot}, (FENE and Lennard-Jones potential) as function of real time at chromosomal scales, for different stable states. In case of ordered closed state, there are two different relaxation regimes in the dynamical folding process: first, the binders randomly aggregate onto the polymer; then, they rearrange to form an ordered structure. The 3D snapshots (bottom only binders, top also polymer) at different time points help to visualize the ordering transitions. Figures adapted from (Annunziatella et al., 2016).

2.3 Fitting experimental data

In this Section, we show that, despite its simplicity, by using stable conformations predicted by the SBS model it is possible to recapitulate average contact properties of the chromosomes, across three orders of magnitude, i.e. from sub-Mb to chromosomal scale. In order to compare our model results against Hi-C data, we reasoned that a single chromosome is likely to be a mixture of a variety of different folded regions, including for instance eu- and heterochromatin domains, which can dynamically change from cell to cell according to functional purposes (Nagano et al., 2013; Stevens et al., 2017). Yet, the stable spatial conformations of such regions must belong to one of the folding classes determined by polymer physics (pure states), at least in a first approximation. To model such

a scenario, we considered a mixture polymer system composed of different chain segments, each folded in one of the given thermodynamics states identified above (**Figure 2.6, Panel a**).

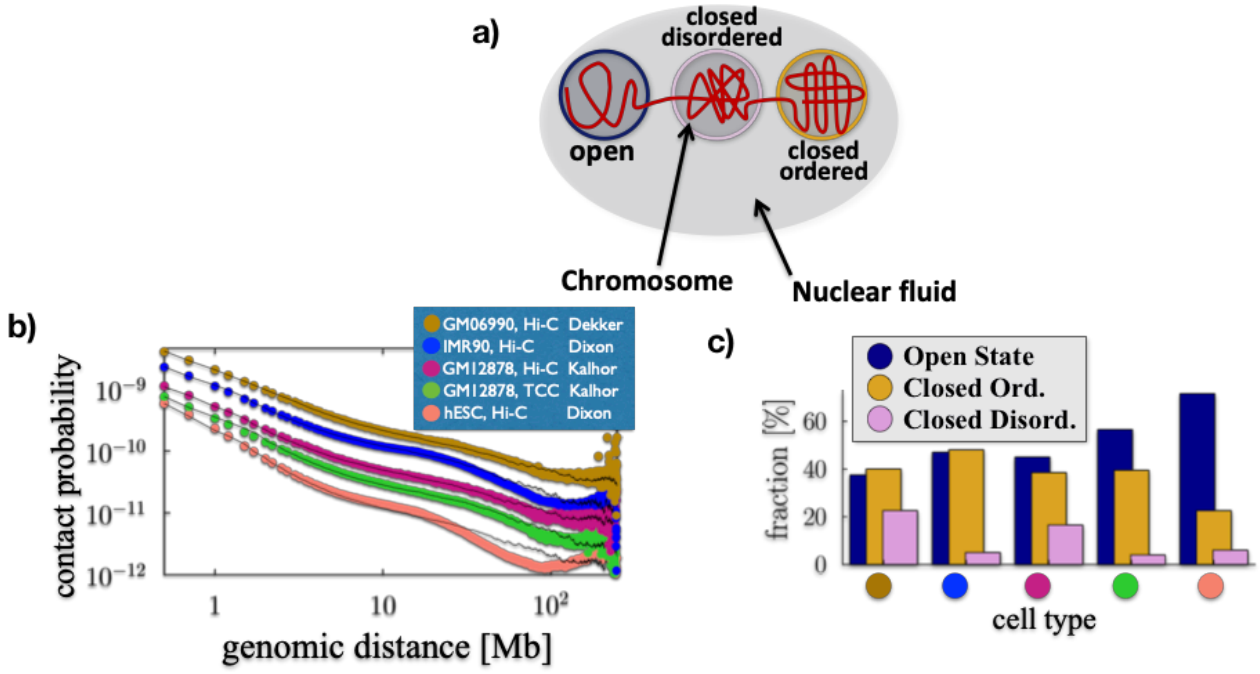

Figure 2.6: Chromatin is a mixture of regions folded in different thermodynamics states.

a) In our model, a chromosome is a mixture of differently folded regions, each belonging to one of the stable thermodynamical classes predicted by polymer physics. The average pairwise contact probability, in this approach, can be determined by the relative abundances of the states in the mixture, as each state has a fixed, specific pairwise contact probability. **b)** By such a mixture model, genome-wide average contact frequencies across human cell types, from different experiments, can be fitted from the sub-Mb to chromosomal scales. **c)** Each chromosome has a different chromatin composition, with hESC (orange circle) more open than differentiated cells, such as IMR90 (blue circle). Figures adapted from (Annunziatella et al., 2018; Chiariello et al., 2016)

For testing the biological significance of such a model of chromatin, we compared its predicted pairwise contact probability, $P(s)$, with available Hi-C, TCC and in-situ Hi-C contact frequency data (Dixon et al., 2012; Kalhor et al., 2012; Lieberman-Aiden et al., 2009; Rao et al., 2014). In our mixture model, $P(s)$ is just a linear combination of the contact probabilities of the pure states, independently derived above (**Section 2.3**). It only depends on the relative abundances of the states in the mixture (and on a scale factor used to map bead sizes into genomic separations). We find the mixture of pure states best describing experimental observations by fitting genome-wide average pair contact data as a function of the pair genomic separation s. This fit is done by use of the Least Square Method (LSM) as follows: we compute the model predicted contact probability of a mixture of open (coil) and closed (globule) states using the corresponding contact probabilities, independently derived

from the MD simulations of the homopolymer chain. At the end, by LSM we find the composition of the mixture of open and closed states that minimizes the distance between the predicted *P(s)* and the one derived from Hi-C data. We find that such a model can fit genome-wide averaged data (**Figure 2.6, Panel b**) and single chromosome data (**Figure 2.7, Panel a, c**) over approximately three orders of magnitude in genomic length, from 0.5 Mb to chromosomal scales, across a variety of different cell types and experimental techniques.

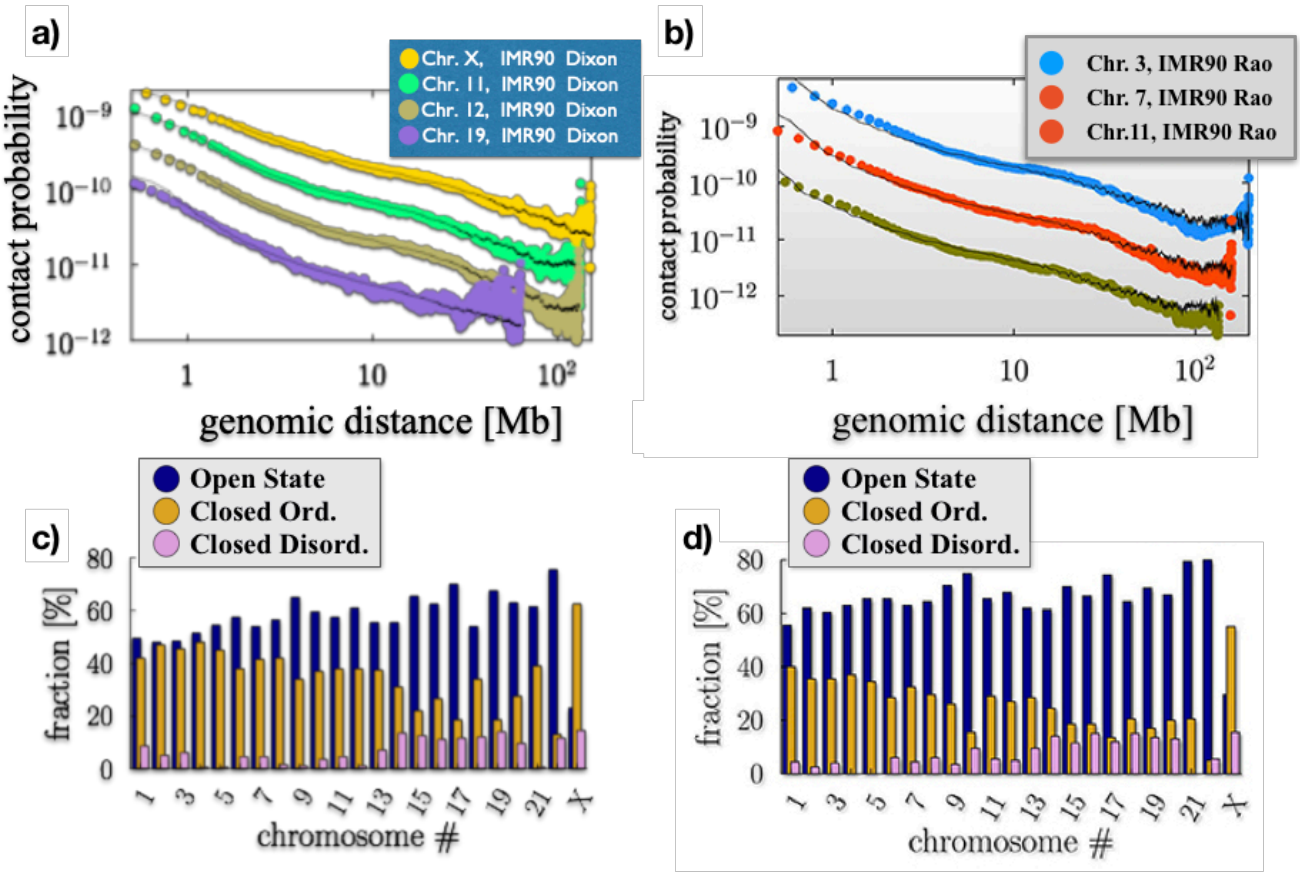

Figure 2.7: Fitting chromosomes trend of average contact frequency using different stable conformation.
a-b) Single chromosome experimental data from IMR90 cells can be explained, in the same way discussed above, using different dataset (Dixon et al., 2012 and Rao et al., 2014, respectively). **c-d)** The different composition of coil/globule states for each chromosome in IMR90 cell line, calculated from the two datasets above. Both datasets show similar results, for instance chromosome X formed mostly of closed regions, whereas gene rich chromosomes, e.g., chr.19, are up to 70% open. Figures adapted from (Chiariello et al., 2016, 2017).

Additionally, our approach also returns the mixture composition that best describes the given data (**Figure 2.6, Panel c; Figure 2.7, Panel b,d**). We find that different cell types have varying fractions

of open state chromatin: human embryonic stem cells (hESC) have the highest one, around 75%, while differentiated cells (e.g., human fibroblast cell IMR90) have values closer to 50%, in agreement with expectations. Different techniques (Hi-C v.s. TCC) give overall similar results in the same cell type: for instance, in GM12878 (lymphoblastoid) cells the fraction of closed ordered chromatin is 40% in both Hi-C and TCC data, yet the other states have a slightly different balance in the two cases. Different chromosomes can have very different compositions: in IMR90 cells, for instance, chromosome X is typically very compact (~70%) with a prevalence of the closed states. In general, the shorter the chromosome the higher is its open fraction. For example, chromosome 1 is only 50% open; chromosome 11 or 12 are 40% in the closed-ordered conformation, with less than 5% in the disordered state; the gene rich chromosome 19 is one of the less compact (>60% open), with the ordered and disordered closed states present in a 3/2 ratio. In brief, the mixture composition reflects the distribution of different folding domains along the chromosomes in the different cell types, across their thermodynamics states.

2.4 Self-interacting domains and hierarchical organization

The intricate pattern of interactions emerging from Hi-C contact matrices shows that chromatin at different scale has a very complex structure (see previous **Chapter 1**), and a more complicated model needs to be introduced to investigate the chromatin organization besides the average contact probability. For this reason, in this Section we introduce a slightly more complex model, where two different type of binding sites are introduced (we call it block-copolymer model). We show that the SBS model explains the biological mechanisms behind the formation of topologically associated domains, as introduced in (Dixon et al., 2012), and the hierarchical organization of higher-order structures, which has been shown to be a key feature in the mammalian genome organization (Fraser et al., 2015; Lieberman-Aiden et al., 2009).

First, we show that the SBS model can explain the formation of self-interacting domains. For that reason, we introduce two different types of binding sites (colored in red and green), adequately distributed along the polymer, with relative cognate binders (equally colored). In our MD simulation, we use a polymer of N=1000 beads, so each sub-polymer is 500 beads long, and we sampled concentrations c and interaction energies E_{int} to cover the three thermodynamic stable states we identified in the homo-polymer case (see, **Section 2.3**). In order to calibrate the length scales in this case, we consider the typical genomic length where chromatin is organized in the A/B compartments, which is one order of magnitude lower than the chromosomal length (**Section 1.4**). If we consider a genomic region with an overall length of 10 Mb, by using **Eq. 2.6** we find an estimation of σ equal to 64 nm. Additionally, by assuming a viscosity of 2.5 cP, we find a time-scale $\tau = 0.003$ s.

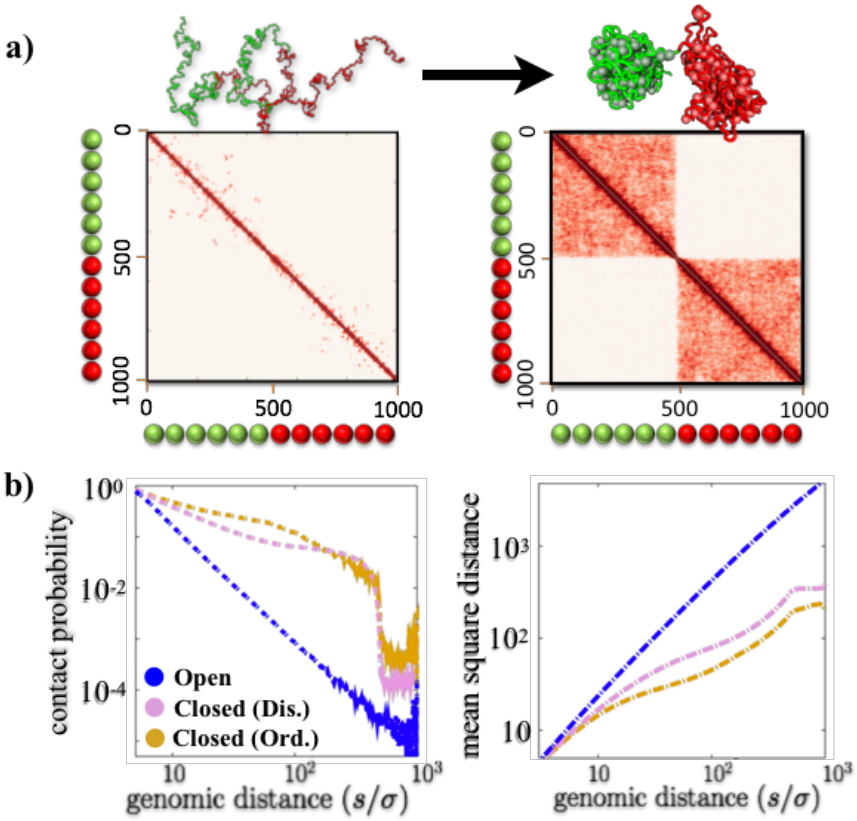

Figure 2.8: SBS model to explain TADs formation.

a) Pairwise contact frequency matrix of block-copolymer model where two types of sites (red and green), at coil (on left) and globule (on right) state. At equilibrium globule state, two different self-interacting domains are formed, which do not have any interactions between each other, as shown by 3D snapshot on top. **b)** Pairwise average contact probability $P_C(s)$ and Mean Squared distance $R^2(s)$ vs the contour separations of polymer bead pairs. We show the trend for the different polymer stable state, each colored with a different color. In the globular states (ordered and disordered), Pc(s) and $R^2(s)$ have apparent crossovers around the domain boundaries that are visible at a genomic distance equal to s=N/2. Figures adapted from (Annunziatella et al., 2016).

We find that, in the globule state, two different domains spontaneously form, which are respectively composed of red and green beads (**Figure 2.8, Panel a**). To quantitatively characterize the equilibrium states for such a system, we computed the average pairwise contact probability P_C and mean squared distance R^2, as function of genomic distance s (as discussed in **Section 2.2**). Additionally, we also computed the contact frequency matrices, which we have achieved by just generalizing the approach for computing the probability P_C. As expected, the contact matrices show that, in the globule state, two different self-interacting domains spontaneously arise from the model (**Figure 2.8, Panel a**). The same conclusions can be drawn by looking at $P_C(s)$ and $R^2(s)$, which show

apparent crossovers around the domain boundaries, at a genomic distance s = N/2 (**Figure 2.8, Panel b**).

Next, we investigated the mechanisms underlying the self-assembly of topological domains. In that case, we considered the case of a block-copolymer where different types of beads (red/green) are alternated in two pairs of blocks along the polymer chain (each block is then formed by 250 beads). Each polymer block can fold in the conformational states discussed for the homopolymer. As similar beads in different blocks can also interact with each other, the long time contact matrices have a more complex, chessboard-like pattern (**Figure 2.9, Panel a**), corresponding to a hierarchical organization of higher-order structures deriving from intra- and inter-domain interactions. The four different domains formed at steady state are also visible in the contact probability P_C and the mean square distance R^2: both, in fact, have apparent crossovers around the domain boundaries that are visible at a genomic distance equal to s = N/4, s = N/2 and s = 3/4N (**Figure 2.9, Panel b, c**). As discussed in previous **Section 1.4**, the presence of blocks in chromatin organization has a key role in gene expression, as in case of A/B compartments. In our view, chromosomal structures discovered in Hi-C data, such as TADs and meta-TADs, and their differential re-wiring across tissues and cell types, emerges naturally by specialization of the involved molecular factors under general mechanisms of polymer physics. We also explored some additional, possibly functional consequences of the self-assembly of domains. As TAD boundaries have been associated with biological markers and, more specifically, to an insulating role, we focused on how they affect the physical distance of pairs of sites differently positioned relative to them. Within our toy block-copolymer model, we focused on pairs of sites with the same contour separation: we considered two cases where the pair is located symmetrically or asymmetrically with respect to a domain boundary (**Figure 2.9, Panel d bottom**). Interestingly, we found that the block boundary can have a simple symmetry-breaking effect: in the closed phases, the sites of the symmetrically positioned pair have a larger physical distance than the asymmetric pair (p-value = 0), whereas in the open phase no difference is recorded. (**Figure 2.9, Panel d top**)

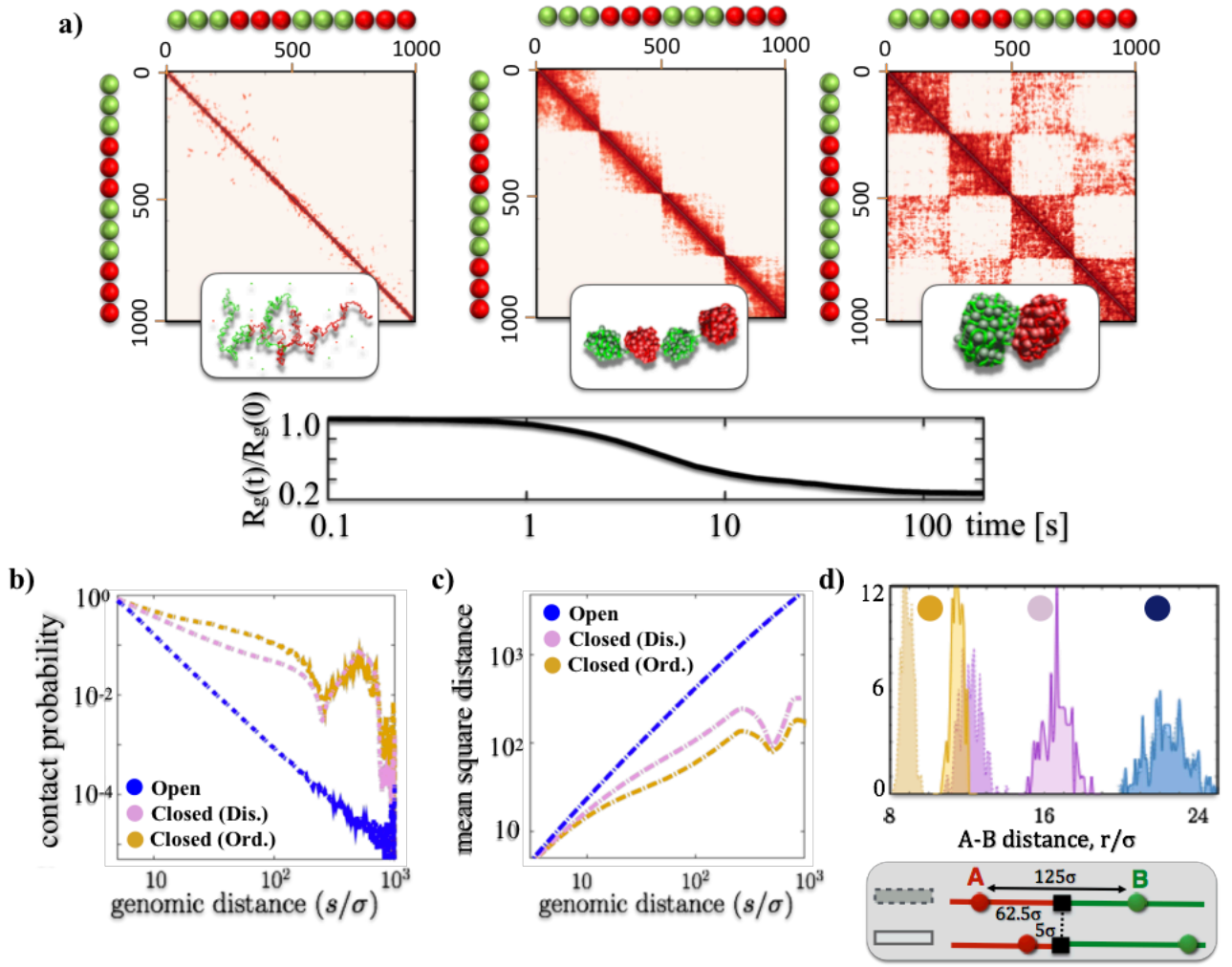

Figure 2.9: SBS model to explain the TADs formation.

a) Block co-polymer model where we introduced two different types of binding sites (red and green) and cognate binders (equally colored). Here, four consecutive blocks along the polymer obtained by alternating two colors. The folding of block-polymer during simulation time is monitored by gyration radius, and the three panel shows the contact matrices at different time of simulation time. **b)** Pairwise average contact probability P_C and **c)** mean squared distance $R^2(s)$ vs the contour separations of polymer bead pairs. The trend for the different polymer stable state are shown, each colored by a different color. In the globular states (ordered and disordered), Pc(s) and $R^2(s)$ have apparent crossovers around the domain boundaries that are visible at a genomic distance equal to s=N/4, s=N/2 and s=3/4N. **d)** Pairs of sites with the same contour separation, differently positioned across a block boundary (see bottom panel), have the same average physical distances, r (dimensionless units), in the open phase. Yet, in the closed states, the symmetry is broken by their different position relative to the boundary as the two pairs have a different physical distance, as seen from the corresponding distributions of r (globule ordered, disordered and coil state respectively). Figures adapted from (Annunziatella et al., 2016; Chiariello et al., 2016).

