Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXXIV

Coordinatore: prof. Salvatore Capozziello

Polymer physics models to unveil the mechanisms of chromosome 3D organization at the single-molecule level

Settore Scientifico Disciplinare FIS/02

Dottorando
Mattia Conte

Coordinatore
Prof. Mario Nicodemi

Anni 2019/2022

CONTENTS

Illustration from the cartoonist John C. Chase. Image taken from the paper:
BMC Biophys **4,** 8 (2011). https://doi.org/10.1186/2046-1682-4-8

Short abstract: Human chromosomes have a complex 3D structure as genes and their regulators located far along the chain have to physically interact. Such an architecture is crucial to define the fate of a cell by establishing active and silenced genes. I investigated models of polymer physics of chromosomes to understand the mechanisms whereby distal DNA sequences recognize and interact with each other to shape the folding of our genome and its functions. Results of my work have been published, e.g., in *Nature Communications*, 11, 3289 (2020), *Nature Methods*, 18, 482 (2021), *Nature Structural & Molecular Biology*, 28, 152–161 (2021) and *Nature Genetics*, 53, 1064–1074 (2021).

INTRODUCTION

Twenty years ago, the Human Genome Project allowed the completion of sequencing of the human genome, i.e., the identification of the entire sequence of DNA letters (technically called "bases") making up our genome. However, just as a list of automobile parts does not reveal us how a car engine works, the complete DNA sequence does not tell us how the system directs its crucial functions, for instance how it controls and modulates the activity of the genes. Indeed, the regulation of activation or repression of genes and many other vital functions have been shown to be not merely controlled by the information contained in the 1-dimensional (1D) DNA sequence, as they are encrypted somehow in the way our chromosomes are folded in 3D space of the cell nucleus [1–5]. As a striking example, our genes often need to contact specific DNA regulatory sequences to be activated, which, however, are in most cases located at significant genomic distances from them (e.g., millions of bases away). Such a long-distance gene control can thus be established only through the folding of the genome in 3D space [2,4], whose disruption can lead to gene misexpression thereby causing important diseases, such as congenital disorders and cancers [6–12].

One of the major challenges nowadays is deciphering the fundamental, underlying rules of the game: how do distal elements (e.g., genes and regulators) find each other inside the dark crowded nuclear environment? What is the physical mechanism shaping those interactions? What are the molecular principles driving chromosome 3D structure? Answers to these big questions are starting to emerge in recent years thanks to novel powerful technologies from molecular biology, which are providing more and more precise quantitative data on the structure of the cell nucleus, and to the development of theoretical models from physics, which are in turn gradually unveiling the mechanisms orchestrating genome 3D structure.

From the experimental side, innovative methods in the last decade collectively propelled the study of 3D genome into a new quantitative era. Those technologies, such as Hi-C [13], SPRITE [14] and GAM [15], allowed for the first time to generate complete maps of genome interactions at the kilobase resolution (i.e., the typical length scale where interactions between genes and their regulators occur), revealing chromosome conformation with unprecedent detail. For instance, a major breakthrough emerging from the data is that our genome is self-organized into a formidable 3D architecture at multiple levels: at the lowest scales, it is partitioned into highly conserved self-interacting domains, called TADs (topologically associating domains) [16,17], which can also associate with each other into higher-order structures to form a much more entangled hierarchy of

domains-within-domains (e.g., “meta-TADs”) extending across genomic scales up to the range of entire chromosomes [18]. Additionally, recent advancements of microscopy-based techniques, such as super-resolution multiplex FISH imaging approaches, are enabling to visualize chromosome 3D structure with nanometer-scale precision in individual nuclei (i.e., in single DNA molecules) [19–23]. Those methods have revealed, for example, that TADs exist as physical 3D globular structures in single-cells, albeit they display a high degree of heterogeneity from cell-to-cell. In other words, if we take a high-resolution picture of our genome in thousands of individual cells of the same type, the same chromosomal regions will appear as different blob-like 3D conformations broadly varying from one to the other cell [21]. Again, what is the physical mechanism leading to such organization? How are those globules established? What is the origin and nature of the single-cell structural variability? Those are some of the challenging and fascinating questions arising in modern biology that we can try to tackle, and hopefully answer to, by using the quantitative methods of physics.

From a physics point of view, DNA is basically a long polymer molecule in which we expect general principles from polymer physics and statistical mechanics to apply. Indeed, different mechanisms of folding have been recently proposed and investigated by models relying solely on fundamental physical processes [9,24,33–42,25,43–48,26–32] or via computational approaches [49,50,59–61,51–58]. Within this dynamic research context, we developed principled models of interacting polymers from statistical mechanics to investigate the mechanisms of genome folding and contact formation. We proposed a basic scenario where physical spatial proximity (i.e., a “contact”) between distal DNA sites (for instance, a gene and its regulator) results from attractive interactions mediated by diffusing cognate particles, corresponding to different molecular species (such as biological Transcription Factors and proteins), that can bridge those sites. We transposed this folding picture into a simple polymer physics model, called the “Strings and Binders” (SBS) model [25,38,43,45], where a chromosome filament is represented as a self-avoiding polymer chain along which different binding sites are located for cognate diffusing bridging molecular binders. As dictated by polymer physics [62], the system folds in just a few conformational classes that correspond to its thermodynamic phases. The SBS model (as well as others within the same universality class) envisages, for example, two main folding classes: the “coil” state, where the polymer is randomly folded because of prevailing entropic effects, and the “phase-separated globule” state, where distinct globules spontaneously self-assemble along the chain by the interactions of cognate binding sites. Upon increasing the number (or affinity) of binders, the system switches from one to the other state via a phase transition mechanism of polymer phase-separation [38]. By determining the thermodynamics phases of the system, which can be predicted by physics, we can derive the full ensemble of 3D conformations where it spontaneously folds into and therefore test how reliable is the folding mechanism envisioned by the model by comparing its predictions against real experimental data.

