Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXIX

Coordinatore: prof. Salvatore Capozziello

Polymer physics and structural organization of chromosomes in mammalian genomes

Settore Scientifico Disciplinare FIS/02

Dottorando

Andrea Maria Chiariello

Tutore

Prof. Mario Nicodemi

Anni 2014/2017

Tesi dottorato/ PhD Thesis

Polymer physics and structural organization of chromosomes in mammalian genomes.

A work by Andrea Maria Chiariello

Università degli Studi di Napoli Federico II

Contents

Introduction

The spatial architecture of chromosomes in the cell nucleus is very complex, and it is intimately linked to important functional purposes (Misteli, 2007; Lieberman-Aiden *et al.*, 2009; Dekker *et al.*, 2013, Tanay & Cavalli, 2013; Bickmore & van Steensel, 2013), such as regulation of gene transcription and expression. Yet, the structure and the folding mechanisms remain not fully understood. In the last decade new technologies, the Chromosome Conformation Capture (3C) based methods (Dekker *et al.*, 2013), have been developed and allow to investigate the three-dimensional spatial folding of chromosomes in an innovative quantitative way. These methods, such as the Hi-C technique, have opened the way to mapping genome contacts at genomic scale and are revealing that the genome of mammalian cells is characterized by a complex 3D organization, with extensive long-range functional interactions (Lieberman-Aiden *et al.*, 2009). In mammals, chromosomes occupy distinct territories (Cremer&Cremer, 2001) and have preferred positions depending on cell type and transcription activity (Misteli, 2007; Tanay & Cavalli 2013; Bickmore & van Steensel, 2013). Within chromosomes, the genome is folded into a sequence of domains, called ”topological associating domains” or briefly TADs (Dixon *et al.*, 2012; Nora *et al.*, 2012), in which segments of DNA interact frequently with each other. Such domains are approximately 0.5-1Mb long, and result to be comparatively conserved between mice and humans. TADs are actually only one level of a more complex, hierarchical organization of higher-order domains (metaTADs) starting from the sub Mb and extending up to chromosomal scales (Sexton *et al.*, 2012; Philips-Cremins *et al.*, 2013; Fraser *et al.*, 2015). Furthermore, chromatin interaction have fundamental biological roles, as the control of the gene activity with the formation of loops between regulatory regions and genes. The disruption of this interaction network can alter the regular activity of the complex and produce effects directly on the fenotype (Spielmann & Mundlos, 2013; Lupianez *et al.*, 2015). To better understand the genome-wide contact data produced with these new experimental techniques and to clarify the mechanisms shaping the chromatin spatial 3D organization, polymer physics models have been introduced (Chiariello *et al.*, 2016; Fudenberg *et al.*, 2016; Tiana *et al.* 2016; Sanborn *et al.*, 2015; Nicodemi & Pombo, 2014; Giorgetti *et al.*, 2014; Jost *et al.*, 2014; Brackley *et al.*, 2013; Barbieri *et al.*, 2012; Rosa & Everaers, 2008; Marenduzzo *et al.*, 2006; Sachs *et al.*, 1995). Such models try to identify the key physical elements involved in these fundamental, still largely not clarified biological processes, in a highly interesting and stimulating research field where Physics and Biology get in touch.

This PhD thesis work has been conceived in this intellectual framework. It consists of a detailed description of results and conclusions from the projects that we have followed during our path in the Physics Department of University of Naples Federico II, under the supervision of Professor Mario Nicodemi, in the group of Complex Systems. Many results have been published in collaboration with the Epigenetic Regulation and Chromatin Architecture group directed by Prof. Ana Pombo, at Max Delbruck Centre For Molecular Medicine (Berlin), the Biochemistry group directed by Professor Josee Dostie at McGill University (Montreal), the Human Genetics group directed by Professor Colin Semple, University of Edinburgh. Other projects are currently work in progress in collaborations with the Development and Disease Group directed by Professor Stefan Mundlos, at Max Planck Institute for Molecular Genetics (Berlin), and the Genome Biology group directed by Professor Jim Hughes, at Oxford University.

In Chapter 1, we try to highlight the importance of the genome spatial organization, and recall very briefly some concepts necessary to the comprehension of this research activity, as the Chromosome Conformation Capture (3C) techniques, the interpretation of the genome interaction data and the relationship between spatial organization and cell functionality. Then, we review the polymer models currently proposed to describe the genome three-dimensional architecture. In Chapter 2, we present some results about the genome spatial structure from the study of Hi-C data in a mouse cell differentiation system, and we show that the chromosomes are organized into complex structure of domains-within-domains (metaTADs) linked to the genome function regulation. Next, in Chapter 3 we focus on a more physical topic, that is the employment of polymer models as a tool to quantitatively explain the information contained in the Hi-C interaction data. In particular, we introduce a new thermodynamic phase to fit the long-range contact profile of chromosomes; then we try to schematically model the hierarchical structure of the DNA, and finally we present a theoretical study of the multiple co-localization contact landscape. In Chapter 4, we introduce a more sophisticated polymer model. We show how we are able to reconstruct the 3D genome structure with very high accuracy, and several real loci are studied in detail. We will show the potentiality of these methods as tool to predict effects of genome alterations on the spatial structure and to capture the conformational rearrangements during cell differentiation.

Chapter 1

The problem of the spatial organization of the genome

The spatial organization of the genome is a very complex problem. Many studies and experimental techniques have been developed to better understand how DNA is spatially organized in the cell nucleus and how such organization affects the genome functions, as transcription and gene regulation. Now, we are able to explore in a deeper way this interesting and fascinating problem using recent molecular biology technologies and novel computational methods. In this chapter, far from being exhaustive and complete about this huge topic, we briefly review some recent, very important, results that are crucial in this research field and that will help the comprehension of our research activity described in the following chapters. In Section 1 we recall very elementary concepts of molecular biology (however, we do not enter into the biochemical details about the DNA molecular nature); in Section 2 we discuss the fundamental technologies that allow to quantitatively investigate the spatial architecture of the genome (in particular we focus on the Hi-C experimental technique); in Section 3 we report the recent results obtained by analyzing the interaction data provided by these experimental methods, and we describe the emerging scenario about how chromatin appears to be organized in the nucleus; finally, we review the most recent polymer physics models that aim to quantitatively describe and reconstruct the three-dimensional structure of the genome. The results described in this chapter have been introduced and mostly discussed in important papers from Lieberman-Aiden *et al.*, 2009, Dekker *et al.*, 2013, Fraser *et al.*, 2015.

1.1 Genome, chromosomes and chromatin

Within the cell of eukaryotic organisms the filament of DNA is associated with a variety of proteins that pack DNA in a compact structure. In addition to these proteins, called histones, there are also many proteins that bind the DNA and are required for many biological purposes, as gene expression, DNA replication, DNA repair and DNA recombination. The complex of DNA and proteins is known as chromatin. Chromatin exhibits a complex spatial organization. Precisely, chromatin is organized in a set of different structural entities called chromosomes, occupying distinct spatial regions that are indicated as chromosomal territories (CT) and are clearly visible with microscopy techniques (Figure 1, Cremer&Cremer, 2001). The genomic length (i.e. the number of base pairs composing the genome) and the number of chromosomes depends on the considered species. For instance, human genome consists of approximately 3.2×10^9 base pairs (bp, in molecular biology notation) and it is distributed over 23 chromosomes. The majority of eukaryotic cells are diploid, i.e. they contain two copies of each chromosome. The complexity of chromatin folding problem results even more evident if we consider the linear length of the total human genome, that is about 2m, included in a nucleus having a diameter of approximately 10÷15μm. This compaction level is achieved through an efficient interaction between DNA and proteins. Histones are responsible for the first and most basic level of chromosome packing, called nucleosome, that is a protein-DNA complex. Each individual nucleosome consists of a structure of eight histone proteins (two molecules each of histone H2A, H2B, H3 and H4) around which a double-strand of DNA is wrapped. The length of DNA associated with each nucleosome is 147 base pairs. This structure is called nucleosome core particle. Each nucleosome core particle (which is about 11nm) is separated from the next by a filament of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. On average, nucleosomes repeat at intervals of about 200 nucleotide pairs. So, since human genome has 6.4×10^9 bp, it consists of about 30×10^6 nucleosomes. This structure is known as ”beads on a string” (where the ”bead” is the nucleosome and the ”string” is linker DNA) organization. Within a chromosome, it is possible to classify the chromosomal regions into two categories: euchromatin and heterochromatin. DNA in both types of chromatin is packaged into nucleosomes. Heterochromatic regions are composed by nucleosomal DNA that shows a high degree of compaction, while euchromatic nucleosomes are much less compacted. The high level of compaction reduces the accessibility of the DNA contained in these regions, which are therefore associated with a very low level of transcription.

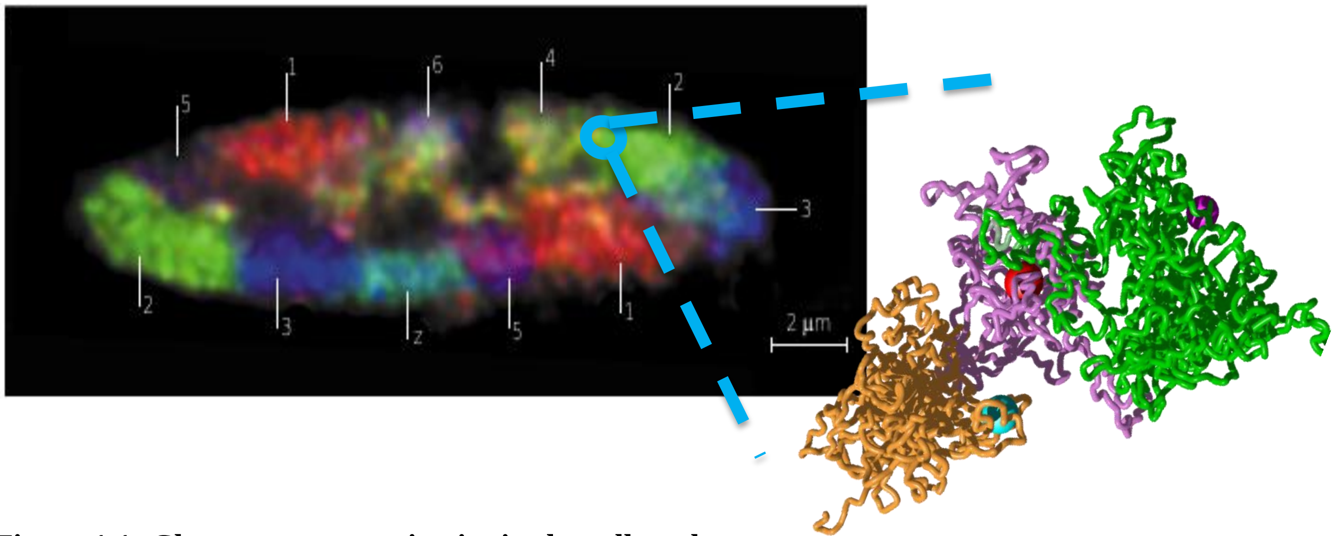

Figure 1.1: Chromosomes territories in the cell nucleus

In this microscopy image, chromosomes are represented by colored spots, and occupy distinct region in the nucleus, forming the chromosomal territories (CT). The image shows a mid-plane section of chicken fibroblast cells (figure adapted from Cremer&Cremer, 2001). Within the chromosomes, at much lower length scales, chromatin exhibits a very complex organization, as shown by the polymer cartoon, and reconstructing its 3D structure is an open problem in molecular biology.

1.2 The Chromosome Conformation Capture (3C) based techniques

During the past decade, a series of molecular and genomic approaches been developed and can be used to study three-dimensional chromosome folding with unprecedented accuracy. These methods are all based on the chromosome conformation capture (3C). They allow the determination of the frequency with which any pair of loci in the genome is in close enough physical proximity (in the range of 10÷100nm) to become crosslinked (i.e. the pair can be bound by some molecule). It is schematically shown in Figure 1.2, Panel A. First, the cells in the population are crosslinked with formaldehyde to covalently link chromatin segments that are in close spatial proximity. Next, chromatin is fragmented by restriction digestion (as HindIII or NcoI). Crosslinked fragments are then ligated to form unique hybrid DNA molecules. Finally, the DNA is purified and analyzed. The experimental steps just described are common to all 3C methods. The difference among the specific methods is how the ligation product is detected (Figure 1.2, Panel B). The most common methods are the 3C, 4C, 5C and HiC (see next subsections).

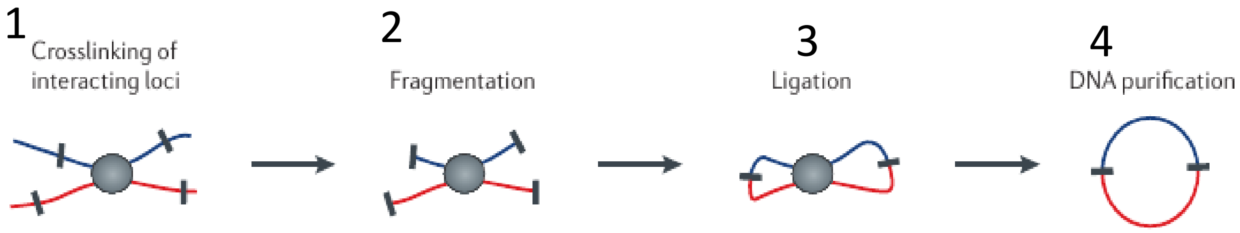

Figure 1.2: The Chromosome Conformation Capture (3C) techniques

Schematic representation of the 3C methods. First, the chromatin is crosslinked (step 1) with formaldehyde in order to capture pair of loci close in space, then they are fragmented (step 2), ligated (step 3) and purified (step 4). Panel B: (figure adapted from Dekker *et al.*, 2013)

The data generated from the 3C techniques

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

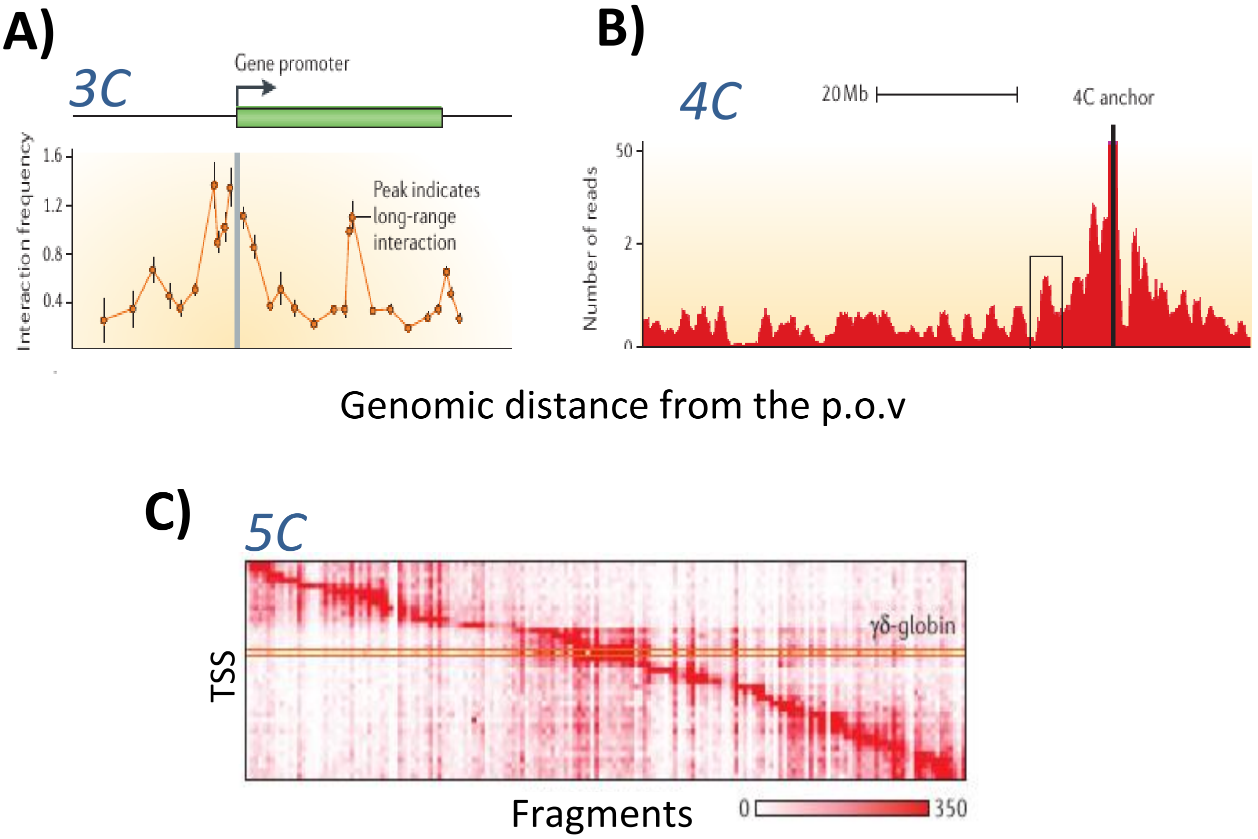

Figure 1.3: Examples of 3C, 4C and 5C data sets.

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

The Hi-C method

The Hi-C method (Lieberman-Aiden *et al.*, 2009) is the first genome-wide adaptation of 3C and include a further step in which, after restriction digestion, the staggered DNA ends are filled in with biotinylated nucleotides (Figure 1.4, Panel A). The resulting DNA sample is composed by ligation products of chromatin that were in spatial proximity in the nucleus, with biotin at the ligation junction. This facilitates selective purification of ligation junctions that are collected in a Hi-C library and then directly sequenced along the genome, producing a list of interacting fragments. Then, data are organized in a genome-wide contact matrix, obtained by dividing the genome into windows (indicated as *loci*) of fixed length (in the first version, this length was 1Mb=1000000bp long). This important parameter defines the Hi-C data resolution. Each bin of the matrix x_{ij} contains the number of ligation products between the locus *i* and locus *j*. So, the extracted information is the contact frequency of any pair of loci *i* and *j* in the chromosome, that is obviously directly related to the chromosome spatial architecture. In Figure 1.4, Panel B, is reported an example of interaction matrix for an entire chromosome, in different cell lines at 50Kb resolution. As the resolution increases, i.e. the size of the partitioning window is reduced, the matrix size increases. Since the Hi-C technique is able to detect interactions between any two loci in the genome, to each chromosome is associated its contact matrix (*Cis* data). Interactions between loci belonging to different chromosomes are also detected (with a much lower frequency), and are organized in *Trans* contact matrices. In all this work, we will focus only on *Cis* contact matrices.

Interpretation of the data and the normalization problem

It is important to consider what kind of information the Hi-C method (and all the 3C-based methods) produces. It gives the relative frequency in the cell population by which two loci *i* and *j* are in close spatial proximity. Anyway, the method does not give information about the nature of the contact, not distinguishing functional from non-functional associations and it does not reveal the mechanisms producing their co-localization. Spatial proximity can be the results of contacts mediated by protein complexes that bind them or can be the result of indirect co-localization to the same subnuclear structure (as the nuclear lamina). Furthermore, co-localization can be due to random collisions between distant regions of chromatin in the nucleus, due to the chromosome flexibility. Also, the exact 3D structure of a specific region is highly variable from cell to cell, even if they are in the same differentiation stage. Each ligation product due to an interaction represents a contact involving a pair of loci in a single cell in the population. Thus, Hi-C (all 3C-based) interaction frequency data represent the fraction of cells in which pairs of loci *i* and *j* are in spatial proximity at the time the cells are fixed. The final value contained in the matrix bin x_{ij} represent the sum of interactions over a large cell population, and in each cell chromosomes conformation is determined by many different factors that act on the chromatin polymer and make the structure highly variable.

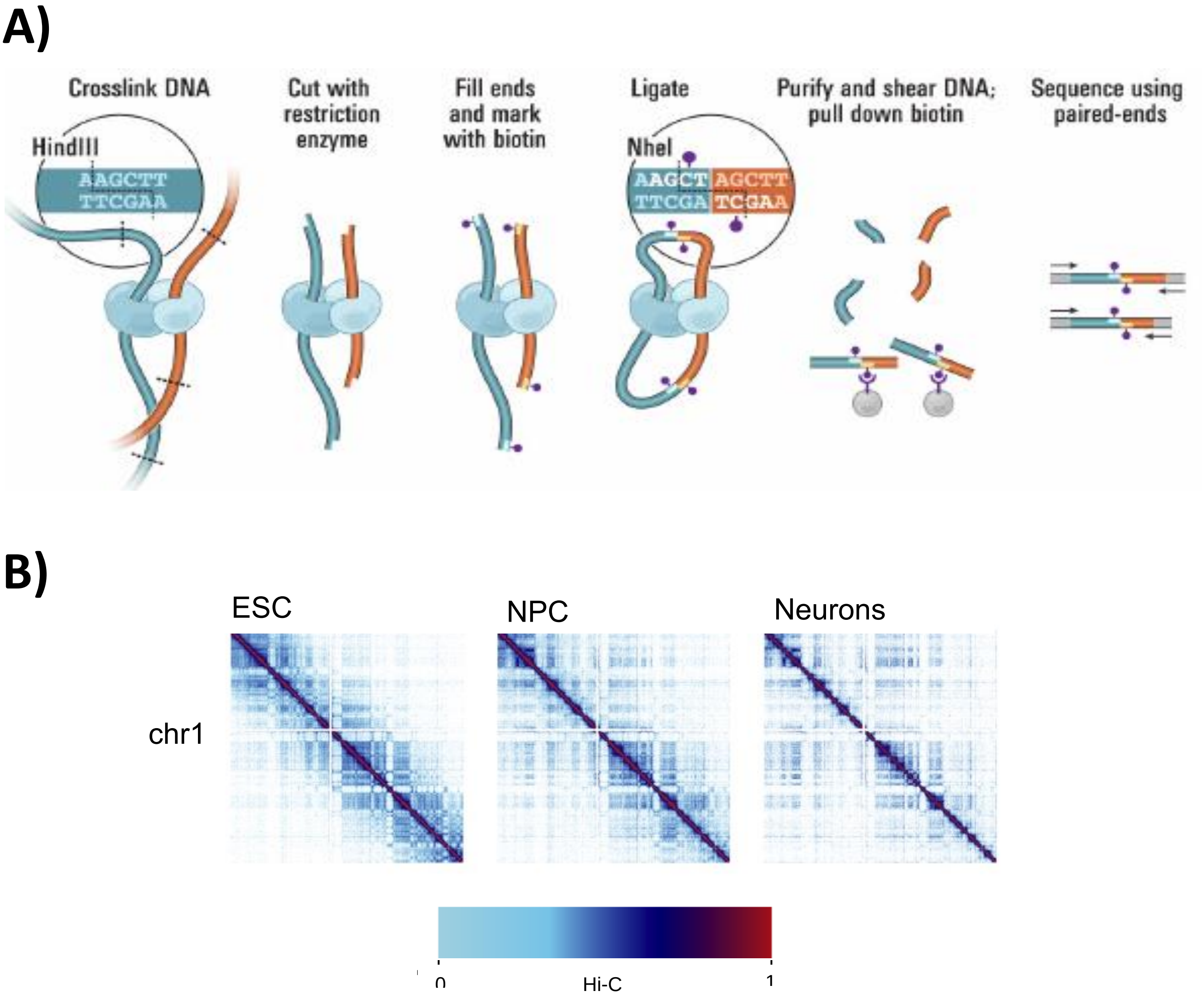
A)
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends
B)
ESC
NPC
Neurons
chr1
0
Hi-C
1

Figure 1.4: The Hi-C method

Panel A: Schematic representation of the Hi-C experimental procedure. The labeling with biotin allow to efficiently detect the ligated fragments (figure adapted from Lieberman-Aiden *et al.*, 2009). **Panel B:** Example of Hi-C data output. Data are collected in a 2 dimensional heat map (as in the 5C case). Hi-C data from the entire mouse chromosome 1, at 50Kb resolution. Data from Fraser *et al.*, 2015 (used in Chapter 2).

1.3 Genome structure from chromatin interaction data

A/B compartments in the genome

Based on the analysis of the pattern contained in the Hi-C matrices (through a principal component analysis, Figure 1.5, Panel A), each chromosome can be partitioned in two classes, named A and B compartments (Lieberman-Aiden *et al.*, 2009, Rao *et al.*, 2014). Two regions in the same compartment are enriched in interaction while two regions belonging to different compartments are depleted in interaction. (as confirmed also by FISH experiments). Compartments are considerably large regions of chromatin, having a characteristic size of 5÷10 Mb, and alternate along the chromosomes. A compartment is typically associated with euchromatin, since it is less compact and correlates with gene presence, higher expression and accessible chromatin, while B compartment has higher interaction values (Lieberman-Aiden *et al.*, 2009). This is in agreement with the known presence of open and closed chromatin in the nucleus (Figure 1.5, Panel B)

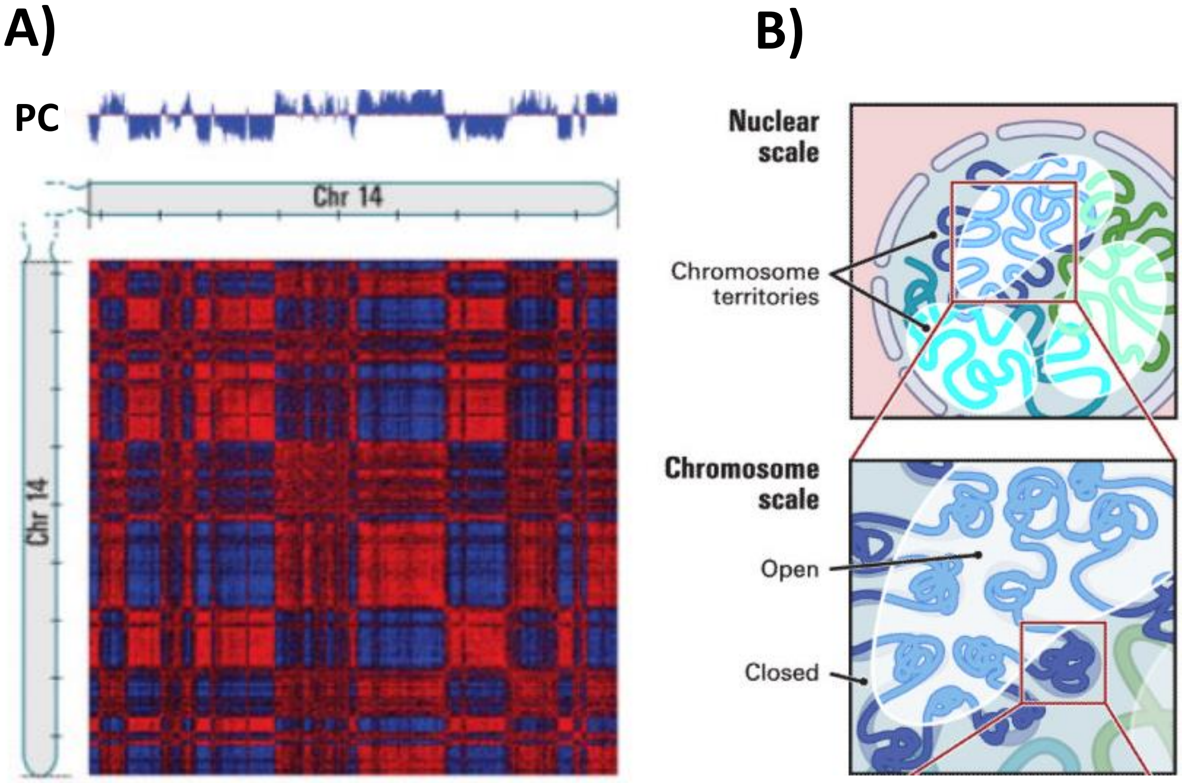

Figure 1.5: A and B compartments

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

The discovery of Topological Associating Domains (TADs)

Since the 5C and Hi-C methods were introduced, interaction data have been analyzed to identify structural properties of chromatin. In the previous subsection, the A and B compartment have been discussed and they allow to classify the genomic regions according to their Hi-C interaction profile. Nevertheless, other levels of organization and structural units have been discovered. In particular, a common feature among several organisms (from *drosophila melanogaster* to mouse and human) is the existence of discrete regions, much smaller than compartments (previous subsection), where chromatin is marked by a high level of interaction. To indicate such domains, various names have been used in literature, as topological domains (Dixon *et al.*, 2012) and topological associating domains or, briefly, TADs (Nora *et al.*, 2012). As standardly used in literature now, we will use the latter in the following. On a 5C or Hi-C matrix, TADs appear to be as squares of high intensity along the diagonal (Figure 1.6, Panel A). From the structural point of view, this correspond to the fact that distinct loci located in the same TAD tend to interact with higher intensity than two loci located in two different TADs (Figure 1.6, Panel B), and FISH experiment confirm such scenario (Dixon *et al.*, 2012). TADs are found to be a universal building blocks of chromosomes, as both mouse and human are composed by more than 2000 domains, covering almost all the genome. Furthermore, they are conserved between different species (Dixon *et al.*, 2012). Their typical size (approximately 0.5÷1 Mb) is much smaller than the A and B compartment, and they can be active or inactive. To identify TADs several computational algorithms exist (Dixon *et al.*, 2012, Rao *et al.*, 2014, Fraser *et al.*, 2015). The mechanism that regulates the formation of TADs is still not clearly understood, and polymer models have been proposed to quantitatively describe it (Barbieri *et al.*, 2012, Brackley *et al.*, 2013, Sanborn *et al.*, 2015, Fudenberg *et al.*, 2016, Chiariello *et al.*, 2016).