References

Allen, M.W., and Tildesley, D.J. (1989). Computer Simulation of Liquids (Clarendon Press).

Annunziatella, C., Chiariello, A.M., Bianco, S., and Nicodemi, M. (2016). Polymer models of the hierarchical folding of the Hox-B chromosomal locus. Phys. Rev. E *94*.

Annunziatella, C., Chiariello, A.M., Esposito, A., Bianco, S., Fiorillo, L., and Nicodemi, M. (2018). Molecular Dynamics simulations of the Strings and Binders Switch model of chromatin. Methods *142*, 81–88.

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 and binders switch model. Proc. Natl. Acad. Sci. *109*, 16173–16178.

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., De Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *24*, 515–524.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. J. Vis. Exp.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Bishop, M., and Michels, J.P.J. (1985). The shape of ring polymers. J. Chem. Phys. *82*, 1059–1061.

Bohn, M., and Heermann, D.W. (2010). Diffusion-driven looping provides a consistent provides a consistent framework for chromatin organization. PLoS One *5*.

Brackley, C.A., Taylor, S., Papantonis, A., Cook, P.R., and Marenduzzo, D. (2013). Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc. Natl. Acad. Sci. *110*, E3605–E3611.

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

Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R., and Marenduzzo, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. Phys. Rev. Lett. *119*.

Chiariello, A.M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M. (2016). Polymer

physics of chromosome large-scale 3D organisation. Sci. Rep. *6*.

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. Front. Neurosci. *11*.

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

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852–852.

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.

De Gennes, P.G. (1979). Scaling concepts in polymer physics. Cornell university press. Ithaca N.Y.,.

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

Grosberg, A.Y., Nechaev, S.K., and Shakhnovich, E.I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. J. Phys. *49*, 2095–2100.

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

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kremer, K., and Grest, G.S. (1990). Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. J. Chem. Phys. *92*, 5057–5086.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

Mirny, L.A. (2011). The fractal globule as a model of chromatin architecture in the cell. Chromosom. Res. *19*, 37–51.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

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

Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G., and Onuchic, J.N. (2016). Transferable model for chromosome architecture. Proc. Natl. Acad. Sci. *113*, 12168–12173.

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

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 219683.

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

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

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Sehgal, N., Fritz, A.J., Morris, K., Torres, I., Chen, Z., Xu, J., and Berezney, R. (2014). Gene density and chromosome territory shape. Chromosoma *123*, 499–513.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Chapter 3: Modeling real loci by polymer physics

In the previous Chapter, we showed that a very simple model (the homo-polymer model) describes with high accuracy the average behaviour of folding dynamics of entire chromosomes in a while range of genomic lengths, from sub-Mb up to whole chromosomal scale. Furthermore, we showed that a little more complex model, where two different bead types are introduced (block-copolymer model), can describe the formation of both self-interacting domains and higher order structures (as appear in experimental data), which spontaneously occur in the self-assembly process.

In this Chapter, we will show how the SBS model can be further improved to reconstruct the three-dimensional spatial organization of real genomic regions (in genetics called loci) at higher resolution scales, and describe the biological mechanism of chromatin folding. To this aim, we generalized our model introducing along the polymer different types of colored beads, each one interacting only with its cognate type of binders. In **Section 3.1**, we briefly describe PRISMR, a machine-learning algorithm we developed to estimate the minimal number of different bead types and their position along the polymer. Taking as input the experimental contact matrix (Hi-C data), PRISMR returns as output the polymer best describing the experimental data, without a-priori assumptions and no additional or tunable parameters. Once we obtain the best polymer chain, in **Section 3.2** we show how, by Molecular Dynamic simulations of the SBS model, we can generate an ensemble of 3D conformations for specific loci, which recapitulate experimental data with very good agreement. Importantly, from the polymer conformations, we can also access additional and independent structural information, which cannot be available from Hi-C, such as the physical distribution of two loci or multi-way contacts. As examples, we report our results for *Sox9* and *HoxB* loci (Annunziatella et al., 2016; Chiariello et al., 2016), while for sake of brevity we will not discuss other loci we investigated, like the *7q11.23* (Chiariello et al., 2017) and the *Bmp7* (Chiariello et al., 2016) loci. In **Section 3.3** and **Section 3.4**, as further application of our method, we investigated how 3D architectures change at different time points of cell differentiation (from mouse undifferentiated ESC-J1 to differentiated Cortex cell) for the *HoxD* locus (Annunziatella et al., 2018, submitted), and in two different cell tissues lines for the *Pitx1* (Forelimb and Hindlimb) locus (Kragesteen et al., 2018), trying to explain the link between structural changes and biological functionality of cells.

All the results are discussed in our articles (Annunziatella et al., 2016; Bianco et al., 2017, 2018; Chiariello et al., 2016), while results about the *Pitx1* locus have been developed in collaboration with Stepan Mundlos's research group at Max Plank Institute in Berlin (Kragesteen et al., 2018). The

results for the *HoxD* locus (Annunziatella et al., 2018, submitted), instead, have not been published yet and represent one of the current research projects of the group. For the sake of simplicity, most of the Chapter, including figures, paragraphs and sentences are adapted from these papers.

3.1 Generalization of the SBS model: PRISMR method

Here, we briefly describe the PRISMR method (Bianco et al., 2018), and how, by exploiting the information contained in Hi-C contact matrix, we can estimate the polymer best describing the experiments.

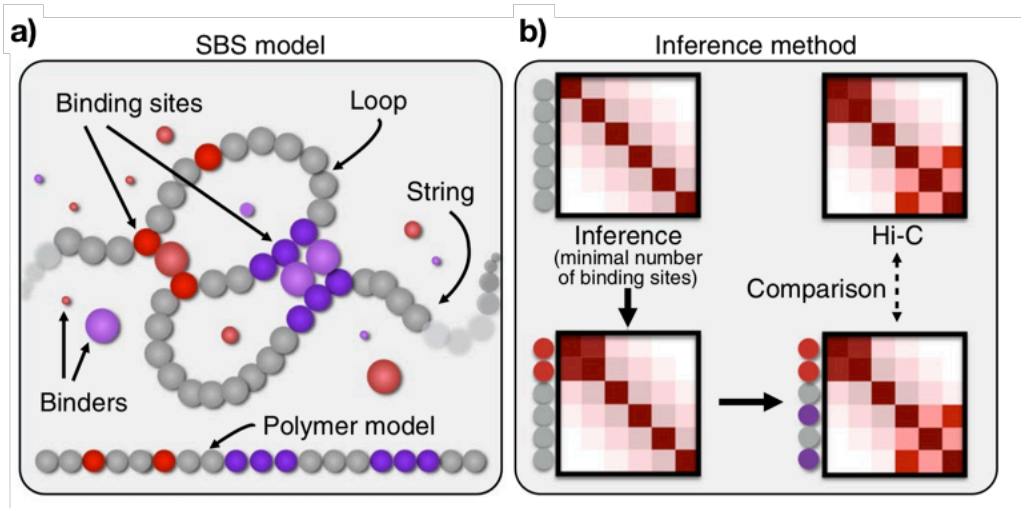

Figure 3.1: The PRISMR method for inference of molecular binders driving chromatin folding.

a) In the SBS model, chromatin is represented as a chain of beads interacting with molecular binders, in which the different types of binding sites (and their cognate bridging molecules) are visualized in different colors. **b)** PRISMR is a machine learning algorithm which, starting from experimental contact matrices, infers the distribution of binding sites best describing the input data. Figure adapted from (Bianco et al., 2018).

PRISMR (polymer-based recursive statistical inference method) is a polymer physics method, recently developed by our group, based on the SBS model. It allows to infer the minimal factors that shape chromatin folding and its equilibrium 3D structures, without a priori assumptions and with no additional or tunable parameters (Bianco et al., 2018). PRISMR employs a standard Simulated Annealing (SA) procedure and uses a cost function that includes the distance between the input Hi-C and the model predicted contact matrix, and a Bayesian term (a chemical potential) to penalize overfitting. The output of PRISMR are: the minimal number of colors n and of binding sites r inferred (we fixed n=r in described cases), and their position along the polymer (**Figure 4.1, Panel a-b**). All the details of the procedure are discussed in (Bianco et al., 2018). Once we find by PRISMR the

"optimal" model, we run Molecular Dynamics (MD) simulations and derive its contact matrix without any approximation. Finally, we use Pearson correlation (or distance-corrected Pearson correlation, see below) to assess the similarity between the MD derived model contact matrix and Hi-C data.

3.2 Modeling real loci

In the following Section, we aim at understanding whether our model can explain the folding dynamics of specific, real genomic regions, rather than the average features of chromosomal conformations. By using the "best" polymer inferred from the PRISMR algorithm (**Section 3.1**), we generate an ensemble of possible configurations predicted by running Molecular Dynamics simulations. These polymer configurations allow, besides reproducing the average contact matrices, to access additional and independent structural information, such as the physical distance between two loci (e.g., enhancer-promoter), the presence of multi-ways contact or investigating chromatin shape.

3.2.1 Methods for Molecular Dynamics simulations

In our MD simulations, we generalized the SBS model to accommodate different types of binding sites, introducing special beads which can only interact with cognate binders having the same dimension. The interaction between bead and binders, as discussed in **Chapter 2**, is modelled by attractive LJ potential (**Eq. 2.3**), where we fix the interaction range $r_{int} = 1.5\ \sigma$ and vary the interaction energy E_{int} and the binder concentration c in order to drive the polymer in each different thermodynamic state predicted by the model, i.e. in coil and globule states (see **Chapter 2** for more details).

By running MD simulations, we generate an ensemble of different equilibrium configurations under physical laws. Over these configurations, we compute the average contact matrices as follows: first, we compute the average contact matrix for coil/globule state separately, considering two beads i and j in contact if they are of the same color and if their physical distance is less (or equal to) $\lambda\sigma$. Finally, to take into account the effects of cell population heterogeneity, i.e., the possibility that the locus could be in different states (coil/ globule) in cellular environment, we considered the contact matrix of the coil/globule mixture that maximized the Pearson correlation coefficient, r, with the corresponding experimental data (i.e., Hi-C). The dimensionless threshold λ can be different from case to case, but we choose that optimizes the final results. However, to check the robustness of our approach we also considered different threshold values and contacts between beads of different colors, and for all the cases considered we find very similar results.

To improve our comparison between experimental and predicted matrices, we have also introduced the Pearson correlation, *r'* (Bianco et al., 2018), which takes into account the effect of genomic distance in contact maps. 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. Unlike Pearson correlation coefficient *r*, the coefficient *r'* is zero in the random case (obtained by bootstrapping diagonal in contact matrices) (Bianco et al., 2018).

Snapshot of polymer configurations are produced with POV-RAY (Version, 2004), and the coordinates of each bead are interpolated with a smooth third-order polynomial splice curve.

3.2.2 *HoxB* locus

As first case, we focused on the *HoxB* locus, a genomic region around the *HoxB* genes, in mouse Embryonic Stem Cell (mESC), where Hi-C experimental data are available (Dixon et al., 2012).

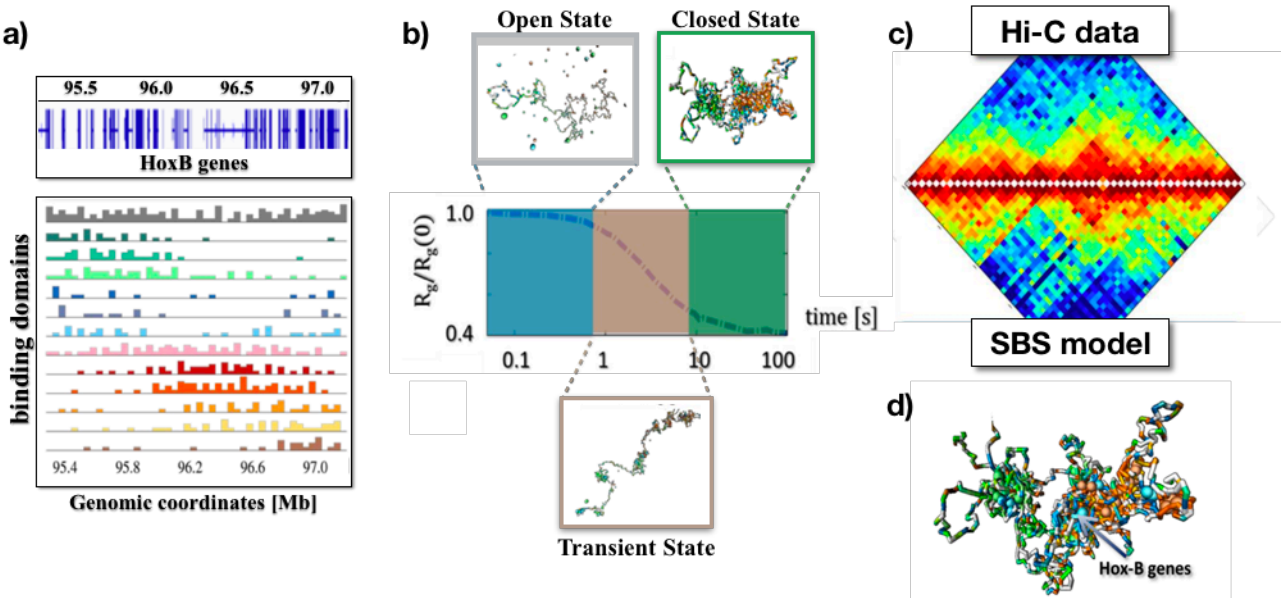

Figure 3.2: Folding mechanism of HoxB locus in mouse ESC-46C cells.

a) On the top, a gene-dense genomic region surrounding *HoxB* genes; on the bottom, the distribution of *n*=12 different binding sites, as computed by the PRISMR algorithm. Each histogram shows the abundance of the corresponding color over the genomic sequence. **b)** The folding dynamics of the *HoxB* locus proceeds gradually passing through an intermediate transient state to a completely folded polymer. **c)** Comparison between the average contact matrix from experimental Hi-C (top) and the matrix predicted from the MD simulations (bottom). **d)** Snapshot of the *HoxB* locus in the globule state, where three different domains emerging from this picture (as reflected in the coloration of the polymer green-light blue-orange) are shown. The *HoxB* genes cluster is positioned in the central part of the locus. In this case, the binders are shown in the 3D structure. Figure adapted from (Annunziatella et al., 2016).

The locus we considered locus is 1.92Mb long (chr11:95280000-97200000, mm9), binned at 40 kb resolution, and centered around the *HoxB* genes cluster (**Figure 3.2, Panel a top**) (Annunziatella et al., 2016). By application of PRISMR, we found that n=12 different bead types (each visually represented by a different color, sorted for position of center of mass in **Figure 3.2, Panel a bottom**) are needed to best describe the folding dynamics of the locus. The polymer we used is a chain made of N = 576 beads, and each elementary bead contains about 3.3 kb of genome. We generated an ensemble of 10^2 independent polymer configurations, each starting from a SAW (coil) configuration (as described in **Section 2.1**). To reach the equilibrium state (as monitored by the gyration radius), we let the system evolve up to 2.5×10^8 time steps. The folding dynamics, as monitored by the trend of the gyration radius, has a hierarchical nature as visualized in **Figure 3.2, Panel b**, where also 3D snapshots of the locus at different dynamical stages have been shown. (Here, snapshot have been produced by using VMD software (Humphrey et al., 1996)). Next, starting from configurations generated by MD simulations, we computed pairwise average contact matrix fixing as threshold λ = 8. We find that the coil/globule mixture that best describes Hi-C data in this case is 72% coil and 28% globule state, with a Pearson correlation coefficient equal to r = 0.95 between model prediction and experimental data (**Figure 3.2, Panel c; Section 3.2.1**). A typical configuration of the locus in the globule state is represented in **Figure 3.2, Panel d**, where we highlighted the position of the *HoxB* genes cluster, which swings between two different interaction domains.

3.2.3 *Sox9* locus and molecular nature of the binding domains

Next, we further tested our polymer models in explaining the details of folding of specific genomic regions, focusing on a 6 Mbs region around the *Sox9* gene (chr11:109000000-115000000, mm9) in mouse embryonic stem cells, which includes gene rich areas as well as gene deserts (Chiariello et al., 2016). *Sox9* is an important gene that plays a key role in sexual development, and the genomic mutations involving this gene are often linked to the skeletal malformation syndrome and to autosomal sex-reversal (Franke et al., 2016). By using PRISMR, we estimated the minimal arrangement and different types of binding sites that best reproduces the Hi-C contact matrix available for mESC-J1 cells, at 40 kb resolution (Dixon et al., 2012). Such a method returns n=15 different interacting bead types, visually represented by different colors in **Figure 3.3, Panel a**, ordered left-to-right according to the location of the domain center of mass.

As before, we generalized our polymer model to accommodate different types of binding sites (colors) and their cognate molecular binders, and we run MD simulations with a polymer made up of N = 2250 beads, each one containing about 2.67Kb of genome. By using **Eq. 2.7**, we estimate the physical bead size σ to 26nm and, assuming as a reference a viscosity of 2.5cP, the time unit τ to

0.0002s. In our simulations we sampled values of the total binder concentration, c, ranging from zero to 215n*mol/l*, and we varied the interaction energy from zero (open state) to E_{int} = 16k_BT. The concentration and the interaction energy employed for the results discussed in **Figure 3.3** are c = 194 n*mol/l* and E_{int} =12 k_BT, corresponding to a globule polymer state. We generated an ensemble of up to $5x10^2$ independent polymers, each starting from a SAW configuration and equilibrated as described in **Section 2.2**. Analogously, the binders are initially placed in random positions in the simulation box. To reach equilibrium, we ran the simulations up to 10^9 MD time steps.

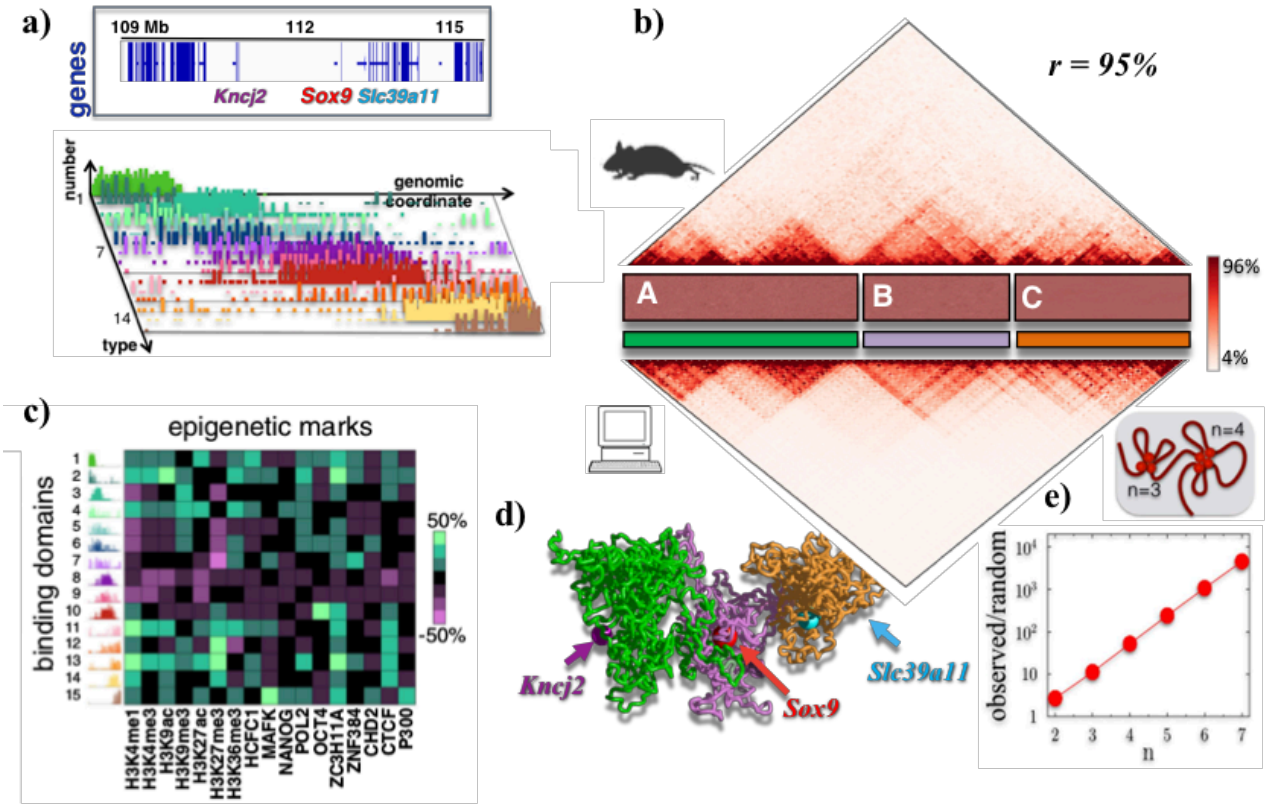

Figure 3.3: Three-dimensional reconstruction of Sox9 region by SBS model.

a) Distribution of binding sites for region including the Sox9 gene (chr2: 109-115Mb). **b)** Average contact matrices computed using 3D polymer structures (bottom), compared to Hi-C contact frequencies (top). The agreement is quantified by Pearson correlation coefficient *r*. **c)** Matrix showing Pearson correlation between the binding sites distribution and the signal of some epigenetic marks. There is no one-to-one matching but each binding site is a combination of different epigenetic marks. **d)** Snapshot showing a typical 3D polymer conformation of globule state at equilibrium. Three different interacting domains can be identified, as expected looking at contact matrices in panel b). **e)** 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 co-localization of multiple functional regulatory regions. Figures adapted from (Bianco et al., 2017; Chiariello et al., 2016).

To test our MD results, we first computed the contact matrix average over all 3D conformations. We find the best results with $\lambda = 10$ and a mixture of coil/globule state equal to 64% and 36%. The derived pairwise contact frequency matrix has a Pearson correlation *r* with Hi-C data of 95% (**Figure 3.3, Panel b**), and a distance corrected correlation *r'* equal to 68% (**Figure 3.4**), supporting the view that our polymer model captures a relevant component of the molecular mechanisms determining the folding of *Sox9*. Next, to infer the specific molecular nature of the predicted binding sites (model colors) and their cognate binding factors, we crossed the information on their position/type with epigenomic databases of chromatin features available for Sox9 region in mESC. By integration of such data, we can identify known and new candidate factors driving folding and responsible for its regulation. The heatmap in **Figure 3.3, Panel c** illustrates the correlation coefficient between the genomic positions of binding domains and chromatin features from ENCODE (Dunham et al., 2012). Each binding domain appears to have an epigenetic barcode that is a unique combinatorial pattern of epigenetic features. For instance, some domains are characterized by active marks and Polymerase II (Pol II), others by more repressive marks. CTCF correlates strongly with many of these domains, yet others are not linked to it. This proves that additional architectural factors, beyond CTCF, play a key role in folding.