The research presented in this thesis is framed within this recent and highly interdisciplinary scientific field at the border between physics and the latest findings of DNA molecular biology. All the studies and investigations here discussed have been conducted in the last three years in the physics department of University of Naples “Federico II”, under the supervision of Professor Mario Nicodemi in the research group of Complex System. Most of the results reported in this work are

published and, in some instances, involved international collaborations, e.g., with the Epigenetic Regulation and Chromatin Architecture group directed by Prof. Ana Pombo at Max Delbruck Centre for Molecular Medicine (Berlin) or research groups within the US 4D Nucleome (4DN) Research Project. In particular, in the 4DN consortium, we collaborated with Professor Bing Ren at Ludwig Institute for Cancer Research and University of California (UCSD, San Diego) to understand, based on our polymer physics approach and computer simulations, the role of specific molecular factors (e.g., CTCF proteins) in promoting long-distance gene regulation and shaping chromosome structural patterns. The results of these studies are not discussed here for brevity and have been recently published in *Nature Structural & Molecular Biology*, **28**, 152–161 (2021) and *Nature Genetics*, **53**, 1064–1074 (2021).

In the present work of thesis, we aim to tackle two major research lines, respectively examined and conveyed in our published papers *Nature Communications*, **11**, 3289 (2020) and *Nature Methods*, **18**, 482 (2021): ***i)*** identifying the molecular mechanisms shaping chromosome folding at the single DNA molecule level based on the predictions of polymer-physics-grounded approaches, such as the SBS model; ***ii)*** using the polymer conformations predicted by the theory as a simplified, yet fully controlled, reference system where to assess, for the first time, the intrinsic limitations and advantages of some of the most powerful technologies nowadays available to probe genome structure, thus providing a blueprint in designing novel experiments. A brief outline of the thesis follows below.

The thesis is divided in four chapters. In **Chapter 1**, we try to briefly introduce the reader through the emerging complex picture of chromosome spatial organization. Starting from basic definitions of DNA molecular biology, we next outline innovative recent technologies, such as Hi-C, GAM and SPRITE, for measuring genome-wide chromosome physical interactions. We focus on some of the major findings of those experiments, such as the discovery of TADs and other functional structures. We also examine recent strides from microscopy-based approaches, which allowed to measure at the few nanometers scale the physical 3D structure of chromosomes at the single-molecule level (i.e., in individual single-cells). Finally, we briefly review two popular recently proposed physical mechanisms of chromosome folding. In **Chapter 2**, we focus on the class of polymer models based on thermodynamic phase-separation. Particularly, we introduce with all details the SBS model, in which physical contacts between distal sites along the genome are achieved via diffusing cognate molecular binders that can bridge those sites via mechanisms of equilibrium polymer thermodynamics. We describe a statistical inference method (PRISMR [9,38]), which combines machine learning and polymer physics, to infer the model binding sites. Single-polymer 3D conformations are generated by performing massive Molecular Dynamics (MD) simulations up to reach stationarity. The details of the PRISMR algorithm and the MD implementation of the model are discussed in dedicate sections. To exemplify the typical accuracy of our models, we discuss an application to real data, i.e., the modeling of the *HoxD* gene region, which is particularly interesting because it is involved in limb development, and we highlight the power of the SBS model in explaining the formation of complex structural patterns as observed in the experiments. In **Chapter 3**, which is based on the results of our paper [38], we investigate the physical mechanisms of folding at the single-molecule level. We use the polymer 3D conformations of the SBS model to make

predictions on genome structure that we validate against independent microscopy (i.e., multiplex FISH imaging) single-cell data available in different human cell types [21]. We show that chromosome folding is controlled at the single-molecule level by a thermodynamic mechanism of polymer phase-separation and also provide a theoretical rationale to explain the origin of the observed single-cell structural variability. Next, we investigate how single-molecule 3D structures are affected upon removal of the cohesin complex, which is a known chromosome organizing factor, and our model predictions are consistently validated against single-cell imaging data. Finally, we explore the steady-state time dynamics of the polymer conformations predicted by the theory and discuss how contact specificity can be achieved in the stochastic nuclear environment. In **Chapter 4**, which summarizes the results published in [40], we show a different application of polymer physics models, which can indeed be used not only to make sense of experimental outcomes, but also to test the quality of the data measured by experimental technologies. To this aim, by using a validated ensemble of SBS single-polymer 3D structures as benchmark, we perform the first quantitative comparison of the performance of three different DNA powerful technologies, currently employed in the field, i.e., Hi-C, SPRITE and GAM. Our analyses, fully detailed in the sections of the chapter, clarify how state-of-the-art genome technologies faithfully represent chromosome 3D structure, highlighting the different experimental conditions where each of them is most effective.