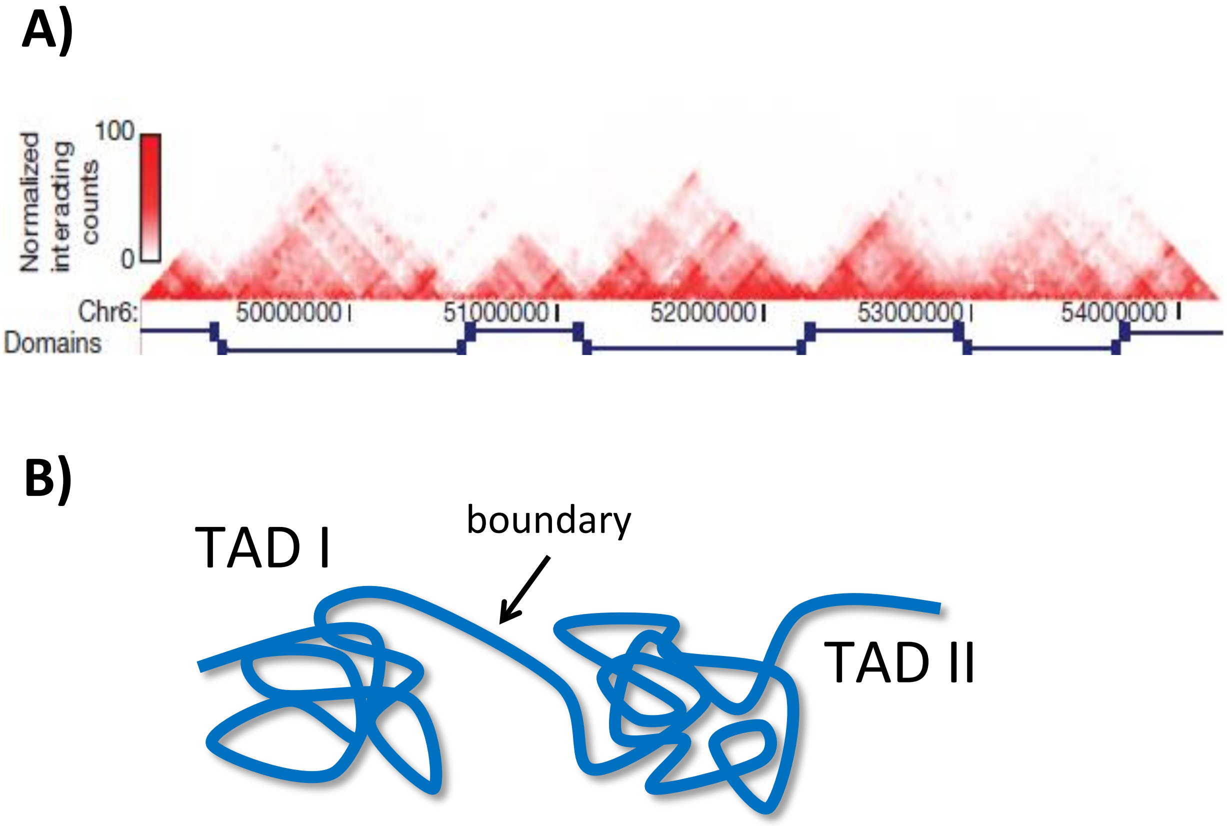

Figure 1.6: Topological Associating Domains

Panel A: Hi-C interaction frequencies for a region of chromosome 6 in mouse embryonic stem (ES) cells (Figure from Dixon *et al.*, 2012). The domains appear as squares along the diagonal of Hi-C matrix (here represented as a triangular matrix). **Panel B:** Schematic representation of chromatin spatial organization in TADs. Each TAD correspond to a high intensity square block in the Hi-C matrix.. Loci belonging to the same TAD interact more frequently than loci in different TADs. See Chapter 3 for a possible (polymer physics based model) mechanism of TAD formation.

Further research developments

The results reviewed in this chapter represent only a limited part, yet fundamental, of the key points in the recent history of this research field. As the technology quickly evolves, more sophisticated and refined experiments have been performed, producing better and higher quality data. In this way, very complex and more complete Hi-C datasets are available, with higher resolutions (up to 1Kb, Rao *et al.*, 2014, Dixon *et al.*, 2015) and for an increasing number of tissues and cell lines. In parallel, other experimental technologies have been developed to detect chromatin contacts (Khalor *et al.* 2011; Rao *et al.*, 2014; Beagrie *et al.*, 2017). Furthermore, experiments have been performed to evaluate the impact of chromatin structure alterations on health (as TADs disruption, described in Lupianez *et al.*, 2015, or neoTAD formation, described in Franke *et al.*, 2016), whether or not pathogenic spatial rearrangements occur. These notable works demonstrate the deep relationship between chromatin organization and individual fenotype, and confirm once more the importance of investigating the genome architecture in space. Overall, these more recent results allow to improve our knowledge (far anyway from being complete) about the chromatin organization, contributing to further enrich the scientific landscape about this interesting topic.

Polymer models

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

References

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

Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, Nicodemi M (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A* **109:** 16173-16178

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

Bickmore W, van Steensel B (2013) Genome architecture: domain organization of interphase chromosomes. *Cell* **152:** 1270-1284

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

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

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

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

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren

B (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature* **518:** 331-336

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

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

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

Franke M *et al.*, (2016) Formation of new chromatin domains determines pathogenicity of genomic duplications, *Nature* **538**: 265-269

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

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

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

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

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

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

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

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

Nicodemi M, Pombo A (2014) Models of chromosome structure. *Curr Opin Cell Biol* **28C:** 90-95

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

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

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

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

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

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, Thurman RE, Cheng Y, Gulsoy G, Dennis JH, Snyder MP, Stamatoyannopoulos JA, Taylor J, Hardison RC, Kahveci T, Ren B *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* **515:** 402-405

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665-1680

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

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

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

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

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

Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Dostie J, Bickmore WA (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & Development* **28:** 2778-2791

Chapter 2

The hierarchical organization of mammalian genomes

In this chapter, we will investigate the 3D spatial architecture of the mammalian genome by analyzing the Hi-C data in a neuronal differentiation model from mouse embryonic stem cells (ESC) via neural progenitor cells (NPC) to neurons (Section 1). In particular we find that the genome is organized in a hierarchical structure of domains (that we name ‘metaTADs’), well described by tree diagrams (Section 2.2) and statistically robust (Section 2.3), up to large genomic length scales. Furthermore, the metaTAD organization is correlated with a variety of epigenetic features, indicating a functional role in this particular structure (Section 2.4) and its reorganization is linked to cell differentiation (Section 2.5). Finally we show with a simple polymer physics model how hierarchical folding increases the chromatin packaging efficiency (Section 2.6). All the results contained in this chapter have been published in the paper *‘Hierarchical folding of chromosomes and its reorganization underlies transcriptional changes in cellular differentiation’* (Fraser *et al.*, 2015).

2.1 Dataset, normalization approach and domain definition

Hi-C experiments

This study has been developed in collaboration with the group of Professor Ana Pombo at Max Delbruck Center for Molecular Medicine in Berlin and with the group of Professor Josee Dostie at McGill University in Montreal, which performed the Hi-C experiment and produced the datasets that we analyzed. To explore the long-range chromatin folding during differentiation, three time points were considered: mouse embryonic stem cells (ESC), intermediate neuronal precursor cells (NPC) and post-mitotic neurons (Neurons). For each time point, Hi-C libraries were produced, and then standard normalization process was performed (see next subsection). For each time point, two Hi-C replicates were generated, with NcoI and HindIII restriction enzyme. The results presented here are based on the NcoI

Hi-C dataset. Since we did not work directly to the experimental stage, all the biological and chemical details of the Hi-C experiment, cell culture, neuronal differentiation, preparation of the libraries and sequencing, will not be discussed here. For any further information, please see the reference paper.

Normalization approach

Once the Hi-C libraries are prepared (see previous chapter), data from different sequencing lanes are combined and binned using 50Kb windows so to be converted in raw Hi-C contact matrices. The normalization approach used is the ICE Iterative Correction (Imakaev *et al.*, 2012), to correct the systematic biases in the Hi-C matrices based on equal DNA visibility principle. To compare data between different Hi-C datasets from different cell lines, an additional normalization step was applied. This consists of a background subtraction that takes into account for the biological noise due to random generated Hi-C ligation products. The subtracted part is calculated from *Trans* data distribution (see previous chapter), and it is the average read count plus one standard deviation. This quantity is then subtracted from all interactions, both *Cis* and *Trans*. Finally, the data are divided by factor that accounts for the differences in library depth between samples. This factor is just the total number of reads (*Cis* and *Trans*) left after the background subtraction step.

The matrices show the usual structure of chromosomes into domains of high interactions, reflecting the presence of A/B compartments and TADs. In Fig.2.1, we observe changes during differentiation in the long-range contacts of each chromosome.

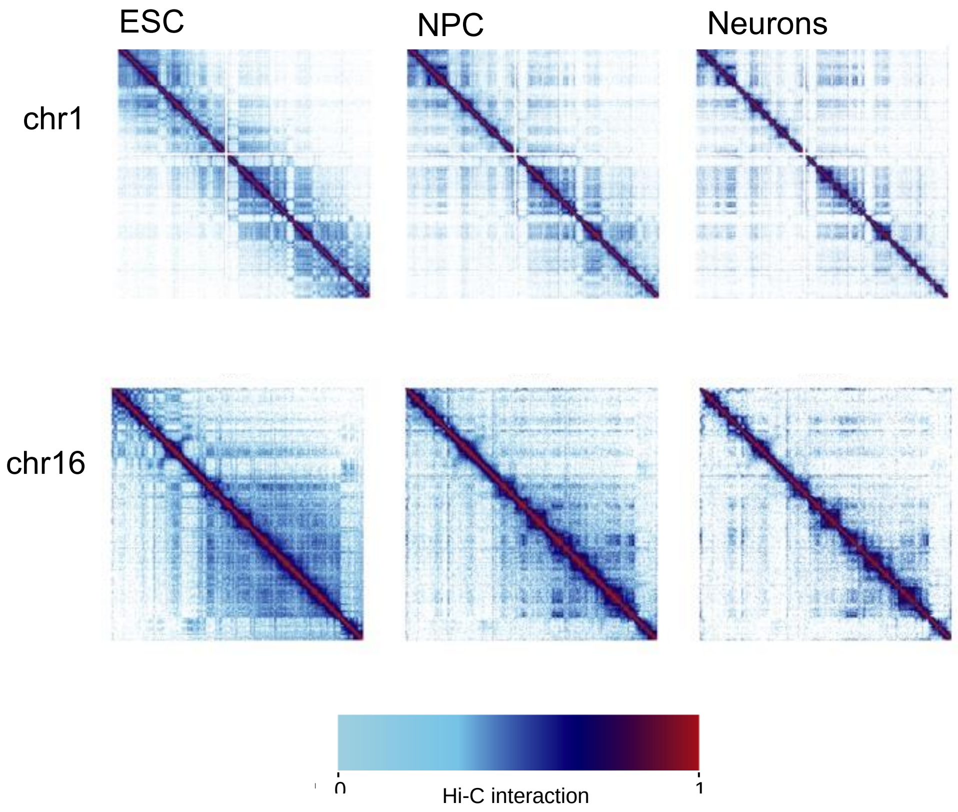

Figure 2.1: Changes in long range interactions Chr1 and Chr16

Complex changes in long-range chromatin interactions of many chromosomes are prominent when plotted during the differentiation time series. To better appreciate the long range contacts, Hi-C interaction data is plotted in log-scale.

TADs identification strategy

To study the higher-order chromosome organization, we first identified the elementary TADs coordinates in the chromosomes for all the three time points, using the Directionality Index (Dixon *et al.*, 2012). This quantity is defined by the relation DI[*i*]=(B-A)/E, where *i* is the bin index, B and A are the read counts upstream and downstream within a window of size L, and E is their average. Since the DI signal becomes approximately independent of L for L>2Mb, we set L=2Mb. We then consider a threshold, α, and we identify the boundaries of TADs where the signal DI is above the given threshold (DI[*i*]>ασ, left boundary) and where DI is below the threshold (DI[*i*]<ασ, right boundary). In previous relations, σ is the standard deviation of the DI signal computed genome wide. For any considered value of the threshold,

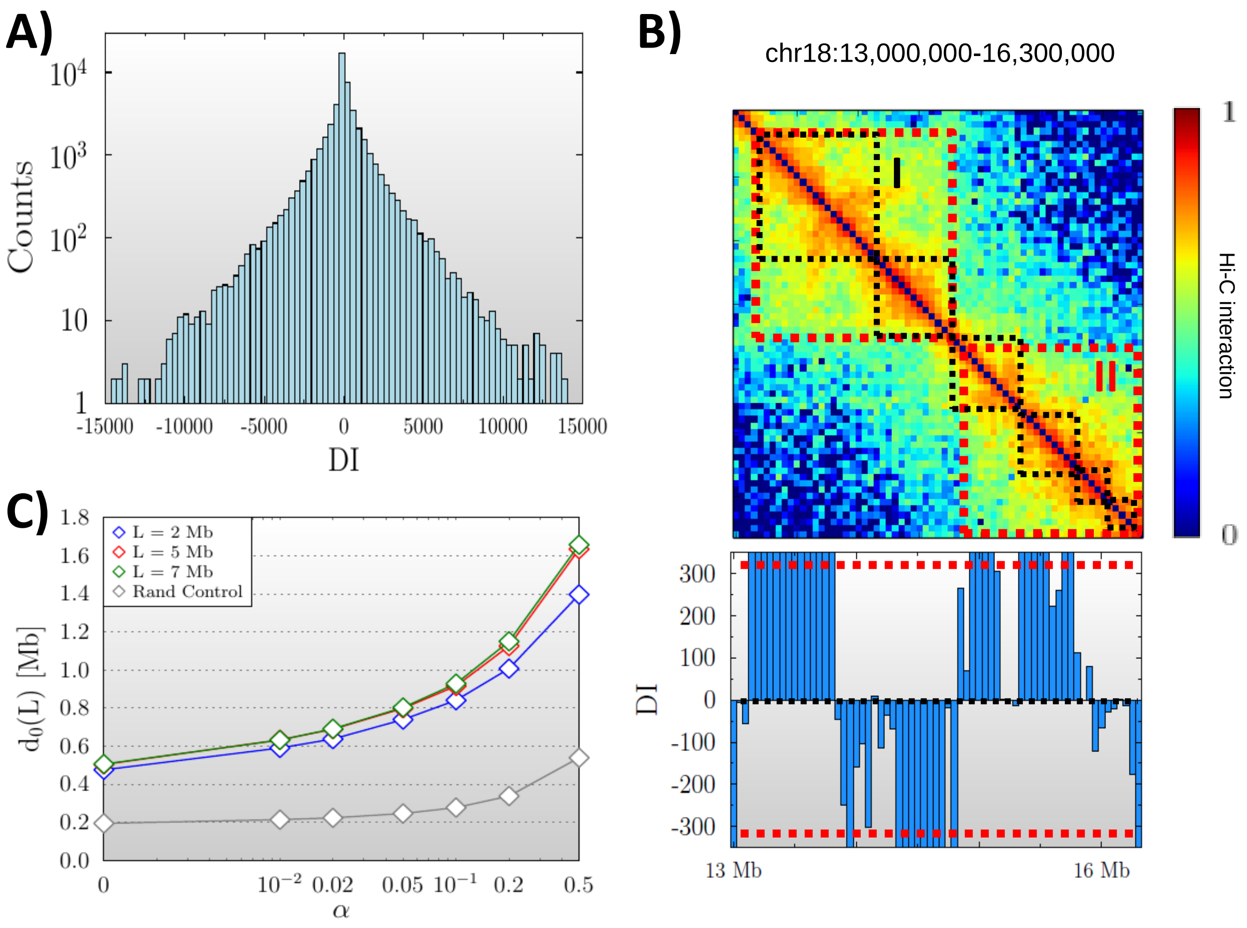

Figure 2.2: TAD identification

Panel A: Distribution of the signal DI genome wide for ESC cell line. **Panel B**: TAD identification with threshold α = 0.0 (black dashed squares) and α = 0.2 (red dashed squares). **Panel C**: Mean TAD size d as function of the threshold α.

the average TAD size d_0 is much larger than the random control case (as illustrated in Figure 2.2, Panel C), even for small α values, where the size is weakly dependent on this parameter. The random control case is an average TAD size where the TAD coordinates are identified on a random control matrix, obtained by bootstrapping the original Hi-C matrix along all the possible sub-diagonals. In this way, we preserve the genomic distance contact profile. This randomization method has been used also in other analysis, that will be presented in following sections.

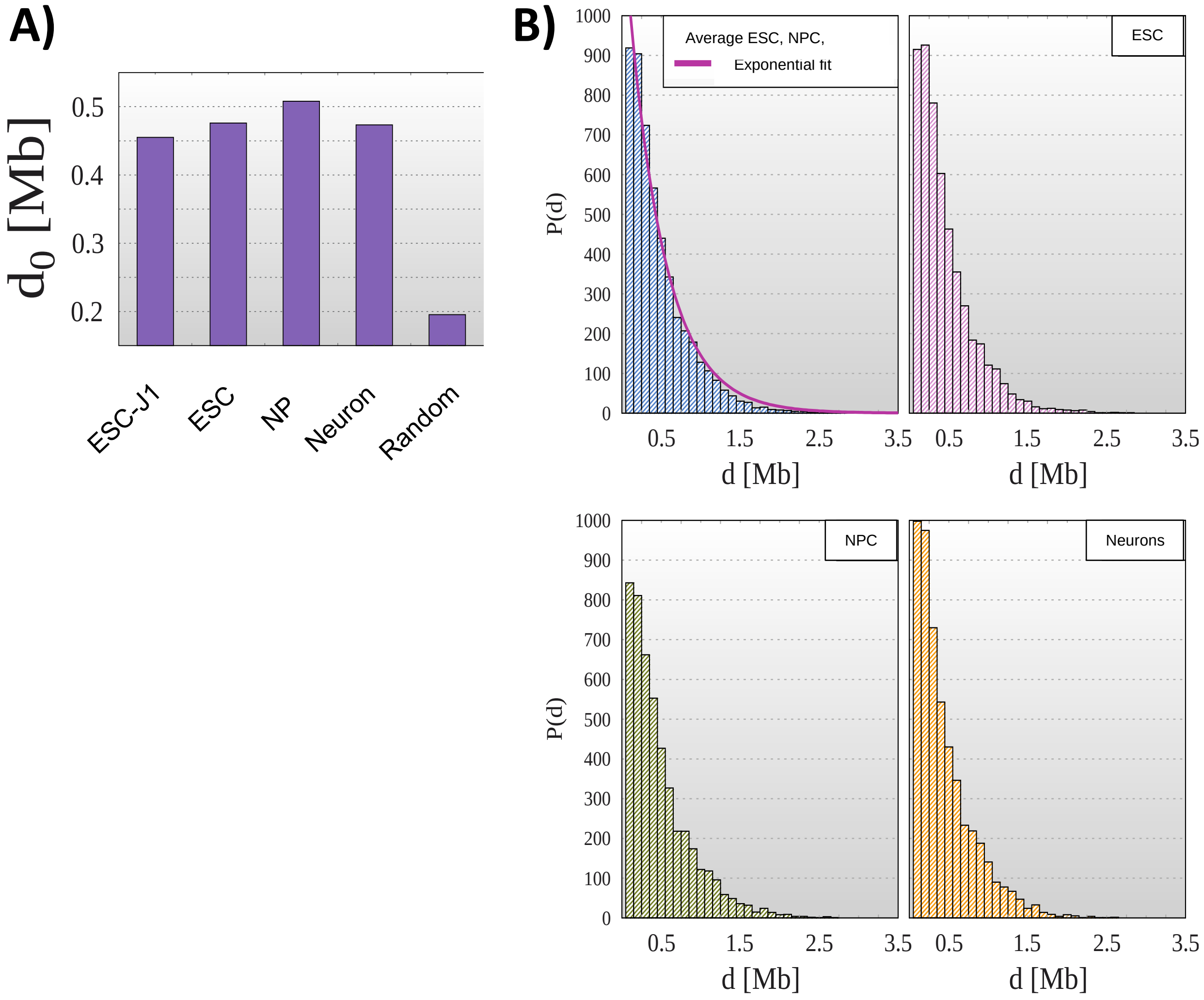

Figure 2.3: TADs size distribution

Panel A: The average TAD size, d_0 (for α = 0 and L = 2Mb), is measured for the three datasets produced in this study (ESC, NPC, Neurons) and for the ESC-J1 published data (Dixon *et al.*, 2012). The random case corresponds to TADs identified in a randomized Hi-C matrix. **Panel B:** The distribution of TAD sizes is roughly exponential in all our datasets.

Effect of the threshold α

To quantify the effect of the threshold α, we consider two values: α=0.0 and α=0.2, and identify the corresponding TADs in our ESC data. Analogously, we apply the same to ESC-J1 published data (Dixon *et al.* 2012), so to compare the results with the original TADs identified with the original HMM approach (Dixon *et al.* 2012, see also Chapter 1). The

TADs obtained from ESC with α=0 have an average size of ~500Kb, which is about half the average size (~1Mb) of the original TADs defined in Dixon *et al.*, 2012, while the TADs with α=0.2 have the same average size. As expected, we find that on average 2 TADs with α=0 threshold overlap with one TAD with α=0.2 or identified by the original HMM. Furthermore, the TADs obtained with α=0.2 overlap almost exactly with the TADs found with the HMM method. These results validate the method that we use here to identify the basic unit (TADs) of the hierarchy.

In the following analysis, we consider the TADs with α=0 as the fundamental domain where we start to investigate the hierarchical structure of the genome. We make this choice so we can explore the hierarchy from the lowest possible level of genomic lengths. Anyway, as we will show, the results are confirmed for other choices of the discussed parameters. The distribution of the basic TAD size is reported in Figure 2.3 for the three time points studied.

2.2 Clustering method

The clustering algorithm

Most chromatin contacts in the Hi-C matrices are found within TADs, but interactions are also detected between specific TADs and they extend up to large genomic scales, as is visually evident from Figure 2.4, Panel A. To quantitatively highlight the higher-order domain organization of chromosomes, we use a simple clustering procedure applied on the Hi-C contact data. For each chromosome, we iteratively select the two most interacting domains and we join them in a new, larger, domain that we name metaTAD. This new domain is then added back to the set of fundamental TADs, and this procedure is repeated up to the whole chromosome length. More precisely, after we identify the fundamental TADs, which is the first level of the hierarchy, we calculate the mean interaction $I_{k,k+1}$ between all the neighbouring domain pairs:

$$(1) \qquad I_{k,k+1} = \sum_{ij} x_{ij} / (b_k - a_k)(b_{k+1} - a_{k+1})$$

where k is the TAD index, x_{ij} are the entries of the Hi-C matrix between the two TADs, b_k and a_k (resp. a_{k+1} and b_{k+1}) are the left and the right boundary coordinates of the TAD with index k ($k+1$). The sum over i runs from a_k to b_k and j from a_{k+1} to b_{k+1} (see Figure 2.4 Panel D, for a schematic representation of this quantity). If the number of domains is n, there are $n-1$ neighbouring TAD pairs (and $n-1$ values for I). We then select the pair with the highest value of the interaction I_{max}, and this pair is joined into one new domain that we call a metaTAD, encompassing both TADs. The list of domains is then updated with the new metaTAD (while its composing subdomains are taken out) and the procedure is repeated iteratively until remains only one metaTAD having the size of the whole chromosome. In this way, we build the entire hierarchy of domains, for all the chromosomes. In Figure 2.4, Panel B, it is shown a pictorial description of the clustering procedure, and in Panel C the result on a real locus having a genomic extension of 5Mb long (ESC cell line, chr2:53000000-58000000).

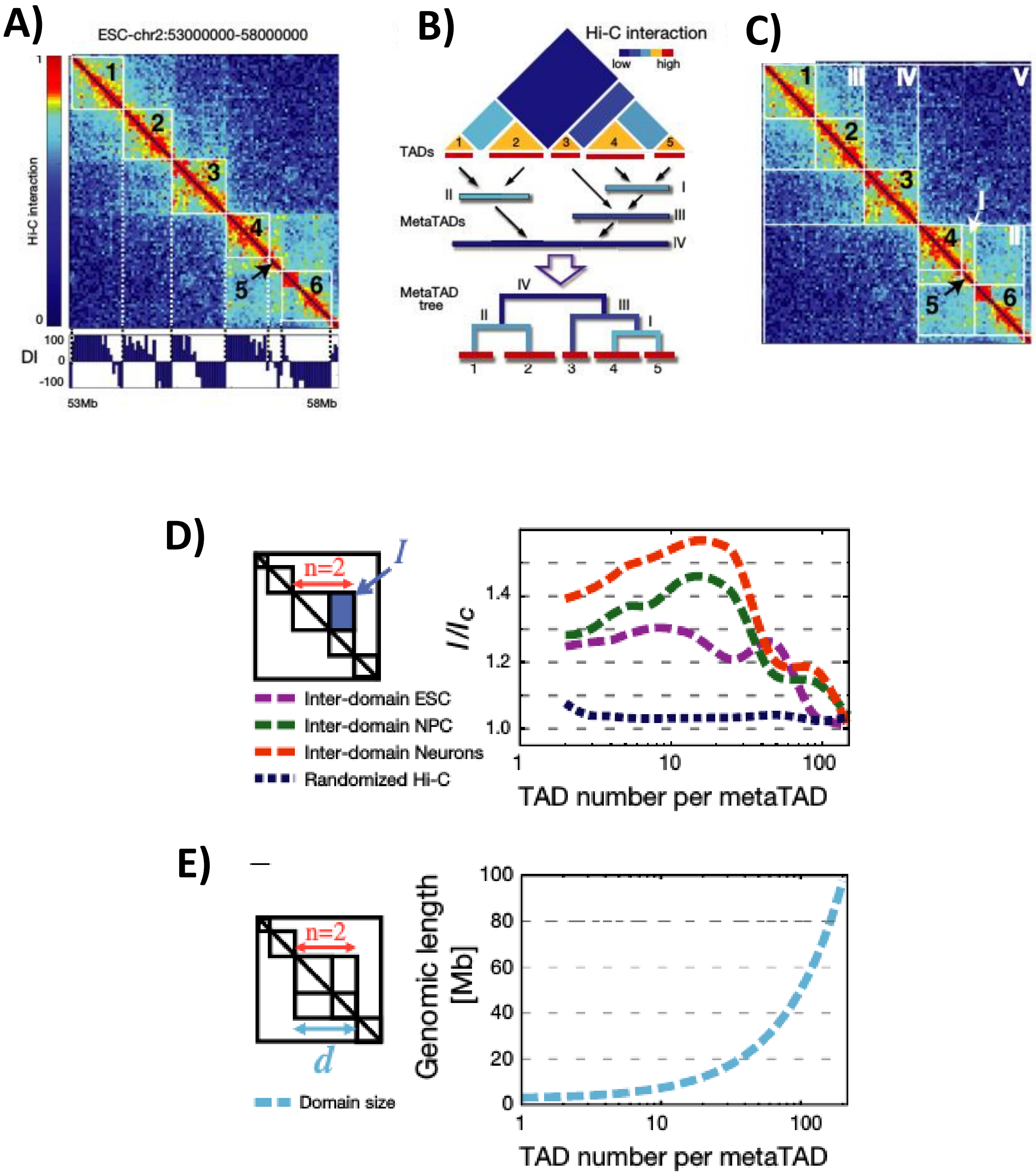

Figure 2.4: Chromosomes are organized in a hierarchy of higher-order domains

Panel A: ESC Hi-C map of chromosome 2, 53-58Mb. The Directionality Index (see previous section) is used to identify the elementary TADs, in this matrix numbered from 1 to 6. **Panel B**: schematic representation of the single-linkage clustering used to identify the hierarchy. **Panel C**: examples of metaTADs (I-V) in the region showed in Panel A. **Panel D:** ratio of the average interaction I and the background control value I_c calculated for the three time point ESC, NPC and Neurons, as a function of the total number of elementary domains included in the metaTAD. In all cases, the curve remains 20% above the random control curve (blue) up to scale of 50Mb (the details about the calculation of I and I_c are given in the next section). **Panel E:** the size of metaTAD size d as a function of the elementary domains that they contain. It is shown that hundred TADs correspond to an average length of about 50Mb.