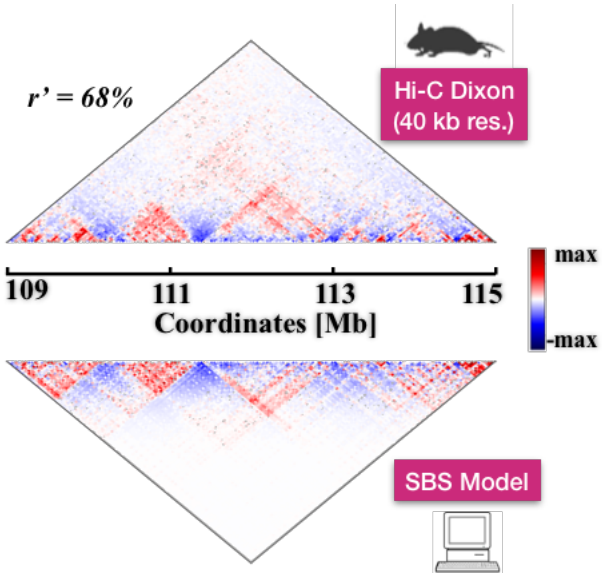

Figure 3.4: Distance-Corrected Matrices

Distance corrected matrices are obtained by subtraction of the average interaction at a given genomic distance. Interestingly, the pattern in the data are still captured by the model after the effects of genomic distance are subtracted: the Pearson correlation coefficient remains high (*r'*=0.68) .

A snapshot of a single typical configuration of the Sox9 locus, in the globule state, is shown in **Figure 3.3, Panel d**, 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, **Section 1.2**) 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 μm v.s. 0.59 μm) as the three genes belong to consecutive regional areas (**Figure 3.3, Panel d**).

The variety of information on *Sox9* and its folding mechanisms that can be inferred from polymer physics extends well beyond the Hi-C pairwise contact data used to infer the model. The self-assembly of the locus architecture from initial open states proceeds hierarchically, as shown for the *HoxB* locus (**Section 3.2**), with early formed local domains folding into larger and larger 3D structures encompassing the entire locus; the order of magnitude of the time-scale of the process is 20 sec. Besides, the *Sox9* locus is marked by many-body contacts that are exponentially more abundant than expected in a randomly folded conformation (**Figure 3.4, Panel e**, error bars within symbol size).

3.3 Impact of 3-dimensional changes in cell regulation

The modulation of three-dimensional spatial organization of chromatin is crucial for regulation of gene expression (**Section 1.4**). During cell differentiation the chromatin structure can be altered by several epigenetics features, such as CTCF or active/repressive histone marks, that make possible the transcription for a specific gene promoter or its inactivation (**Section 1.2**). However, understanding the exact mechanisms involved in genome organization and how this last one changes, during differentiation, is still an open question in modern biology.

In the following, we investigate the changes in chromatin organization for the genomic region flanking *HoxD*, an important genes cluster in mammalian cells that controls the body development of an embryo along the head-tail axis. In particular, we focus on studying the structural variations of *HoxD* locus at two different stages of murine differentiation cells: in embryonic stem cells, where *HoxD* is not activated but in a poised state, ready to be activated (Bernstein et al., 2006), and in Cortex cells, where, conversely, it is completely silent (Annunziatella et al., 2018, submitted). The *HoxD* locus is an example of how the formation of TADs is one important means by which promoters are insulated from enhancer can potentially interact with (**Section 1.4**).

3.3.1 Structural changes of the *HoxD* locus at different stages of differentiation

HoxD genes is a genes cluster involved in vertebrate limb development. As revealed by chromatin contact patterns provided by technologies introduced in **Chapter 1**, the transcriptional activity of the *HoxD* genes is driven by a spatial reorganization of the locus during cell differentiation (Andrey et

al., 2013; Noordermeer et al., 2014). In mouse embryonic stem cells (mESC), the *HoxD* genes are marked by bivalent chromatin states, with both repressive (H3K27me3) and activating (H3K4me3) signatures (Bernstein et al., 2006; Noordermeer and Duboule, 2013; Schuettengruber et al., 2017; Soshnikova and Duboule, 2009). The presence of both epigenetics marks regulates the expression of genes during differentiation, keeping the genes poised and ready to be activated. During mouse embryo development, a collinear activation occurs, as the genes are sequentially turned on according to their genomic position (starting from *HoxD13* up to *HoxD1*), and, correspondingly, a three-dimensional compartmentalization appears, with active genes forming a cluster physically separated from the inactive ones (Noordermeer et al., 2011). Similar complex architectural patterns are also found during limb buds development and in other tissues (Andrey et al., 2013; Noordermeer et al., 2014). The hypothesis has been raised that such 3D compartmentalization has a general functional role, which may help, for instance, the maintenance of the transcriptional states by avoiding contacts between the active and inactive genes, and by restricting the usage of enhancer repertoires during development. However, it is unknown how such a regulatory program is implemented at the single-cell level and, in particular, the folding mechanisms that control contact specificity between genes and regulators at different transcriptional stages.

3.3.2 SBS model of the *HoxD* locus

To investigate those aspects of *HoxD* genes organization, we employ SBS model focusing first on a 7Mb region around the *HoxD* cluster (chr2:71000000-78000000), at 40 kb resolution as in published Hi-C data (Dixon et al., 2012). In this case, the application of PRISMR procedure resulted in a polymer model made of N=2100 beads, including n=12 different types of binding sites in case of mESC and of N=3500 beads, including n=20 different types of binding sites, in case of Cortex cell. By MD simulations, we then derived an ensemble of 10^2 polymer configurations at equilibrium, sampling the total concentration c of binders from zero to 116 nmol/l and the scale of interaction energy between beads and binders equal to $E_{int} \simeq 1\ k_BT$ and $E_{int} \simeq 8.1\ k_BT$, which correspond to the coil and globule conformational states respectively, predicted by polymer physics (**Section 2.2**).
Initially, to test the accuracy of our models, we compared the cell-type specific patterns of Hi-C data (Dixon et al., 2012) in ES and in Cortex mouse cells (**Figure 3.5 and 3.6, Panel a**) against the model pairwise contact matrices derived by our MD simulations (**Figure 3.5 and 3.6, Panel b**), where we set the dimensionless threshold of pairwise contact $\lambda = 9$. As in the other cases, to take into account population effects, the procedure returns the optimal mixture of single molecule structures, in the coil and in the globular thermodynamics state, best describing the population averaged Hi-C contact data (**Section 3.2.1**). Through that approach, we find a mixture of coil/globule states equal to 66%-34%

for ES and 61%-39% for Cortex case. Notably, from our ensemble of MD 3D polymer structures, we derive the "single cell" pairwise contact probability (**Figure 3.5 and 3.6, Panel b; Section 1.3.2**).

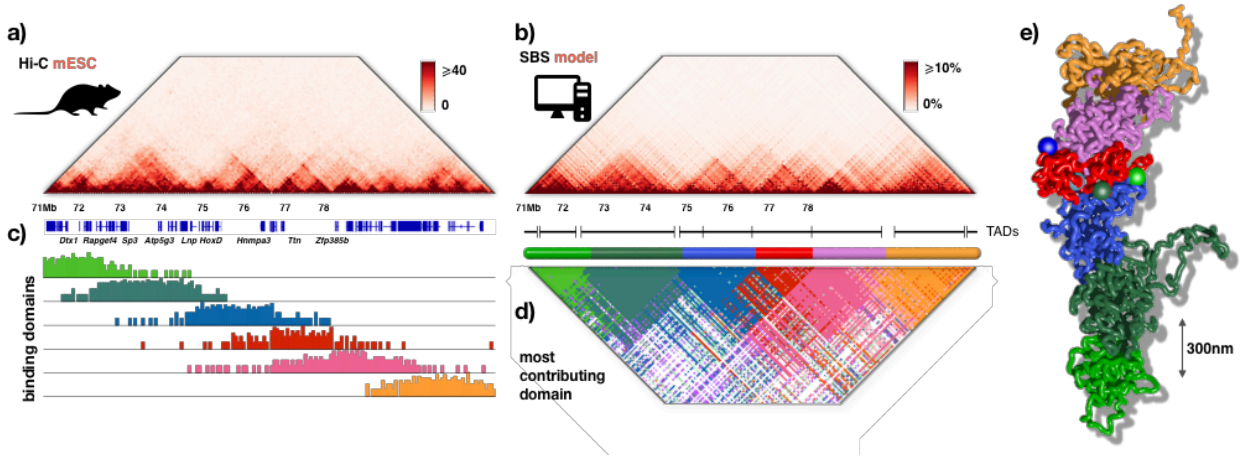

Figure 3.5: In mESC the SBS model describes with good accuracy Hi-C patterns in the extended *HoxD* locus.

a) Hi-C data (Dixon et al., 2012) and **b)** SBS model derived single-cell contact probabilities of the *HoxD* locus in mESC have a Pearson correlation r=0.96 and a distance corrected correlation r'=0.70. **c)** The model envisaged main binding domains are the top contributors to the contact patterns **d)**. The TADs of the locus (black segments) correspond to regions enriched for contacts between a specific type of binding sites. As binding domains overlap, TADs have interactions with each other, especially at their borders. A single-molecule time snapshot visualizes in 3D the structure of TADs (**e**). The bigger colored spheres in the structure highlight the position of the regulatory elements described in Figure 3. Figure adapted from (Annunziatella et al., 2018, submitted).

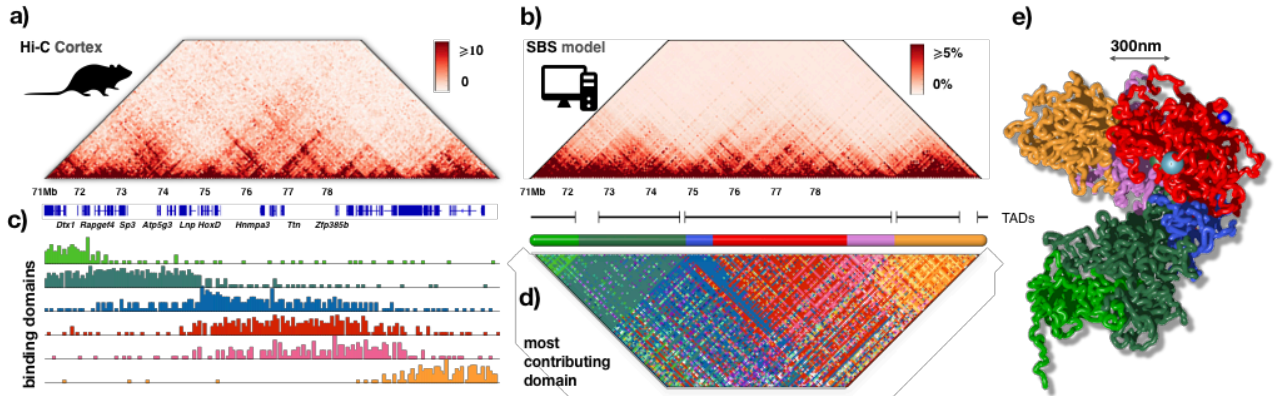

Figure 3.6: In Cortex cells the SBS model describes with good accuracy Hi-C patterns in the *HoxD* locus.

a) Hi-C data (Dixon et al., 2012) and **b)** the SBS model derived single-cell contact probabilities in Cortex have a Pearson and a distance corrected correlation r=0.92 and r'=0.71. **c)** The model main binding domains have broader overlaps in Cortex than mESC, producing interactions across TADs **d)**, as seen in the contact maps. Correspondingly, higher-order structures (meta-TADs) are formed, visible in the single-cell 3D time snapshot of the locus (**e**). Figure adapted from (Annunziatella et al., 2018, submitted).

The Pearson's correlation between model and Hi-C data is r=0.96 and r=0.92 in respectively mES and Cortex. Additionally, to consider the average decay of interactions with genomic distance, we also computed the distance corrected Pearson's correlation, i.e., the correlation between the contact matrices where the average decay is subtracted, which results to be respectively r'=0.70 and r'=0.71. Next, to dissect the origin of the contact patterns of the locus and to provide a principled definition of the otherwise heuristic notion of TAD, we investigated how such patterns arise from polymer physics by the interactions of the model binding sites. In mESC, the model identifies n=12 binding domains, 6 having the highest overlap with the TADs of the locus (**Figure 3.5, Panel c**) and giving the main contribution to the structure of chromatin contacts. **Figure 3.6, Panel d** visualizes the most contributing domain to each pairwise contact, visually illustrating that the TADs in the Hi-C data, identified by (Dixon et al., 2012), roughly correspond to DNA regions particularly enriched by contacts linked each to one of the top binding domains of the model. The binding domains tend to overlap with each other along the DNA linear sequence. Hence, while interactions within the bulk of a TAD are strongly associated with a single main binding domain, contributions from distinct domains overlap at TAD boundaries. That produces inter-TAD interactions and, correspondingly, the apparent blurred patterns at TAD boundaries in Hi-C data. The model also identifies other binding domains, not directly associated to a single TAD, which are more spread over the locus and contribute, in particular, to the weaker, yet non- negligible longer-range interactions across the locus, producing the visible, complex contact patterns. Similar results are found in Cortex cells (**Figure 3.6, Panel c, d**). Here, for visualization purposes, the color given to the binding domains is chosen based on the highest genomic overlap with the corresponding domain in mESC. Interestingly, in Cortex the top binding domains have stronger genomic overlaps with each other with respect to mESC, originating the higher-level of inter-TAD interactions seen in Hi-C data (**Figure 3.6, Panel c, d**). For example, the two TADs flanking the *HoxD* locus in mESC (blue and red, **Figure 3.6, Panel d**), in Cortex tend to intermingle with each other and to fold into a higher-order, meta-TAD structure. Typical representations of *HoxD* locus, shown in **Figure 3.5 and 3.6, Panel a** for mES and Cortex cell respectively, visually recapitulate the features expected from average contact matrix.

3.3.3 Investigating *HoxD* genes locus at single-cell level

Next, to characterize the level of cell-to-cell variability of the 3D structure of the *HoxD* cluster, we measured the distance distribution between *HoxD1* and *HoxD13* genes (**Figure 3.7: Panel a**). In our model, the distance distributions have a bimodal character because the population includes 3D structures in both the open (coil) and compact (globule) thermodynamic state. In figure, we show the distance distribution for the coil (colored in red) and globule polymer states (colored in blue), and the

derived distribution in coil/globule mixture (colored in grey). We find that the average distance is about 350nm in mESC, compatible with a previous, independent measure by FISH (Eskeland et al., 2010), while it slightly decreases in Cortex to 300nm. A closer proximity, however, does not imply more specific contacts because the standard deviation of the distance between, e.g., *HoxD1* and *HoxD13* increases 25% to 170nm from 130nm (**Figure 3.7: Panel a**), highlighting a higher population variability of the architecture in Cortex.

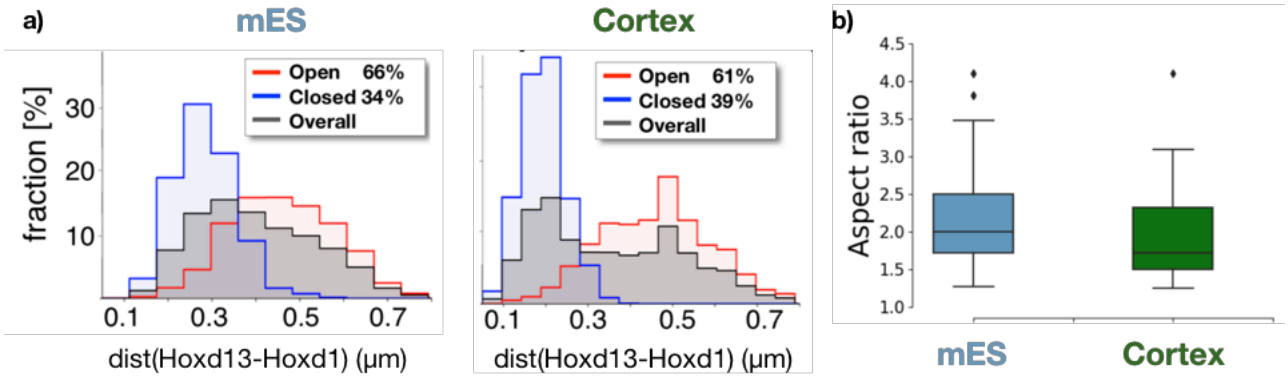

Figure 3.7: The silenced, more compact *HoxD* locus in Cortex has higher cell-to-cell variability and less specific contacts.

a) The model fit of Hi-C data returns that the locus is in an open (coil) state in 66% and 61% of the cell population respectively in mESC and Cortex, and in the closed (globule) state in the rest. The variance to average ratio of the *HoxD1-HoxD13* bimodal distance distribution is around 50%, highlighting a strong cell-to-cell variability in both mESC (left) and Cortex (right). The relative average distance change shows that genes are around 10% closer in Cortex, but the cell-to-cell variability is 25% higher, hinting that contacts are less specific in the more compact, silenced state. **b)** The aspect ratio, A, of the *HoxD* cluster has a mean value A~2.2 in mESC and A~1.9 in Cortex models. Their distributions are statistically different (p-value=0.002, Kruskal-Wallis test). Figure adapted from (Annunziatella et al., 2018, submitted).

Additionally, we computed the aspect ratio A of *HoxD* cluster, as the ratio between the principal axes of inertia $I_3^{1/2}$ and $I_1^{1/2}$ of the coil/globule mixture (here, $I_1 < I_2 < I_3$; for more detail see **Section 2.2.3**). We find that the *HoxD* cluster has on average an ellipsoidal shape, with an aspect ratio, A close to 2 (**Figure 3.7: Panel b**). However, in mES it is statistically higher than in Cortex cells (p-value=0.002, Wilcoxon/Kruskal-Wallis tests), showing that the cluster has a more elongated shape. This is consistent with previous single cell FISH observations in mES and forebrain cells (Fabre et al., 2015).

3.3.4 High-multiplicity regulatory contacts

At last step, we investigated the combinatorial nature of regulatory interactions in the *HoxD* locus, searching for high-multiplicity, many-body contacts. That information is straightforwardly derived within our polymer models but can be currently obtained only at much lower resolution by, say, Hi-

C or GAM experiments (Beagrie et al., 2017; Olivares-Chauvet et al., 2016). In our analysis, we focus on the triplets formed by the promoters and flanking regions, which we find to be strongly gene and cell-type specific organization. For each point of view, labelled with index k, we count a triplet contact if the pairs (i, k), (j, k) and (i, j) are simultaneously in contact, i.e. if their distances r_{ik}, r_{jk} and r_{ij} are all less than (or equal to) a fixed threshold distance $\lambda\sigma$ (here, λ=9 for Cortex and $\lambda=9n_L$ for mES case, where normalization constant n_L is the ratio between gyration radii in coil state, which takes into account the effects of differences in polymer length) and they are all of the same type. We finally normalized over the total number of all possible triplets (i, j, k). We did such analysis for coil/globule states separately and then we averaged over these states as discussed above. We find that many-body contacts are abundant in the system, albeit less frequent than pairwise ones, and statistically significant with respect to the expected random background measured in the only coil state (Wilcoxon test, pvalue<0.001).

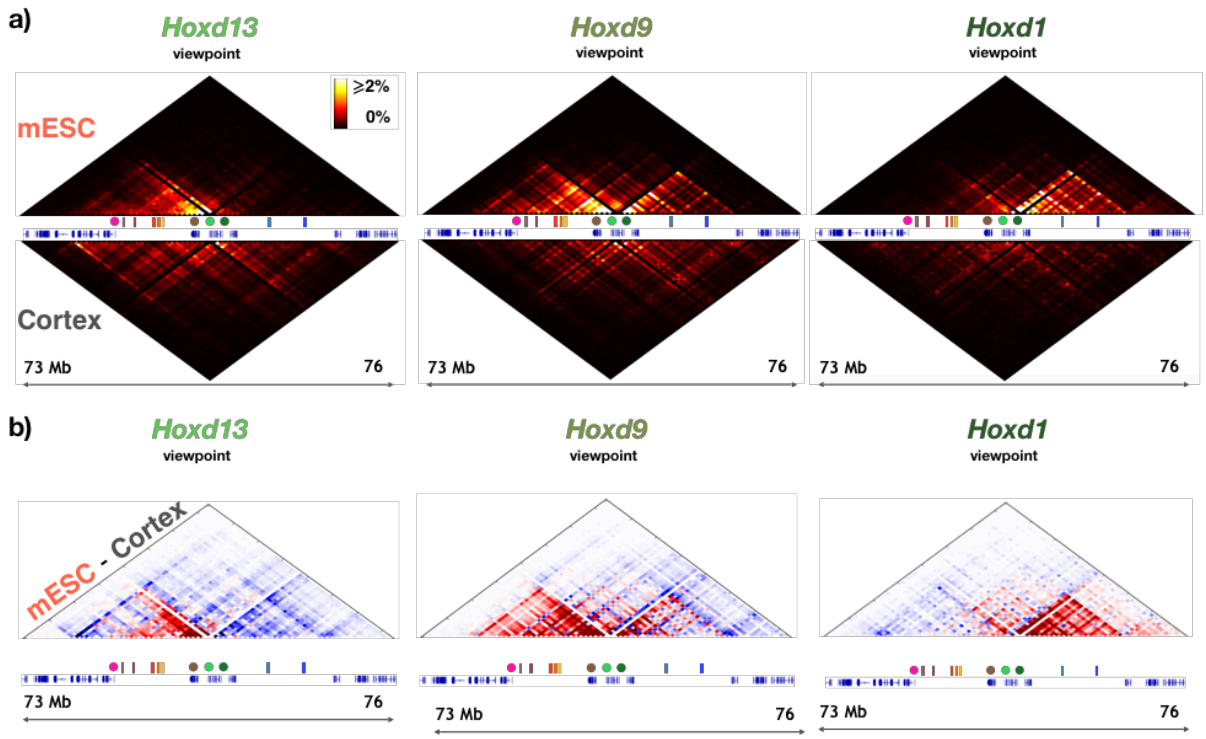

Figure 3.8: Triple contact probabilities of genes and regulators at the *HoxD* locus are gene and cell-type specific.

a) Model derived triplets contact probability from the viewpoint of *HoxD1, HoxD9* and *HoxD13* have a gene and cell type specific compartmentalized structure. In mESC, *HoxD13* and *HoxD1* form triplets especially within respectively centromeric (on left) and telomeric (on right) TADs, while *HoxD9* form triplets with both. Conversely, in Cortex they all share broader interactions within a larger meta-TAD. **b)** Heatmap showing subtractions (mESC-Cortex) to better visualize differences in the two different cell lines. In red, the triplet contacts more frequent in mESC, while in blue the contacts more frequent in Cortex. Figure adapted from (Annunziatella et al., 2018, submitted).