1. Chromosome organization in the nucleus

Powerful technologies from molecular biology are revealing that chromosomes have a complex 3D architecture within the cell nucleus involving an intricate hierarchy of physical genome interactions that serve functional and vital purposes, such as gene regulation. In this Chapter, starting from the naïve textbook view of DNA packaging in the cell nucleus up to reach the latest findings encompassing comprehensive maps of genome-wide interactions, we aim to briefly introduce the reader through the major technological and conceptual breakthroughs of the last decade that are propelling our understanding of chromosome 3D organization. In Section 1.1, we briefly recall some basic concepts of DNA molecular biology and also more recent discoveries such as the formation of chromosome territories. In Section 1.2, we discuss the advent of innovative chromosome conformation capture (3C) technologies [63–65], which are based on high-throughput sequencing to detect chromosome contacts. A particular focus is on the Hi-C method [13], which allowed to generate comprehensive maps of genome interactions at the kilobase resolution (i.e., the typical length scale where interactions between genes and their regulators are established). We also describe two further recent experimental methods, i.e., SPRITE [14] and GAM [15], which even improved many of the Hi-C limitations. In Section 1.3, we highlight the striking findings about chromosome 3D organization emerging from the quantitative data provided by those novel technologies. We will see, for instance, that our genome is partitioned into subsequent large (megabase-sized) functional physical blocks within which genomic interactions are particularly enriched; those domains, called TADs (topologically associating domains) [16,17], can also interact with each other to form higher-order structures, giving rise to a formidable and complex hierarchy of domains-within-domains ranging from the kilobase up to the whole chromosomal scale [18]. In Section 1.4, we describe a different experimental methodology to probe genome structure, which is based on microscopy techniques that allow to measure 3D spatial distances in the targeted genome [19–23]. Specifically, we discuss the recent advancements achieved in this field and focus on a recent super-resolution multiplex FISH (fluorescence in-situ hybridization) imaging approach that enabled for the first time to visualize chromosome 3D structure with nanometer-scale precision in individual nuclei (i.e., at the single-cell level) [21]. Such advances have revealed, for example, that TADs and other structures exist as spatially segregated globular 3D conformations in single-cells, yet with a high degree of stochastic structural variability from cell-to-cell, hence raising questions on the fundamental mechanisms underlying their origin and formation. Models from polymer physics have been developed to quantitively understand the molecular principles controlling genome 3D architecture; to this aim, in Section 1.5, we briefly review two important physical mechanisms of chromosome folding that have attracted much attention in recent years.

1.1 Basic concepts of DNA molecular biology

Far from being comprehensive, this Section briefly provides the reader with the basic terminology of DNA molecular biology. An extensive, fully detailed description of the concepts here summarized can be found in classic textbook, such as Alberts et al. *Molecular Biology of the Cell* and Watson et al. *Molecular Biology of the Gene*, along with the papers and reviews cited below.

1.1.1 A textbook view on DNA packaging in the nucleus

In eukaryotes, i.e., higher organisms (from humans down to the unicellular baker yeast), the genetic information is encoded in DNA (deoxyribonucleic acid) molecules located within the cell nucleus. A DNA molecule is a double helix made of two paired long polymer strands composed of simple units called "nucleotides", which comprise a five-carbon sugar, a phosphate group, and a nitrogenous base. The polymer backbones are made of sugars and phosphate groups joined by ester bonds (covalent bonds with energies $\sim 10^2 K_B T$ at room temperature), while the nitrogen bases of the two separate polynucleotide strands are bound together with hydrogen bonds (energies $\sim$1-10$K_B T$) thus making double-stranded DNA (**Fig. 1.1**, leftmost panel). There are four distinct types of bases in DNA, i.e., adenine (A), cytosine (C), guanine (G), thymine (T), which are bound together via the base pairing rule: adenine (A) only binds the opposite thymine (T), while cytosine (C) only binds guanine (G), with 2 and 3 bonds respectively. Hence, the sequences of the two strands are entirely complementary. The sequence of complementary base pairs along a DNA molecule represents its genetic code, i.e., the information required to build all the compounds (proteins or RNA) essential for life and functioning of the cell.