The tree representation

Once we obtain the hierarchy of domains-within-domains from the procedure above described, the most intuitive representation of this organization is a tree diagram. Indeed, overlaying the hierarchy structure onto the experimental Hi-C contact matrices, as shown in Figure 2.5, gives a visual confirmation that the metaTAD structure matches the patterns contained in the data.

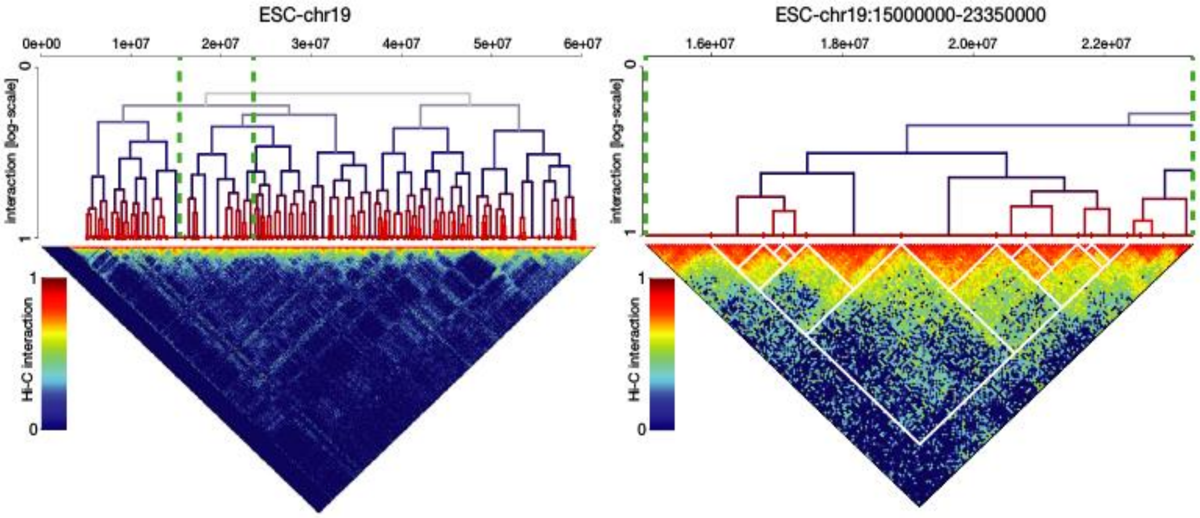

Figure 2.5: The tree representation of chromosome 19 (ESC cell line)

Left panel: full chromosome, right panel: zoomed region. Interaction between elementary domains are not uniform, but occur between specific TADs through specific contacts. It is evident the visual matching between the pattern contained in the data and the metaTAD structure. The interactions are plotted in log-scale.

2.3 Statistical robustness of the tree structure

The *I/Ic* and *J/Jc* control

To give a quantitative measure of the statistical reliability of the identified metaTAD tree, we consider different approaches. First, we test if interactions between metaTADs (fundamental and higher-order domains) are higher than background interactions. In particular, we calculate the average interaction between the domains containing n elementary TADs, $I(n)$. As control, we computed the same average interaction between two neighbouring regions of size equal to the original domains, but randomly located at any other inter-TAD boundary existing at that level of the tree. Mathematically, this correspond to calculate the quantity defined in equation (1) where b_k and a_{k+1} are shifted into new boundary coordinates $b_{k'}$ and $a_{k'+1}$ for two new neighbouring metaTADs, so that $b_{k'} - a_{k'} = b_k - a_k$ and $b_{k+1'} - a_{k+1'} = b_{k+1} - a_{k+1}$. Then, we compute the average $I_C(n)$ over the metaTADs composed by n elementary basic TADs. In this way, we obtain a curve $I/I_C(n)$ to be compared with the random control case. This is achieved by fully repeating the procedure described above (calculation of the metaTAD tree, $I(n)$ and $I_C(n)$) on random control Hi-C matrices obtained by bootstrapping the diagonals (see section 2.1). In this way, we can compare the significance level of the metaTAD hierarchy from the real Hi-C matrices with the hierarchy found on randomized matrices without any higher-order structure. In Figure 2.4, Panel D, we report the results for the three time points. See next subsection for a detailed discussion.

The second quantity that we consider is slightly easier to calculate, and it is based on the inner interaction of a metaTAD containing n fundamental TADs $J_k(n)$:

$$J_k(n) = \sum_{ij} x_{ij}/(b_k - a_k)^2 \qquad (2)$$

where x_{ij} are the values of interaction in the metaTAD k, a_k and b_k are the left and right boundary coordinates in the matrix, the sum over i and j runs from a_k to b_k (see Figure 2.6 Panel E, for a schematic representation of this quantity). Then, by averaging over k, we obtain $J(n)$. Analogously, we calculate the same quantity $J_C(n)$, using the real metaTAD boundaries, from the random control Hi-c matrix defined as usual (bootstrapping approach).

Results for murine ESC, NPC and Neurons datasets.

In Figure 2.6 the resulting $I/I_C(n)$ and $J/J_C(n)$ curves are shown, for the hierarchies found on the mouse ESC, NPC and Neurons cell lines (reported also in Figure 2.4, panel D, with smoothed curves). We find that, in all three time points, the curves remain well above the random control case up to n~80, that corresponds to domains of approximately 40-50Mb. This analysis is a further confirm of a scenario where the genome is organized in a hierarchical structure of domains-within-domains. The error bars reported represent the error extracted from the distribution of the chromosomes, and propagated on the ratio $I/I_C(n)$ and $J/J_C(n)$. We also use a logarithmic binning for n>10, since we span more than two order of magnitude. To give a sense of scale for the interaction scores, in Figure 2.7 is also reported the behavior of $I(n)$ and $J(n)$ as a function of the basic TAD number contained (i.e. n).

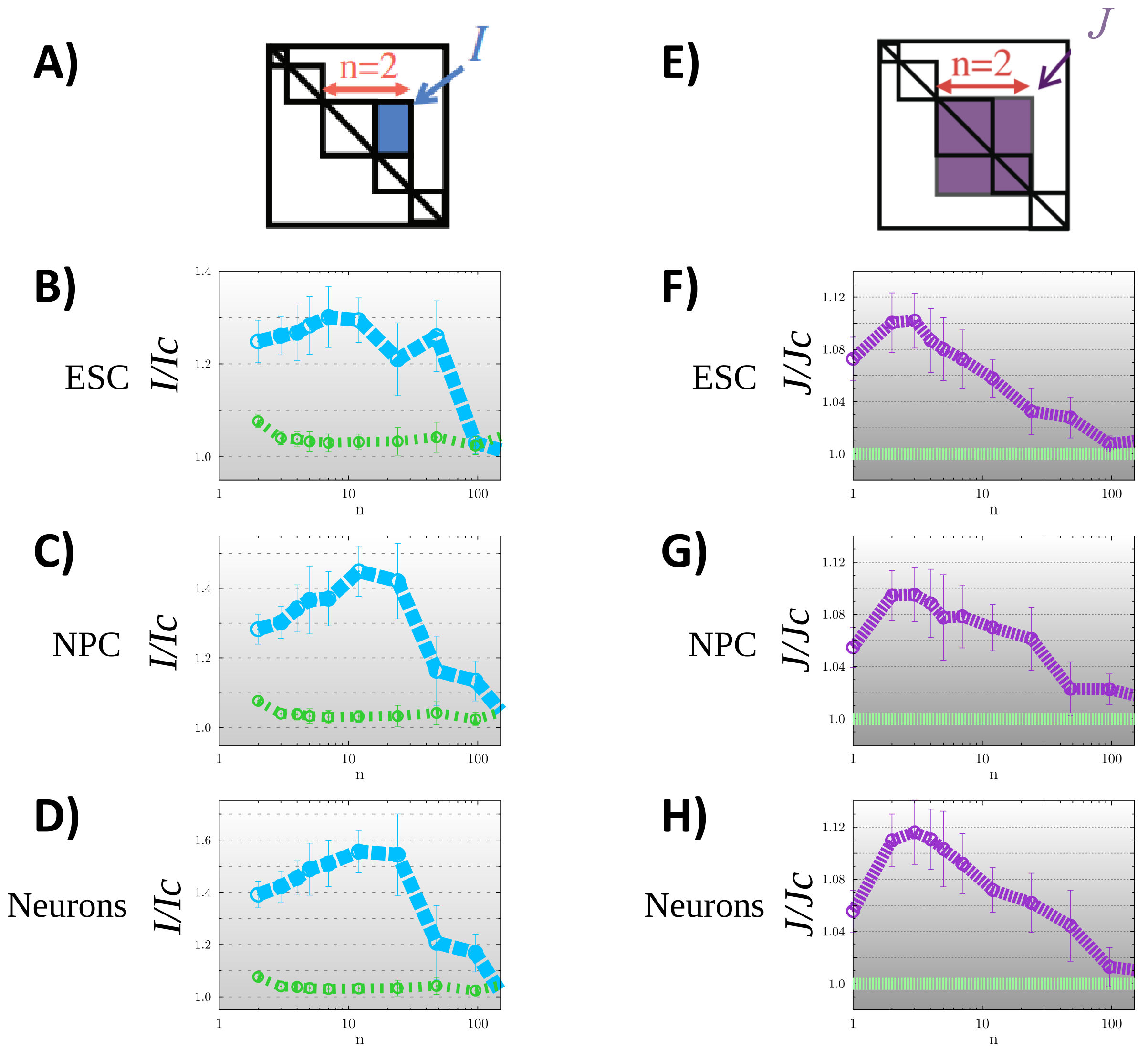

Figure 2.6: *I/Ic* and *J/Jc* controls

Panel A-E: Definition of *I* and *J*. **Panel B, C, D:** The *I/Ic* curves for the three time point. (ESC, NPC and Neurons respectively). In all cases, the light blue curves are well above the random control case (green curves) up to *n*~80-100, that corresponds to genome length approximately 40-50Mb. **Panel F, G, H:** The *J/Jc* curves for the three time points. As in the case of the *I/Ic*, we observe significant differences with the random case (the constant value 1, by definition) up to large length scales.

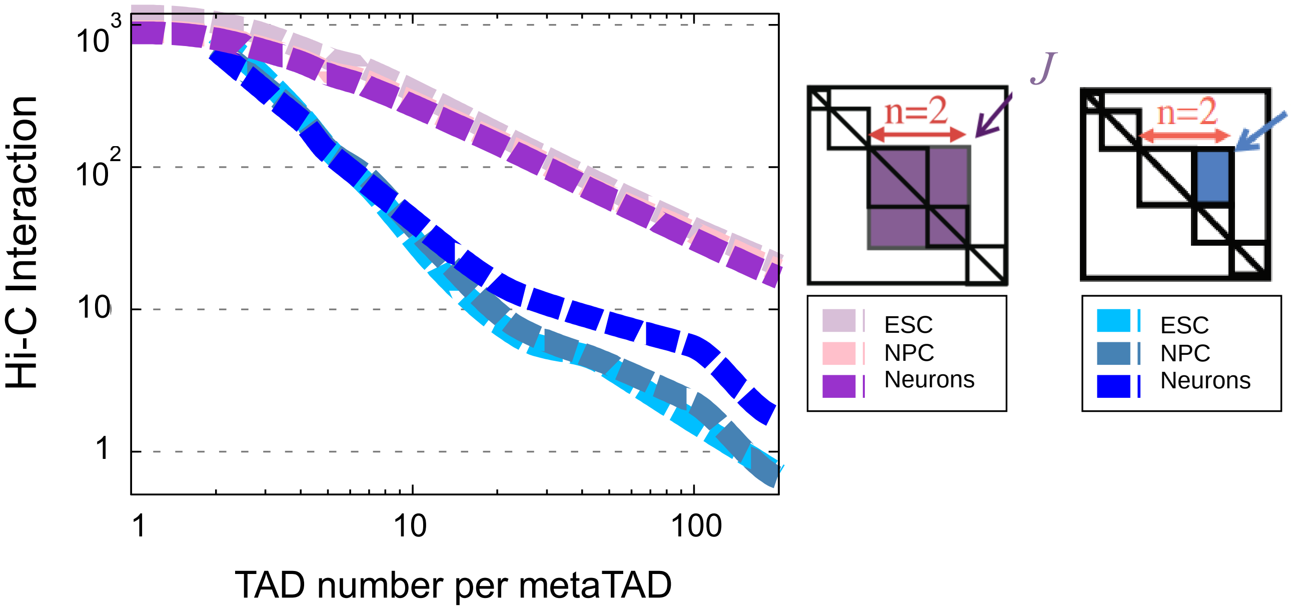

Figure 2.7: Average interaction at each tree level

In the plot is shown the average metaTAD inter-domain *I* (blue curves) and intra-domain *J* (magenta-purple curves) interactions, as a function of the number of elementary domains *n*, in all three time points. Essentially, it is analogous to the curves represented in Figure 2.6, without background normalization. As expected, intra-domain interaction *J* are always above inter-domain interaction *I*.

Results for murine ESC-J1 datasets

To enforce the results discussed above, we repeat the calculation of the *I/Ic* and *J/Jc* control curves using different fundamental TADs and different murine dataset (ESC-J1 from Dixon *et al.*, 2012). In Figure 2.8, we show the curves obtained from trees built starting from TADs identified with α=0.2 in our ESC dataset (Panel A-B) and in the ESC-J1 dataset (Panel C-D). Furthermore, we use also the original fundamental TADs as defined by Dixon with the Hidden Markov Model (HMM) approach, on ESC-J1 dataset. In all cases, both measures are significantly higher than the random control curves (represented in green) up to domains containing about 50 TADs. Since in both cases the average size for TADs is about 1Mb (see previous section), this corresponds to length scales of approximately 50Mb. We can therefore conclude that our findings are independent on the specific method used to identify the basic TADs and independent on the dataset considered.

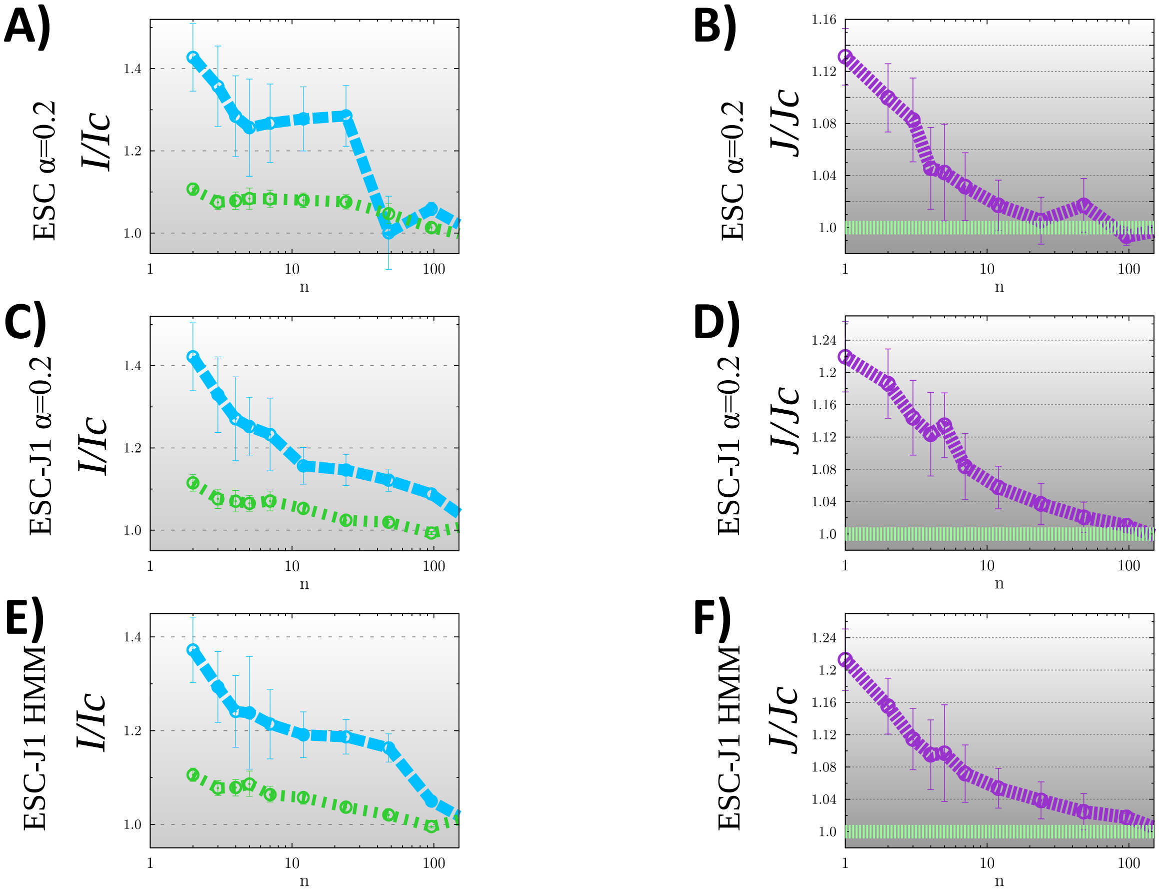

Figure 2.8: General features of metaTAD trees are significant and independent of the algorithm and dataset used.

Panel A-B: I/I_c (blue, A) and J/J_c (purple, B) curves, and the corresponding control cases (in green), using TADs identified with our method for α = 0.2. **Panel C-D:** I/I_c and J/J_c in the original ESC-J1 data from Dixon *et al. (*2012), using TADs identified with our method at α =

0.2. **Panel E-F:** *I/I$_c$* and *J/J$_c$* in the original ESC-J1 data from Dixon *et al.* (2012), using the original HMM TAD classification published in Dixon *et al.* (2012).

Results for human IMR90 and ESC-H1 datasets

To assess if metaTAD hierarchies exist also in other organisms, we analyse human IMR90 and ESC-H1 (human embryonic stem cell) Hi-C data from (Dixon *et al.*, 2012), building the metaTAD trees starting from our fundamental TADs identified with α=0.0. The average TAD size is d_0 = 0.44Mb in human ESC-H1 and d_0 = 0.55Mb in IMR90 cells. As the human Hi-C matrices are lacking the data corresponding to the centromere regions of the chromosomes, we consider separately the two chromosome arms. Accordingly, the random matrices used for the control curves are obtained by excluding the centromeres in the bootstrapping procedure (see previous section). We perform again the robustness analysis computing the *I/Ic* and the *J/Jc* curves, and we find a significant hierarchy up to domains of tens of megabases for both the considered cell lines (Figure 2.9, ESC-H1 in Panel A-B and IMR90 in Panel C-D).

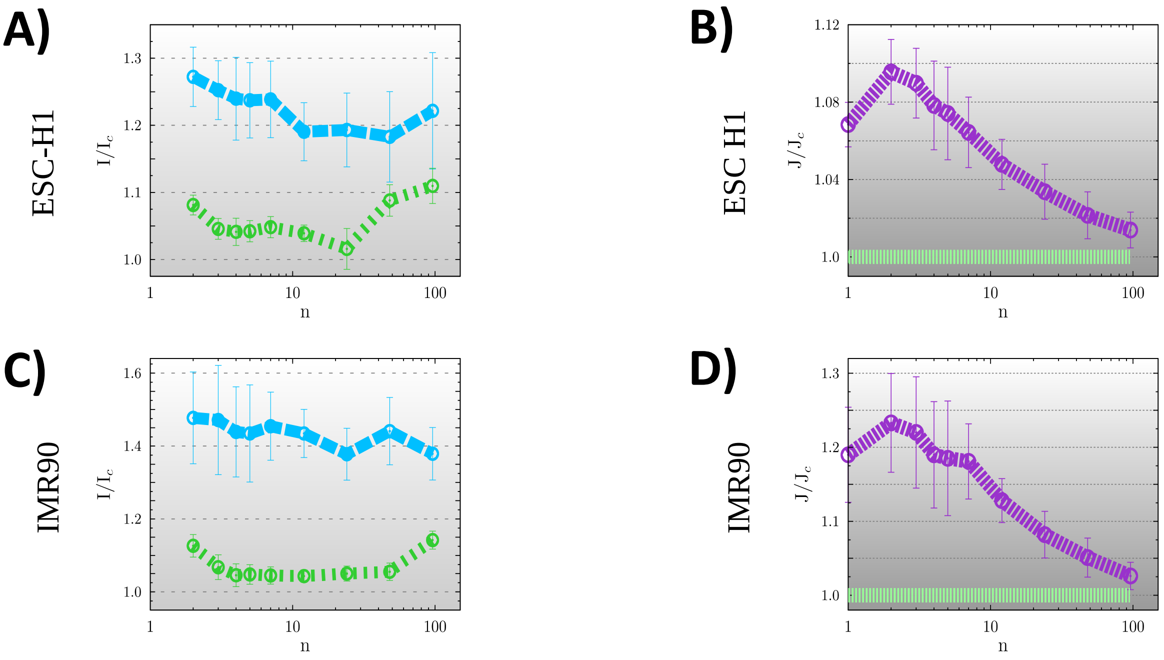

Figure 2.9: MetaTAD trees can be found in human cells.

Panel A-B: I/I_c (blue) and J/J_c (purple) curves, and the corresponding control cases (in green), in human ESC-H1 Hi-C data from Dixon *et al.* (2012). **Panel C-D**: I/I_c (blue) and J/J_c (purple) curves in human ESC-H1 Hi-C data from Dixon *et al.* (2012). In both cases, TADs are identified with our detection method using α=0.

Single cell validation by cryoFISH

To test the existence of long-range interactions across many TADs, with a completely independent method, we present results obtained with cryoFISH experiment (performed in the lab of Ana Pombo in Berlin), that is a fluorescence in situ hybridization approach combining the use of thin section and high-resolution imaging with confocal microscopy (Branco&Pombo, 2006). In this way, cellular architecture is efficiently preserved. We use three probes (*a*, *b* and *c*) covering genomic regions in the mouse genome, located in different TADs on chromosome 2, as shown in Figure 2.10 , Panel A and Panel B. Their genomic separation is 1.5Mb (for *a*-*b*) and 2Mb (for *b*-*c*). We choose these probes since the Hi-C interaction score is much higher between *a* and *b* than *b* and *c*, so we can easily test if it reflects the closer spatial proximity at a single cell level. The results, summarized in Panel C, show that *a* and *b* are significantly closer (average distance 350nm) than the *b*-*c* pair (average distance 587nm). So we can conclude that strong Hi-C interaction between distant TADs translate in a spatial proximity, supporting once more the view that they are organized in a higher-order structure.

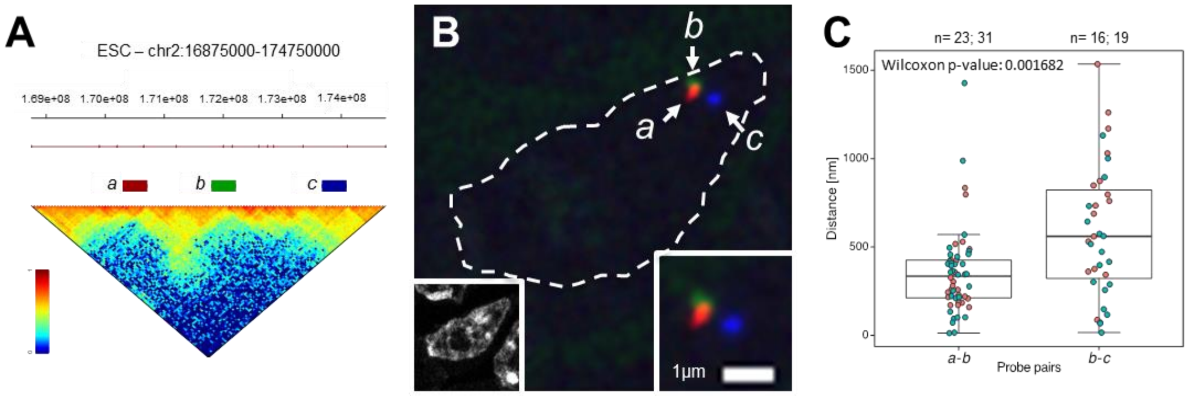

Figure 2.10: Long-range contact are confirmed by cryoFISH

Panel A: To confirm the extent long-range interactions between TADs at the single cell level, cryoFISH experiment was performed with three probes, which cover genomic regions *a* (red), *b* (green) and *c* (blue), separated by 1.5 (*a*-*b*) or 2 (*b*-*c*) Mb, and which belong to different TADs (red line). Hi-C data shows stronger long-range interaction frequency between *a* and *b* than *b* and *c*. **Panel B:** Probes *a*, *b* and *c* in an image on a confocal microscope. Inset on the right shows a magnified region of the same image. Dashed line indicates the nuclear outline. Region *b* often co-localizes with region *a*, but not region *c*. **Panel C:** Distances between regions *a* and *b*, or *b* and *c* were measured and data was collected from two independent cryoFISH experiments (red and green dots); number of distances in each replicate are indicated above the graph. The average distance between the center of the fluorescent signals corresponding to probes *a* and *b* is 350, in contrast with probes *b* and *c*, which are on average separated by 587nm. The ratio of the genomic separation for the two pairs of probes (*b*-*c*/*a*-*b*) is ~1.3, compared with their physical distance ratio of 1.7.

2.4 Correlation of hierarchical organization with epigenetic features

The distances on tree topology

To investigate if the hierarchical structure of metaTAD that we detect has a functional biological role, we study the relationship between the tree organization and several epigenetic features, using our CAGE data and other published datasets. First, we have to define a distance $d(i,j)$ between two loci i and j of the genome that takes into account the hierarchical organization. Since the leaves of our tree are TADs, i and j are TAD indexes. We consider two possible distances: the total height of the smallest branch including i and j, and the smallest number of edges along the tree that connect TAD i with TAD j. Since both distance definitions give similar results, we use the second in the following analysis.

Correlation with the tree topology

Once we have the tree distances between all the possible TAD pairs, we collect the subset of pairs U_d having the same distance d. Then, we compute the Pearson correlation (more precisely, it is the autocorrelation) coefficient $C(d)$ over U_d for a certain biological feature:

$$C(d) = \sum_i (s_{1\mathrm{i}} - \mathrm{E}[s_1])(s_{2\mathrm{i}} - \mathrm{E}[s_2])/(\mathrm{Var}[s_1]\mathrm{Var}[s_2])^{1/2} \quad (3)$$

where the sum is over the N_d TAD pairs in the set U_d, s_{1i} and s_{2i} are the epigenetic signals of the first and second TAD in the pair number i, E[s_1] and E[s_2] are the mean of s_{1i} and s_{2i}, and finally Var[s_1] and Var[s_2] are the variance of s_{1i} and s_{2i}. So, we get the correlation coefficients for a fixed tree distance d.