Notably, we find that the triplets formed by the *HoxD13, HoxD9* and *HoxD1* genes are almost exclusively restricted to their flanking TADs, showing that such multiple contacts are highly selective, even more than pairwise contacts (**Figure 3.8, Panel a**). Triplets are also compartmentalized. For instance, in mESC the triplets formed by *HoxD13* are confined mostly to sites in its centromeric TAD (C-DOM, on left in the figure), whereas those involving *HoxD1* to the telomeric TAD (T-DOM, on right in the figure); *HoxD9* shows an intermediate behavior forming triplets within both flanking TADs. In Cortex, conversely, the patterns of triple contacts from the *HoxD13*, *HoxD9* and *HoxD1* viewpoints are similar to each other, weaker in intensity and broadly distributed within the high-order meta-TAD encompassing the *HoxD* cluster and its flanking TADs. To better visualize differences between mES and Cortex triple contact patterns we produced subtraction matrices of the two cell types (**Figure 3.8, Panel b**). Our results return a picture where the *HoxD* locus is marked by a complex, cell type specific network of high-multiplicity regulatory contacts, where poised *HoxD* genes in mESC interact selectively and combinatorically within their flanking TADs. Conversely, in Cortex, upon silencing, they share unspecific contacts within their larger meta-TAD. That could be the mode of action of compartmentalization to fine tune specific gene activity.

3.4 Investigating tissue-specific interactions by 3D modeling

Most part of cell-type specific interactions between genes and enhancer occurs at sub-TAD scale, but the exact mechanisms regulating these interactions are still unknown (**Section 1.4**). In this Section, we investigate by 3D modeling such a mechanism for regulation of *Pitx1* gene expression, which plays a key role in development of the lower limbs. In particular, we focus on enhancer-promoter interaction driving the activation of the *Pitx1* gene in hindlimb (posterior limb) tissue, which is not activated, instead, in forelimb (anterior limb) tissue. This work has been developed in collaboration with Stepan Mundlos's research group at Max Plank Institute in Berlin, which performed all the experimental part (Kragesteen et al., 2018).

3.4.1 Emerging scenario from cHi-C data

In order to investigate the different chromatin features in *Pitx1* expression, associated with hind- and forelimb tissues, we first look at the cHi-C data, which encompasses a 3Mbp region around the *Pitx1* gene (chr13:54000000-57300000) from mouse at E11.5 stage in both tissues, at 10 kb resolution (Kragesteen et al., 2018). This genomic region is subdivided into different subdomains separated by the regulatory anchors *RAs (RA1* to *RA5)*, with *RA2* representing *Pitx1* promoter, and *RA5* representing Pen (pan-limb enhancer), an enhancer showing transcriptional activity in both limb buds.

As highlighted in subtraction contact matrix in **Figure 3.9**, the region shows some important differences between fore- and hindlimb tissue. In forelimb (**Figure 3.9, Panel a**), there is an increase in interaction between *Pitx1* and repressed *Neurog1* gene. In hindlimb (**Figure 3.9, Panel b**), instead, specific interactions involving *Pitx1* with *RA1*, *RA3* and *Pen* are more frequent. Importantly, in this case, the tissue-specific changes observed in *Pitx1* landscape cannot be due to differential binding of CTCF, which has no major differences at any of the regulatory anchors were evident.

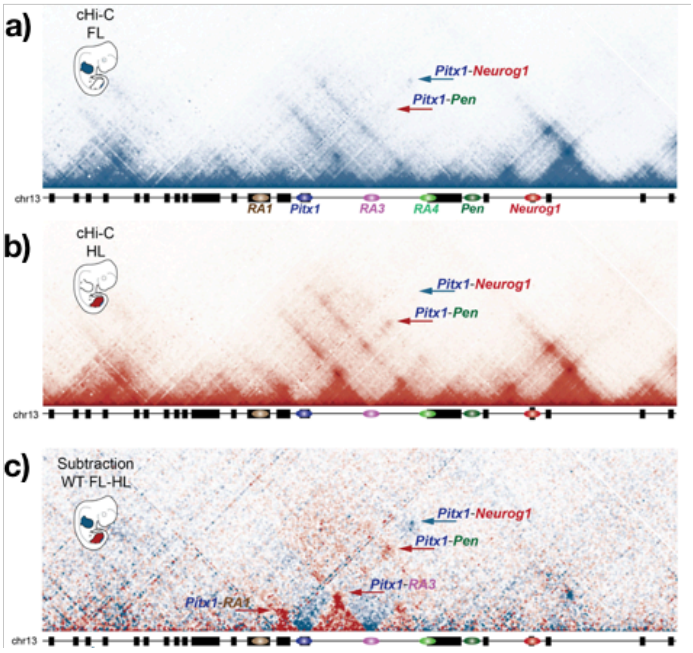

Figure 3.9: Investigation of tissue-specific interactions involving *Pen* and *Pitx1*.

a) cHiC map of forelimb tissue, at E11.5 stage, shows the presence of chromatin subdomains separated by the regulatory anchors. *Pitx1* shows moderate contacts with *RA3*, *RA4*, weak contacts with *Pen* (see red arrow) and a distal interaction with the *Neurog1* region (see blue arrow). **b)** cHiC map of hindlimb tissue at E11.5 display sharper subdomains separated by the regulatory anchors (grid with black boxes). *Pitx1* interacts strongly with *RA3*, *RA4,* and *Pen* (see red arrow), and shows little interaction with the *Neurog1* region (see blue arrow). **c)** cHiC subtraction between wildtype forelimb and hindlimb tissue at E11.5. Chromatin interactions more prevalent in wildtype forelimb and hindlimb tissues are shown in blue and red, respectively. Interactions between regulatory anchors that are more prevalent in forelimbs are indicated with blue arrow (*Pitx1-Neurog1* interaction). Chromatin interactions between regulatory anchors that are more prevalent in hindlimb are shown with red arrows *(Pitx1-RA1*, *Pitx1-RA3, Pitx1-Pen*). Figures adapted from (Kragesteen et al., 2018).

One of the major differences between forelimb and hindlimb is the interaction between *Pitx1* and *Pen* (**Figure 9, Panel c**). In this work, we show that it is possible to induce in forelimb tissue a hindlimb-like structure by inverting a 113 kb genomic region containing *RA4* and *Pen*. This inversion, indicated by $Pitx1^{inv1}$, brings genomically closer *RA4* and *Pen*, and leads a chromatin reorganization that is

nearly similar in both forelimb and hindlimb, as quantified by contact maps (**Figure 3.10**), where the physical proximity between *Pen* and *Pitx1* leads to activation of this latter one. The skeletal alteration observed in *Pitx1*^{inv1} inversion resemble those affected by the Liebenberg syndrome. As a control, a slightly smaller genomic region has been inverted (99 kb, indicated by *Pitx1*^{inv2}), which leaves *Pen* at its original location. *Pitx1*$^{inv2/inv2}$ embryos had a normal skeleton (not shown here, (Kragesteen et al., 2018)) and did not show ectopic expression of *Pitx1* in forelimbs, thus confirming the direct effect of the *Pen* element and its position on the mis-expression of *Pitx1* in *Pitx1*$^{inv1/inv1}$.

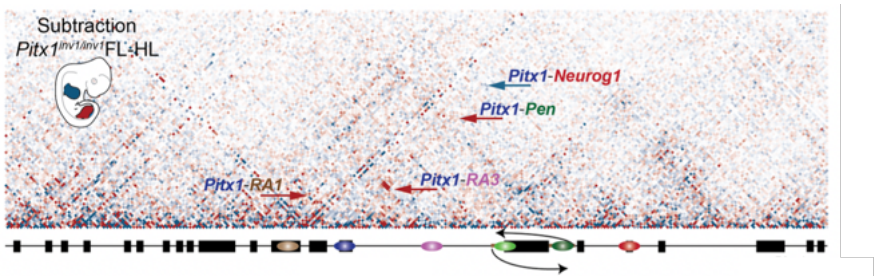

Figure 3.10: *Pitx1*$^{inv1/inv1}$ forelimb induces a hindlimb structure.
Subtraction maps of *Pitx1*$^{inv1/inv1}$ forelimb and hindlimbs. The subtraction map denotes the high similarity of 3D chromatin structure between forelimb and hindlimb tissue in comparison to wild-type animals.

3.4.2 Three-dimensional investigation of *Pitx1* landscape by 3D modeling

To obtain a three-dimensional characterization of the *Pitx1* locus, we employ our SBS model to infer the corresponding 3D organization starting from two-dimensional cHi-C data. In each of the studied cases, forelimb (FL) and hindlimb (HL) tissues, and forelimb *inv1* inversion, the specific SBS models for the *Pitx1* locus were established by the PRISMR algorithm, that finds the minimal number of different types of binding sites (n=14) and their arrangement along the chain (**Figure 3.11, Panel a, b**; **Panel 3.13, Panel a**) returning the best agreement between the corresponding cHi-C data (**Figure 3.11, top Panel c, d**; **Figure 3.13, top Panel b**) and the equilibrium pairwise contact map derived by the polymer model (**Figure 3.11, bottom Panel c, d**; **Figure 3.13, bottom Panel b**). To derive an ensemble of the model equilibrium 3D conformations we implemented Molecular Dynamics computer simulations, focusing on a broad genomic sequence encompassing the mouse *Pitx1* regulatory region to avoid boundary effects and, next, we focused on chr13:55600000-56650000 (mm9 assembly). Based on cHiC interaction data we used a polymer chain of N=1785 beads and we run MD simulations, where we set a molar concentration of binders c equal to 135 nmol/l, and the scale of the bead-binder interaction energy E_{int} equal to 1.0 k_BT and 8.1 k_BT, corresponding respectively to the coil and globule conformational state of the polymer. The initial configurations

evolve up to $5x10^8$ time-steps to approach stationarity, as measured by the plateauing of the gyration radius and of the mechanical energy and confirmed by the polymer scaling exponents. An ensemble of at least 10^2 different equilibrium configurations is derived by MD for each of the considered cases. The size σ of each bead of the polymer chain is approximately 17 nm (**Eq 2.7**).

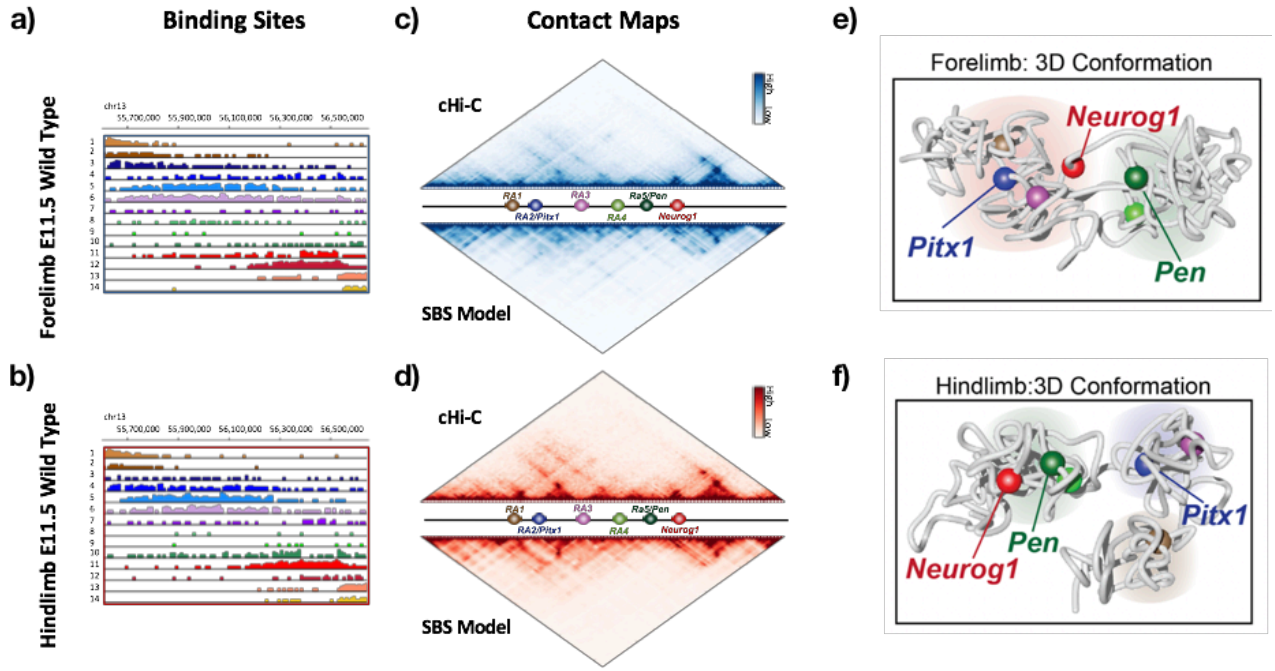

Figure 3.11: Tissue-specific 3D chromatin conformation can be reconstructed by SBS model.

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

Starting from our ensemble of 3d configurations, we compute the average contact matrices, fixing a dimensionless threshold λ=8. To take into account heterogeneity effects (as discussed above), we considered a coil/globule mixture and we find that 80%-20% mixture well describes all cases. The MD model v.s. cHiC Pearson correlation coefficient, r, is 0.98 in FL WT, 0.98 in HL WT, and 0.97 in the *inv1* forelimb case (**Figure 3.11, Panel c, d**; **Figure 3.13, Panel b**); whereas the distance-corrected correlation, r', is 0.84 in FL WT, 0.82 in HL WT, and 0.74 in the *inv1* FL case (here strong outliers above 90th percentile are excluded).

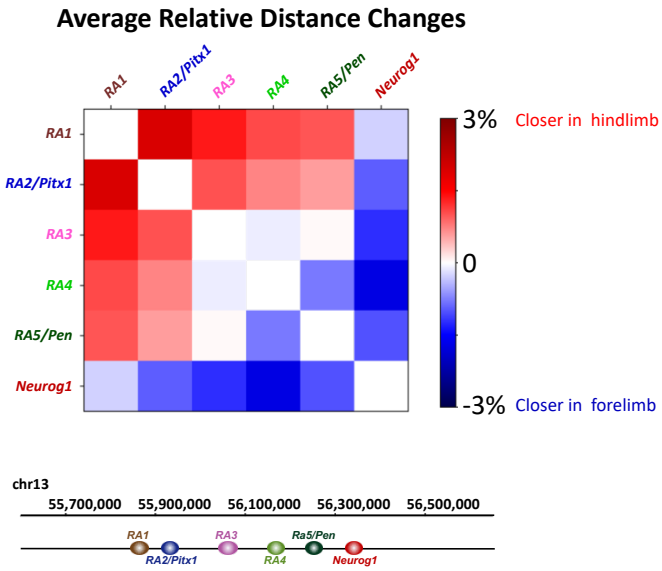

Figure 3.12: Enhancer and promoter interaction can be investigated by their physical proximity.

Heatmap showing relative changes in physical distances between forelimb and hindlimb 3D structure as measured from polymer conformations. Figure adapted from (Kragesteen et al., 2018).

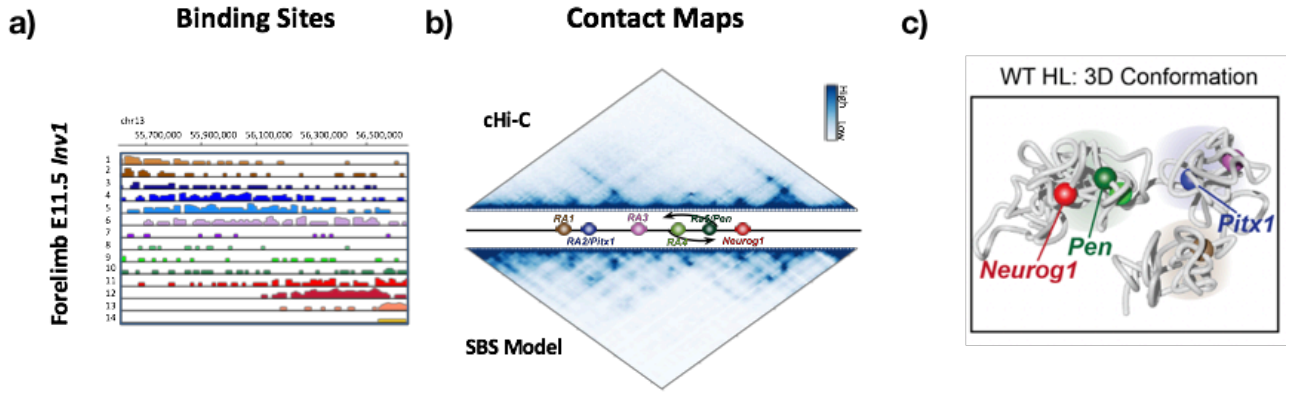

Figure 3.13: 3D chromatin changed caused by genomic mutations are well described by SBS model

a) Histograms displaying the position and abundance of n=14 different types of binding sites along the genome, in *Pitx1*$^{inv1/inv1}$ forelimbs at E11.5. **b)** Comparing of cHiC (above) against SBS model (below) derived contacts maps show high similarities. The Pearson correlation, r, and the genomic distance corrected Pearson correlation, r', between the cHiC and SBS matrices are r=0.97 and r'=0.74. **c)** A representative 3D-structure of the locus in *Pitx1*$^{inv1/inv1}$ forelimbs, selected from the ensemble of 'single-cell' model derived conformations. Figures adapted from (Kragesteen et al., 2018).

Next, to capture the structural differences in the *Pitx1* locus, we measured the physical distances between the regions of interest. The relative distance changes, shown in **Figure 3.12**, are the ratio

(d_{FL}- d_{HL})/ d_{FL} of the distances in FL and HL (resp. d_{FL}, d_{HL}) among *Pitx1* and its key regulatory regions (*RA*s) averaged over the discussed state mixture. Our results confirm the picture we supposed looking at experimental data. Interestingly, we find that the ensemble of thermodynamic stable structures in case of FL tissue shows two different chromatin hubs, one containing *Pitx1* together with *RA3* and *Neurog1*, and another containing *Pen* and *RA4*. The physical proximity between *Pitx1* and repressed *Neurog1* gene leads to *Pitx1* silencing (**Figure 3.11, Panel e**). On the other hand, in hindlimb case, chromatin structures show three different hubs, each containing respectively 1) *RA1*, 2) *Pitx1* and *RA3* and 3) *RA4* with *Pen* and *Neurog1* (**Figure 3.11, Panel f**). Although *Pitx1* and *Pen* residing in different hubs, they are physically closer than forelimb case, while *Neurog1*, that is positioned on opposite face of the hub, is farther from *Pitx1* and *Pen*. Modeling of $Pitx1^{inv1}$ locus from experimental data revealed a 3D conformation strictly similar to hindlimb wild-type conformation, where the three different hubs reappear (**Figure 3.13, Panel c**). Here, a single representative configuration of the *Pitx1* locus in the globule state is shown for each different cell type; to better visualize the relative positions of *Pitx1* and its *RA'*s, a coarse-grained version of the simulated polymer is pictured.

In conclusion, by 3D modeling analysis, we show that *Pitx1* expression is controlled by *Pen* enhancer. In hindlimb, where *Pen* and *Pitx1* are physically closer (**Figure 3.12**), the gene is activated. In forelimb, instead, the gene is inactivated, and indeed *Pen* and *Pitx1* result spatially separated. Structural variant in forelimb can convert the inactive conformation into an active conformation, by bringing close *Pen* and *Pitx1*. These induce an aberrant *Pitx1* expression in the forelimb and a hindlimb-like structure, causing partial arm-to-leg transformation in mice and humans.

References

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science (80-.). *340*.

Annunziatella, C., Chiariello, A.M., Bianco, S., and Nicodemi, M. (2016). Polymer models of the hierarchical folding of the Hox-B chromosomal locus. Phys. Rev. E *94*.

Annunziatella, C. , Bianco, S., Andrey, G., Chiariello, A.M., Esposito, A., Fiorillo, L., Prisco, A., Conte, M., Campanile, R., Nicodemi, M. (2018). *Single-molecule conformations of the HoxD locus in mouse ES and Cortex cells*. Cell Reports. (*submitted)*

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326.

Bianco, S., Chiariello, A.M., Annunziatella, C., Esposito, A., and Nicodemi, M. (2017). Predicting chromatin architecture from models of polymer physics. Chromosom. Res. *25*, 25–34.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

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

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. Front. Neurosci. *11*.

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

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A., et al. (2010). Ring1B Compacts Chromatin Structure and Represses Gene Expression Independent of Histone Ubiquitination. Mol. Cell *38*, 452–464.

Fabre, P.J., Benke, A., Joye, E., Nguyen Huynh, T.H., Manley, S., and Duboule, D. (2015). Nanoscale spatial organization of the *HoxD* gene cluster in distinct transcriptional states. Proc. Natl. Acad. Sci. *112*, 13964–13969.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schï¿½pflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A.M., Jerković, I., et al. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. *50*, 1463–1473.

Noordermeer, D., and Duboule, D. (2013). Chromatin Architectures and Hox Gene Collinearity. Curr. Top. Dev. Biol. *104*, 113–148.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The

dynamic architecture of Hox gene clusters. Science (80-.). *334*, 222–225.

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., and Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. Elife *2014*.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296–300.

Schuettengruber, B., Bourbon, H.M., Di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57.

Soshnikova, N., and Duboule, D. (2009). Epigenetic temporal control of mouse hox genes in vivo. Science (80-.). *324*, 1321–1323.

Version, P. (2004). POV-Ray Reference. Am. J. Surg. *187*, 114–119.

4. Predicting Structural Variants effects on chromatin architecture

In the previous Chapter, we showed that by using the String&Binders Switch model we can explain folding mechanisms of real, specific genomic regions at high-resolution scale and reconstruct their three-dimensional spatial organization. This is made possible by combining PRISMR, a machine-learning algorithm that uses information from the Hi-C experimental data, and Molecular Dynamics simulations of the SBS model.

In this Chapter, we show a stringent test for our polymer physics model. We will show how the SBS model can predict the structural changes of a DNA locus caused by genomic rearrangements along its genomic sequence. As revealed by recent studies, indeed, the 3D architecture, and especially the TAD structures, can be disrupted by genomic rearrangements, called structural variants (SVs), such as deletions, duplication or inversion of specific genomic regions. SVs can result in a re-wiring of enhancer-promoter contacts, and lead to gene mis-expression and disease (**Section 1.4**). Until now, the only chance to estimate ectopic interactions was performing extensive 3C-based experiments (**Section 1.3**). In this scenario, PRISMR represents a valid approach to predict, *in-silico*, such interactions, thereby providing a tool for analyzing the disease-causing potential of SVs.