In cell nuclei, DNA is divided into linear filaments called "chromosomes". For example, human cells are diploid, as they include two copies ("alleles") of each chromosome; overall, they comprise 23 pairs of chromosomes (46 in total), formed of 6.4×10^9 base pairs (bp), which would stretch the distance between the earth and the sun and back again about 100 times if aligned one after the other [66]. Similarly, a common striking estimation is that the linear length of human genome is roughly 2m while the nucleus in which it is constrained has a diameter of 10-15μm, that is equivalent to wrap 40km of a thread into a tennis ball. As a result, DNA must be tightly packed within the cell nucleus and, to orchestrate such peculiar folding, numerous specialized DNA-binding structural proteins are involved. The complex of DNA and the proteins that organize it is called "chromatin". As known in textbook biology [67,68], chromatin is organized into different layers of increased compaction (**Fig. 1.1**). At the first level, DNA wraps around nucleosomes, which form the basic 11nm-wide repeating unit of eukaryotic chromatin. In a nucleosome, 146 bp of DNA are wound around an octamer core complex made of eight proteins called histones (they are named H2A, H2B, H3 and H4 and two molecules of each histone form the octamer). Nucleosomes provide roughly a 7-fold reduction of DNA linear length, resulting in a "beads-on-a-string structure" as they appear as beads located along a string of loosen linker DNA. At the next step of compaction, nucleosomes are packed on the top of each other by another specific protein, known as histone H1, which condenses chromatin into a 30nm fiber. However, over the years there has been a great deal of speculation concerning this structure, as the 30nm fiber has been observed only in-vitro [69], i.e., in experiments with controlled and artificial conditions, casting doubts on its existence in real cells [24,70,71]. From this organizational level on, the details of folding remain still uncertain. In a traditional and ultra-simplified picture, a much greater compaction occurs via different proteins that further compress and arrange chromatin into higher-order fibers up to the scale of the entire chromosome (**Fig. 1.1**). In the following (see Sections 1.2-1.4), well beyond the textbook scenario described here, we will discuss very recent technological advancements from molecular biology that are allowing to probe

higher-order chromatin structure with unprecedent detail, revealing that chromosomes are folded into a complex, non-random 3D architecture within the cell nucleus involving a hierarchy of extensive long-range, functional interactions.

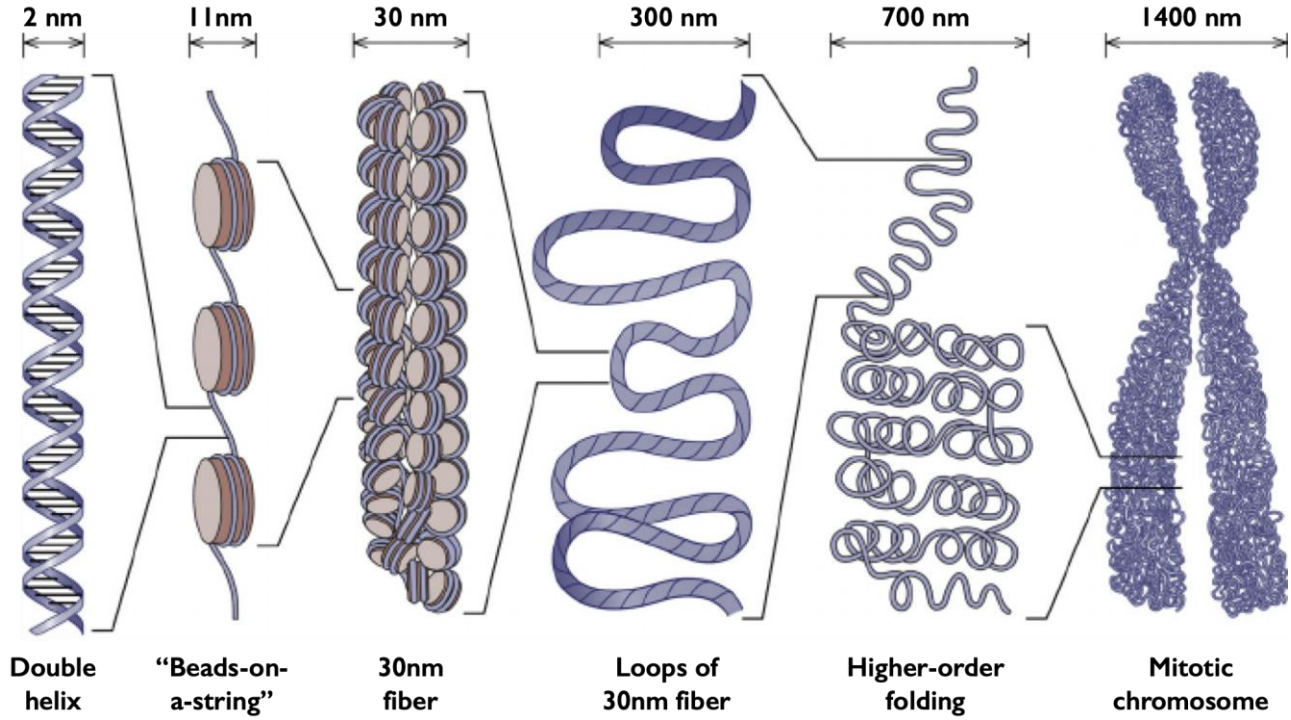

Figure 1.1: Packaging of DNA involves different length scales and degree of compaction. DNA is wrapped around a histone octamer to form nucleosomes, which are connected by stretches of linker DNA. This basic repeating unit is folded into a fiber-like structure of about 30nm in diameter. Those fibers are further compacted into higher-order structures, which, however, remain largely elusive in the classical textbook view, to form chromosomes (in the figure is sketched a chromosome during cell division, which has a typical hourglass-shaped structure). Adapted from [72,73].