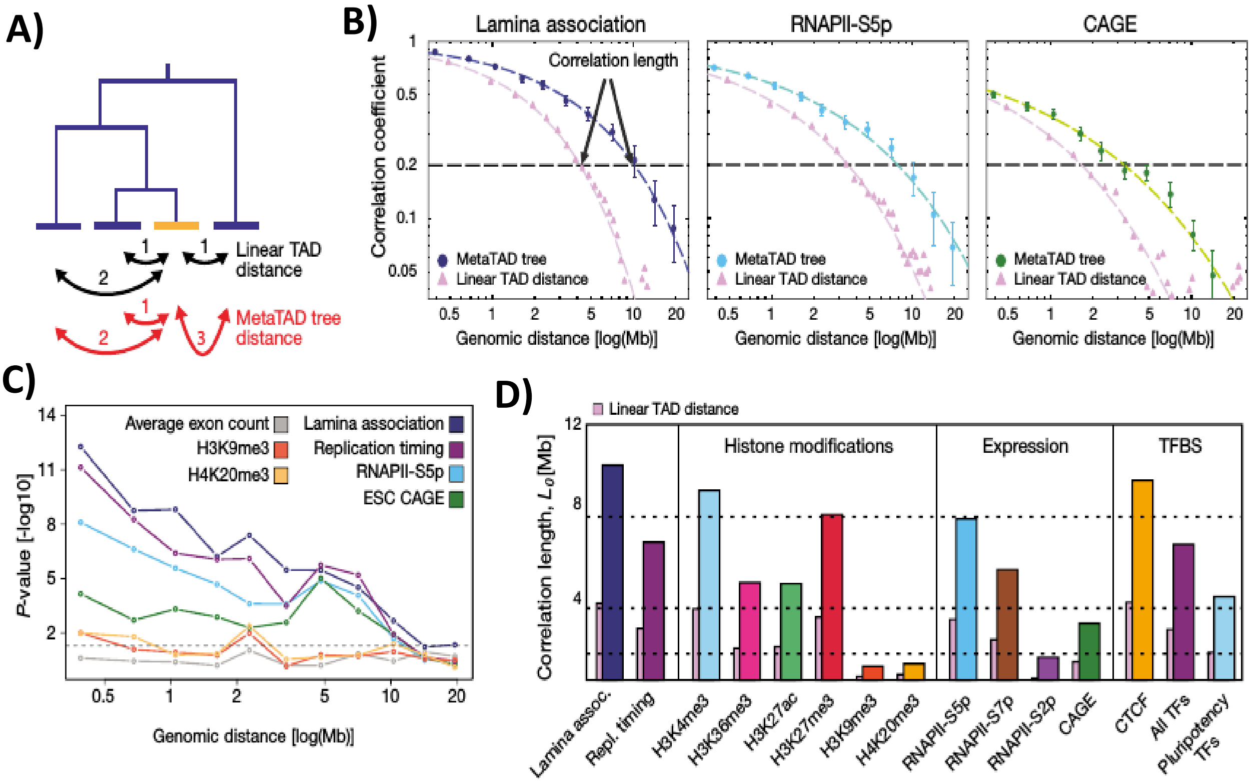

Figure 2.11: **metaTAD tree organization correlates with genomic, epigenomic and expression features**

Panel A: the diagram represents the difference between linear and metaTAD tree distance (number of edges along the tree minus one) for a given TAD (yellow) relative to other TADs (blue) in the same tree. **Panel B**: correlations over the tree extend up to genomic scales of tens of Mb (filled circles) and are significantly stronger than those observed in linear genomic sequence (filled triangles). The horizontal dashed line indicates a 20% correlation coefficient. **Panel C**: statistically significant differences are observed in the correlations measured across the metaTAD tree and across random neighbour trees constructed from the same linear array of TADs (horizontal line P-value = 0.05). Heterochromatin marks H4K20me3 and H3K9me3 levels do not correlate with the tree structure above what is expected from linear genomic distance. **Panel D**: CAGE data, different epigenomic features and pluripotency transcription factors binding sites (TFBS) have different average correlation lengths.

Relation between tree structure and linear genomic distance: the correlation length L_0

In order to compare the correlation curve $C(d)$ along the tree with the equivalent correlation curve that we obtain simply considering the linear genomic distance between two generic

TADs (i.e. without considering the hierarchical structure), we need to convert the tree distance d into genomic length $s(d)$. This is achieved by the following relation:

$$s(d) = \sum_i (c_{2i} - c_{1i}) / N_d \tag{4}$$

where c_{1i} and c_{2i} are the genomic coordinates in the middle of the TAD in pair i contained in the set U_d (see Figure 2.14, genomic distance s as a function of the tree distance d). In this way, we are able to convert the function $C(d)$ in the function $C(s)$. Generally, we find that this function decays as a stretched exponential law, shown in the plots as a dashed line:

$$C(s) \sim \exp - (s / s_0)^{\beta} \tag{5}$$

The parameters of the fit depends on the epigenetic feature considered. In Figure 2.11, Panel B, we report the correlation curves $C(s)$ for the some epigenetic signals in ESC. To give an estimate of the length scale involved in the correlation pattern, we also define a correlation length L_0 as the genomic distance where the stretched exponential equals 0.2 (represented as a horizontal dashed line in all the correlation plots). In Figure 2.11, Panel D, we show the barplot with all the correlation lengths for the epigenetic tracks analyzed (ESC cell line), and compare them with equivalent quantities calculated on the curves . In Figure 2.12, we report the correlation curves for all the epigenetic tracks studied in the three time points ESC, NPC and Neurons.

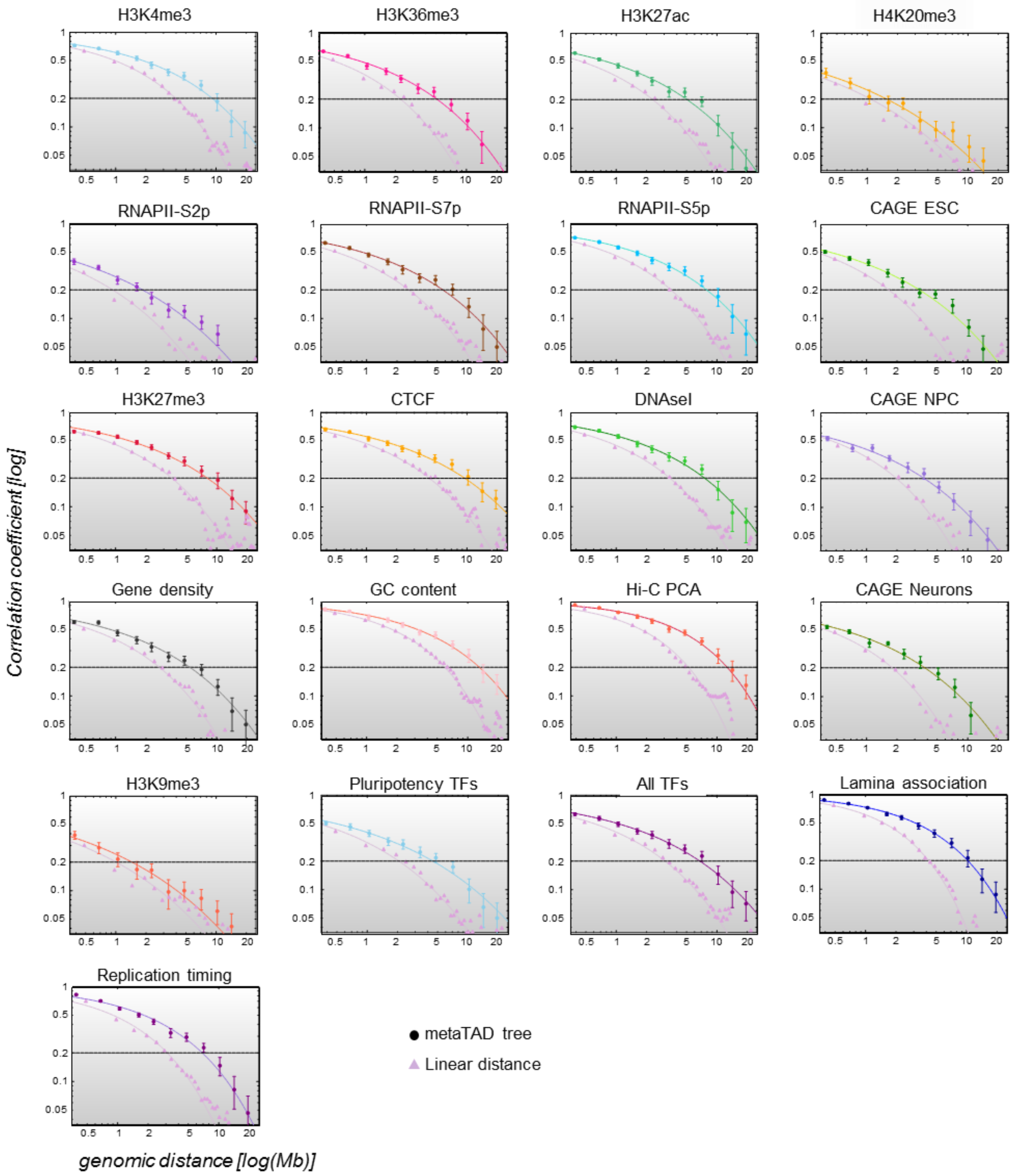

Figure 2.12: Correlation of genomic, epigenomic and TF features along metaTAD trees is much stronger than along linear genomic distances.

Pearson correlations of epigenetic and CAGE data, and transcription factor (TF) binding sites over the metaTAD hierarchy (filled circles, upper lines; transformed to average genomic distance; both axes are logarithmic) is much larger than the same correlation measured over the linear genomic distance of TADs (filled triangles, lower lines). Superimposed lines are stretched exponential fits.

Statistical significance of the tree correlations coefficients

To quantify the significance of the correlations that we find between the metaTAD tree and the epigenetic features, we repeat the procedure above described, for random neighbor trees. Such trees are built according to this procedure: we start from the real linear sequence of elementary TADs, then we randomly join TAD pairs without considering their interaction. So, we obtain a tree where each TAD has the 50% probability to be joined with his left neighbour or his right neighbour. Note that the random trees have a high degree of similarity with the original tree when we consider the TAD distance distribution. Once we produce the random trees, we proceed with the calculation of the correlation curve, and , as before, we convert the tree distances into genomic distances in order to compare with the real correlation curves. The statistical significance is given by the p-values (Figure 2.13, one-sided Wilcoxon rank sum test) between the real and the random neighbour trees.

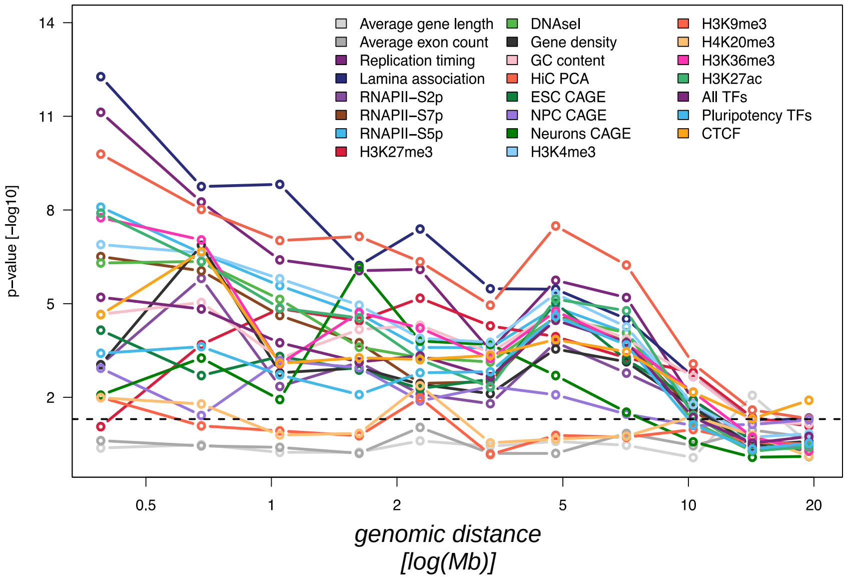

Figure 2.13: Correlation of features along the metaTAD hierarchy in comparison with random neighbor trees

To measure significance, we computed one-sided Wilcoxon tests of difference in median between the real and random neighbor tree correlations and found the former to be significantly stronger up to large distances, approximately 10 Mb, for a number of features;

horizontal line, p-value=0.05. Average gene length and exon count per TAD were used as controls.

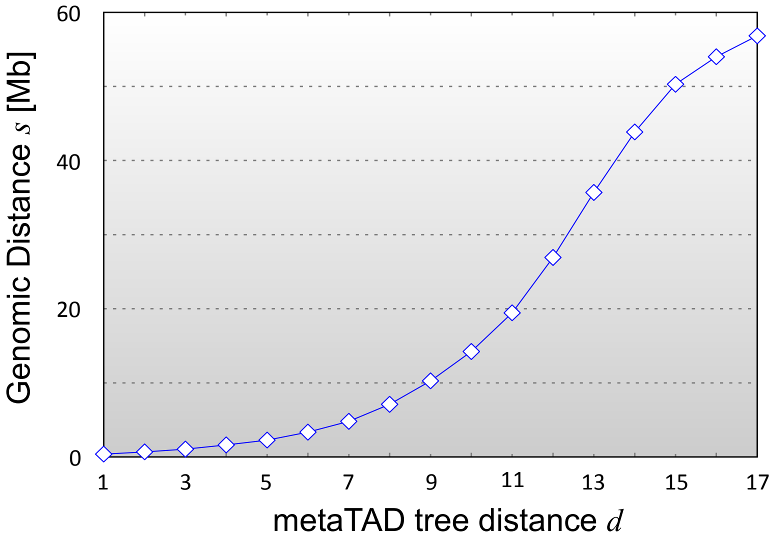

Figure 2.14: Relation between the genomic distance and the metaTAD distance

The plot shows the average genomic distance corresponding to a given metaTAD tree distance over the tree. Data are averaged over the three cell types of this study.

Analysis of epigenetic features association with TAD and metaTAD boundaries

The boundaries of TADs are enriched for specific genomic features (Dixon *et al.*, 2012; Nora *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Moore *et al.*, 2015). The hierarchy of metaTADs identifies different types of domain boundary, comprising boundaries that connect two TADs at the lowest metaTAD levels up to boundaries that separate higher-order metaTADs containing large blocks of TADs. We measured the enrichment of chromatin features across TAD and metaTAD boundaries genome-wide, such as RNAPII and CTCF occupancy, and promoter activity measured by CAGE in ESC, NPC and Neurons. Interestingly, we found that features previously observed as significantly enriched at TAD boundaries are even more strongly enriched at higher-order metaTAD boundaries (corresponding to genomic lengths of 10–40 Mb, Figure 2.15), consistent with important functional roles of the metaTAD organization. In this analysis, we focus only on metaTADs with size between 10Mb and

40Mb. We consider a genomic region centered in the TAD or metaTAD boundary and extended 450Kb in each direction.

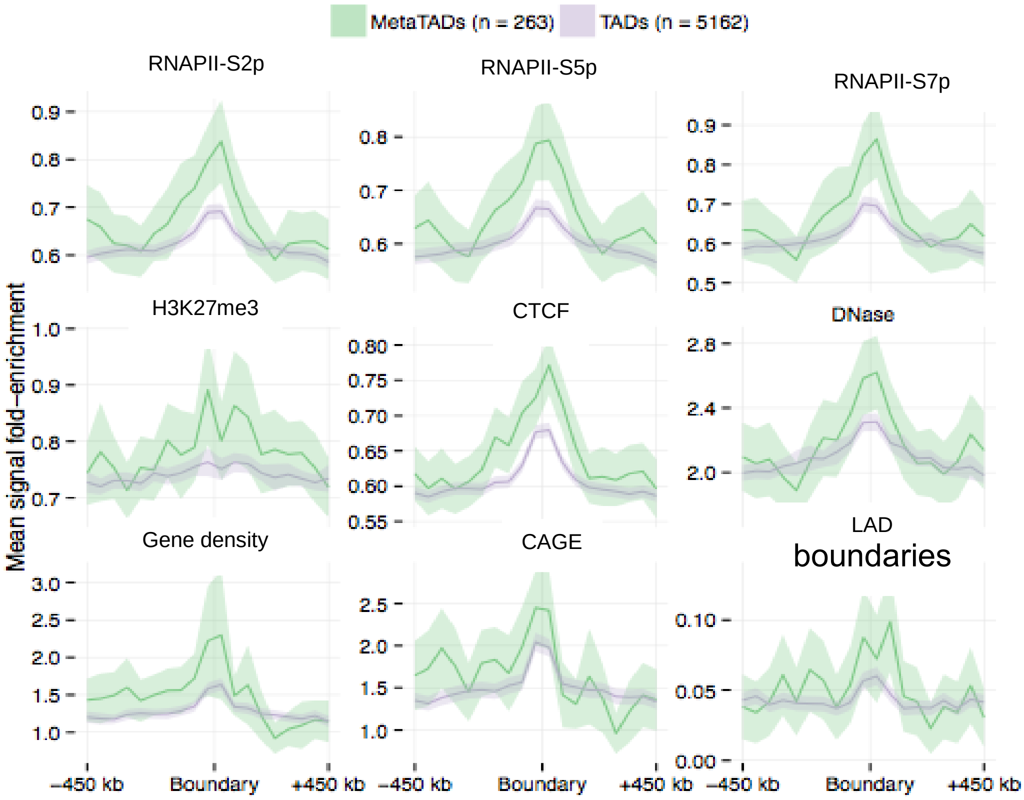

Figure 2.15: Comparison of enriched features at TAD and metaTAD boundaries.

Genome-wide profiles of epigenomic features and gene densities averaged over all TAD and over only large metaTAD (10 – 40 Mb) boundaries (ribbons show 95% confidence intervals of the mean). The enrichment of most features is significantly increased at metaTAD boundaries when compared to TAD boundaries. ‘CAGE’ represents CAGE-defined active TSS.

2.5 The evolution of the tree during cell differentiation

In this section we analyze the structural rearrangements in the metaTAD organization occurring during the differentiation process, and we try to understand whether a relationship exists between cell functionality and topological spatial reorganization.

The global similarity measure

The structural reorganization of the metaTAD hierarchy during the differentiation is evident if we compare the trees of single chromosomes in the different time points. The tree topologies are compared by using measures of structural changes. Precisely, we use the cophenetic correlation (Sokal & Rolhf, 1962), a global and general measure of tree similarity. This is obtained by calculating the Pearson correlation coefficient between the cophenetic matrices associated to each chromosome, for two specific time points. The cophenetic matrix consists simply of the distances between all the leaves of the tree. For simplicity, we take all the tree length branches equal to 1. In Figure 2.16, Panel A, it is shown the comparisons for the cophenetic coefficients, and it emerges that Hi-C based trees are more similar among each other than to random trees. We consider two types of random control trees: the first type is the random neighbour tree, already described in the previous section, and the second type is the totally random tree, without any constraint in the joining procedure (i.e. the nearest neighbours constraint). The coefficient between two sets of total random trees is practically zero, while it is 0.49 between two sets of random neighbour trees. In the latter case this is expected since at each level the probability to be joint with the left or right nearest neighbour is ½, and the resulting overall conservation level is about 50%. Analogous results are obtained if other similarity measures (Robinson-Foulds distance measure) are used, but we will not discuss here the details.

Results

The comparison between Hi-C based trees for the three time points is implemented in the following way. First, we consider only elementary TADs having conserved boundaries, then we calculate the cophenetic matrices and coefficients as described in previous subsection. We consider only the conserved TADs since in this way we use the same number of domains (leaves) and we compare the same genomic regions across different time points. TADs are defined conserved if the left and right boundaries coincide within a given tolerance (here we set 200Kb) in the cell lines being compared. As shown in Figure 2.16, Panel A, we obtain that the cophenetic coefficient is around 81-84%, which is well above the random level (see previous subsection). In Figure 2.16, Panel B, the coefficient for all the chromosomes in the three transitions is reported, and it is evident that the degree of reorganization is highly dependent on the time point and on the chromosome considered. So, for instance, we find that chromosomes 4, 6 and 19 have a low similarity between ESC-NPC but are highly conserved between NPC-Neurons. Overall, we can summarize that the metaTAD structure has a degree of structural reorganization (about 20%) against a general background of conservation.

The local similarity measure

In order to quantify the degree of local structural change of a conserved elementary TAD during differentiation we developed a measure reflecting the level of reorganization at a particular transition. Precisely, for a fixed conserved TAD, we consider the other nearest conserved TADs according to the tree distance in the considered time point. The nearest conserved TADs are those having a distance less than the conserved TAD with the third shortest distance. If we identify n neighbour conserved TADs, we evaluate the tree change using the following quantity:

$$\text{Degree of change } (i) = \sqrt{\sum_j^n \frac{(x_{A,j} - x_{B,j})^2}{n}} \qquad (6)$$

where $x_{A,j}$ and $x_{B,j}$ are the distances along the tree between the conserved TAD i and conserved TAD j in the time point A and B. A conserved TAD is in a region of tree conservation if the associated z-score is > 0, and analogously it is in a region of tree changes if the z-score is < 0. In Figure 2.17, Panel A, it is shown an example of tree topology change for chromosome 6, in the transition ESC-NPC, where the regions conserved are highlighted in green and the region not conserved are highlighted in red, with an intensity depending on the level of change degree. Overall, the most local tree changes occur in the transition ESC-NPC than in the transition NPC-Neurons, as showed in Figure 2.17, Panel B.

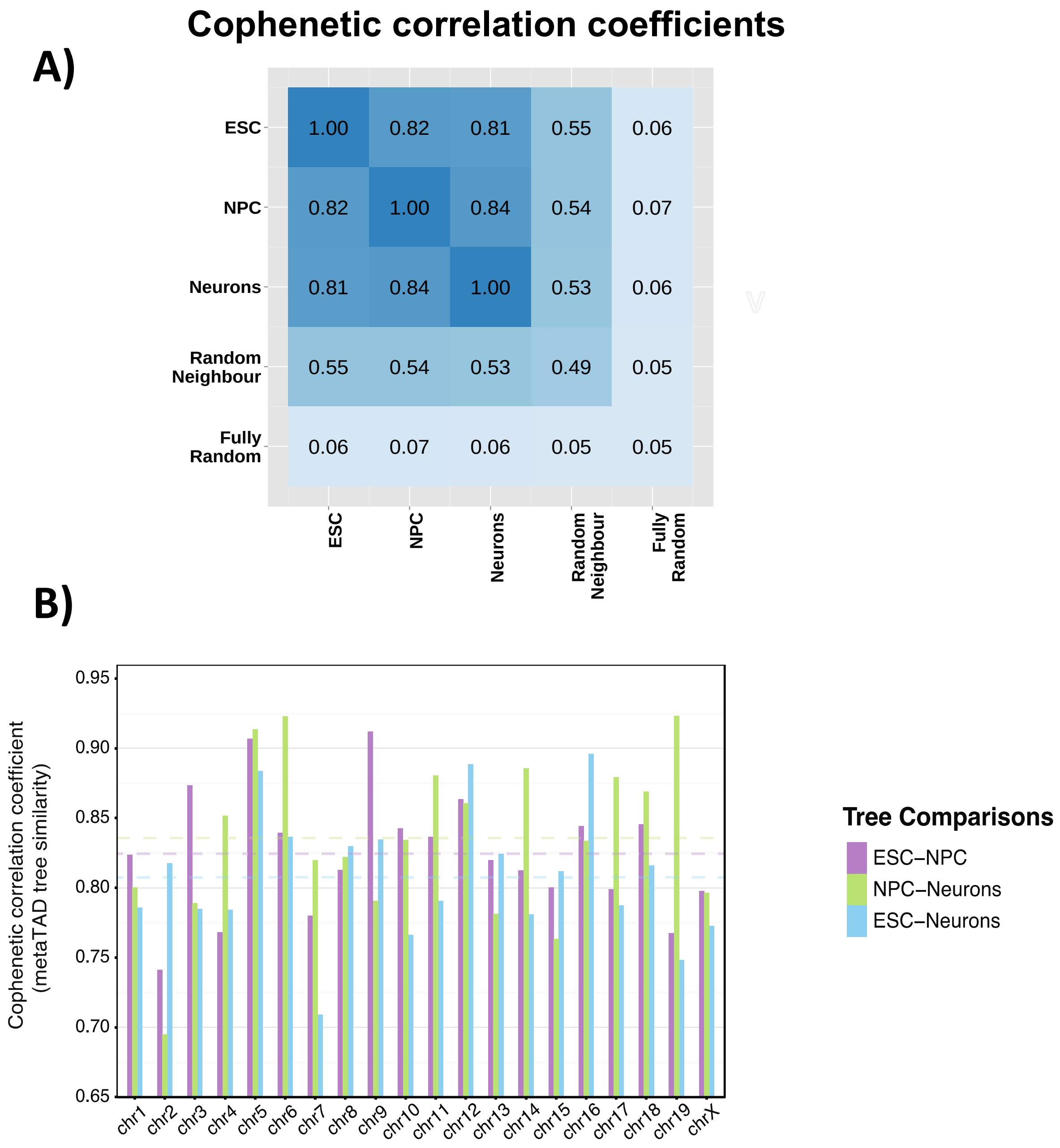

Figure 2.16: Reorganization of metaTAD trees across differentiation (global measure)

Panel A: Cophenetic correlation between conserved metaTAD trees. Comparisons with two different sets of 100 random tree models are also shown: the random neighbour tree and the fully random tree models. Shown are genome-wide averages over all chromosomes. MetaTAD trees are more similar to each other than to random trees. On average, the topology of the NPC metaTAD trees is close to that of Neurons, both of which are approximately different to the ESC metaTAD trees. **Panel B**: Cophenetic correlation coefficients comparing metaTAD trees per chromosome show different levels of tree restructuring in each differentiation time-point transition. Dashed horizontal lines represent the average value for all autosomal chromosomes

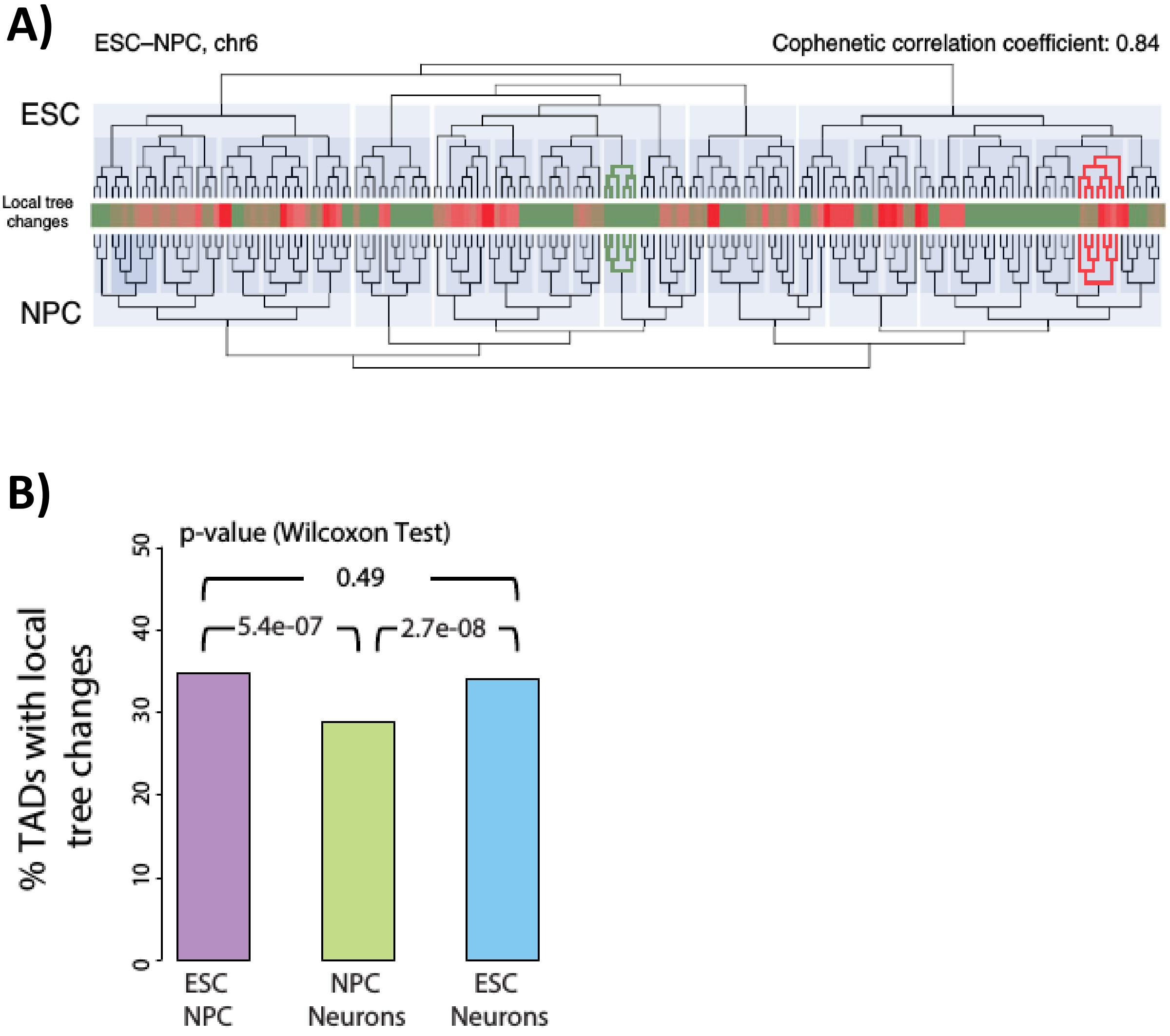

Figure 2.17: Reorganization of metaTAD trees across differentiation (local measure)

Panel A: Tree topology of chromosome 6. The central heatmap reflects the degree of reorganization during the transition ESC-NPC. In green are highlighted the conserved region, in red the not conserved regions. **Panel B:** Percentage of TADs undergoing a local tree change (threshold = 1). In the transition ESC-NPC there are the most tree changes.