In the following, we will focus on a set of SVs involving *EPHA4*, a gene associated with limb malformations (Lupiáñez et al., 2015), across four different mouse and human cell lines. The work we will discuss was developed in collaboration with group of Prof. Stepan Mundlos at Max Plank Institute in Berlin, which performed the experimental part. In **Section 4.1**, we give an overview of the experimental dataset we used in our investigation. In **Section 4.2**, we apply our model to describe the *Epha4* locus for wild-type case using all four datasets. In **Section 4.3** and **Section 4.4** we compare the model predictions, in mouse and human cell respectively, in case of SVs, against to capture Hi-C data from mouse limb buds and patient-derived fibroblasts.

Most of the material presented in this Chapter, including figures, paragraphs and sentences, is taken literally from the paper (Bianco et al., 2018), which I co-authored. For sake of brevity, we do not discuss other results about SBS/PRISMR predictions, shown, e.g., in (Chiariello et al., 2016) and (Annunziatella et al., 2018, submitted).

4.1 Epha4 locus: studied datasets

In order to test the predictive power of PRISMR method, when genomic structural variations of wild-type (WT) genomic sequence are present (**Figure 4.1**), we chose the *Epha4* locus, a genomic region having a key role in limb development and associated to types of limb malformation (Lupiáñez et al., 2015).

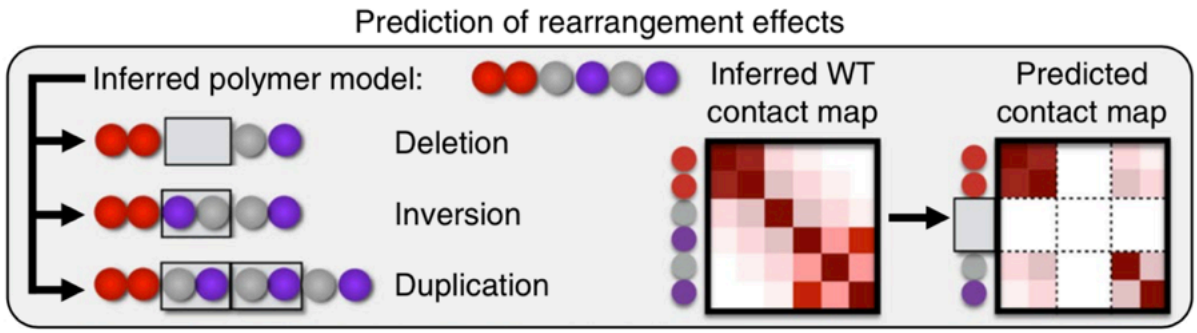

Figure 4.1: SBS model to predict effect of structural variants.
By informing the model inferred from wild-type (WT) data (**Section 3.1**) with a given rearrangement, the effects of genomic mutations on folding can be predicted from only polymer physics without any fitting parameters.

This work has been developed in collaboration with Stepan Mundlos's group at Max Plank for Molecular Genetics in Berlin, which performed capture Hi-C data and produced the datasets that we used for the modeling part. In a previous study the same group, by using 4C technology (**Section 1.3**), showed that structural variants (SVs) at the *Epha4* locus (such as deletions, duplications and inversions), cause distinct phenotypes: the alteration of chromatin organization of the locus can cause rewiring of enhancer-promoter contacts and then lead to gene mis-expression (**Section 1.4**). To this work, new and more complete cHi-C experiments have been performed from E11.5 mouse limb buds and human skin fibroblast. In addition, we also used already published Hi-C dataset from CH12-LX murine and IMR90 human cells (Rao et al., 2014).
The murine *Epha4* locus discussed is a 6 Mb long region around the *Epha4* gene (chr1:73000000-79000000, mm9). In this case, we employed in situ Hi-C data from CH12-LX cells using already published data (Rao et al., 2014) and we performed cHi-C experiments in E11.5 mouse limb buds, at 10 kb resolution. The studied human *Epha4* locus in skin fibroblasts is 5.77 Mb long (chr2:218320000-224090000, hg19); in that system, we produced our cHi-C data at 10 kb resolution. We also studied the *Epha4* locus in human IMR90 cells, where we used previously published in situ

Hi-C data at 10 kb resolution (Rao et al., 2014); the considered locus is 8 Mb long (chr2:217000000-225000000, hg19).

Next, to test the model predictions, new experimental cHi-C dataset was used from mouse limb buds carrying homozygous structural variants. By term homozygous, we mean that both alleles at the *Epha4* locus show the same mutation on the homologous chromosomes. The mutations we considered are: 1.6 Mb deletion, *DelB* (chr1:76388978-78060839, mm9), a 1.5 Mb deletion, *DelBs* (chr1:76388978-77858974, mm9), and a 1.1 Mb inversion, *InvF (*chr1:74832836-75898707, mm9).

A structural variant is, instead, 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 (present on only one chromosome) 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 (chr2:221276849-223021152, hg19, similar to mouse *DelB*), a 900 Kb duplication, *DupP* (chr2:219875536-220789199, hg19) associated with polydactyly and *IHH* activation, and a 1.4 Mb duplication, *DupF* (chr2:219713606-221090946, hg19) associated with syndactyly and *WNT6 gene* activation (Lupiáñez et al., 2015).

All the datasets have been normalized applying the Knight and Ruiz (KR) normalization (Knight and Ruiz, 2013). The KR normalization is 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. However, in the following, we will overlook at all the biological and chemical details of the cHi-C experiment, biological samples, preparation of the libraries and sequencing since we did not work directly to the experimental stage.

4.2 PRISMR models of the murine *Epha4* locus

The different wild-type Hi-C datasets we introduced in the previous Section, are shown in **Figure 4.2** and **Figure 4.3**, on top. Regardless of the cell or tissue type or the species, we observed a subdivision of the locus in one large TAD, containing only *EPHA4*, in a smaller TAD, containing *PAX3* and *SGPP2*, and in a gene-dense region on the centromeric side, showing no clear TAD structure. Differences were apparent within the *Epha4* TAD that likely reflect cell- and tissue-specific patterns of interaction and gene regulation (Kragesteen et al., 2018; Phillips-Cremins et al., 2013).

Toward developing predictive models of the architecture of the *Epha4* locus across different cell types, we applied PRISMR to all four Hi-C datasets. In the studied murine *Epha4* locus, the algorithm returns n=21 in both the published in situ Hi-C data of CH12-LX cells (**Section 3.1**) (Rao et al., 2014), and in our limb tissue cHi-C data. In the considered human *Epha4* locus, PRISMR finds n=16 and in the published in situ Hi-C data in human IMR90 cells (Rao et al., 2014), and n=24 in the cHi-C data

in human fibroblast produced in this study. The model contact matrices, derived by full-scale MD simulations of optimal polymer found by PRISMR (**Section 3.2**), are similar to the original Hi-C data, not only recapitulating the global TAD conformation of the locus, but also capturing cell-specific intra-TAD organization: the Pearson correlation, r, and distance-corrected correlation coefficient, r', range up to $r = 0.95$ and $r' = 0.69$ (**Figure 4.2, Table 4.1**).

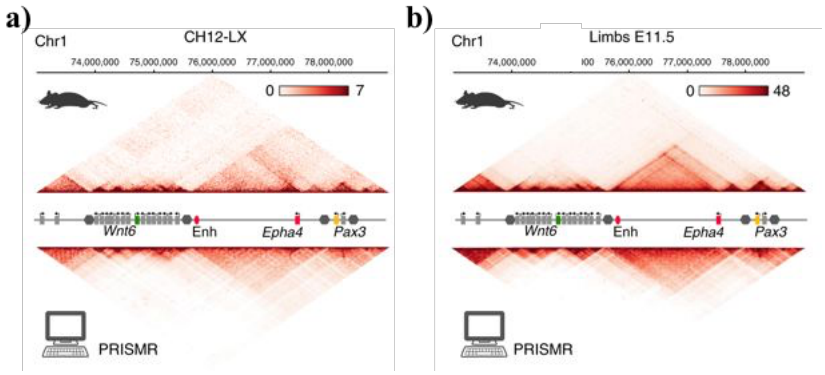

Figure 4.2: PRiSMR recapitulates 3D conformation at the *EPHA4* locus on mouse cells.

a) Published Hi-C data (CHR12-LX) (Rao et al., 2014) and **b)** capture Hi-C data (Limbs E11.5), (Bianco et al., 2018) compare well with the contact matrices derived by PRISMR/SBS. Their Pearson correlations, r, and distance-corrected Pearson correlation coefficients, r', are comparatively high: r = 0.91, $r' = 0.56$ in CH12-LX; and $r = 0.94$, $r' = 0.60$ in limb tissue E11.5 (**Table 4.1**). We show schematically genes with rectangles, TAD boundaries with hexagons, and enhancers. Additionally, relevant genomic elements are highlighted with colors and corresponding names. Figures adapted from (Bianco et al., 2018).

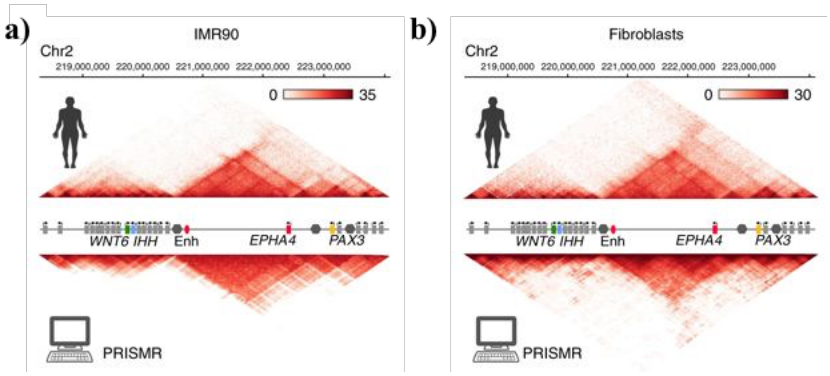

Figure 4.3: PRISMR also recapitulates 3D conformation at the *EPHA4* locus on human cell lines.

a) Published Hi-C data (left) (Rao et al., 2014) and **b)** capture Hi-C data (right) (Bianco et al., 2018). compare well with the contact matrices derived by PRISMR. Their Pearson correlations, r, and distance-corrected Pearson correlation coefficients, r', are comparatively high: $r = 0.92$, $r' = 0.64$ in IMR90; $r = 0.93$, $r' = 0.69$ in human fibroblasts (**Table 4.1**). Figure adapted from (Bianco et al., 2018).

4.2.1 Statistical significance and robustness of identified 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 minimal distribution of binding sites in CH12-LX cells (i.e., the best polymer model) found for n=21 with the distribution of binding sites 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. For the *Epha4* locus, we find that the in-situ Hi-C contact matrix in CH12-LX cells has a Pearson correlation coefficient equal to r=0.95 with the PRISMR predicted contact matrix in the n=21 case. An 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. Similar results are found in the other studied datasets.

To check the level of randomness inherent to the binding domains identified by PRISMR, we compared their overlap (see overlap definition **Section 4.2.4**) 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 case, 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 by bootstrapping. We found that the average overlap between domains within 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.

Taken together our results support the view that the optimal polymer identified by PRISMR and its binding domains are far from random and robust to changes in the parameters of the algorithm.

4.2.2 Epigenomic barcoding of binding domains in *EPHA4* locus

As an initial step to investigate the molecular nature of the factors contributing to define the different types of binding sites ('colors') envisaged by PRISMR, we derived their epigenomic barcode. In murine erythroleukemia CH12-LX cells, chromatin data are available from the ENCODE project database (Dunham et al., 2012) that we use to characterize the binding domains identified by PRISMR, as discussed for *Sox9* locus case (**Section 3.2**). We crossed the information about their genomic positions with a number of published chromatin features, such as histone modifications and transcription factors. Specifically, for each binding domain ('color') and for each chromatin feature we calculated the Pearson correlation coefficient between the number of binding sites of that domain at 10 kb resolution and the number of called peaks present in those 10 kb wide bins (by at least a base pair) as identified by the *bedtools coverage* tool (**Figure 4.4**) (Quinlan, 2014; Quinlan and Hall, 2010).

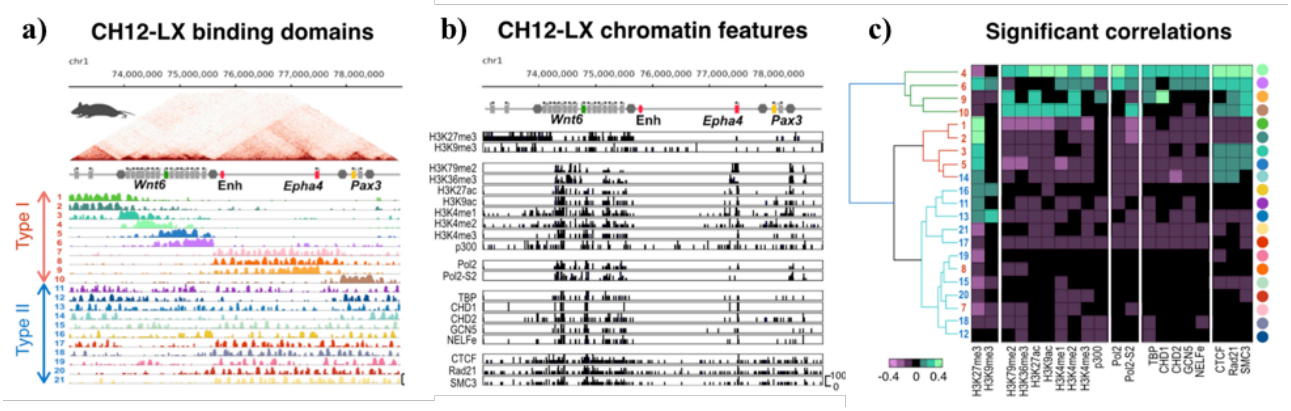

Figure 4.4: Epigenetic barcoding of the binding domains envisaged by PRISMR in the Epha4 locus of CH12-LX cells.

a) In the Epha4 locus of CH12-LX murine cells, PRISMR method envisages n=21 statistically significant (Wilcoxon's rank-sum test; $P < 1.1 \times 10^{-7}$) different binding domains, each represented with a color. **b)** Signal, from ENCODE dataset, for different chromatin features for the Epha4 region. **c)** Matrix with the statistically significant Pearson correlation coefficients of the different binding domains (shown in Panel a) with the ENCODE signals (shown in Panel b). The domains have been clustered according to the similarity of their epigenetic barcode. Figures adapted from (Bianco et al., 2018).

Afterwards, in order to find only statistically 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 colors do not correspond to single molecular factors, as each usually correlates with a combination of different marks. Finally, a hierarchical clustering was performed on the significant correlations matrix by using the *Python SciPy* clustering package (Oliphant, 2007). From the clustering analysis, a non-trivial relationship emerges between binding domains and epigenetic features. For instance, we find that Type-I binding domains (**Figure 4.4, Panel a bottom; Section 4.2.4**) can be broadly subdivided into 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 (Nora et al., 2017; Schwarzer et al., 2017). 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., 2018), MLL3/4 (Yan et al., 2017), Active/Poised Pol-II (Barbieri et al., 2017), etc. Additionally, many other colors have no significant correlation with CTCF/Cohesin. Type-II colors (**Section 4.2.4**) can also be subdivided 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. The statistical meaning of the anti-correlations found between some of the types of binding sites (colors) and some histone modifications is that the presence at specific genomic sites of one histone modification coincides with the absence of the other.

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 (Hug et al., 2017; Kubo et al., 2017; Kundu et al., 2018; Yan et al., 2017).

4.2.3 The 'PRISMR + CTCF' method

Next, to explore the roles of various factors to the folding patterns detected by PRISMR simulations of cHi-C data, we considered the architectural protein CTCF, a DNA-binding transcription factor thought to facilitate the formation of chromatin loops (**Appendix A**)(Fudenberg et al., 2016; Nora et al., 2012; Rao et al., 2014; Sanborn et al., 2015). Notably, some binding site types identified by PRISMR correlated with CTCF (**Figure 4.4, Panel b, c**; **Section 4.2.2**). Although PRISMR does not exploit prior information on binding sites and factors, to test its reach we considered a variant of the model in which we included previous knowledge about the location of CTCF binding sites in the locus, which were added to interact with an additional type of binder that bridges opposed (forward or reverse) CTCF binding sites (see **Section 4.2.4**, for more detail).

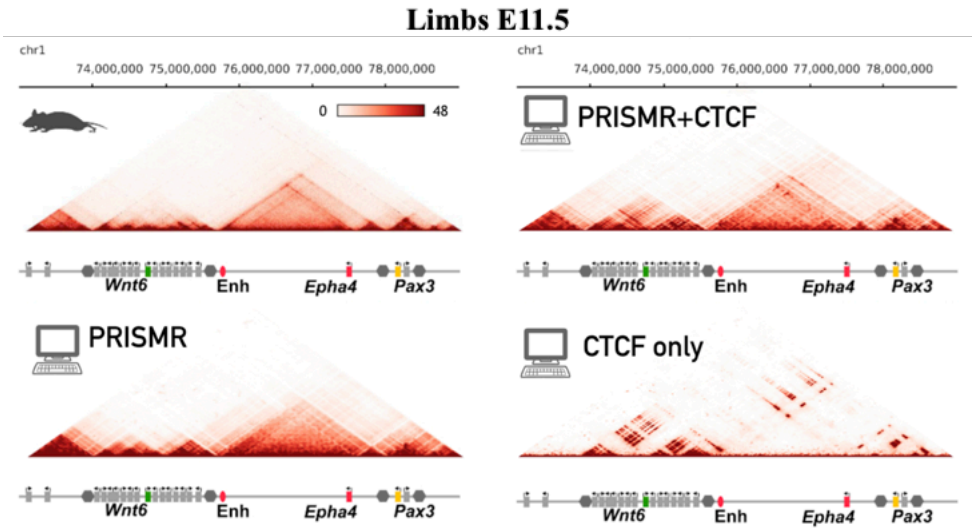

Figure 4.5: Comparison of original PRISMR model, with the model that also includes prior knowledge on CTCF and with a model with only CTCF.

The figure shows the contact matrices from cHi-C data in mouse E11.5 limb bud tissue and from three different models. The bottom left panel reports the results derived by MD simulations of PRISMR: 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 includes a-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). Figures adapted from (Bianco et al., 2018).

In limb tissue E11.5 cells, for instance, this variant (named 'PRISMR + CTCF') had correlations with Hi-C data similar to those of the initial model: it improved the visualization of the large *Epha4* TAD, mainly by strengthening the loop anchors characteristic for CTCF-associated loops, but it also resulted in additional contacts in the neighboring gene-dense region that were not present in the original cHi-C data (**Figure 4.5, top right**). Conversely, a model with only CTCF (named 'CTCF-only') can describe some of the loops seen in the data, but poorly captured the global contact patterns of the *Epha4* locus (**Figure 4.5, bottom right**), resulting in a lower correlation coefficient ($r' = 0.05$). These results indicate that other factors besides CTCF were 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.

4.2.4 Computational details

In this Section, we briefly discuss some details about the implementation of 'PRISMR' and 'PRISMR + CTCF' models by Molecular Dynamics simulations and about statistical analysis of our results.

Characterization of the identified binding domains

The different n binding domains identified by PRISMR in the best polymer 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:

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

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 (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 (see **Section 4.2.1**), we used their overlaps with TADs: 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.

MD simulation details

To model the mouse wild-type (WT) *Epha4* locus (**Figure 4.2**), we used the same parameters for murine CH12-LX cells and for E11.5 limbs. In our MD simulations we use an SBS polymer chain with N=12600 beads. The corresponding genomic content per chain bead is s_0 = L/N= 476bp. The physical diameter of the bead σ is approximately estimated by assuming a chromatin compaction factor of 50bp/nm (i.e., an intermediate value between the 30 nm fiber and the naked DNA (Bohn and Heermann, 2010)), so we obtain σ≈10nm. 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 8.2 k_BT, corresponding to the polymer coil and globule state respectively (**Section 2.3**). The coil-globule transition is identified by collapse of the gyration radius, R_g, of the polymer, from the SAW predicted value of the coil state to the much lower value in the globule state (**Section 2.2**). To approach stationarity, our simulations run up to 10^9 timesteps. Our ensemble averages span up to $4x10^2$ independent runs for each set of system parameters. The same parameters have been used to simulate the model derived from our murine E11.5 cell cHi-C data (**Figure 4.2**).

To model the human wild-type *Epha4* locus in human skin fibroblast cells (**Figure 4.3**), we used a polymer made of N=13848 beads, so the genomic content results s_0=417bp and σ=8.3nm. As before, 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 range explored of the total binder concentration, *c*, was from zero to *c*=300*nmol/l*, and the interaction energy used is the same than in the mouse cases above. Finally, our polymer models of the *Epha4* locus in IMR90 human cells, were made of N=12800 and interaction energy and concentrations were in the same range of the other simulations.

Contact matrix analysis

To extract the average pairwise contact frequency matrices of the polymer model, we proceed as discussed **Section 3.2.1**. For all the mentioned datasets, we computed the matrices with the parameter λ ranging from 2 to 10, and we find the mixture composition that maximizes the correlation coefficient between the model predicted and experimental contact matrices. We find similar results in all cases. For example, in the model of the murine wild-type *Epha4* locus in mouse CH12-LX cells, we find a 89%-11% open-closed mixture and a correlation coefficient *r*=0.91 with in-situ Hi-C data.

The corresponding contact matrices are shown, e.g., in **Figure 4.2**. As Hi-C matrices are computed from sequencing reads, a scale factor must be used in the comparison. Analogously, in the case of our human fibroblast cell data (**Figure 4.3**, 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 (**Table 4.1**). Since our cHi-C experimental are from heterozygous mutants for the human fibroblast cells, the corresponding simulated contact matrices are equally averaged with the simulated healthy control.

'PRISMR + CTCF' Model

To investigate the performance of our model in the case of murine E11.5 cell cHiC data, we tested it against models where previous knowledge on a known, important chromatin organizer such as CTCF is taken into account (**Section 4.2.3**).

First, we considered a variant of our PRISMR model (named 'PRISMR+CTCF') where CTCF peaks are added and are supposed to interact with an additional type of binders that can only bridge opposed (forward/reverse) CTCF sites. Specifically, 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 (Andrey et al., 2017) 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 by using the FIMO tool in the MEME Suite online software (Grant et al., 2011) to identify the best matching peak, within the considered 10 kb 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.

To speed up MD simulations of this model (i.e., PRISMR 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. To explore the effects of the initial condition, we also considered, for WT and inversion, an independent second ensemble of starting conformations where we placed springs also between forward CTCF sites and their second left and right nearest neighbor reverse binding sites. After the initial state is equilibrated by MD, the springs are removed and replaced by the specific CTCF binders, until equilibrium is reached, as discussed in the section on our MD simulations (**Section 2.1**). The contact matrices are then computed and averaged over the two ensembles (**Figure 4.5**). We find, as expected, that the novel extended model slightly improves

the visualization of loops in the *Epha4* TAD but does not improve the global comparison against cHi-C data (**Table 4.1**).