1.1.2 Genes and regulators engage in long-range physical contacts

Along the DNA sequence are located the genes, i.e., specific DNA segments carrying the genetic information to produce proteins and other compounds (e.g., RNAs), which are the building blocks essential for the vital cell functions. Genes are transcribed by polymerase enzymes into RNA (a single-stranded ribonucleic acid), which by use of the genetic code is then translated into proteins. Such an ordered scheme whereby the information from genes is processed to final proteins is known as the "central dogma" of molecular biology. Humans have roughly 22000 protein-coding genes, which strikingly comprise less than 2% of our genome length. The remaining 98% of our DNA is non-coding, i.e., it is not made of genes that encode proteins. For that reason, as far as back as the 1960s those non-coding regions were collectively termed as "junk DNA" because regarded to serve no purpose. Such a picture is dramatically changed, as we understand nowadays that much of the DNA not used to encode proteins is crucial to regulate the activity of the genes. In fact, not all the genes of the genome need to be expressed (i.e., transcribed to produce a protein) altogether during the life of the cell, as some may be required only at specific moments or because of external stimuli and conditions; also, in multicellular organisms each cell is specialized in tissue-specific functions, thus

only the subset of genes controlling those functions must be expressed. Hence, how does the non-coding DNA regulate such a complex and highly diverse gene expression?

Over the last decade much has been learned about how this is achieved, revealing the critical role of the spatial organization of the chromosomes (see, e.g., excellent reviews on the topic such as [1–5]). For instance, based on microscopy experiments and other biochemical techniques, it has been found that gene activity is often modulated by regulatory non-coding elements that can be located from few kb up to as much as several Mb (Mb=10^6 bp) away from the gene promoter (a promoter is a control DNA sequence typically located within 1kb upstream of the starting site of gene transcription). Examples of regulators are enhancers and repressors, which are involved, respectively, to activate and inhibit gene expression. One mechanism through which distal regulatory elements, e.g., enhancers, can control and activate genes that are located far away in the genome is based on long-range physical interactions that bring the two parties into close spatial proximity (**Fig. 1.2**). If the human genome is regarded only as a linear 1D string, it is hard to imagine mechanisms whereby an element could regulate a target gene located million bases away along the sequence. Therefore, an essential component of the long-distance regulation is the third dimension of chromosome conformation, as chromatin needs to fold into specific 3D structures to achieve physical proximity between regulators and target genes (**Fig. 1.2**). Experimental evidence indeed supports this scenario, pointing to direct molecular association as a means for long-range 3D communication [74–77]. Also, other mechanisms of transcriptional control have been observed where, for instance, physical communication between genes and regulators is not direct yet mediated by diffusible proteins (such as Transcription Factors and many others) that can facilitate or hinder gene expression [2,78–82].

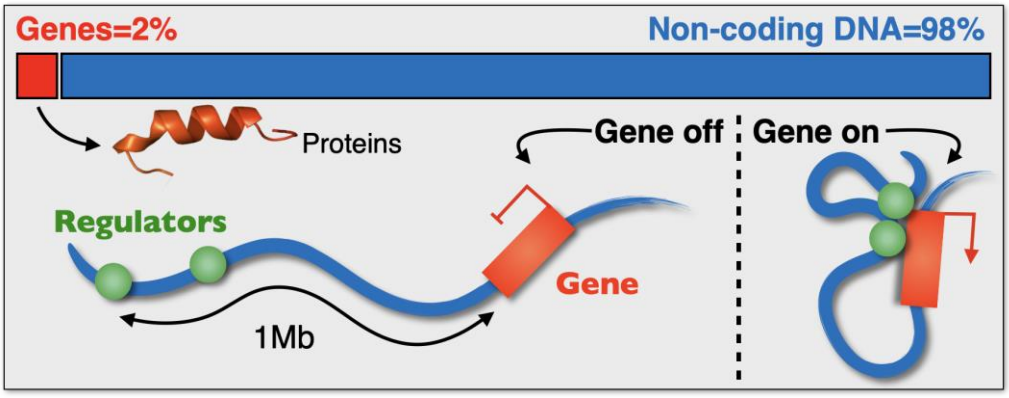

Figure 1.2: Gene regulation requires a complex 3D chromosome organization. Less than 2% of our genome length comprises genes, i.e., specific DNA sequences that encode proteins and other functional compounds. The remaining 98%, regarded decades ago as "junk DNA", performs many crucial tasks for the life of the cell, such as gene regulation. In those non-coding regions are located specific sequences that can regulate gene activity by coming into close spatial proximity with gene itself. One striking feature of gene regulation in mammalian genome is that regulatory elements can be distant from their target gene (e.g., a million bases away as shown in the figure), hence requiring DNA to loop in 3D space to enable gene regulation.

Taken together, those findings have revealed that while genetic information is written into the DNA 1D sequence, the regulation of activation or repression of genes (as much as many other functional

activities) is controlled by the 3D architecture, i.e., by the way our genome is folded in 3D space of the cell nucleus. Major challenges now, as we will discuss in the next chapters, include deciphering the molecular mechanisms shaping chromosome 3D structure.