2.6 A polymer physics model for hierarchical organization

In this last section we will present a simple, possible mechanism that regulates the hierarchical folding of the genome and the formation of metaTADs using a polymer physics model. The model used is the strings and binders switch (SBS) model (Barbieri *et al.*, 2012), already introduced in Chapter 1. In the next chapter, we will introduce another possible mechanism, based on the same model, for hierarchical folding.

The red-green-blue model

To model the formation of a higher order domain, we consider a polymer with three types of binding sites, which can be visualized with three different colors (red, green and blue). The red and the green sites are positioned along the first and the second half of the polymer respectively (Figure 2.18, Panel A), while the blue sites are interspersed with them (Figure 2.18, Panel B). Floating binders interact with the binding sites and the polymer folds (as for the binding sites, three types of binder exist in system, each interacting with its cognate type). When equilibrium is reached, the blue binding sites induce the higher-order interaction between the two distinct red and green domains (which can be seen as elementary domains or TADs). From the biological point of view, the blue binding sites are expected to contribute to the correlations between epigenetic features observed at metaTAD scales.

Packaging efficiency

To measure the effect on the polymer packing resulting from the mechanisms just proposed, the volume of the whole system is compared with the volume of its subparts (red, green and blue separately). In particular, the interaction mediated by the blue binders reduces the distance between the red and green domain and the polymer volume decreases, resulting in an increase of the packaging efficiency of 50% (Figure 2.18, Panel C). Of course, other mechanisms can be considered to promote higher-order domain folding, as we will show in the next chapter.

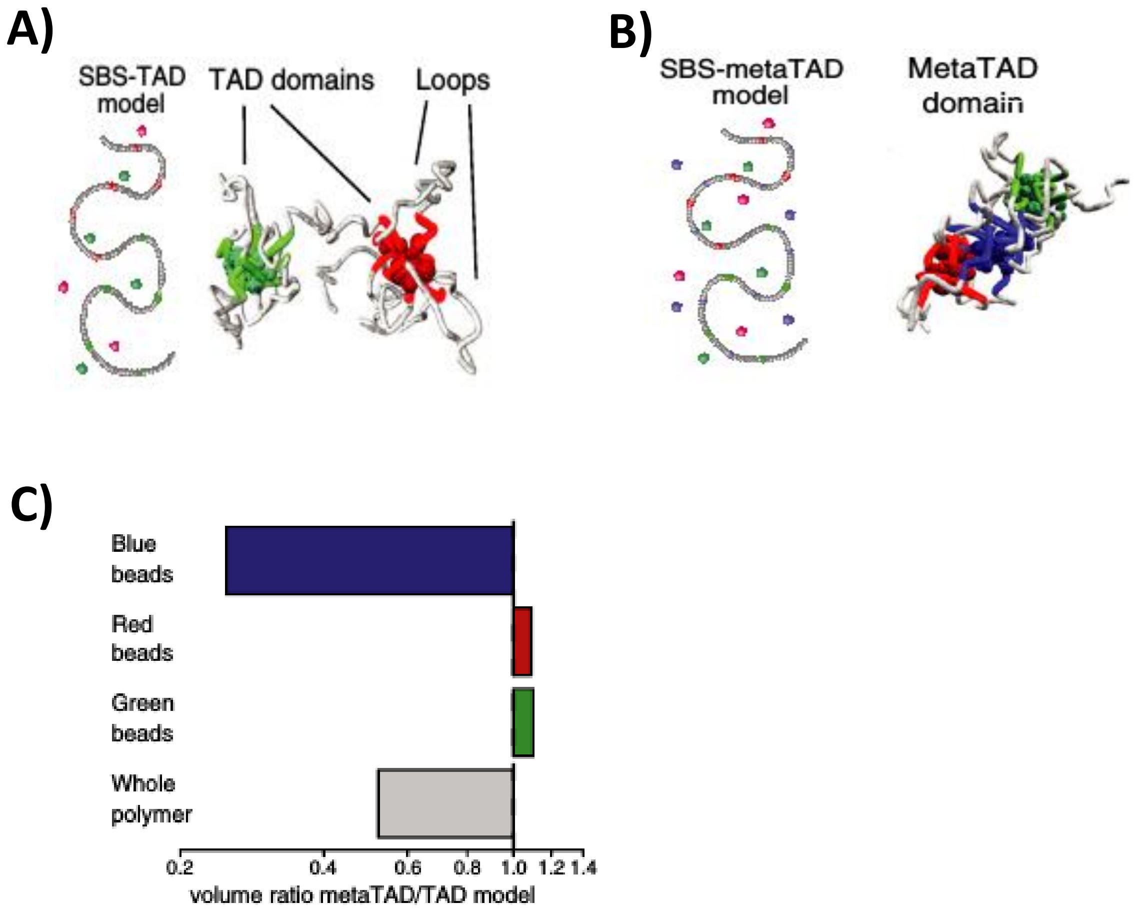

Figure 2.18: Mechanisms of higher-order domain formation

Panel A: The red and green binding sites interact with their binders and form two distinct domains. **Panel B:** The blue binding sites mediate the interaction between the red and green domains and form a higher-order domain. **Panel C:** The packaging efficiency is increased of 50% when compared to the non-hierarchical model (metaTAD/TAD model).

References

Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, Nicodemi M (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A* **109:** 16173-16178

Bickmore W, van Steensel B (2013) Genome architecture: domain organization of interphase chromosomes. *Cell* **152:** 1270-1284

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

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

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106-1117

Crutchley JL, Wang XQ, Ferraiuolo MA, Dostie J (2010) Chromatin conformation signatures: ideal human disease biomarkers? *Biomarkers in Medicine* **4:** 611-629

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

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

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

Ferrari KJ, Scelfo A, Jammula S, Cuomo A, Barozzi I, Stutzer A, Fischle W, Bonaldi T, Pasini D (2014) Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Molecular Cell* **53:** 49-62

Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jorgensen M, Dimont E, Arner E, Schmidl C, Schaefer U *et al.* (2014) A promoter-level mammalian expression atlas. *Nature* **507:** 462-470

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* **6:** e245

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

Jaeger I, Arber C, Risner-Janiczek JR, Kuechler J, Pritzsche D, Chen IC, Naveenan T, Ungless MA, Li M (2011) Temporally controlled modulation of FGF/ERK signaling directs midbrain dopaminergic neural progenitor fate in mouse and human pluripotent stem cells. *Development* **138:** 4363-4374

Sokal RR, Rohlf FJ (1962) The Comparison of Dendrograms by Objective Methods. *Taxon* **11:** 33-40

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P (2006) CAGE: cap analysis of gene expression. *Nat Methods* **3:** 211-222

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

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

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553-560

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

Moore BL, Aitken S, Semple CA (2015) Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biol* **16:** 110

Nicodemi M, Pombo A (2014) Models of chromosome structure. *Curr Opin Cell Biol* **28C:** 90-95

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

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

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

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, Thurman RE, Cheng Y, Gulsoy G, Dennis JH, Snyder MP, Stamatoyannopoulos JA, Taylor J, Hardison RC, Kahveci T, Ren B *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* **515:** 402-405

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665-1680

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551-1555

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

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

Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7:** 542-561

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

Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Dostie J, Bickmore WA (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & Development* **28:** 2778-2791

Yu H, Gerstein M (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A* **103:** 14724-14731

Chapter 3

Understanding the mechanisms of chromosomes folding: polymer physics models

In this chapter, we will show how polymer physics can explain some aspects of the spatial structure of chromosomes in the cell nucleus. First, we will briefly discuss the Strings and Binders Switch Model (SBS, Section 3.1), originally presented in the work by Barbieri *et al.* (2012), then we will discuss its Molecular Dynamics (MD) implementation and the resulting phase diagram, which shows a novel thermodynamic stable state (Section 3.2). We will show how, with few parameters, we are able to recapitulate the average behavior of the contact probability in a very large range of genomic lengths (Section 3.3). Also, we will describe the theoretical multiple contact profile, which recently have been discovered to be very important (Olivares-Chauvet *et al.*, 2016, Beagrie *et al.*, 2017) for genome architecture and regulation (Section 3.4). The results presented in the present chapter have been published in Chiariello *et al.* 2016, and Annunziatella *et al.*, 2016.

3.1 The Strings and Binders Switch (SBS) model in MD

The model

In the SBS model (Nicodemi&Prisco, 2009, Barbieri *et al.*, 2012), a chromatin filament is modeled at a coarse-grained level as a classical Self-Avoiding-Walk (SAW) of beads, that can interact with binders floating in the surrounding environment. The beads interact with the diffusing molecular binders though an attractive potential with interaction energy E_{int}. The binders have a concentration c, and can bridge the beads of the chain and fold spontaneously the polymer (Figure 3.1).

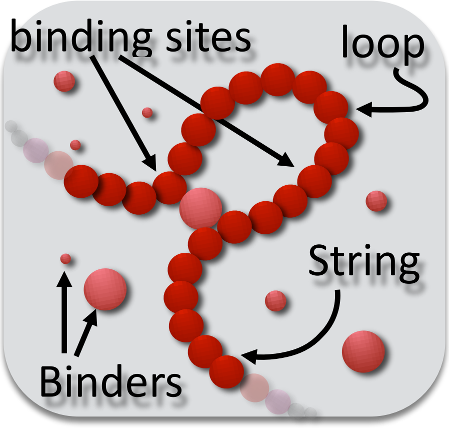

Figure 3.1: The SBS model

The Strings&Binders (SBS) model is a Self-Avoiding (SAW) chain of beads interacting with molecular binders having a concentration, *c*, and a binding affinity E_{int}.

The MD implementation: Dynamics

In our simulation, we use the open source software LAMMPS (Large Atomic Molecular Massive Parallel Simulator (Plimpton, 1995)). The chromatin filament is represented as a polymer composed by N beads. The binders are single particles that move randomly in the environment. Both beads and binders are subject to Brownian motion, which is mathematically described by the well-known Langevin equation (Kremer&Grest, 1990):

$$m\frac{d\vec{v}\,(t)}{dt} = -\zeta\vec{v}\,(t) + \vec{f}(t) - \nabla \mathrm{V} \qquad (2)$$

where *m* is the mass of the generic particle, *v*(*t*) the particle velocity, *V* is the potential acting on the particles (see next subsection) and *f*(*t*) stochastic random force that takes into account the thermic fluctuation of the environment. The friction coefficient ζ is related to the viscosity of the solvent η from the Stokes relation ζ=3πησ. As usual in MD simulations, we work in dimensionless units (Kremer&Grest, 1990). So, we set the diameter of the polymer bead σ equal to 1 (for simplicity, we do the same for the binder diameter). The diameter fixes our length unit. Analogously, we set the mass of the particle *m* equal to 1. The energy scales are measured in k_BT, where the Boltzmann constant k_B is 1 and the temperature T is 1. For the dynamics, we set ζ=0.5 (Kremer&Grest, 1990; Rosa&Everaers, 2008; Brackley *et al.*, 2013). The Langevin equation is integrated using the Verlet algorithm (Plimpton, 1995). All these settings are standard choices and are well described in Kremer&Grest, 1990. The simulation box, having boundary periodic conditions, has a linear size D, that is as large as the gyration radius of a SAW with the same number of beads (D $\propto$ $N^{0.588}$) . Physical units will be obtained once we fix the length scale and other parameters of the system (see next section).

The MD implementation: Potentials

The SBS model in MD is implemented through the potentials definition. The potential energy V(x), of a particle having a position x, has three components. Between two consecutive beads of the chain there is a potential that models a finitely extensible non-linear elastic (FENE, Kremer&Grest, 1990) spring:

$$(3) \qquad V_{FENE} = -0.5KR_0^2 \ln\left[1-\left(\frac{r}{R_0}\right)^2\right]$$

where R_0 is the maximum extension of the spring (otherwise the argument of the logarithm becomes negative), K is the strength of the spring. We set R_0=1.6sigma and K=30k_BT/σ^2 (Kremer&Grest, 1990; Brackley *et al.*, 2013). To take into account for excluded volume effect between any two particles, there is an hard-core repulsive force $V_{hard}\ (r)$, modeled by a shifted Lennard-Jones (LJ) potential:

$$(8) \qquad V_{hard}\ (r) = \begin{cases} 4\left[\left(\frac{\sigma_{\mathrm{b-b}}}{r}\right)^{12} - \left(\frac{\sigma_{\mathrm{b-b}}}{r}\right)^{6} - \left(\frac{\sigma_{\mathrm{b-b}}}{1.12}\right)^{12} + \left(\frac{\sigma_{\mathrm{b-b}}}{1.12}\right)^{6}\right] & r < 1.12 \\ 0 & otherwise \end{cases}$$

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

$$(9) \qquad V_{int}\,(r) = \begin{cases} 4\,\epsilon_{int}\left[\left(\frac{\sigma_{\mathrm{b-b}}}{r}\right)^{12} - \left(\frac{\sigma_{\mathrm{b-b}}}{r}\right)^{6} - \left(\frac{\sigma_{\mathrm{b-b}}}{r_{\;int}}\right)^{12} + \left(\frac{\sigma_{\mathrm{b-b}}}{r_{\;int}}\right)^{6}\right] & r < r_{\;int} \\ 0 & otherwise \end{cases}$$

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

$$(10) \qquad E_{int} = \left|4\,\epsilon_{int}\left[\left(\frac{\sigma_{\mathrm{b-b}}}{r_{int}}\right)^{6} - \left(\frac{\sigma_{\mathrm{b-b}}}{r_{int}}\right)^{12} - \frac{1}{4}\right]\right|$$

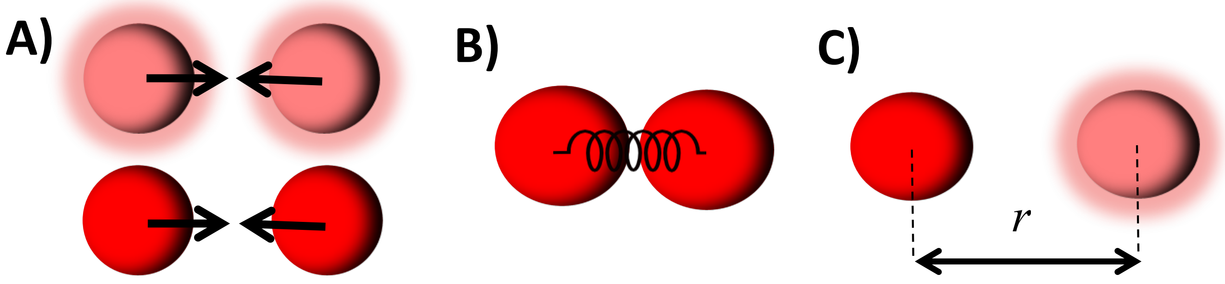

Figure 3.2: Schematic representation of the three potentials

Panel A: Between any two particles, there is a purely repulsive LJ potential, defined in equation (8), necessary to model the excluded volume effects. **Panel B:** Between two consecutive polymer beads, the bond is a finite extensible non-linear elastic (FENE) potential, defined in equation (7). **Panel C:** Between bead and binders the interaction is modeled with an attractive, shifted Lennard-Jones potential, defined in equation (9).

Initial states

As in all MD simulations, the initial state is very important (Rosa&Everears, 2008). In all our simulations the polymer is initially prepared in a random SAW configuration, while the binders are randomly located in the simulation box. To produce an initial random SAW configuration we use the following standard approach (Kremer&Grest, 1990): we generate a random walk chain, with a customary script written in Python, where the distance between two consecutive beads is equal to the average length of an equilibrium SAW chain under the FENE potential above described (i.e., 0.97σ). Then, to remove overlaps between beads and binders, we let the system equilibrate, for some timesteps, where the hard-core LJ repulsion is replaced by a soft potential $V_{soft}(r)$ (Kremer&Grest, 1990; Brackley *et al.*, 2013):

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

where the factor *A* increases linearly in time. The scaling properties of the polymer are then measured to check that the stationary SAW state is attained. Finally, the chain is simulated under the FENE potential. and its scaling properties checked again. MD techniques are very powerful and standard methods to investigate molecular structures, and are broadly used to study the folding processes and conformational properties of other completely different, yet very important, molecules like proteins (Di Carlo, Minicozzi *et al.*, 2015).

3.2 The Phase Diagram

In this section, we will apply the model just discussed in the previous section to a generic region of the genome (i.e. a chromosome) and will extract quantitative information about its structure. Given the genomic length L of the region to be modeled, the corresponding genomic content (i.e. the number of bases) per bead is s_0=L/N, where N is the number of beads forming the chain. Here, we consider polymer made of N=1000 beads. Since a typical chromosome have a genomic length of approximately L=100Mb, each bead contains s_0=L/N=100Kb. To model the average chromosome behavior, we use the simplest SBS polymer, where each bead can interact equally with all the binders in the environment. In the following chapter, we will increase the complexity of the model by introducing a specificity ('colors') in the interaction between beads and binders.

Thermodynamic conformational classes: the coil-globule transition

In this model, the control parameters are the interaction energy E_{int} between bead and binders and the binders concentration *c*. As known from polymer physics, there is a coil-globule folding transition, highlighted by a sharp drop of the gyration radius (that is the order parameter of this transition) when crossing the theta point in the phase diagram. The coil state is characterized by small values of E_{int} and *c*, i.e. when the binders do not succeed in forming stable loops, and the polymer remains open (as in a SAW, Figure 3.5, Panel B, light blue box). On the contrary, in the globular state the polymer is in a closed configuration, occupying a very small fraction of the open state volume (Figure 3.5, Panel B, red box).

Thermodynamic conformational classes: the order-disorder transition

We identify also a new phase transition, occurring in the polymer globular phase, where the binders undergo an order-disorder transition, despite that they do not interact directly with each other. Two states exist: at low energies or concentrations, the binders form a disordered aggregate attached to the chain, while at high energies, with a sufficiently high concentration, they form an ordered aggregate. The phase diagram is summarized in Figure 3.3. Such thermodynamic stable states are expected to play an important role in the chromatin organization. The different nature of these configurations is visually evident in Figure 3.5, Panel D, and will be discussed in detail in the next subsection.

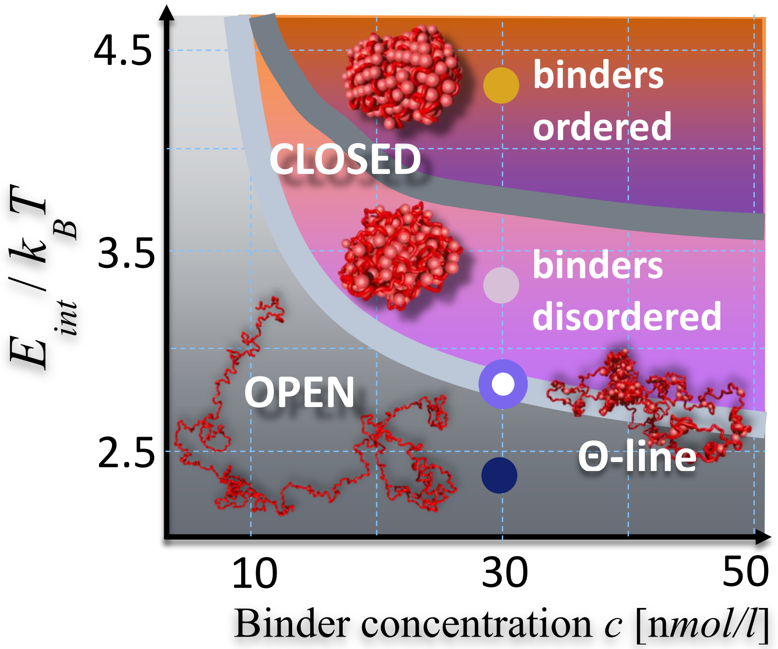

Figure 3.3: The Phase Diagram

Polymer physics dictates that the model stable architectural classes correspond to the different phases of its phase diagram: the polymer is open and randomly folded or, above its Θ-point transition, closed in more compact conformations; in the closed state, its binders can have a transition from a disordered to an ordered arrangement. Conformational changes can be sharply controlled (switch-like) by phase transitions driven by increasing c above threshold, e.g., by up up-regulation of the binder genes, or by chemical modifications of their binding sites, acting on E_{Int}. The values reported in the axes are those used in the simulations (see “Mapping MD units into physical units” subsection).

Order parameters of the transitions

To identify the values of and c which give the transitions discussed above, we proceed as follows. The coil-globule transition is identified (Barbieri *et al.*, 2012) by studying the gyration radius R_g associated to the polymer, defined by the following relation:

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

where m_i and r_i are the mass and the position of the i-esim bead respectively, M is the total mass of the polymer and r_{CM} is the position of the center of mass of the polymer. Essentially, this is a measure of the average linear size of polymer. The order-disorder transition is highlighted by two quantities associated to the configuration of the binders: the pair distribution function $g(r)$ and the structure factor function $S(k)$. They are defined as follows (Allen&Tildesley, 1987):

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

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

where $\rho = N_b/V$ is the concentration of the binders attached to the polymer, δ is the Dirac delta function. The structure factor $S(k)$ is basically the Fourier transform of the pair distribution function. It is almost flat when the binders are in a disordered configuration, while it is characterized by sharp peaks when the binders are in a ordered configuration (Figure 3.4). In our study, we consider as order parameter the ratio $S(k^*)/S_{MAX}$, where k^* is the position of the second peak in the $S(k)$ function and S_{MAX} is a normalization constant equal to the maximum value of $S(k^*)$ among the different studied cases, so to have a quantity normalized between 0 and 1. Such order parameter have a sharp jump at the order-disorder transition (Figure 3.5). Analogous results are obtained if other peaks of $S(k)$ (for instance the first peak or the third peak) are taken.

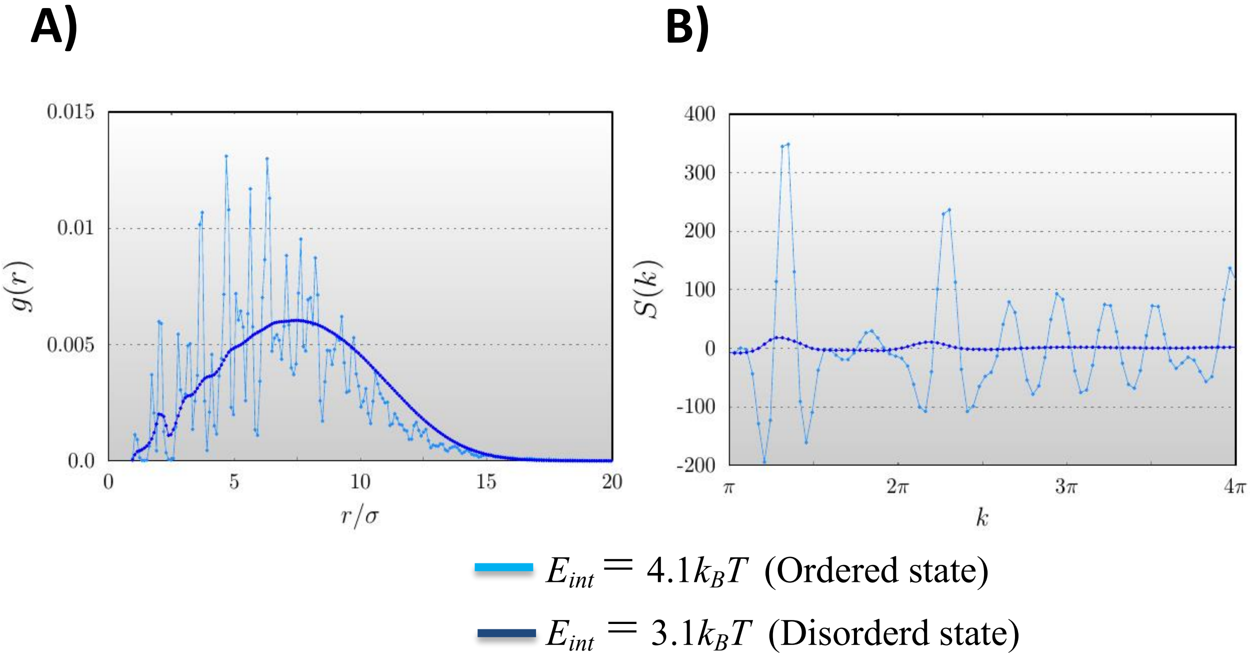

Figure 3.4: The pair function distribution *g*(*r*) and the structure function *S*(*k*)

Panel A: An example of pair distribution function $g(r)$ defined by equation (12), in the closed state. In the disordered state (blue curve, interaction energy E_{int}=3.1k_BT) it is characterized by a smooth behavior, while in the ordered state (light blue curve, interaction energy E_{int}=4.1k_BT) it has several sharp peaks. **Panel B**: The structure factor $S(k)$, i.e. the Fourier transform of the $g(r)$ function (eq. (13)), is practically flat for the disordered state, while it has sharp peaks in the ordered state.