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 'colors' 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. Such 'CTCF only' model can describe some of the 'loops' seen in cHi-C data (i.e., the contact peaks at TAD vertices (Rao et al., 2014)), but poorly captures the global contact patterns of the *Epha4* locus (its distance-corrected Pearson correlation with cHi-C is r'= 0.05, **Figure 4.5, bottom right**).

Taken together, our analyses show that albeit our PRISMR approach does not require *a-priori* information on binding factors, it can recapitulate Hi-C data and previous biological knowledge about important chromatin organizers such as CTCF. Instead, a minimal model considering only CTCF is unable to explain the broader pattern of contacts seen in cHi-C data.

4.3 PRISMR Model predictions on mouse cells

To test whether PRISMR can predict the effects of homozygous SVs on chromatin folding, we investigated three previously reported variants at the *Epha4* mouse locus (Lupiáñez et al., 2015): a deletion (*DelB*) encompassing a large part of the *Epha4* TAD and the telomeric TAD boundary (associated with brachydactyly due to misexpression of *Pax3*), a slightly smaller deletion (*DelBs*) that leaves the TAD boundary intact (no misexpression, no phenotype), and a balanced 1.1-Mb inversion (*InvF*) that causes misexpression of *Wnt6*. We implemented these mutations in polymer models of the wild-type E11.5 limbs (**Figure 4.6, Panel a; Figure 4.7, Panel a**) and CH12-LX (**Figure 4.8, Panel a**) cells inferred by PRISMR and re-ran the ensemble of folding conformations to derive an average locus contact matrix. For E11.5 limb tissue, we tested both the PRISMR model (**Figure 4.7, Panel a**) and the 'PRISMR + CTCF' (**Figure 4.6, Panel a**) version with the addition of CTCF sites (**Section 4.2.4**).

To identify the regions of statistically significant ectopic interactions in each predicted rearrangement, we subtracted each mutant matrix from the wild-type matrix (**Figure 4.6, Panel b; Figure 4.7, Panel b; Figure 4.8, Panel b**), as described in next **Section 4.3.2**. Although the studied locus is populated by more than 40 genes, our matrices predicted that only certain regions, containing a limited number of genes, would display changes in the interaction profiles. For example, in the larger deletion (*DelB,* **left columns**) including the *Epha4* TAD boundary, we identified new contacts that predicted 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 (*DelB*s, **middle columns**), which leaves the *Epha4-Pax3* boundary intact. Moreover, virtual 4C analysis derived from our predictions showed that the enhancers–*Pax3* ectopic interaction was diminished, consistent with the absence of *Pax3* activation in these mutants (**Figure 4.6, Panel c; Figure 4.7, Panel c; Figure 4.8, Panel c**).

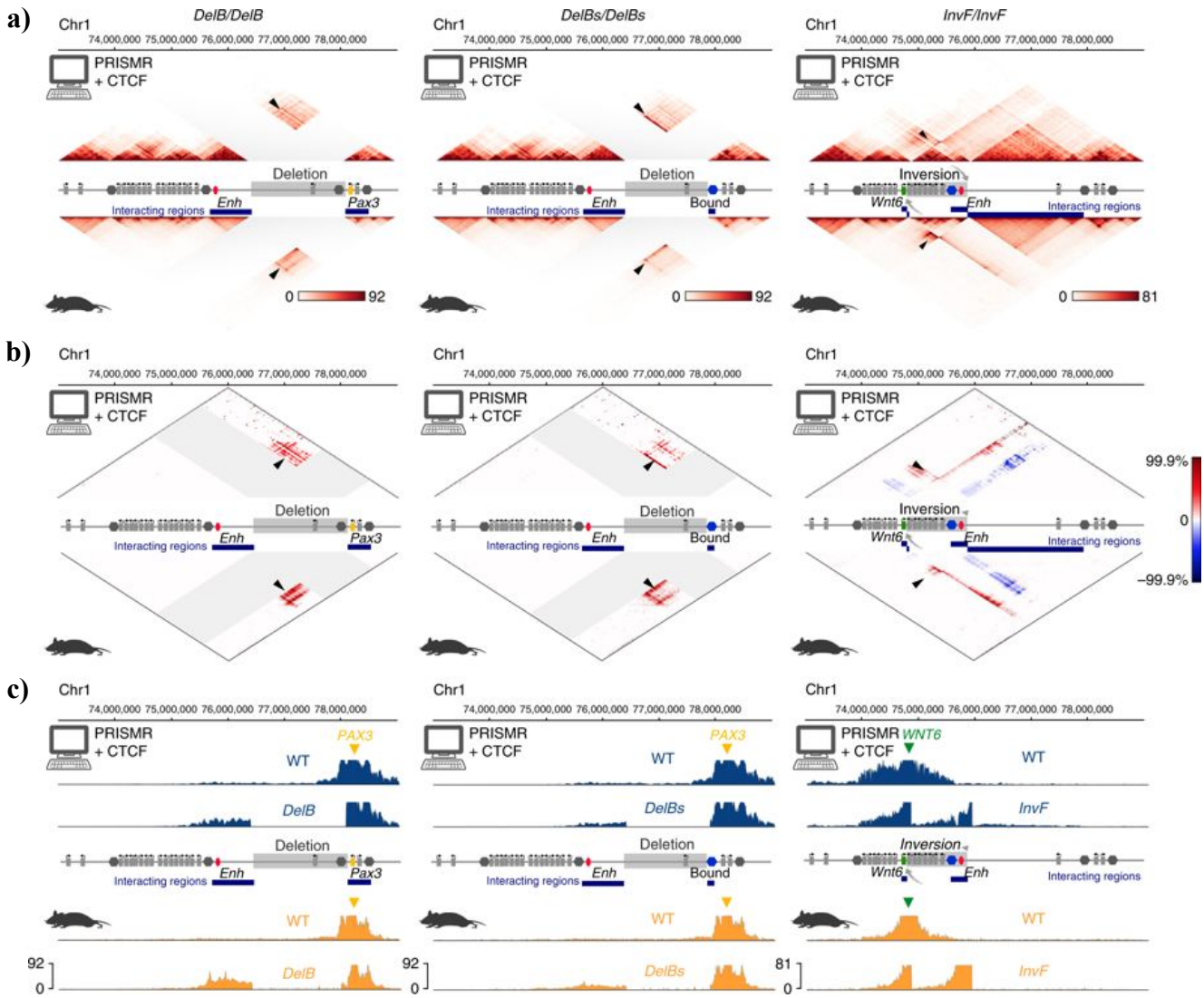

Figure 4.6: 'PRISMR+CTCF' predicts the effects of mouse homozygous structural variants on chromatin architecture.

a) Contact matrices from model predictions derived from WT data (top) and cHi-C experiments performed in E11.5 limb buds from mouse mutants. *DelB/DelB*: PRISMR prediction on a 1.6-Mb homozygous deletion affecting *Epha4* TAD and *Epha4–Pax3* boundary (Pearson correlation $r = 0.95$, distance-corrected Pearson correlation $r' = 0.41$). Note the increased interaction between remaining *Epha4* and *Pax3* TADs (arrowhead and blue bars). *DelBs/DelBs*: 1.5-Mb deletion affecting *Epha4* TAD but not the *Epha4–Pax3* boundary ($r = 0.95$, $r' = 0.50$). Note the increased interaction between remaining *Epha4* TAD and *Epha4–Pax3* boundary (blue hexagon). *InvF/InvF*: 1.1-Mb homozygous inversion ($r = 0.95$, $r' = 0.60$) (**Table 4.1**). Note the increased interaction between enhancer and *Wnt6*

regions. An additional region at the centromeric inverted position (containing *Wnt10a*) gains interaction with *Epha4* TAD. The centromeric *Epha4* boundary retains functionality despite inversion (blue hexagon). **b)** Subtraction maps (WT and mutants) from predictions and cHi-C data. Top: threshold gain (red) and loss (blue) of interaction is displayed (absolute differences > 2 s.d.). Ectopic interactions are indicated (arrowheads and blue bars). **c)** Virtual 4C plots derived from predictions and cHi-C data from viewpoints on the respective phenotype-causing genes. *DelB/DelB*: note increased interaction of the *Pax3* promoter with the remaining *Epha4* TAD, including enhancer cluster in both prediction and experimental data. *DelBs/DelBs*: the *Pax3* promoter interacts less frequently with *Epha4* TAD compared to *DelB/DelB* mutants. *InvF/InvF*: increased interaction between *Wnt6* gene and *Epha4* enhancer cluster. Figures adapted from (Bianco et al., 2018).

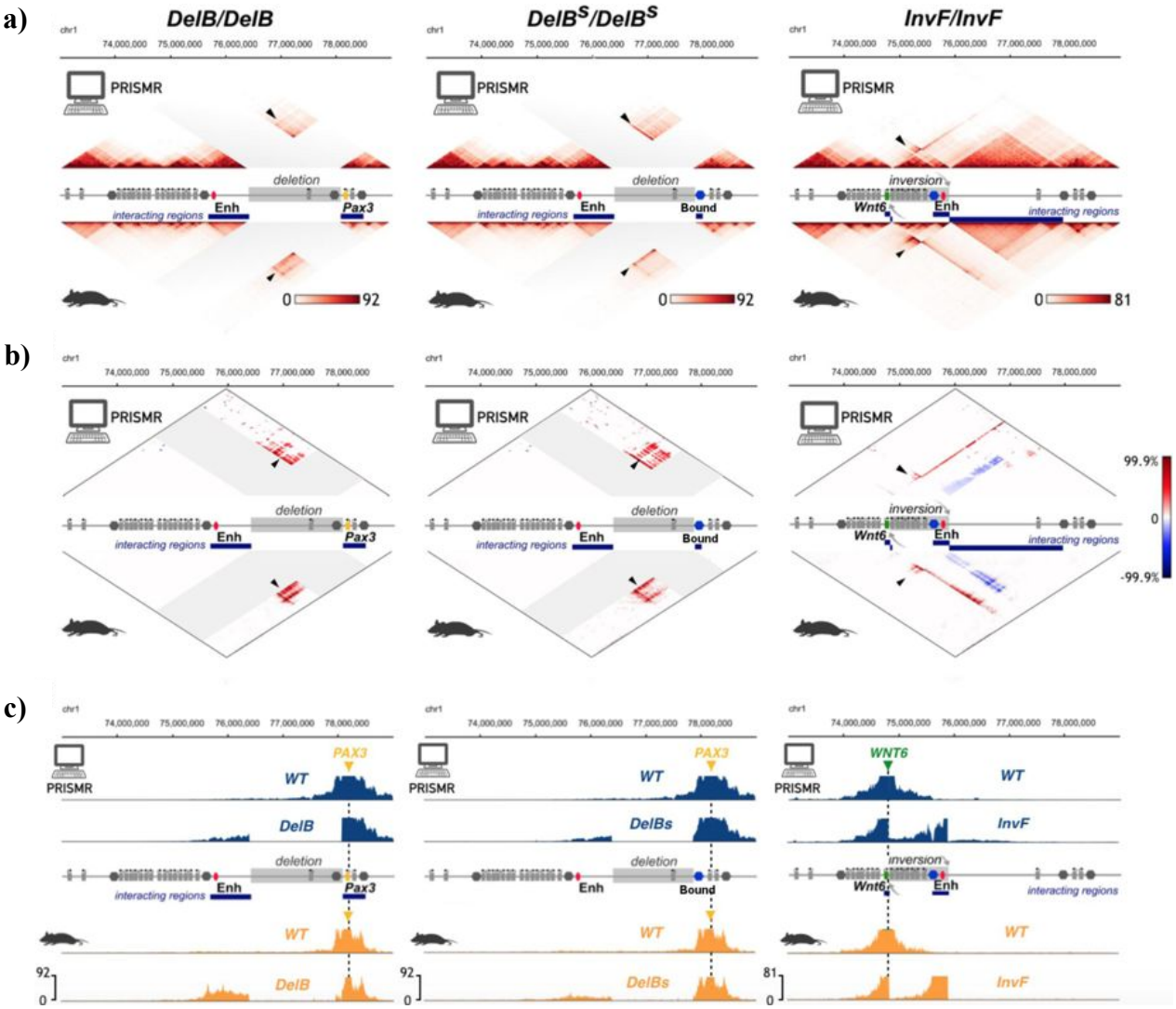

Figure 4.7: 'PRISMR' predicts the effects of mouse homozygous structural variants on chromatin architecture.

As in Figure 4.6, for PRISMR model predictions based on our capture Hi-C obtained from E11.5 limb, without a-priori CTCF peaks analysis. The Pearson correlation are: r=0.94, r'=0.50 (*DelB);* r=0.95, r'=0.55 (*DelBs);* r=0.93, r'=0.52 (*InvF*) (**Table 4.1**). Figures adapted from (Bianco et al., 2018).

The inversion (*InvF,* **right columns**) was predicted to result in a rearrangement of the genomic content of the two adjacent TADs with interaction hotspots between *Epha4* enhancers and a gene-dense region (three genes affected) that would be consistent with the ectopic *Wnt6* activation reported previously. 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 as a consequence of genomic rearrangements, identifying a small number of genes whose regulation might be directly affected.

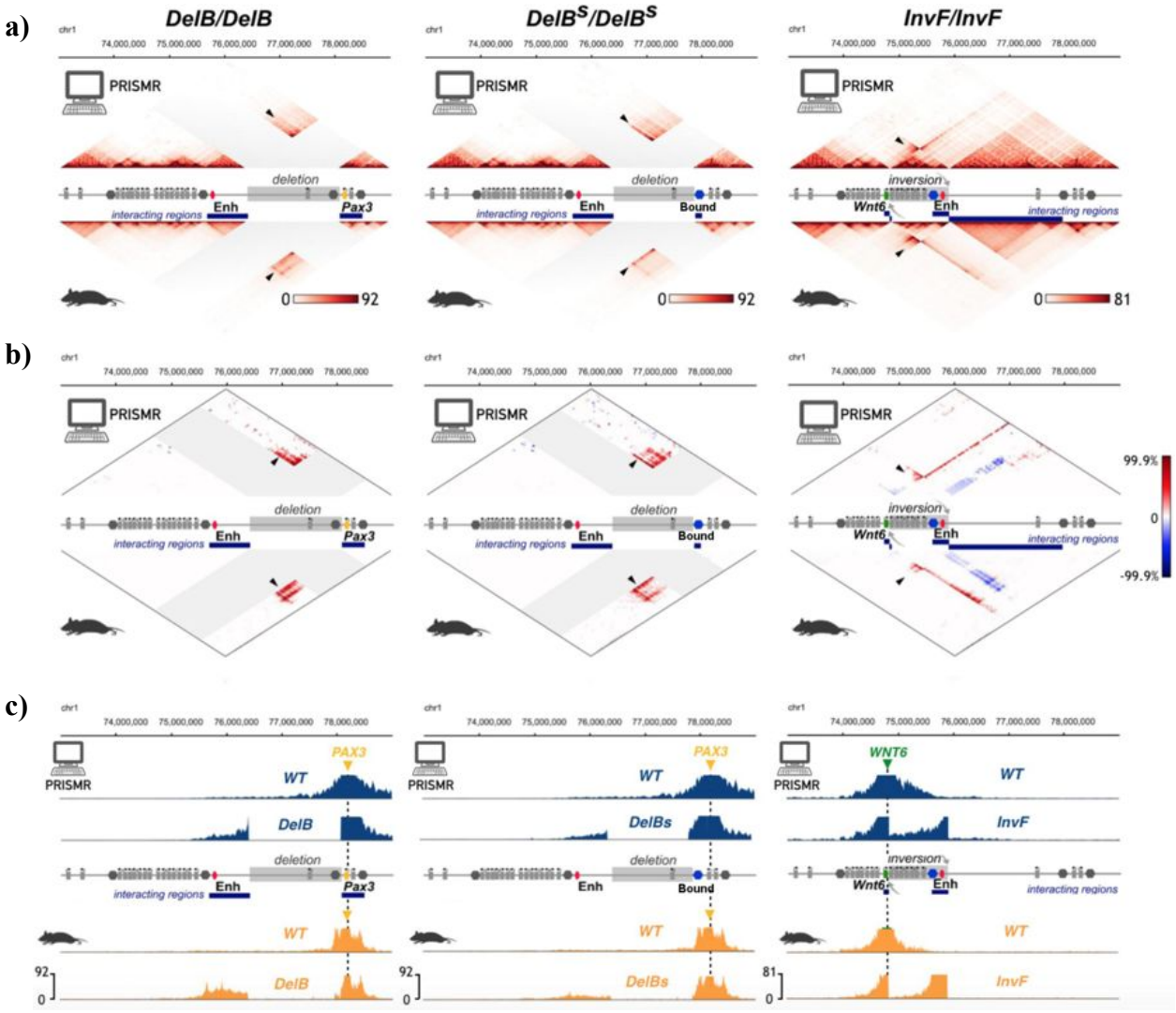

Figure 4.8: PRISMR model based on mouse wild-type CH12-LX Hi-C data predicts the effects of homozygous structural variants on chromatin architecture

As in Figure 4.7, for PRISMR model predictions based on in-situ Hi-C data from (Rao et al., 2014) in CH12-LX cells. Figures adapted from (Bianco et al., 2018).

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 homozygous mutations (**Figure 4.6; Figure 4.7**). The new dataset showed the same regions of ectopic interaction and displayed a noticeably high agreement with both PRISMR and 'PRISMR + CTCF' predictions, not only across the entire locus, but also when the regions of ectopic interaction were compared.

Our results confirmed that the larger deletion in *DelB* mutant led to a fusion of the *Epha4* and *Pax3* TADs, not occurring in the smaller *DelBs* mutation, in which the TAD boundary remains intact. In the inversion, ectopic contacts were observed between *Wnt6* and the *Epha4* enhancer region, which facilitated *Wnt6* activation as previously observed in vivo (Lupiáñez et al., 2015), and between a region at the centromeric breakpoint and the entire *Epha4* TAD. Notably, the observed ectopic interaction was interrupted by the *Epha4* centromeric boundary, which, although inverted, appeared to retain its functionality. Hence, deletions and inversions that include boundary elements can result in fusions or reorganization of TADs, respectively.

4.3.1 3D conformations of the polymer models of the *Epha4* locus and its mutations

Using 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 **Figure 4.9** 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, by using POV-Ray software (Version, 2004). 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 **Figure 4.9** 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 in **Section 4.3**. 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.

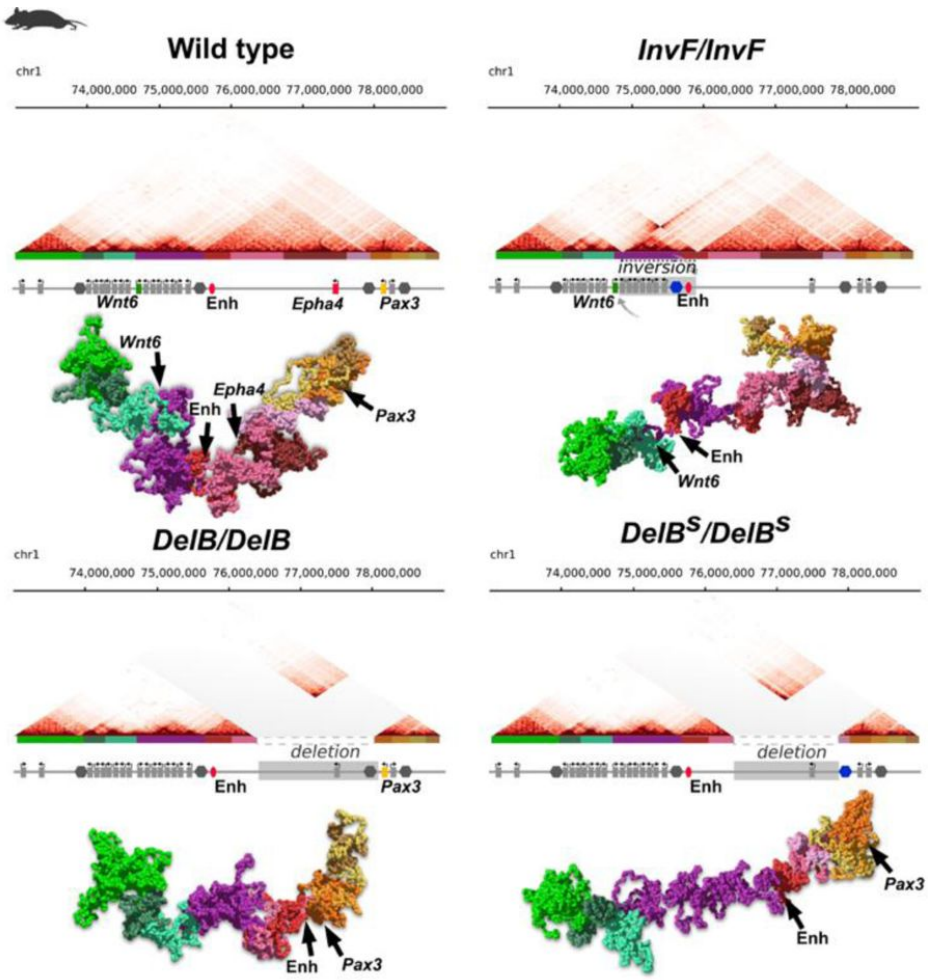

Figure 4.9: PRISMR predicted 3D conformations of the Epha4 locus in murine CH12-LX cells

Top-left: the PRISMR model based on published Hi-C data in murine CH12-LX cells recapitulates (Pearson correlation r=0.91, distance-corrected Pearson correlation r'=0.56) the experimental pairwise contact matrix (see also **Figure 4.2**). 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. 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. Figures adapted from (Bianco et al., 2018).

4.3.2 Statistical analysis details

In this Section we will discuss in more detail the statistical analysis used to quantitatively estimate the comparison between the predictions of PRISMR method and the experimental capture Hi-C data, discussed in this **Section 4.3** and, next, in **Section 4.4**.

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 (**Figure 4.6, Panel b; Figure 4.7, Panel b; Figure 4.8, Panel b**) 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 interactions, 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 96^{th} percentile in the cases where the data are marked by strong outliers, as in the human deletion data, discussed in next **Section 4.4**. 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. Since the mouse cHi-C matrices are homozygous mutants, the data corresponding to the deleted genomic segments are not represented. The experimental subtraction matrices were computed on the raw experimental cHi-C data.