1.1.3 Epigenetic modifications and the histone code

Regulatory elements can be typically identified by their high evolutionary sequence conservation or by specific-associated "epigenetic" marks. Epigenetics is the study of heritable changes in gene activity that arise in the absence of alterations in the DNA sequence; examples are chemical modifications of DNA or interactions of DNA with molecular factors that do not alter the underlying nucleotide sequence. Those modifications have been shown to play a role in controlling gene transcription. For instance, methylation (i.e., the addition of a methyl group) to cytosine at CpG sites (regions of the genome where cytosine is followed by guanine) are frequently observed at gene promoters with repressive functions. Histones are also subject to chemical modifications, particularly their tails can be modified by acetylation, methylation, phosphorylation, ubiquitylation, and many other processes. Those modifications can generate synergistic or antagonistic interaction affinities for chromatin-associated proteins, which in turn dictate dynamic transitions between transcriptionally active or silent chromatin states by affecting the DNA chemical accessibility. Thus, the combinatorial nature of histone modifications has been thought to reveal a "histone code" (see, e.g., [83]) linked to activation or repression of transcription. For example: H3K4me3 (histone H3 lysine 4 trimethylation) is associated with promoter regions, H3K4me1 (histone H3 lysine 4 methylation) and H3K27ac (histone H3 lysine 27 acetylation) mark enhancer regions, H3K36me3 (histone H3 lysine 36 trimethylation) signals transcribed regions, H3K27me3 (histone H3 lysine 27 trimethylation) repressed or poised regions, and so on with many other examples. The above picture further illustrates the complexity of genome regulation in eukaryotes: chromatin states are controlled by a combination of different factors and the epigenetic signature of chromatin 3D structure can help to understand its functional meaning.

1.1.4 Discovery of chromosome territories

The recent progress toward the comprehension of the structure of chromosomes grows out of discoveries made in the 1980s concerning their spatial organization at the nuclear scale (1-10µm). While during cell division chromosomes are duplicated and highly compact into a typical hourglass-shaped structure, early researchers proposed two main models regarding the way in which they were likely organized within the nucleus when cells are not dividing [84]: *i)* the first model, originally proposed by Carl Rabl in 1885 and known as the "chromosome territory model", envisages the scenario where the DNA of each chromosome occupies a defined volume of the nucleus and only overlaps with its immediate neighbors; *ii)* the second or "spaghetti" model, by contrast, considers a different picture where chromosomes are randomly located in the nucleus, therefore being largely intermingled and entangled with each other. The key experiment to discriminate between the two models was eventually realized in the early 1980s by Thomas Cremer, a German cell biologist, and his physicist brother, Christoph Cremer [66,85,86]. They reasoned that the two models made very distinct predictions upon inducing, for instance, an external perturbation on cell nuclei. Specifically,

they argued that if chromosome territories existed, then a laser beam oriented onto a particular area of the nucleus would have branded only a few chromosomes; on the contrary, if the spaghetti random model were correct, then the laser pulse would have struck many more chromosomes. The experiment showed that only a few chromosomes per cell were damaged by the laser light, hence strongly supporting the chromosome territory model. The existence of those territories was directly confirmed a few years later by the development of the fluorescence in situ hybridization (FISH) technique in which fluorescently labeled probes complementary to a specific chromosome are used to visualize a given chromosome in the nucleus. These experiments, by allowing a direct visualization, demonstrated that chromosomes exist in the nucleus as distinct entities, each occupying a space (typically roughly spherical in shape and 2–4μm in diameter [87]) separate from other chromosomes (**Fig. 1.3**).

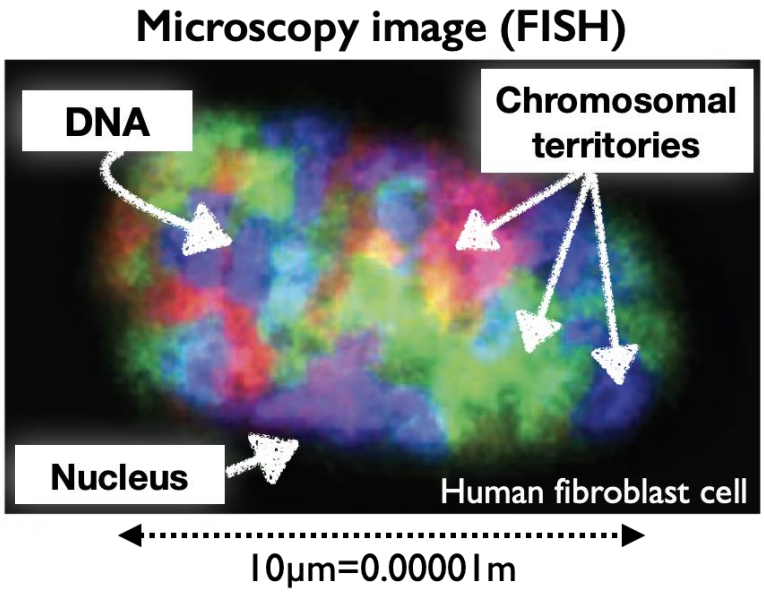

Figure 1.3: Chromosomes form discrete and distinct territories at the nuclear scale. Advanced microscopy techniques have disproved in the last decades a longstanding view whereby chromosomes are randomly positioned in the cell nucleus, like "spaghetti" in a bowl. This image from microscopy FISH experiments individually colors the chromosomes in the nucleus of a human fibroblast and shows that they occupy distinct nuclear regions called "chromosomal territories". Hence, chromosomes fold into a complex 3D structure and their arrangement within the nucleus is far from random. Adapted from [66,86].