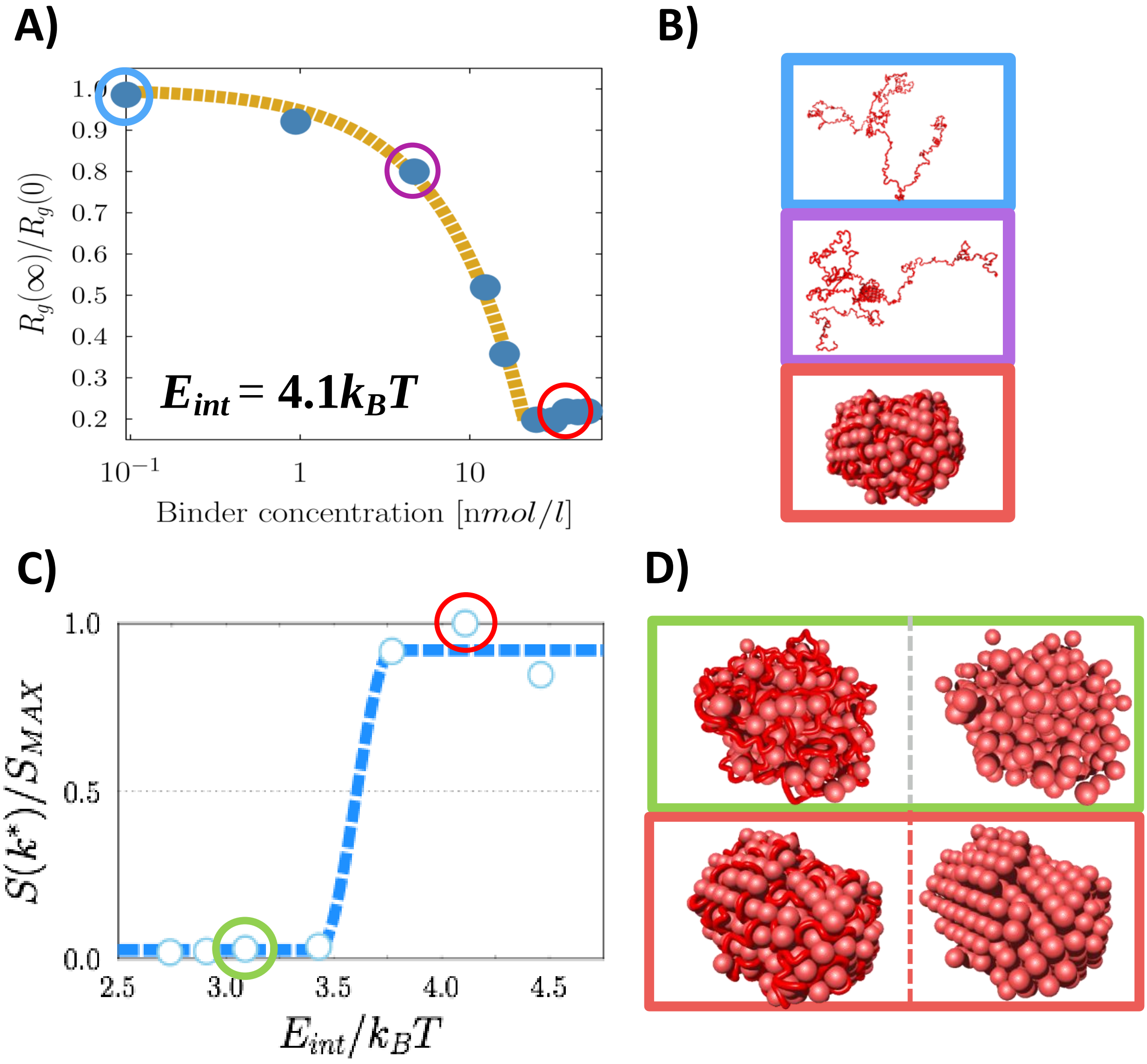

Figure 3.5: The order parameters of the transitions

Panel A: The gyration radius of the SBS polymer, R_g, signals its coil-globule transition point as a function of the concentration of binders. **Panel B**: Three different configurations at different concentration. **Panel C**: The structure factor $S(k)$ peak marks the order-disorder transition in the arrangement of the binders around the folded polymer. **Panel D:** The binders in disordered configuration (E_{int}=3.1k_BT, green box) and in an ordered configuration (E_{int}=4.1k_BT, red box).

Polymer folding dynamics

The stable conformational classes discussed in the previous subsections are equilibrium states. Starting from a completely SAW configuration, for each choice of the system parameters (i.e. concentrations *c* and interaction energy E_{int}), we consider the evolution of the polymer gyration radius and the total potential energy (FENE and LJ potentials) as a function of time (Figure 3.6, Panel A and Panel B). The values of the order parameters are calculated only for configurations taken from the last part of the dynamics, when the polymer is completely folded and no more (or in a negligible amount) binding events occur. The finer details of the dynamics depends on the particular values of *c* and E_{int}, as shown in Figure 3.6, Panel B, when comparing the green curve (disordered state) with the orange curve (ordered state).

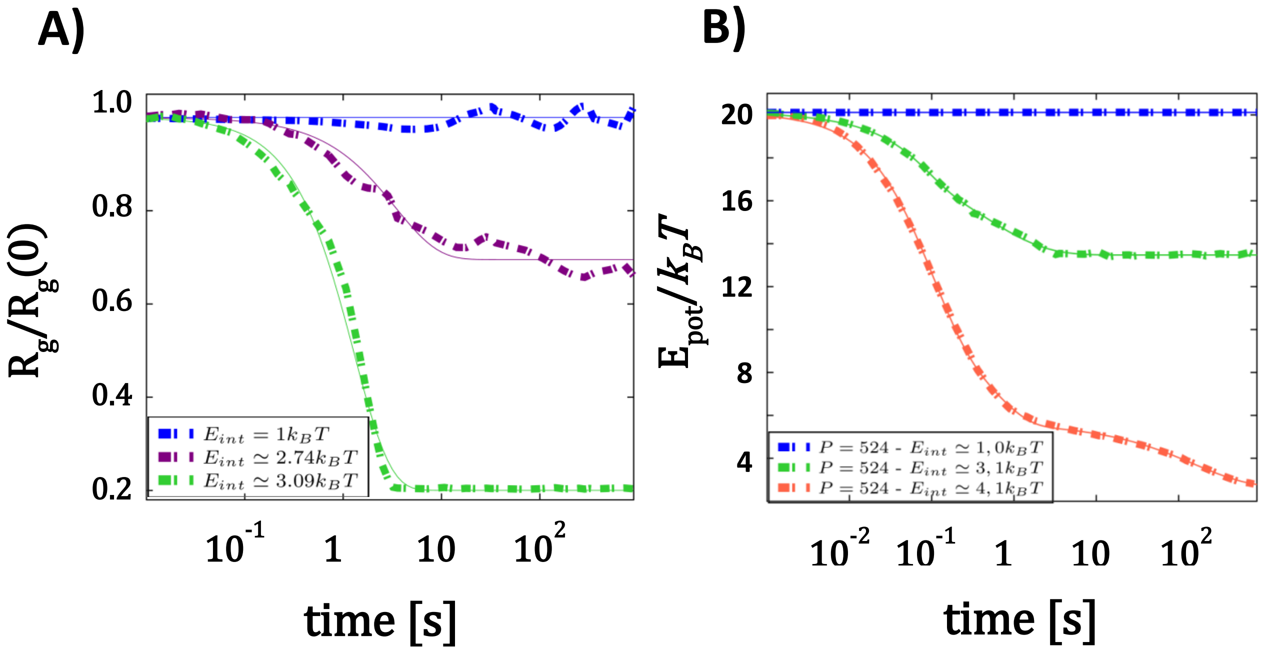

Figure 3.6: Polymer folding dynamics

Panel A: The system dynamics is monitored by the gyration radius (relative to its initial value) as function of time. When the energy is sufficiently high (E_{int}=3.09k_BT, disordered state, green curve), the polymer folds and the coil-globule transition occur after a transition time of 5÷10s. **Panel B:** Potential energy (FENE and LJ) as function of time. In the disordered state (green curve) the transition time is 5÷10s (as for the gyration radius). In the ordered state case (E_{int}=4.1k_BT, orange curve) there is the first transition after 5÷10s (coil-globule), then the second transition occur at ~100s (order-disorder transition). In all cases, the concentration *c* is above the transition threshold.

3.3 The mixture model fits large scale chromatin contact profile

To characterize the folding state of our polymer model, we calculated the pairwise contact probability $P_c(s)$ of beads separated by a given genomic distance s. Its behavior depends on the state of the system (Figure 3.7, Panel A). In the open state the probability decreases as a power law with s, i.e. $P_c(s) \sim s^{-\alpha}$, where the exponent α is about 2.1, as predicted by polymer physics (de Gennes, 1979). At the theta point, the exponent becomes 1.5, while in the globular state the probability has different shapes depending on whether the system is in the disordered or in the ordered state. In the former, it has an asymptotic plateau, with the power law exponent equal to 0, in the latter it decreases with an observed exponent 1.0. Analogously, the mean square distance between bead pairs $R^2(s)$ has a complementary behavior to the $P_c(s)$ function, as shown in Figure 3.7, Panel B, so it depends on the thermodynamic state of the system. The properties just discussed are general features of this kind of systems. The finer details of the polymer configurations depend anyway on other aspects, like the position of the binding sites on the chain, the presence of 'inert' neutral sites and confinement. Different distributions of binding sites would produce, for instance, different types of 'rosette-like' globular conformations. In the following we will present polymer models which by use of appropriate binding sites positioning can describe very accurately the three-dimensional structure of real loci. Furthermore, off-equilibrium, unstable conformations are also expected to be encountered in real chromosomal regions, in particular during changes in the folding state.

The mixture model

To compare our very simple model with the Hi-C data, we suppose that the chromosome is a mixture of differently folded regions, where some loci can be more compact than others, like eu- and heterochromatin (see Chapter 1). The regions can change their conformation from cell to cell following functional purposes (Nagano *et al.*, 2013). At a first approximation, the conformation of each region must belong to the stable thermodynamic states (Nidocemi *et al.*, 2009; Barbieri *et al.*, 2012) previously identified, as schematically represented in Figure 3.8, Panel A. So, we consider a linear combination of the different contact probability profiles represented in Figure 3.7, Panel A.

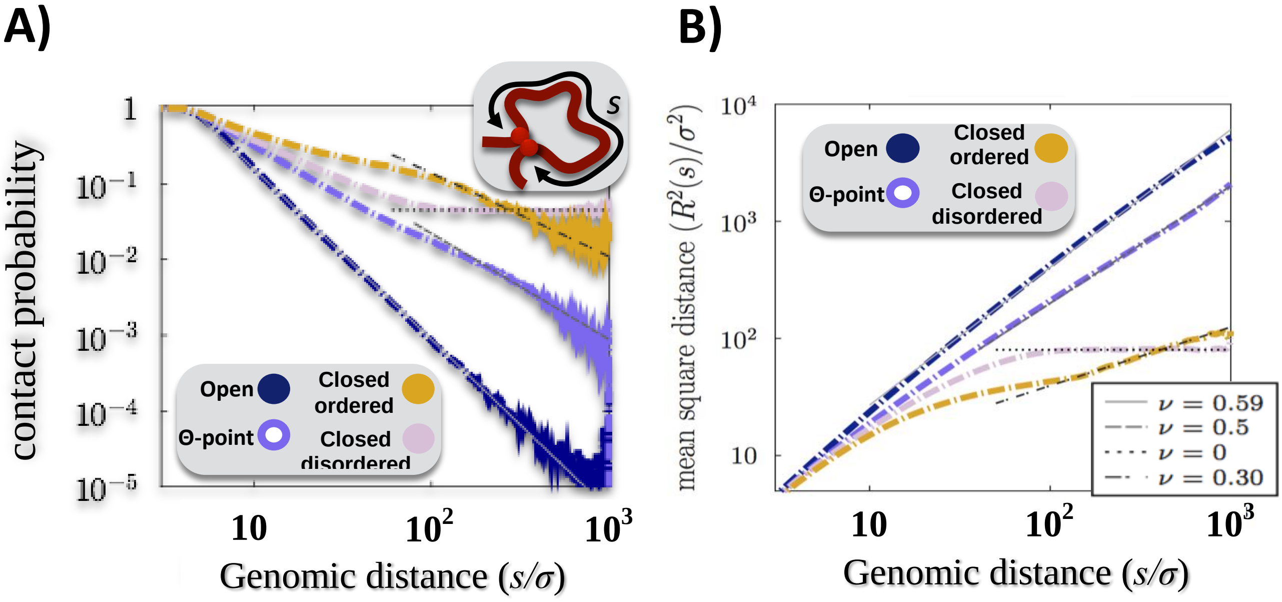

Figure 3.7: Contact probability $P_c(s)$ and mean square distance $R^2(s)$

Panel A: The contact probability as a function of the contour distance *s* (i.e. genomic distance), in the thermodynamic phases. **Panel B**: The mean square distance of two generic sites having a contour distance *s* along the polymer.

This combination depends on the relative abundances of the states in the mixture and on a scale factor necessary to map the bead size into genomic distances. Interestingly, we find that the model is able to fit the experimental contact probability data over very large length scales, from the sub-mega base scale up to the whole chromosome length. This results is valid for genome wide averaged data (Figure 3.8, Panel B) and for single chromosomes data (Figure 3.8, Panel C). Furthermore, we use data obtained from different experimental techniques (Hi-C, TCC and *in-situ* Hi-C), and the results are similar. By fitting the data experimental data we obtain the percentages of open and closed state that best describe the chromatin certain cell line (averaged over all the chromosomes), or the percentage that best describe the chromatin for a fixed chromosome. We find different results depending on the cell type: in the human embryonic stem cells (hESC, from Dixon *et al.*, 2012), the open state is approximately 75%, while in the differentiated cells as IMR90 fibroblast (data from Dixon *et al.*, 2012), this value is approximately 50%, as expected. If we consider contact probabilities extracted from data obtained from different experimental techniques (Hi-C vs TCC, data from Dekker and from Kalhor *et al.* respectively), the fit gives similar results, with a closed ordered state of 40%, but a slightly different balance between the other states. For a fixed cell type, we register a quite wide variability of these fractions among the different chromosomes, as shown in Figure 3.8, Panel E, for IMR90 cell type. For instance, chromosome X is very compact with a 75% of closed ordered state, while chromosome 1 has a 50% of open state. Generally, the percentage of open state decreases with the chromosome size, while the closed disordered phase increases, even though it represents a very small fraction (always less than 20%).

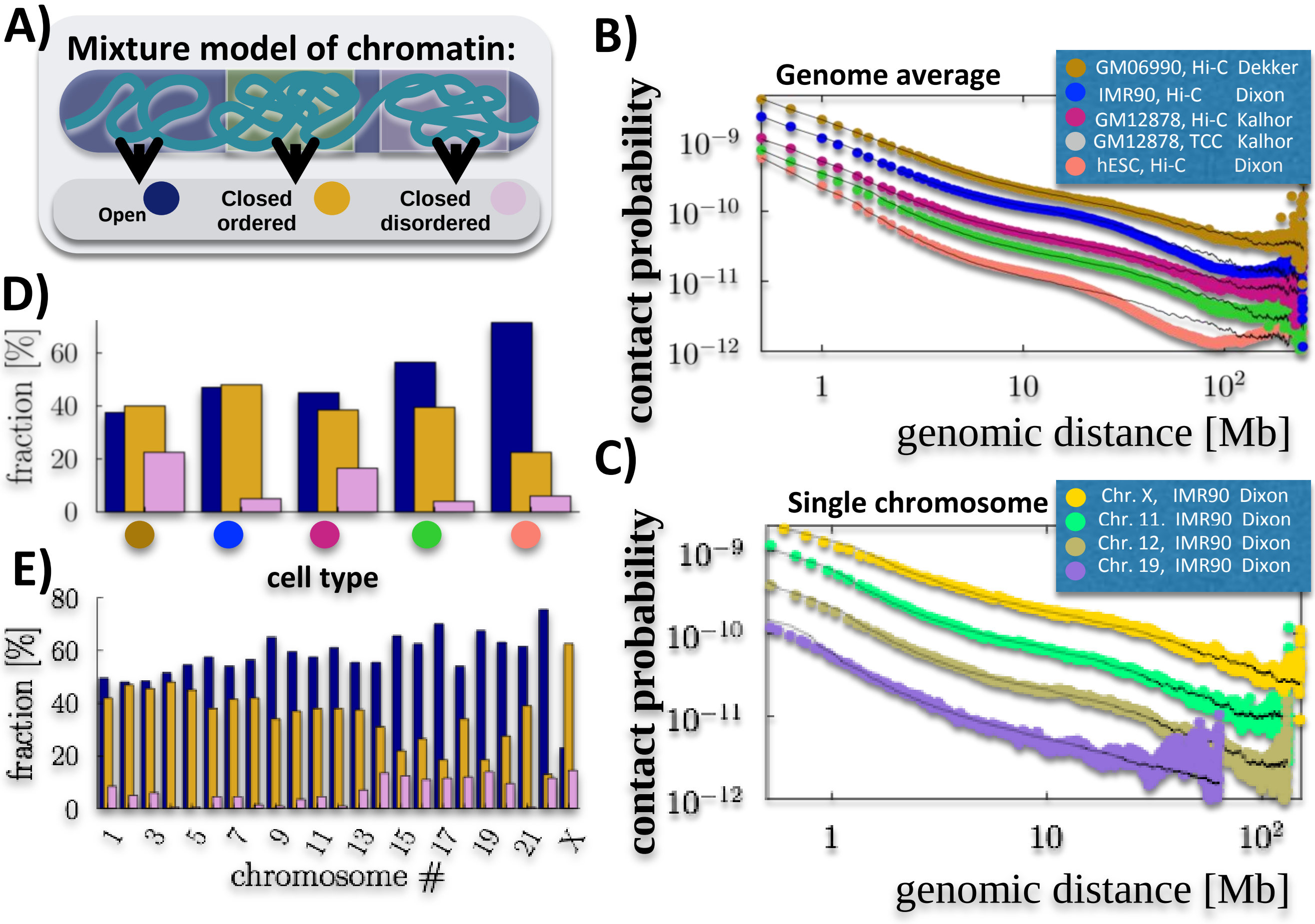

Figure 3.8: Chromatin is a mixture of regions folded in different thermodynamic states.

Panel A: We model a chromatin filament as a mixture of differently folded regions, each belonging to one of the stable conformational classes. In this view, the average pairwise contact probability is only determined by the relative abundances of the states in the mixture, as each state has a fixed, specific pairwise contact probability. **Panel B**: Genome-wide average contact frequencies across human cell types, obtained from various experimental techniques, can be fitted from the sub-Mb to chromosomal scales by such a mixture model. **Panel C**: Single chromosome data (here from IMR90 cells) can be similarly explained. **Panel D**: Different cell types have a different chromatin composition, with hESC (orange circle) more open than differentiated cells, such as IMR90 (blue circle). **Panel E**: Within a given cell type (here IMR90, as in Panel C) distinct chromosomes have also a different compositions, with chromosome X formed mostly of closed regions, whereas gene rich chromosomes, e.g., chr.19, are up to 70% open.

Method to fit the experimental data

The fit of genome-wide Hi-C average pairwise contact data as a function of the pairwise genomic separation is done by use of the Least Square Method (LSM). We compute the model predicted contact probability of a mixture of open and closed states by using the independently derived corresponding contact probabilities from the MD simulations of the homopolymer chain. Then, by LSM we find the composition of the mixture of open and closed states that minimize the distance between the predicted $P_C(s)$ and the one derived from Hi-C data.

Mapping MD units into physical units

To map the dimensionless units used in the MD simulation, we proceed in the following way: given the genomic length L to model, the genomic content in each bead is s_0=L/N. The physical bead diameter is estimated by equalizing the average nuclear chromatin density with the local chromatin density. In this way we obtain the relation $\sigma=(s_0/G)^{1/3}D_0$ (Barbieri *et al.*, 2012), where D_0 is the nucleus diameter and G the total genome length (in base pairs). We consider mouse embryonic stem cells, so we suppose D_0=3.5μm and G=6.5Gb. In the previous case, we set L=100Mb, so it results σ=87nm. The concentrations are estimated by using the relation c=P/VN_A, where P is the absolute number of binders in solution, V is the box volume and N_A is the Avogadro number. Analogously, the time scale τ is estimated by fixing the viscosity η and energy scale ε through the relation $\tau=\eta\ (6\pi\sigma^3/\varepsilon)$. So, by considering η=0.1P at room temperature T=300K, we obtain τ=0.03s. In all the following polymer models, all the physical will be calculated in this way.

The two colors model

The model just discussed have one kind of bead that can interact with all the binders floating in the surrounding environment. Despite its simplicity, it is able to recapitulate the average long-range contact properties of chromosomes. Nevertheless, the real Hi-C matrices have a very complex structure (see previous Chapter 1 and 2), and it is necessary to complicate the model to further investigate the patterns of the experimental data beyond the average long-range contact probability. To this aim, we now consider a block-copolymer, with two types of beads (visually represented by two colors, red and green), that can interact only with a specific kind of binder (red and green, as shown in Figure 3.9, Panel A). We consider as first case a 2-block co-polymer where each block is made of 500 beads, one red and one green, and the entire polymer is made of 1000 beads in total. To give a sense of length scales, we consider scales one order of magnitude lower than the chromosome modeling, which are typical genomic lengths where chromatin is known to be subjected to compartmentalization (Lieberman-Aiden *et al.*, 2009). Thus, we suppose that the region is 10Mb long. To estimate the length scale, we proceed as before and we find that the bead has a diameter σ=64nm. The time step results to be 0.003s, assuming a viscosity of 2.5cP. The concentrations and interaction energies are sampled so to cover the three thermodynamic stable states identified in the homopolymer study. When equilibrium is reached, each block folds in the configurations discussed in the previous subsection, and two stable globular domains are formed. The contact probability $P_c(s)$ and the average square distance $R^2(s)$ associated to this conformation are reported in Figure 3.9, Panel C. It is apparent the crossover at the domain boundary, where the $P_c(s)$ has a sharp drop at s=N/2, and complementary, $R^2(s)$ reaches the maximum value as a plateau. In the second block co-polymer, the distribution of the colors along the polymer consists of four consecutive blocks (red-green-red-green, Figure 3.9, Panel B), each block 250 beads long (as before, the total polymer is composed by 1000 beads). As before, each block can fold in the stable configuration and it forms, at the beginning of the dynamic process, a lower level structure (that can be interpreted as a TAD sequence). These objects correspond to enriched interaction squares along the diagonal of the contact matrix (Figure 3.9, Panel B, central matrix). When equilibrium is completely reached, the blocks of the same color interact, and the result is a hierarchical organization of higher-order structures, which is known to be a feature of the mammalian genome (Fraser *et al.*, 2015). In the contact matrix (Figure 3.9, Panel B, right matrix), such organization is represented by a

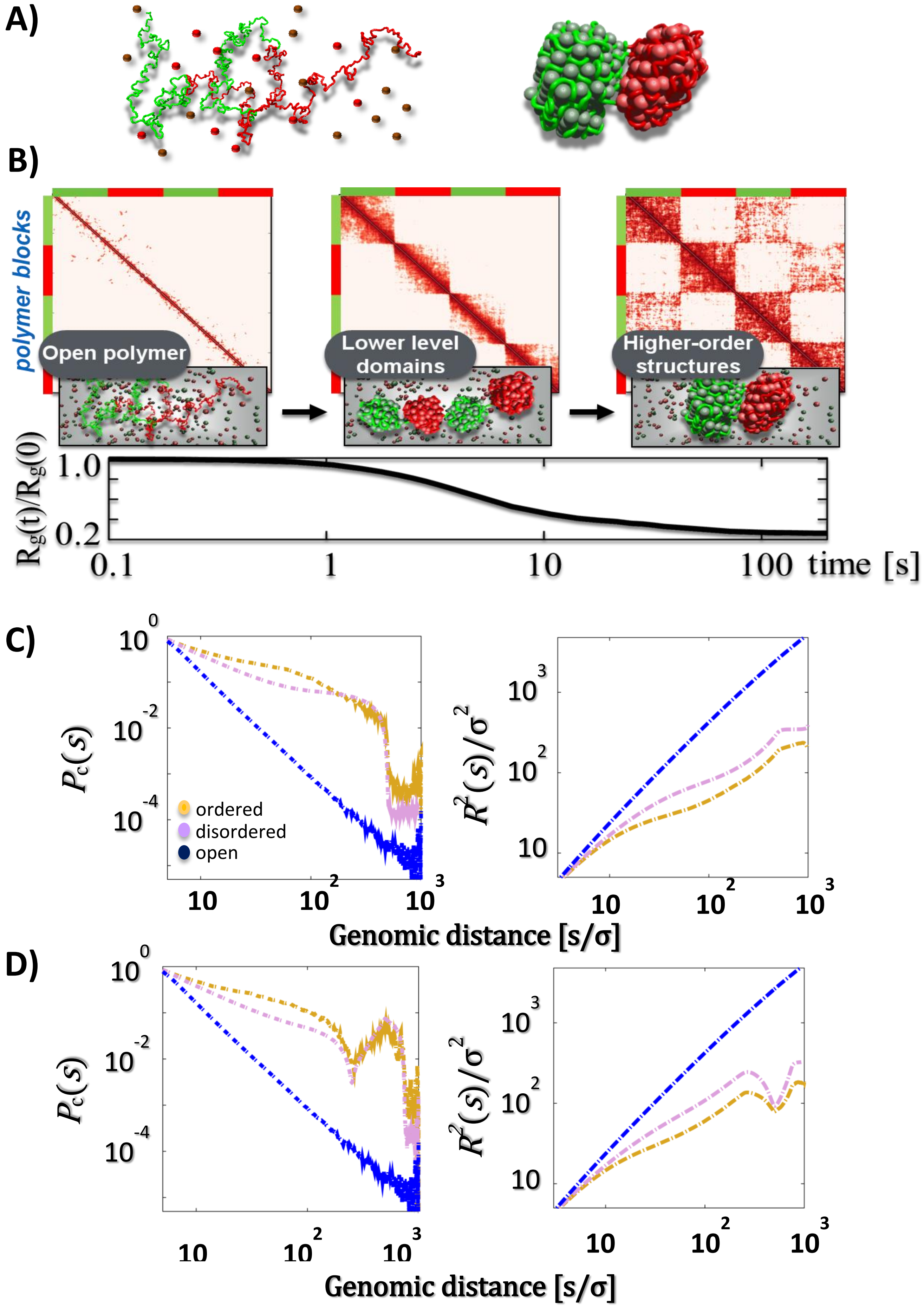
A)
B)
polymer blocks
Open polymer
Lower level domains
Higher-order structures
$R_g(t)/R_g(0)$
1.0
0.2
0.1
1
10
100
time [s]
C)
$P_c(s)$
10^0
10^{-2}
10^{-4}
ordered
disordered
open
10
10^2
10^3
$R^2(s)/\sigma^2$
10^3
10^2
10
10
10^2
10^3
Genomic distance [s/σ]
D)
$P_c(s)$
10^0
10^{-2}
10^{-4}
10
10^2
10^3
$R^2(s)/\sigma^2$
10^3
10^2
10
10
10^2
10^3
Genomic distance [s/σ]

Figure 3.9: hierarchical self-assembly of domains

Panel A: Block co-polymer model made of two types of beads (red and green) interacting with two types of binders. There are four consecutive blocks along the polymer, alternating their color. **Panel B:** the dynamics of the systems is marked by a decrease of the gyration radius, and a hierarchical self-assembly of domains spontaneously occur, as in the corresponding contact matrices (here E_{int}=4.1k_BT). **Panel C:** Contact probability (left plot) and quadratic distance (right plot) in the 2-block co-polymer, for the three thermodynamic stable phases. **Panel D:** Contact probability (left plot) and quadratic distance (right plot) in the 4-block co-polymer (the corresponding contact matrices are in Panel B).

chessboard-like pattern. $P_c(s)$ and $R^2(s)$ are reported in Figure 3.9, Panel D, and they reflect the information contained in the matrix. In fact, we register a sharp drop at s=N/4, then it increases since there are higher order interactions, and then it sharply drops again at s=3N/4. In the framework of our model, such structural features naturally emerge by specialization of the involved molecular factors under the laws of polymer physics.

The symmetry-breaking mechanism in the co-polymer models

An interesting consequence of the self-assembly of domains, that probably can have functional roles, is a symmetry-breaking mechanism occurring in the spatial organization of the loci. In particular, since TAD boundaries have been associated to an insulating role in the cell functionality, we consider the effect of the domains on the physical distance between pairs of sites differently located with respect to the domain itself. Specifically, we consider two pairs of loci having the same genomic distance (i.e. the same contour distance along the polymer). We focus on two cases where the sites can be symmetrically or asymmetrically located with respect to the boundary of the domains. We find that the symmetrical pair have on average a larger spatial distance than the asymmetrical one, while in the open state (i.e. the SAW) no difference is observed. This is found for both closed states (ordered and disordered), as shown in Figure 3.10. The details of the distances and contact matrices computation are given in the next subsection.