In the case of deletions, a part of ectopic contacts just arises because previously distant regions along the genome become flanking (due to the deletion). Yet, the specific pattern of ectopic contacts could be only vaguely guessed by the above argument. In our model, it can be derived in a principled way, also in case of more complex SVs. The identification of the specific pattern of novel contacts is crucial to identify potentially disease causing interactions between single genes and enhancers, beyond the average changes of interactions around the SV expected from topology (e.g., from changes in genomic separations). The effect of inversions and duplications is even less intuitive, but also partially influenced by topology. As previously observed (**Section 4.3.1**), our 3D reconstructions of the duplications show that the ectopic contacts (**Figure 4.9**) are partially produced because the duplicated regions tend to twist back in a loop onto each other. Our algorithm can be straightforwardly extended to model translocations (or insertions) of regions deriving from the same locus: a piece of the polymer model (of the WT case) can be moved to any other location along it and the corresponding novel

contact matrix can be obtained by MD simulations of the mutated model. Insertions could be analogously implemented. 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.

Virtual 4C analysis

In order to better highlight ectopic interactions, we produced virtual-4C plots from the viewpoint of the phenotype causative genes in each mutant in mouse (**Figure 4.6, Panel c; Figure 4.7, Panel b; Figure 4.8, Panel b**) and in next **Section 4.4** for human (**Figure 4.12, Panel b**). 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 normalized the WT matrix by equalizing the number of its reads to the total reads in the mutation, as described in the previous section.

4.4 PRISMR Model predictions on human cells

Finally, we wanted 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.

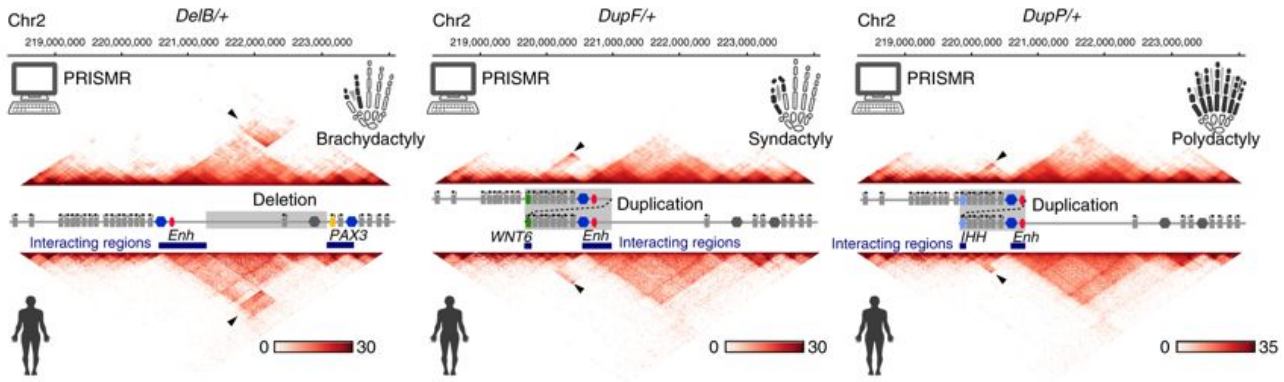

Figure 4.10: PRISMR predicts the effects of human heterozygous structural variants on chromatin architecture on human cell lines.

Contact matrices from model predictions derived from WT data (top) and cHi-C experiments in mutation carrying cultured human skin fibroblasts (bottom). Human phenotypes associated with the rearrangement are indicated on right. *DelB/+* : PRISMR predicts the chromatin effects of a 1.6-Mb heterozygous deletion ($r = 0.93$, $r' = 0.61$). Increased interaction is detected between the remaining *EPHA4* and *PAX3* TADs (arrowhead and blue bars), resulting in *PAX3* misexpression and brachydactyly. *DupF/+* : heterozygous 1.4-Mb duplication ($r = 0.88$, $r' = 0.52$). Increased interaction is detected between *EPHA4* enhancer cluster and *WNT6* regions. *DupP/+* : heterozygous 900-bp duplication ($r = 0.90$, $r' = 0.56$). Increased interaction is detected between *EPHA4* enhancer cluster and *IHH* regions. Figures adapted from (Bianco et al., 2018).

A PRISMR polymer model of the *Epha4* locus, inferred from healthy control human fibroblast cHi-C data (**Figure 4.3**), was employed to predict the effects of SVs on chromatin contact matrices (**Figure 4.10**). To test the model predictions, we used fibroblasts obtained from human patients to perform cHi-C (**Figure 4.10**). We analyzed a 1.6 Mb deletion associated with brachydactyly (similar to mouse *DelB*), a 900-kb duplication (*DupP*) associated with polydactyly and *IHH* activation, and a 1.4 Mb duplication (*DupF*) associated with syndactyly and *WNT6* activation (Lupiáñez et al., 2015).

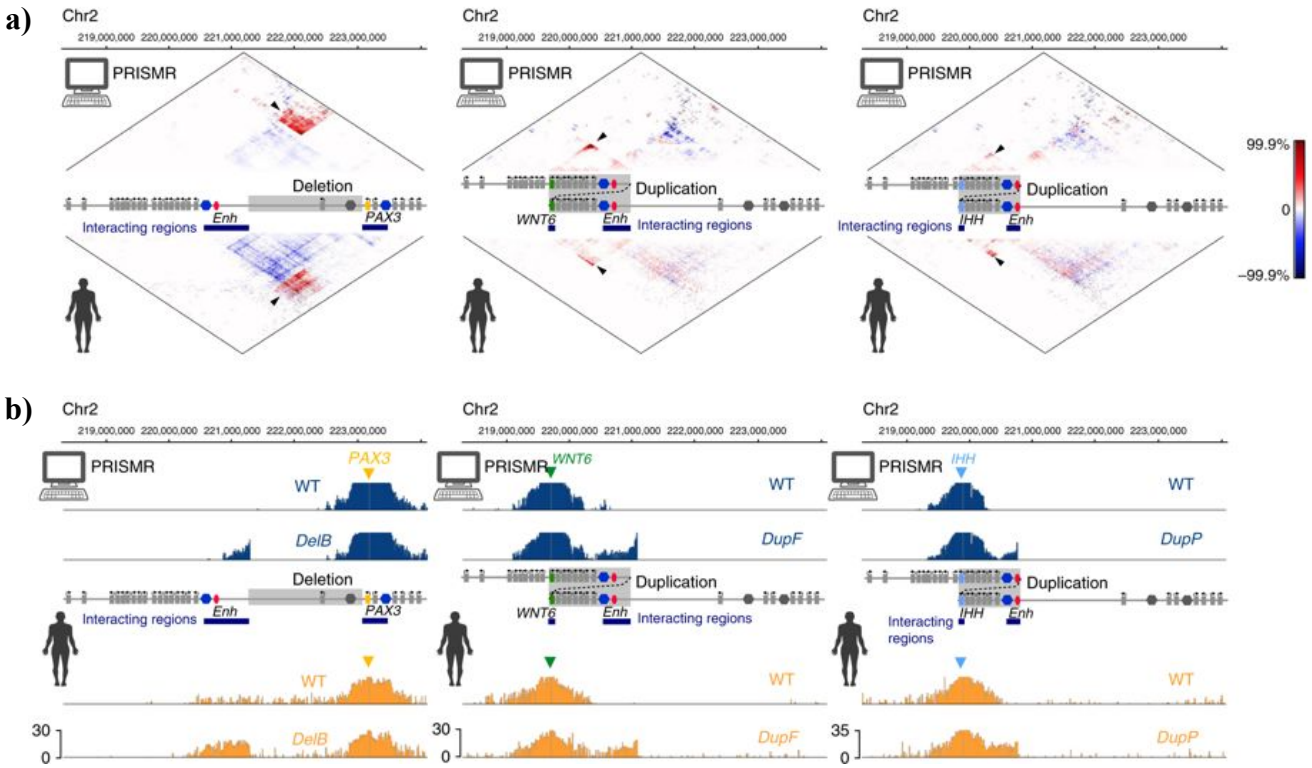

Figure 4.11: Quantification of PRISMR predictions by subtraction matrices and virtual 4C data.

a) Subtraction maps produced (using a healthy control and patients) from predictions and cHi-C data. Above, threshold gain of interaction is displayed in red and loss in blue (absolute differences > 2 s.d.; see Methods). Ectopic interactions between *EPHA4* TAD and genomic regions are indicated (arrowheads and blue bars). **b)** 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 remaining *EPHA4* TAD, including *EPHA4* enhancer cluster in both, prediction and experimental data. *DupF*/ + : note increased interaction of *WNT6* promoter with the *EPHA4* enhancer cluster. *DupP*/ + : note increased interaction of *IHH* promoter with the *EPHA4* enhancer cluster. Figures adapted from (Bianco et al., 2018).

Subtraction maps were computed to identify the precise regions and intensity of significant ectopic interactions (**Figure 4.11, Panel a**; see **Section 4.3.2**) In the brachydactyly-associated deletion, and

even in complex genomic regions with high gene density. Our polymer physics predictions can be used to identify regions of ectopic interactions that can then be scanned for their content, i.e., the presence of genes and enhancers that could interact. Finally, to further quantify the comparison between model predictions and cHi-C, we computed Virtual 4C plots from the viewpoint on the respective phenotype-causing gene (**Figure 4.11, Panel b**; see **Section 4.3.2**). In *DelB*/+ deletion, we find increased interaction of *PAX3* promoter with remaining *EPHA4* TAD, including *EPHA4* enhancer cluster in both, prediction and experimental data. In *DupF*/+ duplication we find increased interaction of *WNT6* promoter with the *EPHA4* enhancer cluster, while in *DupP*/+ increased interaction of *IHH* promoter with the *EPHA4* enhancer cluster has been found.

The derived 3D structures shown in **Figure 4.12** refer to analogous model predictions in human fibroblasts. The 3D snapshots of the duplications illustrate, for instance, that part of the ectopic contacts discussed in **Figure 4.11** is produced because the duplicated segments tend to twist back in a loop onto each other. The color code in the mouse case (**Figure 4.9**) is derived from the one of the human case based on their synteny (as determined by the liftOver tool in the UCSC Genome Browser).

Furthermore, our results indicate that PRISMR can be used in cases where affected tissues or equivalent cell types are not available. Recent advances in high-throughput sequencing have boosted the identification of SVs (Gilissen et al., 2014; Hehir-Kwa et al., 2016; Newman et al., 2015). In this scenario, polymer modeling by PRISMR emerges as a valid approach for predicting pathogenic effects, facilitating the interpretation and diagnosis of this type of genomic rearrangement.

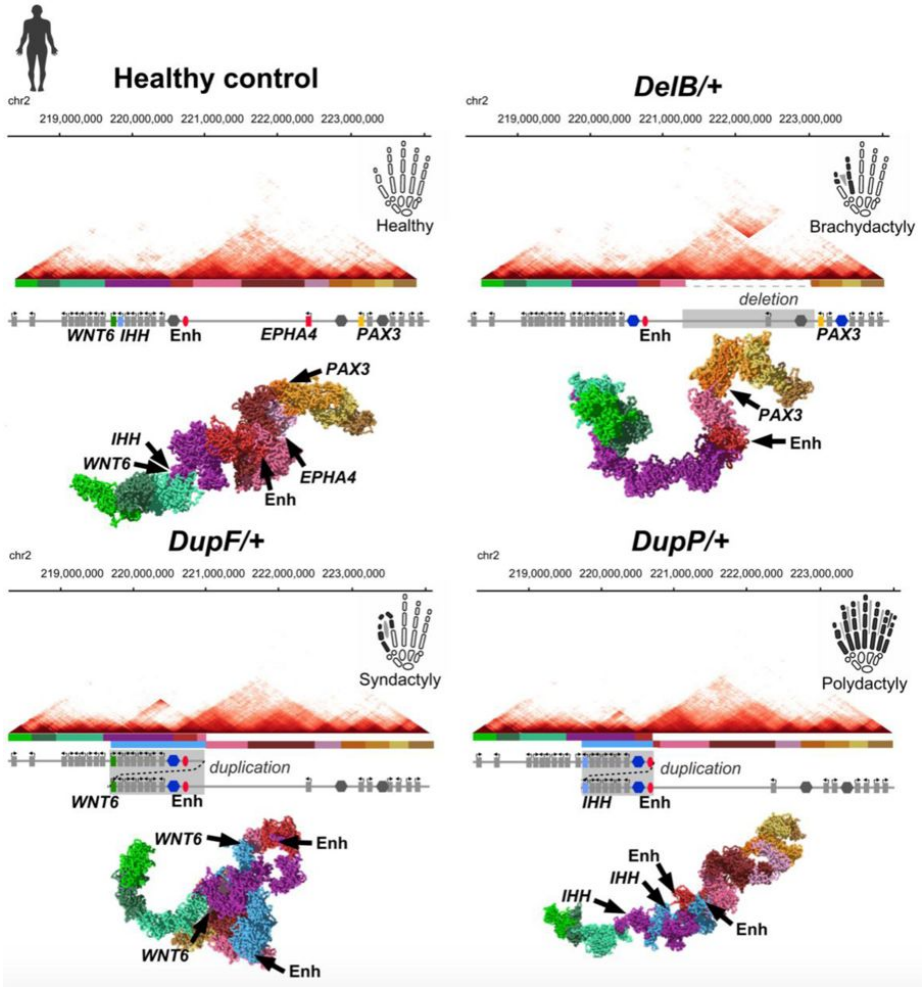

Figure 4.12:

PRISMR predicted 3D conformations of the *EPHA4* locus in human fibroblast cells. Figures adapted from (Bianco et al., 2018).

Dataset	Genotype	Coefficient r	Coefficient r'
E11.5 limbs cHi-C vs. 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
human fibroblasts cHi-C vs. PRISMR	healthy	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
E11.5 limbs cHi-C vs. PRISMR derived from mouse CH12-LX[23]	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
E11.5 limbs cHi-C vs. 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

Table 4.1: **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., 2018).

References

Andrey, G., Schöpflin, R., Jerković, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M., et al. (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome Res. *27*, 223–233.

Annunziatella, C. , Bianco, S., Andrey, G., Chiariello, A.M., Esposito, A., Fiorillo, L., Prisco, A., Conte, M., Campanile, R., Nicodemi, M. (2018). *Single-molecule conformations of the HoxD locus in mouse ES and Cortex cells*. Cell Reports. (*submitted)*

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., De Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *24*, 515–524.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Bohn, M., and Heermann, D.W. (2010). Diffusion-driven looping provides a consistent provides a consistent framework for chromatin organization. PLoS One *5*.

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

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

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

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.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., Van De Vorst, M., Van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

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

Hehir-Kwa, J.Y., Marschall, T., Kloosterman, W.P., Francioli, L.C., Baaijens, J.A., Dijkstra, L.J., Abdellaoui, A., Koval, V., Thung, D.T., Wardenaar, R., et al. (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat. Commun. *7*.

Hug, C.B., Grimaldi, A.G., Kruse, K., and Vaquerizas, J.M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell *169*, 216–228.e19.

Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.

Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A.M., Jerković, I., et al. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. *50*, 1463–1473.

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J., et al. (2017). Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. BioRxiv 118737.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2018). Erratum: Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation (Molecular Cell (2017) 65(3) (432–446.e5) (S1097276517300357)

(10.1016/j.molcel.2017.01.009)). Mol. Cell *71*, 191.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Newman, S., Hermetz, K.E., Weckselblatt, B., and Rudd, M.K. (2015). 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.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22.

Oliphant, T.E. (2007). Python for scientific computing. Comput. Sci. Eng. *9*, 10–20.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army tool for genome feature analysis. Curr. Protoc. Bioinforma. *2014*, 11.12.1-11.12.34.

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

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

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Version, P. (2004). POV-Ray Reference. Am. J. Surg. *187*, 114–119.

Yan, J., Chen, S.-A.A., Local, A., Liu, T., Qiu, Y., Lee, A.-Y., Jung, I., Preissl, S., Rivera, C.M., Wang, C., et al. (2017). Histone H3 Lysine 4 methyltransferases MLL3 and MLL4 Modulate Long-range Chromatin Interactions at Enhancers. BioRxiv 110239.

Conclusions

In this work, we investigated by using polymer physics models a very interesting problem in modern biology: the three-dimensional organization of chromatin in mammalian cell nucleus. Recent studies have shown that 3D structure of genome has a key role in vital biological functions of the cell, and its misfolding is often linked to several human diseases, such as congenital diseases and cancers. However, chromatin structure is currently poorly understood despite being subjected to intense investigation.

In the last decade, new experimental techniques, such as Hi-C, have revealed that chromatin has a complex, hierarchical organization spanning from the sub-Mb scale up to the entire chromosome length. To shed light on this intricate pattern of interactions revealed by experimental data, polymer physics models have been introduced. In this work, we focused on the "String&Binders Switch" (SBS) model, where non-random chromatin conformations are established through specific interaction of chromatin with diffusing DNA-binding molecules, driving folding by formation of loops. As first step, we recapitulated with a very simple model some important features of chromatin organization, such as the large-scale average behavior of experimental data, the mechanisms underlying the self-assembly of TADs and the hierarchical organization of genome, as emerging from Hi-C and FISH data. Next, we generalized the SBS model and we developed an innovative machine learning algorithm, PRISMR, by which we reconstruct, starting from experimental data, the 3D architecture of real genomic regions with high accuracy. This method does not require any a-priori knowledge of the molecular factors responsible for chromatin folding; conversely, our information can be used to infer the nature of key folding factors, by crossing the distribution of binding sites predicted with several epigenomic datasets available. Such epigenetic analysis has shown that different types of binding sites, and their cognate binders, do not correspond simply to single molecular factors associated with chromatin, but, rather, to combinations of different factors. In particular, some binding sites correlate with CTCF, factors known shaping chromatin structure and at the base of other recent polymer models, such as the Loop Extrusion model. As next step, we applied our polymer models to investigate and capture the structural differences of a specific genomic region during different stages of differentiation and in different cell types. In the final part, we showed that our polymer models are able to predict the effects of structural variants in the genomic sequence on the 3D architecture, with a very high accuracy. Therefore, our polymer modeling methods emerge

as a powerful approach to predict pathogenic effects, facilitating the interpretation and diagnosis of this type of genomic rearrangements.
In this work, we analyzed a set of deletions, duplications and inversions. However, we are currently working on the improvement of our method for being extended to model also translocation and insertion of genomic regions deriving from the same locus or from a distinct genomic region. This latter case would require modeling both considered genomic regions or making additional hypothesis on the structure of the polymer to be inserted, for instance by exploiting the epigenetic barcode of that region. For this reason, we are also improving our models to model entire chromosomes and loci at higher resolutions. Such improvements, together with cross-analysis of epigenetics marks seems a promising means to unravel the molecular determinants of chromatin folding. Finally, we also improved our models to be equally applicable to new technologies, such as GAM and SPRITE. In summary, we are following new research lines, not described in this thesis, in order to improve the predictive power of our model and to investigate at a deeper level the several mechanisms involved in genome organization, that are still unknown.

Acknowledgments

We thank the University of Naples Federico II for grant of the Ph.D. program in Physics. We thank all the members of the Complex Systems group at Department of Physics “Ettore Pancini”, and our collaborators in Berlin. We acknowledge computer resources from Scope (University of Naples), CINECA and INFN, which have a fundamental role in our research activity.

Appendix A: Comparing different chromatin polymer models

The principal Chapters showed that the String&Binders Switch model recapitulates Hi-C and FISH data to a high degree (Annunziatella et al., 2016; Barbieri et al., 2012; Chiariello et al., 2016), and it also predicts the effects on 3D architecture when Structural Variants (SVs) in genomic sequence are present (Bianco et al., 2018). This is done without a-priori assumptions and no additional or tunable parameters. Additionally, we showed that SBS model can be improved to take into account CTCF transcription factors, by exploiting prior knowledge of their binding sites along the chromatin. CTCFs are, indeed, known to play an important role in chromatin architecture through the formation of chromatin loops (Fudenberg et al., 2016; Rao et al., 2014; Sanborn et al., 2015).

In this Appendix Chapter, we give an overview of other polymer physics models that, together with the SBS model, have been recently developed to explain the mechanisms behind the chromatin folding (see, **Chapter 2**). In particular, we focus on two of these models: the Loop Extrusion (Fudenberg et al., 2016; Sanborn et al., 2015) and the Slip-Link (Brackley et al., 2017) model. Both these models are based on previous knowledge of CTCF binding sites distribution, and on their motifs orientation. The interaction between CTCF binding sites, that must be oriented in convergent way, is physically mediated by a protein complex, mostly cohesin.

Briefly, in the Loop Extrusion (LE) model, the chromatin domains are formed by an active extrusion process, driven by the presence of protein complex (i.e., cohesin), which generates larger and larger chromatin loops. The extrusion process can halt with a certain probability when cohesin reaches a region enriched of CTCF, generally at TADs boundaries. Such model has been shown to explain some features of chromatin architecture data, such as formation of self-interacting domains and relative enrichments at boundaries of architectural proteins (**Section 1.4**), the preferential orientation of CTCF motifs (Rao et al., 2014), and prediction of CTCF or cohesin depletion (Nuebler et al., 2018). Additionally, an in-silico experimental reconstruction of extrusion loop model has been recently created at force-dependent using ATP energy (Ganji et al., 2018). However, other parallel experiments have shown that the depletion of CTCF can have a minor effect on chromatin organization (Kubo et al., 2017), by underlying that other factors, beyond CTCF, play a key role in chromatin organization. In the Slip-Link (SL) model, alternatively called diffusive LE model, instead, the cis-active extrusion process is replacing by a diffusive process. By such a model, similar results to the LE model can be found (Brackley et al., 2017), by proving that no active motors need to drive the chromatin folding.

In this Chapter, to compare different polymer physics models, we applied them to the same genomic regions. In **Section A.1,** we discuss the Loop Extrusion model, showing that, in some cases, it successfully explains the folding dynamics of chromatin and, in particular, the formation of self-interacting domains. In **Section A.2,** we show that a similar approach where an active motor is replaced by a diffusive process can be equally used. In **Section A.3,** we show that the SBS model can be improved to take into account prior knowledge of the CTCF binding sites (as discussed in **Chapter 4**). Part of the material presented in this Chapter, including figures, paragraphs and sentences, is adapted or taken literally from the paper (Pereira et al., 2018), which I co-authored.