Despite this important advancement, a key limitation of the Cremers' experiment was its spatial resolution (~μm, i.e., comparable with the nuclear size), which hindered to probe chromatin structure at the lower and finer sub-chromosomal scale where, e.g., contacts between genes and their regulators are typically established. In the last decade, major developments have been performed to overcome these challenges by the invention of new revolutionary technologies as discussed in the following sections.

1.2 Mapping genome structure via sequencing-based methods

A crucial breakthrough in chromatin biology was the development of chromosome conformation capture (3C) methods, which marked the beginning of the era of high-throughput sequencing-based

techniques for the investigation of chromosome conformation at the kilobase resolution, i.e., the typical scale where gene-regulator interactions occur. Some of those methods, such as 3C [63], 4C [64], 5C [65], will be outlined in subsection 1.2.1 along with the type of quantitative data that they typically provide. In the subsection 1.2.2, we focus on the Hi-C method [13,88], which has been the first genome-wide adaptation of 3C technologies and is currently one of the most influential and powerful ligation-based method to dissect genome 3D organization. Finally, in the subsection 1.2.3 we discuss two recent ligation-free technologies, called SPRITE [14] and GAM [15], which do not rely on biochemical ligation processes and have been shown to overcome many of the intrinsic limitations of Hi-C experiments, such as the detection of multiway chromatin interactions (i.e., interactions between multiple DNA sites) beyond mere pairwise contacts.

1.2.1 Chromosome conformation capture (3C) technologies

Chromosome conformation capture (3C) technologies identify the pairwise contact frequency between pairs of DNA sites, i.e., the frequency (across a population of cells) with which any pair of sites in the genome is in close enough spatial proximity (say, in the range of 10–100 nm) to become crosslinked (i.e., bound together by chemical molecules such as formaldehyde) [89]. Albeit different 3C methods exist, their protocols have some common steps. In brief: *i)* a population of cells is crosslinked, typically with formaldehyde, to covalently link chromatin segments that are in close physical proximity; *ii)* chromatin is digested, i.e., fragmented, by sonication or via specific restriction enzymes; *iii)* crosslinked fragments are ligated to form unique hybrid DNA molecules that contain the fragments that were close in space in the original cell population; *iv)* DNA is purified and *v)* analyzed [89]. The difference among the distinct 3C-based technologies rely on the specific biochemical essay used to detect the ligation product. The most common technologies are the 3C [63], 4C [64], 5C [65] and Hi-C [13]. Below, we concisely discuss the type of datasets produced as output of those technologies, while all technical, experimental details of their protocols can be found in the referenced papers.

3C and 4C methods generate single profiles of interaction for specific, individual loci. For instance, 3C is particularly suited to provide the long-range (>100kb) interaction profile of a gene promoter or another genomic element of interest (e.g., gene regulators) versus chromatin in genomic proximity ("one-versus-one", **Fig. 4a**); 4C, instead, detects genome-wide contacts formed with a single locus that acts as a fixed "viewpoint" ("one-versus-all", **Fig. 4b**). In both cases, the data can be represented as single tracks that can be plotted along the genome 1D sequence and compared with other interesting genomic features (such as the positions of genes or their regulators). 5C and Hi-C methods are not constrained to a single locus of interest, as they generate matrices of interaction frequencies that can be represented as 2-dimensional heat maps with genomic positions along the two axes. 5C typically identifies in parallel the interactions between two large sets of loci (up to tens of Mb), e.g., between a set of gene promoters and their distal regulatory elements ("many-versus-many", **Fig. 4c**). Hi-C, which is the first genome-wide version of 3C, provides a true all-by-all genome-wide interaction map and will be described with more detail in the next subsection.