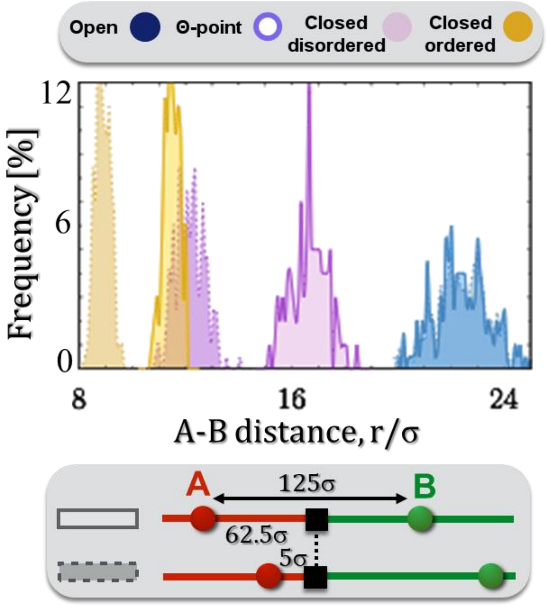

Figure 3.10: Simmetry-breaking mechanism in the physical distances

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

Computational approach for distance distributions

To measure the physical distances between two sites in the block co-polymer model, we consider two loci A and B, belonging to different blocks (A in the red block and B in the green block). In both cases, their contour distance is d=125σ. In the symmetric case, they are equally distant from the boundary of the domain, while in the asymmetric case the site A is located at a distance of 5σ from the domain boundary, and consequently the site B is 120σ from the boundary (so it is well inside the domain). To increase the statistics, we consider also the case where B is located at 5σ from the boundary and A at 120σ. The asymmetric distribution plotted in Figure 3.10 is an average of the two cases. The result is valid also if the distance from the boundary is higher (we checked the case 25σ from the boundary and similar results are found).

Computational approach for contact matrices

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

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

3.4 Multiple contacts interaction landscape

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

Computational approach for the many-body contact

To estimate the average number of many-body contacts involving simultaneous interactions of *k* beads occurring in a given polymer conformation, we count the number of beads n_i that are in contact with the *i*-esim bead within the fixed threshold λ (for this computation, we use as above λ=3.5), and the number of possible combinations of *k* simultaneous contacts that contain the *i*-esim bead, $\binom{n_i}{k-1}$. We average that number over all the beads in the polymer. As normalization factor, we consider the number of total possible many-body contacts of k particles with the i-esim bead, $\binom{N}{k-1}$. In Figure 3.11, Panel A, we show the value of this frequency as a function of the multiplet complexity n, computed in the homopolymer case discussed in the Section 3.2.

The triplet surface

The calculation described in the previous subsection gives an estimate of the many-body contact average probability. A more accurate calculation is made for the computation of the multiple contact profile when the complexity n of the multiplet is 3 (i.e. the triplets). As in the pairwise contacts probability the mathematical object is a 1-dimensional curve (Figure 3.3) and the parameter is the genomic distance s, here we need a 2-dimensional surface and the parameters are the two genomic separations s_1 and s_2 that separates the first bead and the second bead, and the second with the third bead, as schematically represented in Figure 3.11, Panel B, bottom part. As expected, the surface is symmetric, and in the particular case where s_1 or s_2 equal to zero (i.e. two beads coincides) we recover the pairwise contact profile.

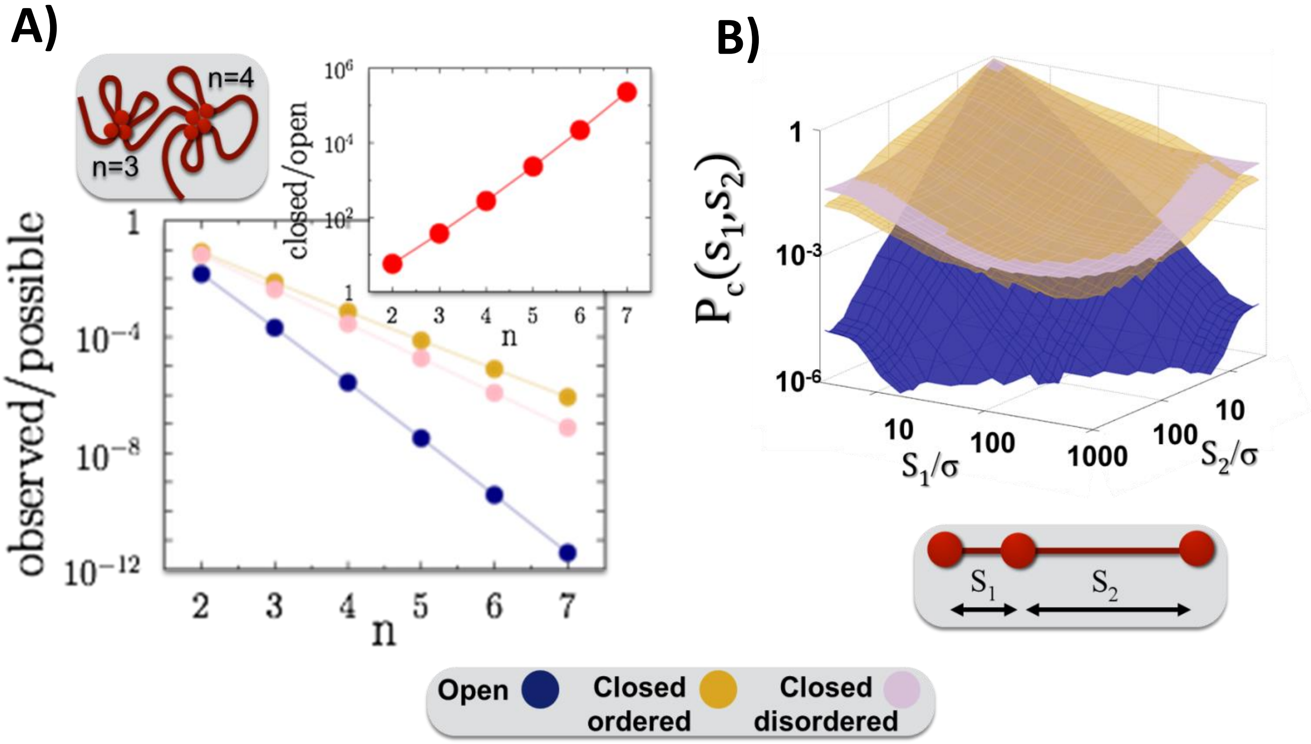

Figure 3.11: The theoretical multiple contact interaction profile

Panel A: The plot shows the frequency, of observing n sites in simultaneous physical contact (normalized by the number of possible combinations of n sites) along the SBS homopolymer discussed previously. The top-left inset shows the ratio in the compact-disordered and open states. **Panel B:** The plot shows the contact probability of bead triplets at different contour separations, $P_c(s_1, s_2)$, along the SBS homopolymer in its different thermodynamics phases.

Importance of multiple contacts

Multiple interactions are currently not detected by Hi-C methods, yet our model highlights that they are likely to be an abundant structural component of chromatin, as is emerging from new researches in the field (Olivares-Chauvet *et al.*, 2016, Beagrie *et al.*, 2017). That hints towards an important functional role of closed chromatin domains where multiple regulatory regions (like enhancers) can loop simultaneously onto a given target (gene promoter) with a much higher probability than in open regions. Taken together our results support a view whereby basic mechanism of polymer folding could play key functional roles in the regulation of the genome by controlling the spatial organization of chromatin.

References

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

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

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

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

23: 197-203.

Bickmore WA, van Steensel B. Genome architecture: domain organisation of interphase chromosomes. (2013) *Cell* **152**:1270-84.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interac- tions. Nature 2012; 485:376-80.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, *et al.*. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**: 381-5.

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

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

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

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

Spielmann M, Mundlos S, *Bioessays* (2013) **35:** 533.

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

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

Annunziatella C, Chiariello AM *et al*. (2016) Polymer models of the hierarchical folding of the HoxB chromosomal locus, *Phys Rev E* **94**: 042402

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

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

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

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

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

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

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

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

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

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E (2014)

Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**: 950-63.

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

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

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

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

Kremer K, Grest GS, (1990) *J. Chern. Phys.* **92**: 5057

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

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

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

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

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

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

Chapter 4

3D reconstruction of the real genome

In this chapter, we will show how the model discussed above can be used to describe and reconstruct the 3D architecture of real loci in the DNA. In the first part of the previous chapter we showed how with a very simple model (homopolymer model), polymer physics is able to recapitulate with a good degree of accuracy the average behavior of the chromosome structure in a wide range of genomic lengths (from the sub-Mb scale up to the whole chromosome scale). Next, we introduced an extended model, and this was done with the introduction of just a second bead type (the red-green polymer models), in order to explain other aspects of the chromatin architecture and to highlight mechanisms that could possibly have important functional roles: the existence of domains, the symmetry-breaking in the distance distribution, and the hierarchical structure contained in the experimental Hi-C contact matrices, occurring in a spontaneous self-assembly process. Now, we try to capture the finer spatial structure, in the deepest possible way, of specific region of real genomes. To do this, we generalize the model by introducing a multicolor polymer, where each color can interact only with its cognate type of binder. The number of required colors and their positions along the sequence depends only on the structural features of the considered locus, which are assumed to be entirely contained in the experimental Hi-C matrix. The method used to obtain these information will be briefly discussed in Section 4.1 (anyway, the details of the procedure is object of another publication and will not be discussed here). Next, we will show how from real experimental data we will be able to reconstruct with high accuracy the 3D structure of real loci. As first application, in Section 4.2 we will presents the results about the modeling of the *Sox9* and *Bmp7* loci, containing very important genes for the cell functionality (Franke *et al.*, 2016); then, in Section 4.3 we will model the *Xist* locus, which is another very important region (Nora *et al.*, 2012, Giorgetti *et al.*, 2014), and we will apply the model to predict the effect of a structural variant (precisely, a deletion, Giorgetti *et al.*, 2014). Finally, in Section 4.4, we will present the results about the modeling of *HoxB* and *HoxD* loci. In particular, in the case of *HoxD* locus we will study, as further application of our method, the same region in two different time points during the cell differentiation (mouse undifferentiated ESC-J1 and differentiated Cortex), and try to visually understand how the differentiation process affect and change the spatial structure of the locus in order to allow its functions. The results presented in the first three sections have been published in the papers Chiariello *et al.*, 2016, Annunziatella *et al.*, 2016 and Bianco *et al.*, 2017. The results presented in the last section have not been published yet and represent one of the current research projects of the group.

4.1 The generalized SBS model

Method used to obtain the binding site position

To identify the binding domains for the models of the studied loci, we use a Simulated Annealing Monte Carlo procedure to locate the minimal arrangement of binding sites and types (colors) that, based only on polymer physics, best explains the experimental contact matrix. Our method employs a standard Simulated Annealing scheme 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 the procedure is the number of colors required and the position along the polymer. So, the experimental Hi-C matrix is the starting point of the method, and it is used to extract information that cannot be directly obtained from them, as the 3D structure and physical distances. The size of the investigated regions in this chapter is order of magnitude smaller than the whole chromosome scale. Further details of the procedure will not be discussed here.

4.2 The *Sox9* and *Bmp7* loci

The *Sox9* locus

As first application of the model we consider a 6Mb sequence around the *Sox9* gene (chr11:109000000-115000000, mm9), that is a very important locus linked to congenital diseases (Franke *et al.*, 2016), including gene rich regions and gene desert regions, as shown in Figure 4.1, Panel A, upper part. The Hi-C datsets used to infer the polymer are published from Dixon *et al.*, 2012, mouse ESC-J1 cell line, at 40kb resolution, and are shown in Figure 4.1, Panel B. The experimental data used have been normalized following standard precedures (Yaffe&Tanay, 2011). The distribution of binding sites is made of 15 different colors, and it is represented in Panel A of Figure 4.1. As expected, the binding domains tend to overlap with the different TADs existing in the locus, but they also overlap with each other and produce interactions between TADs, giving the hierarchical structure (metaTADs) visible in the original experimental matrix (Figure 4.1, Panel B, top matrix). Once we obtain the optimum arrangement of the binding sites along the polymer, we perform MD simulations to recostruct the 3D structure of *Sox9*. To test the accuracy of our structures, we compute from the ensemble of configurations formed in the dinamics process the contact maps and compare it with the experimental data. In the framework of the SBS model, we consider separately the open phase (i.e. the SAW conformational class) and the closed phase (i.e. the equilibrium phase after the complete folding of the polymer). Then, we seek the open-closed mixture that maximizes the Pearson correlation coefficient between model inferred and Hi-C data.

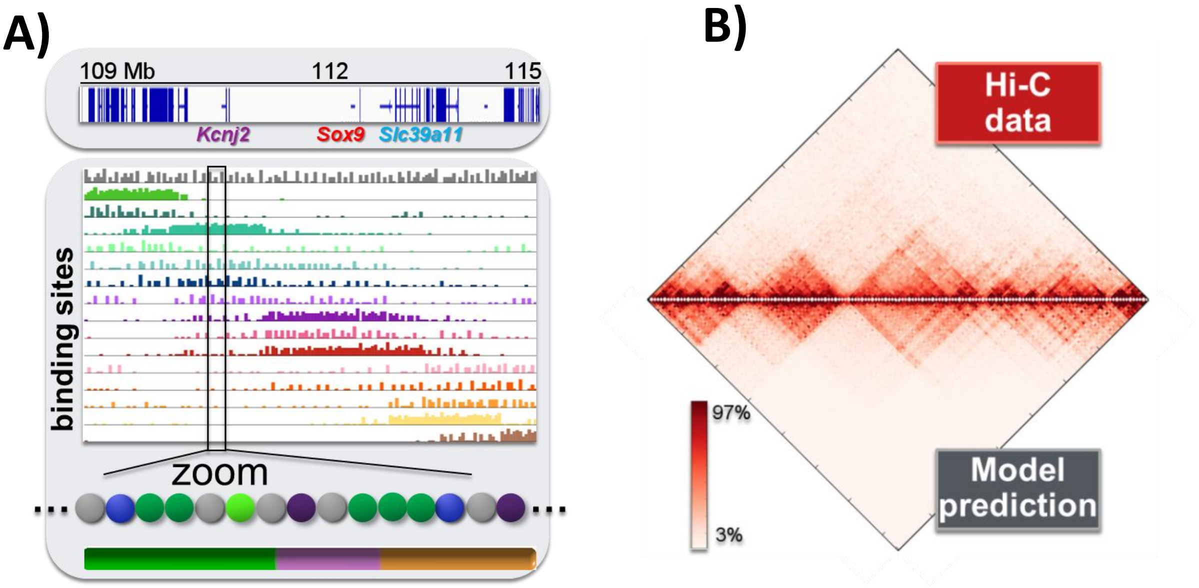

Figure 4.1: The *Sox9* locus is captured by polymer physics

Panel A: In the top, the *Sox9* locus in mouse ESC-J1 cells, with some genes represented; in the bottom, the SBS polymer that best reproduces the experimental Hi-C contact matrix is made of 15 different binding domains distributed as shown, and highlighted in the zoom. Each histogram is the abundance of a certain color over the genomic sequence. The schematic color scheme reported (linear green-thistle-orange bar) reflects the relative abundance of that color in each region. **Panel B:** The model pairwise contact frequency matrix (bottom) compared to the experimental Hi-C data (top). The Pearson correlation coefficient r is 0.95.

Features of the *Sox9* locus

Here, we find that the *Sox9* locus is made of 64% of open and 36% of closed state. The result of the whole process returns a simulated contact matix very similar to experimental data, and it is represented in Figure 4.1, Panel B, bottom matrix. The Pearson correlation coefficient between Hi-C data and simulated contact frequency matrix is 0.95, proving that our model captures relevant features of the mechanisms deteremining the folding of *Sox9*. We consider the transcription starting sites (TSS) of three fundamental genes of the locus, *Sox9*, *Kcnj2* and *Slc39a11* and compute the physical distances. Interestingly, we find that the *Sox9* and *Kcnj2* TSS, having a genomic separation *s*=1.72Mb, have an average physical distance d=1190 nm, while the *Sox9* and *Slc39a11*, having a genomic separation of *s*=0.46Mb (four times smaller) have an spatial distance d=590nm, so the two pairs are proportionally closer, as they belong to consecutive regional areas. As shown in Figure 4.2, Panel A, the self-assembly of the locus spatial structure starts from a totally random SAW initial state (open confromational class) and proceeds hierarchically, passing through early local domains folding into larger and larger domains that cover the whole locus. In Figure 4.2, Panel B, is shown a snapshot of a single typical 3D configuration, obtained from the dinamics, when the polymer is fully equilibrated, i.e. when it is in the closed state. Here, we represent the relative positioning of *Sox9*, *Kcnj2* and *Slc39a11* across its different higher-order domain organization.

Simulation details of the *Sox9* polymer model

To model at higher-resolution the 3D structure of the *Sox9* locus in mESC-J1, we use a chain made of N=2250 beads. Since the region to model is L=6Mb long, the elementary bead of the polymer has a genomic content of L/N=2.67Kb. The size of the bead is thus 26nm, as follows from the calculation described in the previous chapter. In this case, the MC procedure returns a polymer with 15 different interacting bead types. Each type interact only with its specific binder. This interaction is modeled by an attractive Lennard-Jones potential, with an interaction distance between bead and binders $\sigma_{\text{bead-binder}}$=1σ and the cutoff range r_{int}=1.5σ (as before, σ is the diameter of beads and binders). The time unit is 0.0002s, obtained by assuming an enviromental viscosity of 2.5 cP. The binder concentration is sampled in the range that allow to explore the main thermodynamic stable states (open and closed disordered). In particular, in Figure 4.3 the simulations were performed with c=194nmol/l. The interaction energy regulating parameter is ϵ_{int}=12$k_B T$, in the same notation of equation (9) and (10) reported in Chapter 3. The 3D structure presented in Figure 4.2, Panel B, is obtained from the polymer dynamics, in the equilibrium closed phase. Mathematically, it is a smooth curve described by a third order polynomial spline passing through the centers of each polymer bead. The color scheme used (green-thistle-orange) in Figure 4.2 is chosen to reflect the pattern contained in the experimental Hi-C matrix, where three main domains (interestingly not coincident with the TADs identified in Dixon *et al.*, 2012) are visually evident.

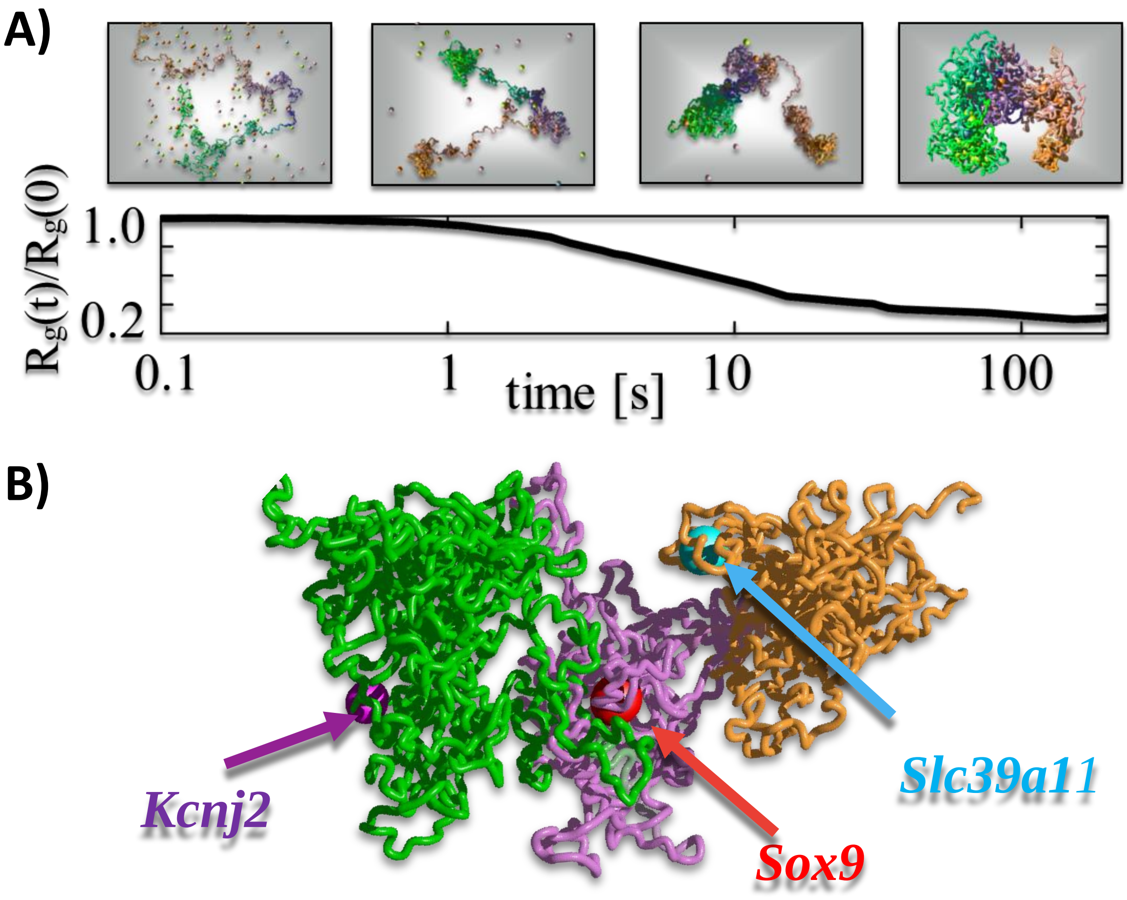

Figure 4.2: 3D reconstruction of the *Sox9* locus

Panel A: The *Sox9* folding dynamics from a completely random SAW configuration, passing through intermediate states with local domains formation. The dynamics state is monitored by measuring the gyration radius $R_g(t)$ at a genetic time step t relatively to its initial value $R_g(0)$.

Panel B: A snapshot of the 3D structure in the closed state, with the position of TSSs of the relevant genes, (*Sox9*, *Kcnj2* and *Slc39a11*) highlighted. The color scheme used is the same pictorially reported in Figure 4.1, and reflects the abundance of the color in each region.

An alternative color scheme for *Sox9*

Importantly, our polymer models can be used to derive any information on the folding of interesting loci genome-wide, beyond Hi-C pairwise contact data. For instance, the same snapshot of the full 3D structure of the *Sox9* locus in Figure 4.2, along with a comparison with its average contact matrix and TADs, is shown in Figure 4.3 with an alternative color scheme that follows the coordinates of the original TADs identified in Dixon *et al.* 2012. For example, it is visible that TAD D (red) has a complex internal 3D structure, with a large part of it mostly associated to TAD E. Additionally, the dynamics of the interactions between the genes within the locus and their regulators can be derived.

Contact maps

The contact maps presented in this chapter, are computed following the approach described in the previous chapter, with a variant where only contacts between beads of the same type are considered. In the case of *Sox9*, the parameter for the interaction threshold is set to λ=10 (in the same notation of Chapter 3), since the resolution in this case is much higher than in chromosome wide simulations. The same approach is used for the models presented in the following sections, with similar values for the λ parameter.

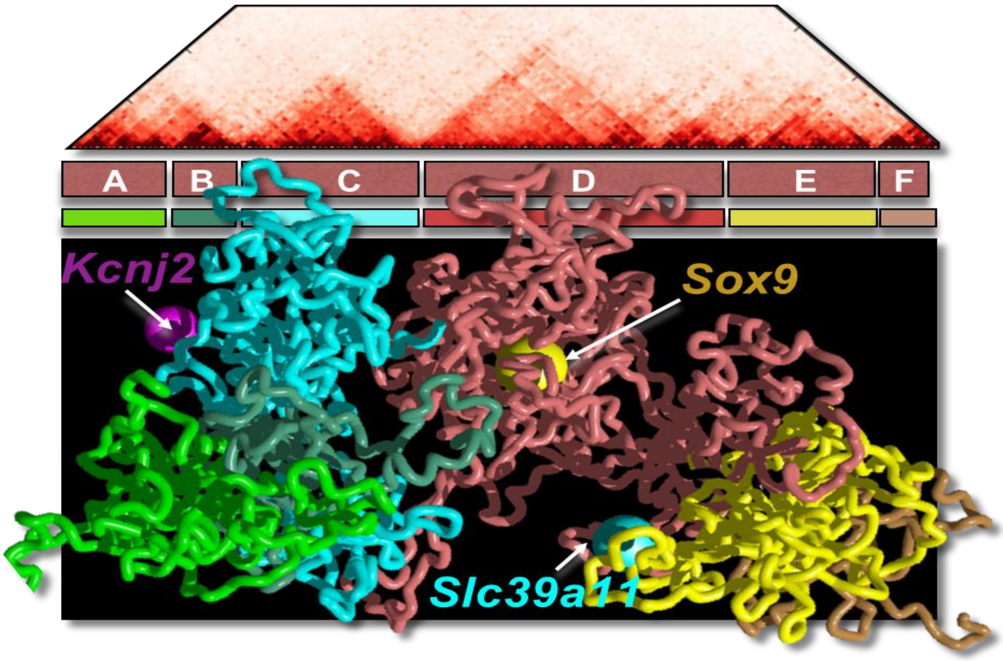

Figure 4.3: 3D reconstruction of the *Sox9* locus, alternative color scheme

The same configuration of the full 3D structure of the *Sox9* locus in mESC-J1 represented in Figure 4.2, with a color scheme reflecting the TADs position (from Dixon *et al.* 2012). It allows to visually interpret the patterns seen in its Hi-C map and the relative organization and interactions of TADs.

The *Bmp7* locus

In this subsection we apply the method just described above to another locus, to test the general validity of our approach and consider it as a powerful tool to reconstruct and visualize the 3D architecture of real loci in the genome. In particular, we focus on the *Bmp7* locus (chr2: 171090000-173430000), a region approximately 2Mb long around the *Bmp7* gene, which is very important in tissue development. We use Hi-C data published in Fraser *et al.*, 2015, in mouse ESC-46C cell line. The experimental data resolution is 30Kb, and the normalization procedure used is described Chapter 2 (ICE iterative correction and background subtraction). Our method returns a contact matrix very similar to the experimental data, with a Pearson correlation coefficient r=0.95 between model and data, as shown in Figure 4.4. This results is even more important if we consider that in the *Sox9* case we used experimental data normalized in a completely different way (see the *Sox9* subsections). In this way, we enforce the validity of the method, that results to be unaffected by the underlying experimental data treatments. Actually, in the next section we will also use data produced with a different technique (5C), so the approach results valid even in case we use data derived from different experimental method.

Simulation details of the *Bmp7* polymer model

We consider a region 2.34Mb long around the *Bmp7* gene. The MC procedure returns a polymer composed by 11 colors and made of N=858 beads. The parameter used in the MD simulations (potentials, interaction energies and concentrations) are the same used for the modeling of the *Sox9* locus. The 3D structures are produced as previously described.

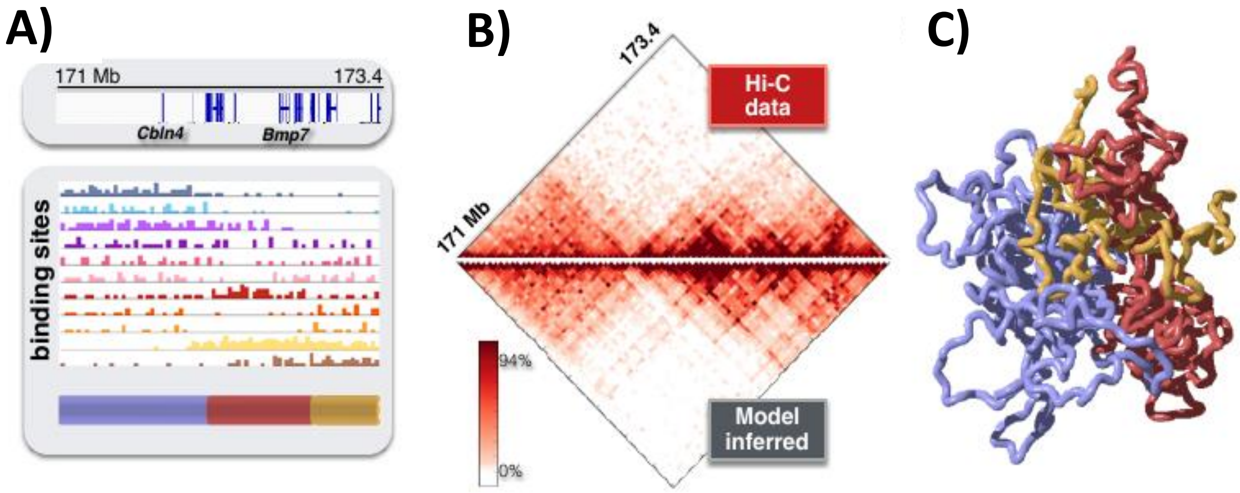

Figure 4.4: The *Bmp7* locus

Panel A: In the top, the *Bmp7* locus in mouse ESC-46C cells, with some genes represented; in the bottom, the SBS polymer that best reproduces the experimental Hi-C contact matrix is made of 10 different binding domains distributed as shown. Each histogram is the abundance of a certain color over the genomic sequence. The color scheme used reflects the abundance of the color in the considered region. **Panel B:** the model pairwise contact frequency matrix (bottom) compared to the experimental Hi-C data (top). The Pearson correlation coefficient r is 0.95. **Panel C:** a snapshot of the 3D structure in the closed state, at equilibrium after a dynamical process starting as usual from a completely random SAW configuration. The color sequence is the same of Panel A, bottom part.