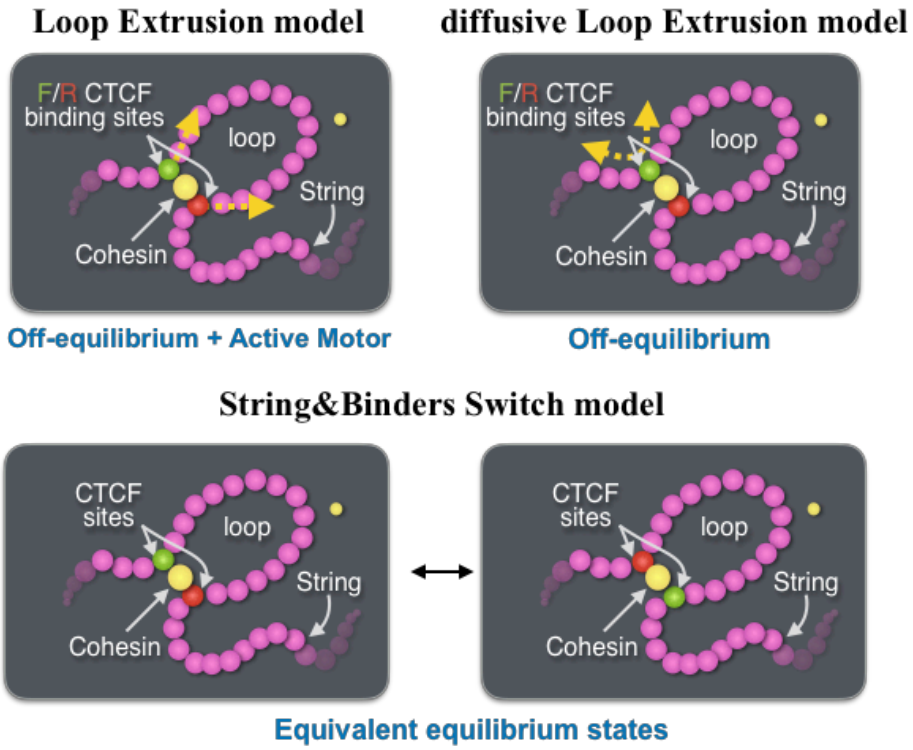

Figure A.1: Schematic representation of different polymer models describing chromatin folding.

We investigate, by using different polymer models, chromatin loops driven by cohesin, which bridges Forward (colored in green) and Reverse (in red) CTCF binding sites. On top left, the Extrusion (LE) model quantifies the off-equilibrium folding scenario where an active motor binds to DNA and actively extrude a DNA loop. The interaction occurs only between CTCF binding sites oriented in convergent way. On the top right, the diffusive Loop Extrusion (dLE) model is a variant of the LE model without active, energy burning mechanisms, where the DNA diffusively slips through a bridging factor. On the bottom, the String&Binders Switch (SBS) model where a chromatin locus can be modeled as an equilibrium polymer conformation. Importantly, while both the LE and the dLE models include only loops where CTCF binding sites are convergent, the SBS model includes both convergent (bottom left) and divergent (bottom right) case.

A.1 Loop Extrusion Model

The Loop Extrusion (LE) Model is a polymer physics model introduced for the first time in (Sanborn et al., 2015), and further investigated in (Fudenberg et al., 2016). In the LE model, loop-extruding factors (LEs), e.g., cohesin, extrude the DNA filament during interphase state forming larger and larger loops, until they halt at binding sites enriched of specific proteins, e.g. CCCTC-Binding factor (or briefly CTCF (**Figure A.1, top left**). The halting process can occur only if CTCF motifs point towards each other. This observation is in accord with high-resolution Hi-C data, which showed that in about 90% of cases CTCF loops are "convergent" (Rao et al., 2014).

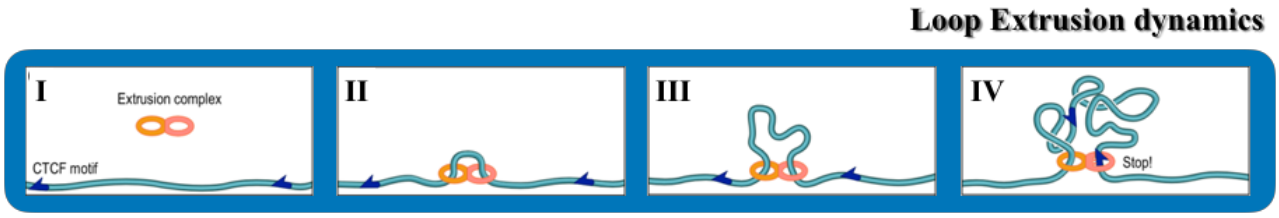

Figure A.2: Loop Extrusion Dynamics.

Schematic description of LE dynamics: **I)** the extrusion complex is initially bound to two consecutive beads (i, j) on the polymer. **II)** Every T timesteps the bond (i.e., extrusion complex) moves from beads (i, j) to monomers (i-1, j+1). **III)** The extrusion dynamics continues, until **IV)** it may interact with oriented CTCF binding site, represented as special bead, halting and fixing with a given binding probability. Figure adapted from (Sanborn et al., 2015).

The LE dynamics is schematically pictured in **Figure A.2**. First, a chromatin loop initially binds by cohesin two consecutive DNA regions (**I**). Next, the two ends of cohesin (we call up- and down-stream binding domains) move in opposite directions respect the genome and thus DNA is extruded through the complex (**II**). The loop continues to grow (**III**) until one end of complex reach a CTCF oriented in opposite way: in that case, the extrusion process can halt with a certain probability (**IV**). In this picture, down-stream binding site can halt at a reserve CTCF binding sites (colored in red in **Figure A.1**) and is unaffected by a forward CTCF binding site (colored in green). Conversely, up-stream binding sites can halt at forward, and not reverse, CTCF binding site.

To implement the Loop Extrusion model in our Molecular dynamics simulations, we followed the same approach described in (Sanborn et al., 2015). For LJ potential, we set $\varepsilon_{LJ} = 1$, $\sigma = 2^{-1/6}$ and $r_{cut} = 2.5\sigma$, in order to have a minimum at $r = 1$ (**Eq. 2.1**). Consecutive beads are bound instead by harmonic potential $V_{HARM} = k_{bond} (l - l_0)^2$, where we set $k_{bond} = 1000\ k_BT$ and $l_0 = 0.71\sigma$. The extrusion complex is modeled by a harmonic bond, with elastic constant set to $k_{bond} = 10\ k_BT$ and rest length $l_0 = 1\ \sigma$. As initial configurations, we used random walks ($l_0 = 3\ \sigma$) statistically minimized, in order to relax

configurations in their minimum of energy. Initially, the complex binds a pair of consecutive beads, randomly chosen, and it slides unidirectionally along the polymer chain every T = 200 dt (here, we set dt = 0.005τ). The extrusion process continues, and the chromatin loop grows, until the complex reaches a CTCF binding sites oriented in opportune way, modeled as a special bead on polymer chain. The interaction allows the complex to halt and to be fixed with some probability. The complexes are subjected to following constrains: since two extrusion complexes cannot pass to each other, when they collide, one complex unbinds; similarly, when a moving complex collides against a halted complex the first unbinds whilst the second remains fixed; when a complex reaches a polymer end, it dissociates (schematic cartoon in **Figure A.3**).

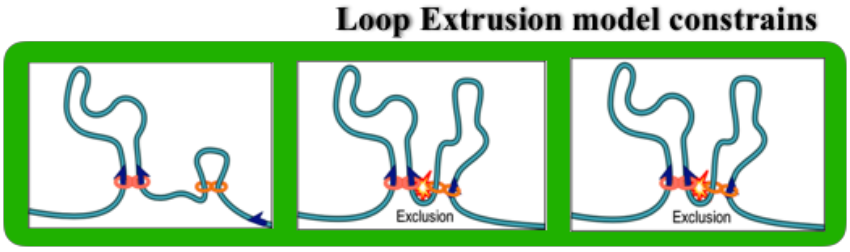

Figure A.3: Characterization of Loop Extrusion Dynamics.

A moving complex which collides with a halted complex, dissociates whilst the halted complex remains unchanged. Instead of the dissociated another random complex is placed between two neighbor monomers. Figure adapted from (Sanborn et al., 2015).

To estimate the oriented binding strengths for a specific genomic region we proceed as follows (Sanborn et al., 2015). First, we compute the halting probability *P* proportional to ChiP-seq CTCF signal *s* (**Section 1.2**), binned at given genomic resolution (**Figure A.4, Panel a**). Then, we identify within each peak the best match to consensus CTCF and we associate the corresponding orientation depending on forward/reverse strand of binding motif (Kim et al., 2007), by using FIMO tool in the MEME Suite online software (Grant et al., 2011). Finally, we orient each halting probability according to the orientation of the nearest CTCF binding site within 5 kb (**Figure A.4, Panel b**).

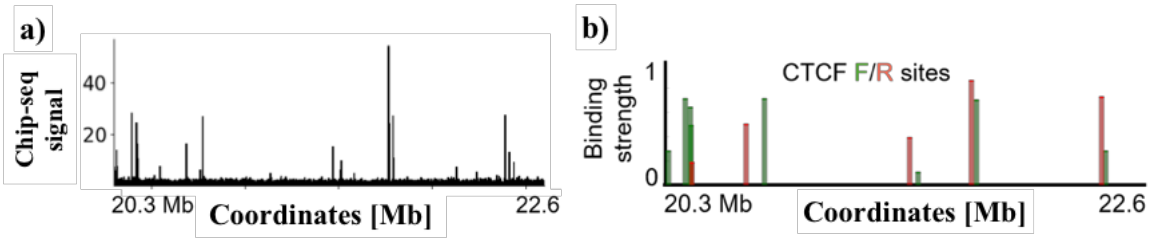

Figure A.4: Oriented CTCF binding strength.

a) Example of Chip-seq signal for chr4:20,3-22,6Mb in GM12878 (Sanborn et al., 2015), from which it is possible to extract **b)** the binding strengths for CTCF binding sites, with relative orientation.

To test our model, we focused on the same genomic region discussed in (Sanborn et al., 2015) (chr4:20.3-22.6Mb, hg19) at lower resolution (4kb resolution) compared to original simulation (1kb), in human lymphoblastoid cell line (GM12878). We considered an ensemble of 10^2 different simulations, where the number of extrusion complexes on region span from 6 to 15, and this is kept constant during the simulation: whether a complex dissociate, in agreement with one of events shown in **Figure A.3**, it is replaced by a new, randomly positioned, complex in order to keep constant complexes concentration. We ran each simulation for $8x10^5\Delta t$, corresponding to $t = 4x10^3\tau$. We computed the averaged contact maps over these configurations every $8x10^5\Delta t$ timesteps, where we considered two beads in contact if their physical distance was $\leq 1.5\sigma$. The Pearson correlation between experimental Hi-C data and simulated contact matrix binned at 4kb is 85% (**Figure A.5, Panel b**).

5.2 Diffusive Loop Extrusion Model

In previous **Section A.1**, we describe the LE model, where an active extrusion process needs to be introduced, although there is no experimental evidence in-vivo. This suggested that the diffusive sliding of cohesin is equally sufficient to reproduce the same results, unless convergent loop restriction is maintained (Brackley et al., 2017; Pereira et al., 2018).

To this aim, we implemented the diffusive Loop Extrusion dynamics (dLE) in MD simulations just generalizing the dynamics of Loop Extrusion model, i.e., allowing the cohesin moving independently in both directions with same probability (**Figure A.1, top right**). Initially, we investigated the diffusive Loop Extrusion (dLE) model by MD simulations for this same region investigated by LE (chr4:20,300,000-22,600,000). Since unlike the LE model, a CTCF binding site could be visited many times for diffusive cohesin complex, we introduced a refractoriness time τ_{reft} for those CTCF sites where a possible bonding event fails (here, we set $\tau_{reft} = 10\tau$). In the light of the active LE model, here CTCF bonding events are such that only convergent CTCF loops are allowed. For this region, we derived an ensemble of 40 different configurations and we let the system evolve up to $10^6\,\Delta t$. The contact matrix has been computed in the same way discussed for LE model (**Section A.1**), and we find a Pearson correlation with experimental data equal to 85% (**Figure A.5, Panel c**).

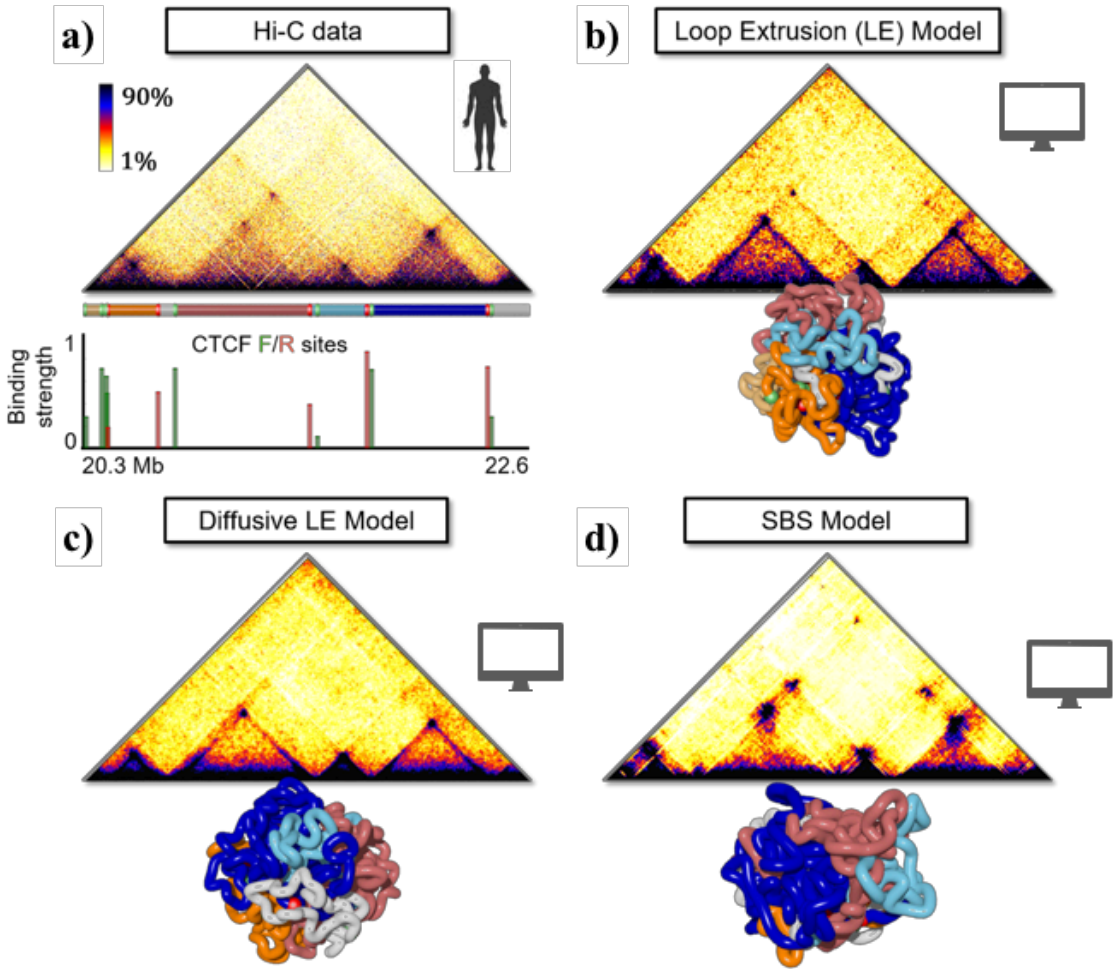

Figure A.5: **Performance of different polymer extrusion models to explain Hi-C data for the human chr4.**

a) *in-situ* Hi-C data for GM12878 cells of the region chr4:20,300,000-22,600,000 (Rao et al., 2014) at 4kb resolution investigated in (Sanborn et al., 2015) and corresponding CTCF binding sites, strength and orientation (green is forward, and red reverse), estimated from Chip-seq signal in **Figure A.2, Panel b**. The colored bar highlights CTCF positions and main polymer interacting regions to help 3-D visualization; Contact maps for the same region obtained by **b)** Loop Extrusion, **c)** diffusive Loop Extrusion and **d)** SBS 3-D chromatin models with interactions between CTCF sites oriented in opposite ways. For each model a typical 3D polymer structure is shown. Diffusive extrusion is as efficient as active extrusion at predicting Hi-C domain boundaries and peaks. Figure adapted from (Pereira et al., 2018).

Next, by dLE we also investigated the folding process for a 10 Mb long region of chromosome 7 in cell line GM12878, at 25kb resolution, for which Hi-C data are available (Rao et al., 2014). Here, we used a refractoriness time $\tau_{reft} = 10\tau$ for the CTCF sites where a bonding event fails. Additionally, in order to allow higher order loops to be explored, we introduce dissociation events where LEs unbind from CTCFs and start diffusing again (with $\tau_{reft} = 4\times10^3\tau$). We found that Molecular Dynamics simulation of diffusive Loop Extrusion can explain most of interaction and TAD or meta-TAD boundaries (**Figure A.6**). Importantly, we used 25 kb resolution that allows to reach “steady state”.

Longer MD simulations could likely allow to recapitulate long range contacts as well, as predicted from 1-D dynamics simplification (Pereira et al., 2018). However, this is beyond our purposes of demonstrating equivalence between LE and dLE.

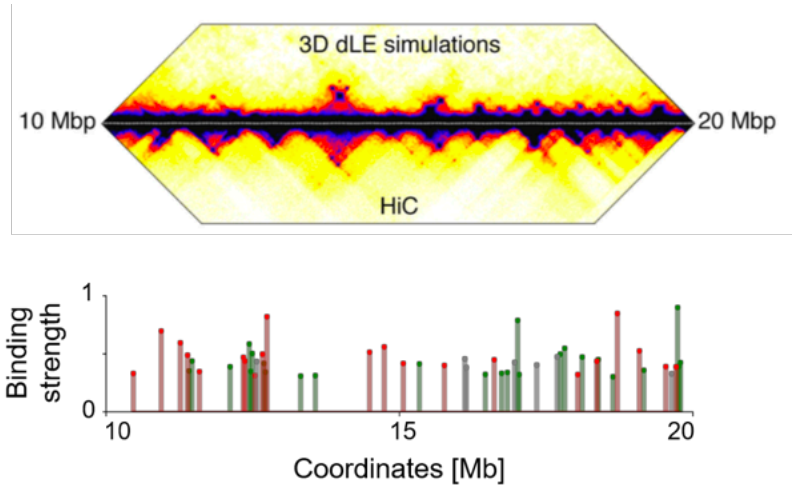

Figure 5.6: **The diffusive Loop Extrusion model of a 10Mb wide, gene poor genomic region.** On the top, *in-situ* Hi-C map from published data (Rao et al., 2014) of the gene poor 10Mb long region chr7:10,000,000-20,000,000 (hg19 assembly) for GM12878 cells, at 25kb resolution. In the middle, the colored bar highlights CTCF positions and main interacting regions to help visualization of the 3D polymer structure shown on the right panel, while the histogram is the CTCF binding strength profile with corresponding orientation at the same resolution. On the bottom, the interaction map derived by the diffusive Loop Extrusion model, showing high agreement with experimental data (Pearson correlation r =0.92). Figure adapted from (Pereira et al., 2018).

5.3 String&Binders Switch Model

We investigated the region discussed in **Section A.1** and **Section A.2** with the "Strings-and-Binders" model (chr4:20.3-22.6Mb, hg19). We performed MD simulations considering a polymer chain made of N=2300 beads, each bead corresponding to 1 kbp. Here, all particles interact by a repulsive Weeks-Chandler-Andersen potential (**Eq. 2.1**) and consecutive beads are connected by a FENE spring (**Eq. 2.2**), while beads and binders interact by an attractive Lennard-Jones potential (**Eq. 2.3**). Chromatin is modeled by a homopolymer, where all beads can interact with the same type of binders (Annunziatella et al., 2016; Bianco et al., 2018; Chiariello et al., 2016). CTCF binding sites were considered according to the approach described in **Section A.1**. CTCF sites interact with an additional type of binders, which bridge CTCFs with opposite orientations (forward – reverse) (**Figure A.1, bottom panel**). Note that, unlike the LE and dLE, the SBS model allows the formation of chromatin loops with CTCF oriented in both convergent and divergent way. As discussed in **Chapter 2**, for such a system there are three possible thermodynamic states depending on the interaction energy and

concentration of the binders – coil, globule disordered, and globule ordered (**Section 2.2)**. As discussed in, the system evolves under Langevin dynamics by MD with an integration timestep $\Delta t = 0.012\tau$. From the 3-dimensional equilibrium configurations in each thermodynamic state we computed averaged contact maps as described **Section A.1** ($r_{int}= 3.5\sigma$). Then, we find the mixture of the three states described above which best describes the locus by maximizing the distance corrected Spearman correlation coefficient between model and experimental data (at 4 kbp resolution). We find the best mixture to be 10% open state and 90% closed state (of which 55% is in the ordered state and 35% in the disordered state). The Pearson correlation coefficient in this case is around 90% (**Figure A.5, Panel d**).

In conclusion, according to results discussed in (Sanborn et al., 2015), the folding dynamics of some genomic regions can be well described by the Loop Extrusion model. Such a model uses prior information about CTCF/cohesin interactions, which form chromatin loops binding together DNA strands, and proves that these interactions have an important role in regulating 3D structure of chromatin. In the cases here analyzed, however, we have shown that no active extrusion processes are needed, and similar results can be successfully found by supposing a diffusive process (as done in diffusive Loop Extrusion model) (Brackley et al., 2017; Pereira et al., 2018) or using our SBS equilibrium model (Chiariello et al., 2016). However, the findings shown in previous Chapters have proved that sometimes CTCFs are not sufficient, and other factors beyond these play a role in chromatin organization, consistently recent developments in the literature (Barbieri et al., 2017; Bianco et al., 2018; Kundu et al., 2018; Pereira et al., 2018; Yan et al., 2017).

References

Annunziatella, C., Chiariello, A.M., Bianco, S., and Nicodemi, M. (2016). Polymer models of the hierarchical folding of the Hox-B chromosomal locus. Phys. Rev. E *94*.

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 and binders switch model. Proc. Natl. Acad. Sci. *109*, 16173–16178.

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., De Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *24*, 515–524.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L.,

Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R., and Marenduzzo, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. Phys. Rev. Lett. *119*.

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

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.

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. Science (80-.). *360*, 102–105.

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

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D.D., Zhang, M.Q., Lobanenkov, V. V., and Ren, B. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell *128*, 1231–1245.

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J., et al. (2017). Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. BioRxiv 118737.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2018). Erratum: Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation (Molecular Cell (2017) 65(3) (432–446.e5) (S1097276517300357) (10.1016/j.molcel.2017.01.009)). Mol. Cell *71*, 191.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L.A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. *115*, E6697–E6706.

Pereira, M.C.F., Brackley, C.A., Michieletto, D., Annunziatella, C., Bianco, S., Chiariello, A.M., Nicodemi, M., and Marenduzzo, D. (2018). Complementary chromosome folding by transcription factors and cohesin. BioRxiv 305359.

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

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Yan, J., Chen, S.-A.A., Local, A., Liu, T., Qiu, Y., Lee, A.-Y., Jung, I., Preissl, S., Rivera, C.M., Wang, C., et al. (2017). Histone H3 Lysine 4 methyltransferases MLL3 and MLL4 Modulate Long-range Chromatin Interactions at Enhancers. BioRxiv 110239.