4.3 The *Xist* locus: predictive power of the model

The *Xist* locus and the dataset analyzed

In this section, we will show how with our model is possible to predict the effect on the spatial organization of a locus generated by a mutation on the real genomic sequence. In particular, we consider the *Xist* locus, an important region containing the *Xist* gene (chrX: 100298000-101373000), schematically represented in Figure 4.5, Panel A, top part. We make this choice because experimental data are available (Nora *et al.*, 2012) for the wild type (WT) and also for a deletion variant (indicated as *ΔXTX* deletion), so we can test directly the results of our simulated predictions with a totally independent dataset. In the wild type case, we use data produced with the 5C technique in male undifferentiated WT mouse ESC-E14 cell line. The fragment based 5C interaction maps have been binned in a 20Kb resolution map using the online tool my5C (Lajole *et al.*, 2009). The corresponding map is represented in Figure 4.5, Panel B, top matrix. In the deletion *ΔXTX*, we also use 5C data from the XO mouse ES cell line (Panel D, top matrix).

Results

Starting from the 5C map at 20Kb resolution, we first obtain the polymer model describing the WT locus, and we find a good agreement between the contact map extracted from the simulations and the experimental data (Pearson r=0.96), as shown is Figure 4.5 Panel B. Next, we implement *in silico* the *ΔXTX* deletion in the same WT polymer, produce new initial and completely independent configurations, and perform again MD simulations, in the exactly same conditions of the WT case. We find that the predicted matrix has a pattern of ectopic interactions compared to the WT case (magenta box in Figure 4.5, panel D, bottom matrix). Interestingly, the predicted pattern is similar to the one reported in the experimental data (magenta box in Figure 4.5, top matrix), with a correlation coefficient r=0.91. After the deletion, in the structures obtained from MD simulation, the yellow regions, close to the deleted part represented in cyan in Figure 4.5, Panels C-E, are spatially repositioned with respect to each other, and contribute to form the ectopic contact between regions sharing the same binding sites.

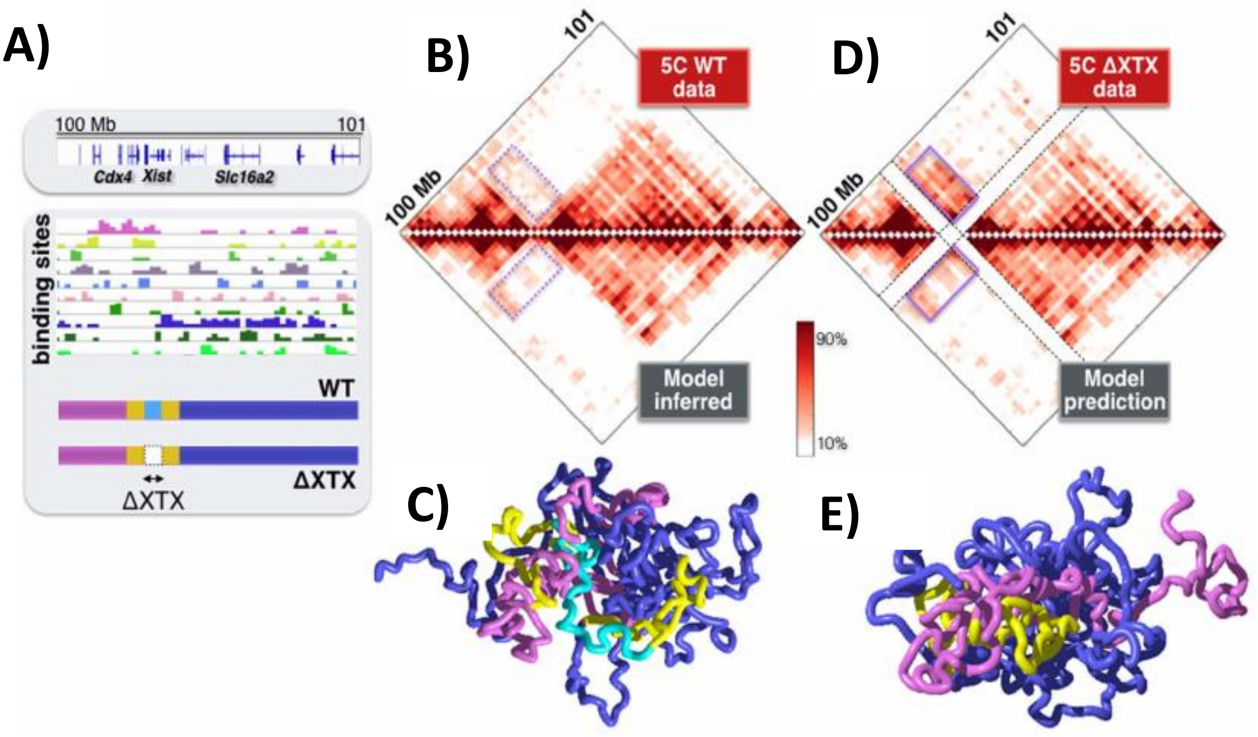

Figure 4.5: The *Xist* locus and its *ΔXTX* deletion

Panel A: In the top, the *Xist* locus in mouse ESC-E14 cells, with some genes represented; in the bottm, the SBS polymer that best reproduces the experimental Hi-C contact matrix is made of 10 different binding domains distributed as shown, and highlighted in the zoom. As usual, each histogram is the abundance of a certain color over the genomic sequence. To help 3D visualization, the color scheme used, reflecting the abundance of the color in the considered region, highlights in cyan the region deleted in Δ*XTX* cells and in yellow the sequences involved in the ectopic interaction. **Panel B:** The model inferred contact matrix (bottom) has a Pearson correlation 0.96 with 5C experimental data (top). **Panel C:** A snapshot of the *Xist* locus in its closed state. **Panel D:** The contact matrix predicted by the same polymer in Panel A after the Δ*XTX* deletion (bottom) reproduces with a high degree of similarity (top, correlation 91%) the ectopic interactions (full line magenta box) compared to the WT data in Panel B. **Panel E:** Visual representation of the ectopic interactions, where the yellow regions come closer in space after the Δ*XTX* deletion, as visible in the 3D structure of the deleted locus (cyan segment in Panel C).

Simulation details of the *Xist* WT and *ΔXTX* polymer models

We consider a region 1.3Mb long around the *Xist* gene. The MC procedure returns a polymer composed by 10 colors and made of N=540 beads. The polymer modeling the *ΔXTX* deletion, is made of N=510, since the deleted part, at 20kb resolution, is 3 bin long on the matrix. Note that from the modeling point of view this is a very small modification. Nevertheless, is produces effects analogous to what happens in the experiment. The parameters used in the MD simulations (interaction energy and interaction range) are the same used for the modeling of the *Sox9* and *Bmp7* loci.

4.4 The *HoxB* and *HoxD* loci: capturing the spatial reorganization during the differentiation

In this section, we will discuss the modeling of the *Hox* loci, which are other fundamental regions of the genome, very important in the embryo development (Andrey *et al.*, 2013). In particular, we focus on the *HoxB* locus (chr11:95280000–97200000) in the mouse ESC-J1 cell line and on the *HoxD* locus (chr2:71000000-78000000). In the latter case, we will study the spatial architecture of the region during the cell differentiation. In particular, we will consider two time points: the mouse ESC cells (not differentiated) and brain cortex cells (fully differentiated, indicated as Cortex in the following). We make this choice since these data are publicly accessible. Using the approach developed and discussed in the previous sections, we will show how it is possible to extract information about the structural rearrangements occurring during the differentiation.

The *HoxB* locus

We consider a region 1.92Mb long around the *Hoxb* gene. The procedure that finds the best polymer is applied to data binned at 40Kb (from Dixon *et al.*, 2012), and the result is showed in the diagram in Figure 4.6, Panel A. The experimental data are reported in Figure 4.6, Panel B, top matrix, with a color scheme different from the previous cases. The simulated contact matrix is showed in the bottom part. Once again, the agreement with experimental data is good, with a Pearson correlation coefficient r=0.95. As the procedure return 12 different type of binding domains, the polymer we use is made of N=576 beads, and the elementary bead contains 3.3Kb. the parameters used for the simulations ($\sigma_{\text{bead-binder}}=1\sigma$, $r_{\text{int}}=1.5\sigma$, $\epsilon_{int}=12k_BT$).

The *HoxD* locus and its differentiation

As in the case of the *HoxB* locus, we use data binned at 40Kb resolution, from Dixon *et al.*, 2012. The considered genomic region is centred around the *HoxD* gene cluster and it is 7Mb long. The corresponding contact matrices are reported in (Figure 4.7, Panel B). As visually clear, the two contact matrices have some discrepancies. In particular, it is possible to observe in the Cortex Hi-C contact map the presence of long range interactions, in the region sourrounding the *HoxD* cluster. Very dstinct domains along the diagonal in mESC cells are, after differentiation, less evident, and tend to interact outside their region, forming a larger square in the central part of the matrix (Figure 4.7, Panel B). The contact matrices obtained from MD simulations, reported in Figure 3.7, Panel A, reproduce with a good accuracy the same behaviour observed in the data. By analyzing the 3D structure from the simulations (Figure 3.8) it is possible to give a structural interpretation of the differences between the cell lines. To this aim, we choose a color scheme reflecting the boundaries of the main domains (highlighted by dashed lines on the matrices in Figure 4.8), clearly visible in the experimental mESC Hi-C matrix, where they appear as red squares on the diagonal. Naturally, to compare the different polymers structures, we use the same color scheme in both cell lines. Interestingly, it emerges that in the polymer describing the mESC cell line the domains are positioned in an approxiamtely linear sequence, without any special interaction between differently colored region, reflecting their particularly high inner interaction. On the contrary, in the Cortex it is possible to observe that the regions located in the central part of the matrix (colored in blue, cyan and thistle respectively) are closer in space, and the overall result is a more globular structure, having probably a signficant functional role related to the different activity of the *HoxD* gene cluster in the two differentiation phases (note that the *HoxD* cluster is placed exactly at the boundary between the blue and cyan domains). All this this observation depict a scenario where the differentiation acts on the structure of the locus in such a way to reshape it from a linear sequence of domains to a “U-like” shaped structure for the *HoxD* locus. Interestingly, this result is consistent with recent findings about the spatial structure and shape of the *HoxD* locus (Fabre *et al.*, 2015).

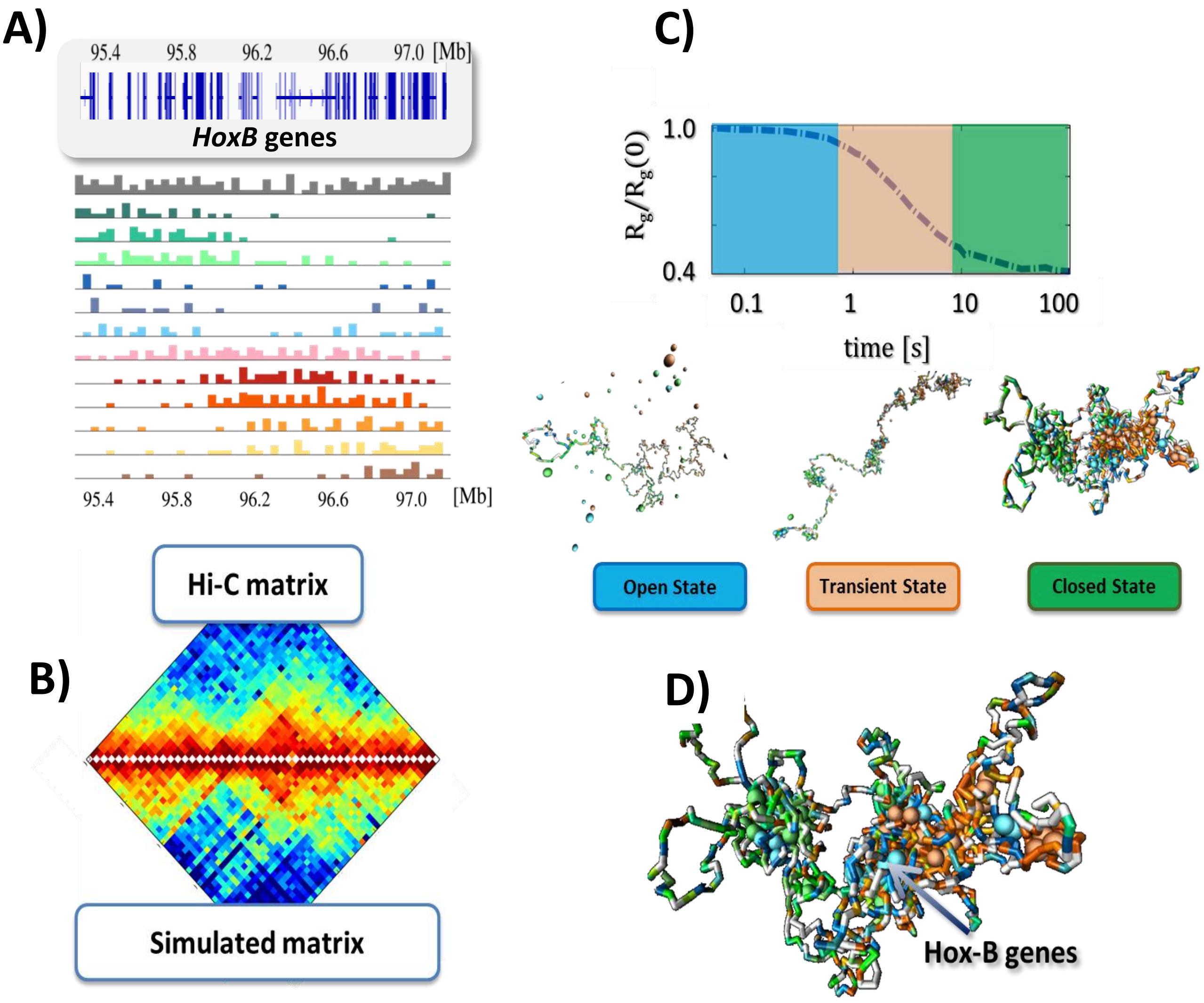

Figure 4.6: the *HoxB* locus and its folding mechanism

Panel A: In the top, the *HoxB* locus in mouse ESC-46C cells, with some genes represented; in the bottm, the SBS polymer that best reproduces the experimental Hi-C contact matrix is made of 12 different binding domains distributed as shown, and highlighted in the zoom. Each histogram is the abundance of a certain color over the genomic sequence. **Panel B:** The model pairwise contact frequency matrix (bottom) compared to the experimental Hi-C data (top). The Pearson correlation coefficient r is 0.95. **Panel C:** The folding dynamics of the *HoxB* locus proceeds gradually passing through an intermediate transient state to a completely folded polymer. **Panel D:** Magnification of the folded state in Panel C. The *HoxB* gene is located in the central part of the matrix. It emerges also an organization in three blocks, reflected in the coloration of the polymer (green-ligh blue-orange). In this case, the binders are showed in the 3D structure.

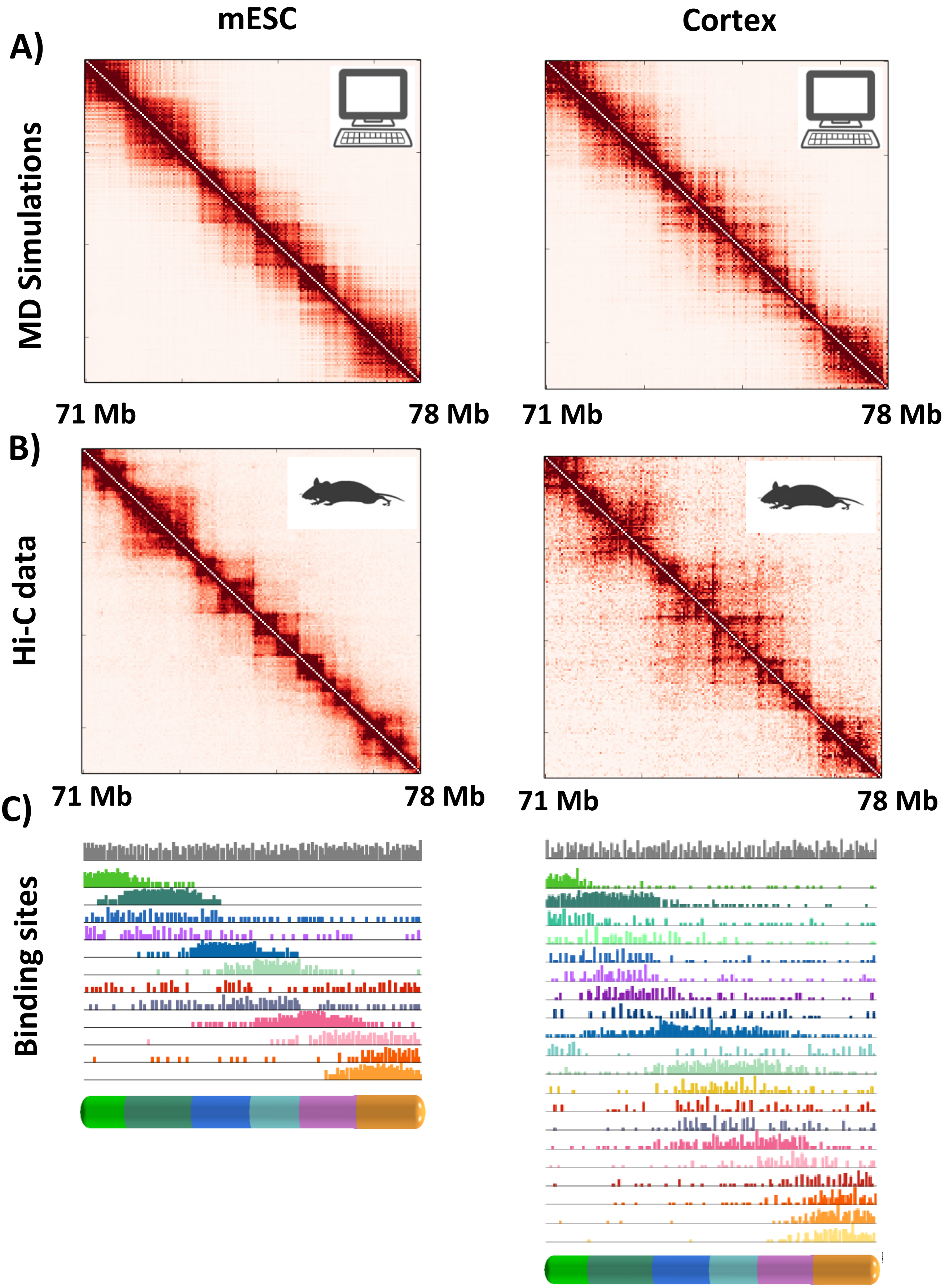
mESC
Cortex
A)
MD Simulations
71 Mb
78 Mb
71 Mb
78 Mb
B)
Hi-C data
71 Mb
78 Mb
71 Mb
78 Mb
C)
Binding sites

Figure 4.7: The *HoxD* locus in mESC and Cortex cell line

Panel A: MD simulated contact matrices for mESC (undifferentiated cells) and Cortex (differentiated cells). As for the experimental data, passing from mESC to Cortex we observe a decrease of the enrichment in the diagonal domains and an increase of long range contacts between such domains. **Panel B:** experimental Hi-C contact maps for mESC (left) and Cortex (right). **Panel C:** the inferred polymers best describing the locus in mESC and Cortex. In mESC, the differently colored binding sites are more localized then in Cortex. Furthermore, to capture the pattern contained in there, the MC procedure gives more binding types in the Cortex case. The color scheme shown in the bottom reflects the binding site distribution along the polymer and allow a comparison between mESC and Cortex.

Simulation details of the *HoxD* polymer models

As the pattern in Hi-C matrices between the mESC data and Cortex data have some evident differences (Figure 4.7, Panel B), we have to use different polymers to take into account such deviations and accurately describe both the cases. So, the MC procedure finds in Cortex cell line (Figure 4.7, Panel C, right diagram) binding domains less localized than in mESC (left diagram), and also extra binding domains necessary for the much more complex contact pattern (Figure 4.7, Panel C). Preciseley, it returns 12 types of binding sites for mESC and 20 types for Cortex. The resulting polymers consist of 2100 beads and 3500 beads respectively. The values used for the set of parameters are analogous to the previous models.

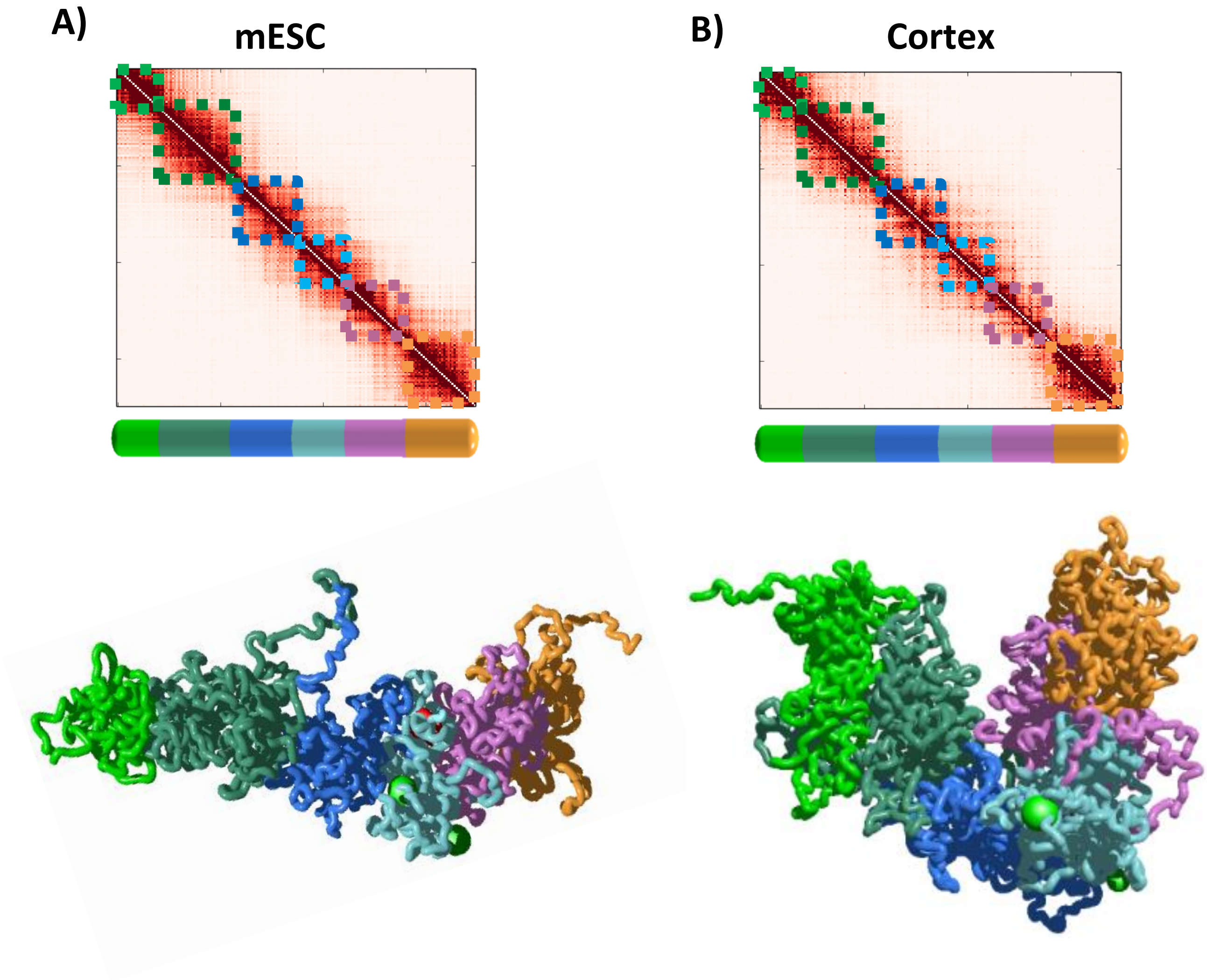

Figure 4.8: 3D structures for the *HoxD* locus in mESC and Cortex cell line

Panel A: 3D structure of the *HoxD* locus obtained from simulations performed from mESC data (undifferentiated cells). To help the visualization, we report also the simulated contact map and the colored domains highlighted by dashed lines. The organization is in compact and localized domains arranged in an approximately linear sequence. **Panel B:** 3D structure of the *HoxD* locus from Cortex data (differentiated cells). In this case we observe a more globular, U-shaped organization resulting from the interaction of the blue-cyan-thistle domains. Interestingly, the *HoxD* cluster is located exactly at the boundary between the blue and cyan domains, suggesting a functional purpose behind such structural rearrangement. Highlighted in green are the *Cns39* and *Cns65* regions, while in red the *Hoxd13* and *Hoxd1* genes.

References

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

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

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

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

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**:376-80;

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**:381-5;

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

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

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

Spielmann M, Mundlos S, *Bioessays* (2013) **35**: 533.

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

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

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

Lajole, BR, van Berkum, NL, Sanyal, A & Dekker J (2009) My5C: web tools for chromosome conformation capture studies. *Nature Methods* **6**: 690-1.

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

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

Annunziatella C, Chiariello AM *et al*. (2016) Polymer models of the hierarchical folding of the HoxB chromosomal locus, *Phys Rev E* **94**: 042402

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

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

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

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

De Santis E, Minicozzi V, Proux O, Rossi GC, Silva KI, Lawless MJ, Stellato F, Saxena S, Morante S (2015) Cu(II)-Zn(II) cross-modulation in amyloid-beta peptide binding: an X-ray Absorption Spectroscopy study *J Phys Chem B* **119**:15813-20

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

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

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

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

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

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

Fabre PJ, Benke A, Joye E, Huynh THN, Manley S and Duboule D (2015) Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. *Proc Natl Acad Sci U.S.A.* **10**: 112.

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

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

Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, Trono D, Spitz F, Duboule D (2013) A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. *Science* **340**: 6137

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

Kremer K, Grest GS (1990) *J. Chern. Phys.* **92**: 5057

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

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

Conclusions and perspectives

In this work we discussed some relevant aspects of chromatin organization in mammalian genome. First, we studied in detail Hi-C contact maps for three time points during mouse cell neuronal differentiation, and we showed that the spatial architecture of the DNA seems to be described by a complex, hierarchical organization starting from the sub-Mb scale up to the entire chromosome length. This conformational behavior is well captured and visualized by tree diagrams. These tree structures are correlated with most of the epigenetic features analyzed, suggesting that such organization has functional purposes for the genome regulation. Furthermore, the tree rearrangements occurring during the cell differentiation are linked to transcriptional state modifications. Next, we used polymer models to investigate quantitatively the properties of genome. At the beginning, we recapitulated with a very simple and essential model some important aspects as, for instance, the long range average behavior of the experimental contact probability and the spontaneous hierarchical folding mechanism. Then, by generalizing the model, we showed some examples of highly accurate 3D reconstruction of real genomic loci. Furthermore, we showed that the model is able to predict with a good degree of accuracy the effect of a variation in the genomic sequence on the 3D architecture, and it is also able to capture the structural differences of a certain genomic region in two different cell lines.

Without any doubt, this last aspect of the work is to us the most promising and exciting because of its potential future developments. New researches lines, not described in this thesis, we are following in order to improve the predictive power of these methods. The goal is to realize a reliable tool, able to investigate at a deeper level the numerous, still unknown, mechanisms involved in the genome organization, and to predict the effects due to variations in the spatial organization of DNA and their impact on the cell functionality.

Acknowledgments

We thank the University of Naples Federico II for grant of the PhD program in Physics. We acknowledge computer resources from Scope (University of Naples), CINECA and INFN, which have a fundamental role in our research activity. We thank all our collaborators in

Berlin and Oxford and all members of the group of Complex Systems at Physics Department “Ettore Pancini” in Naples.