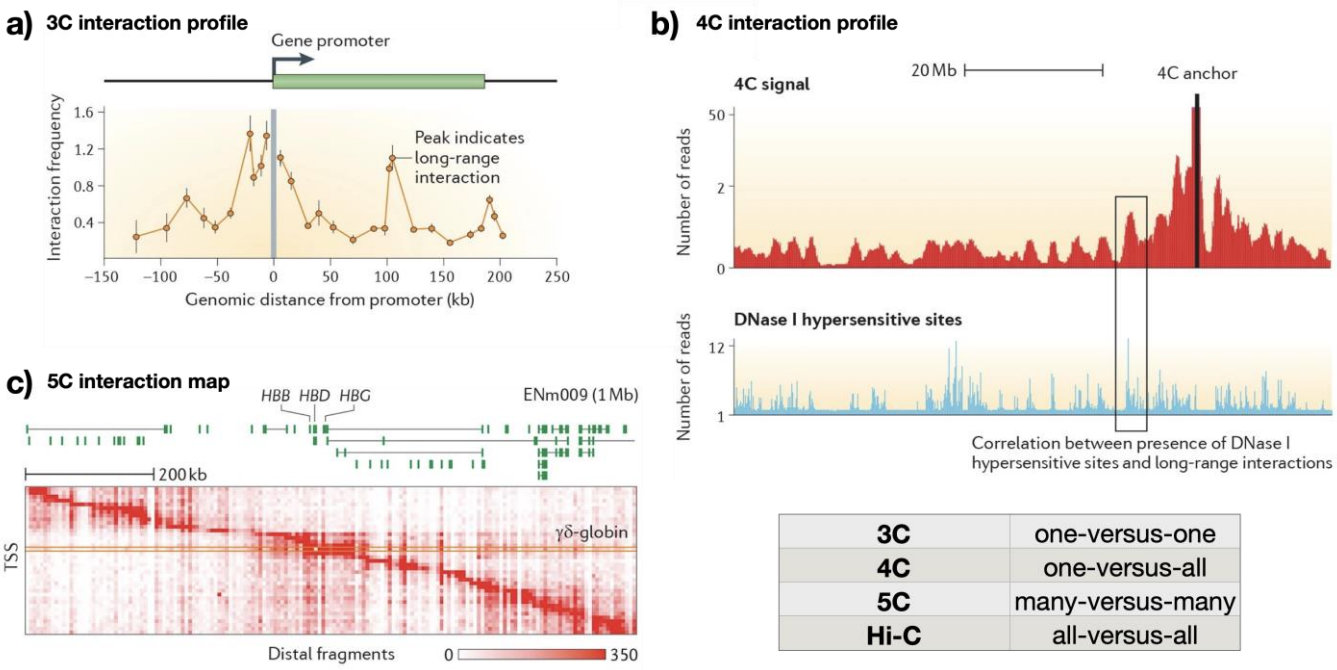

Figure 1.4: Examples of datasets produced by 3C technologies. a) Example of chromosome conformation capture (3C) data for the *CFTR* gene in Caco-2 cells (a human colon adenocarcinoma cell line). On the x-axis is reported the genomic distance in kilobases (kb) from the anchor point (or viewpoint). The highlighted peak indicates a strong long-range interaction involving the considered viewpoint and a 100kb distant site. **b)** Example of 4C data from the mouse genome (top). The 4C track is compared with DNase I hypersensitive sites (bottom), which are hallmarks of gene regulatory elements or genes. The highlighted box indicates a positive correlation between the two tracks, showing that the considered 4C anchor has a strong long-range interaction with a gene or regulatory element in that region. **c)** Example of a 5C interaction map for the ENCODE ENm009 region in K562 (a human erythroleukemia cell line). The different rows contain an interaction profile of a transcription start site (TSS) in the 1 Mb region on human chromosome 11 encompassing the *β-globin* locus. Adapted from [89].

1.2.2 Hi-C method for comprehensive maps of genome-wide contacts

To comprehensively assay chromatin contacts, Hi-C relies upon a specific chain process including as major steps the crosslinking, digestion, biotinylation, ligation and sequencing of DNA fragments [13] (**Fig. 1.5a**). First, a population of cells is crosslinked with formaldehyde molecules, which tightly bridge via covalent bonds DNA sites that are close in space (distance range 10-100 nm), hence "freezing" chromatin contacts at a given time. Then, nuclei are lysed and DNA is digested, i.e., cut into fragments, by a specific restriction enzyme (for instance HindIII in **Fig. 1.5a**). Hi-C resolution is intrinsically linked with the use of specific restriction enzymes, as the median length of a DNA fragment can depend on their type, ranging, e.g., from hundreds of bp to units of kb [90]. Once chromatin from all cell nuclei is reduced to a set of crosslinked DNA fragments, the sticky ends of those fragments are filled with biotinylated nucleotides. That enables the subsequent ligation process, where the biotinylated ends are ligated, i.e., bound together, by DNA-ligase enzymes, resulting in hybrid ligation products made of pairs of DNA fragments that were in spatial proximity (i.e., in "contact") in their original nucleus. Finally, DNA is purified, ligation products are isolated,

collected in a Hi-C library and then sequenced along the genome, thus producing a list of interacting pairs of fragments.

Hi-C data are organized in a genome-wide contact matrix by dividing the genome into windows (also called "bins") of fixed length (**Fig. 1.5b**). That length determines the resolution of the Hi-C experiment and in the first version [13] it was 1Mb (current resolutions are units of kb); as the typical length of a chromosome is hundreds of base pairs, the first Hi-C matrices comprised roughly hundreds of bins. Each entry (*i*, *j*) of the contact matrix is the number of ligation products (i.e., the number of contacts) between the DNA segments *i* and *j* and, by construction, the matrix is symmetric. Therefore, by measuring the contact frequency between any pairs of sites along a chromosome, Hi-C reflects an ensemble-average of the interactions in the original sample of cells and, because contacting fragments are naturally close in space, it provides a proxy to study the spatial organization of the genome and, in particular, interactions occurring between functionally relevant elements (e.g., genes and their regulators). Hi-C contact matrices are typically visualized as 2-dimensional heatmaps, with higher intensity colors indicating higher contact frequencies (**Fig. 1.5b**).

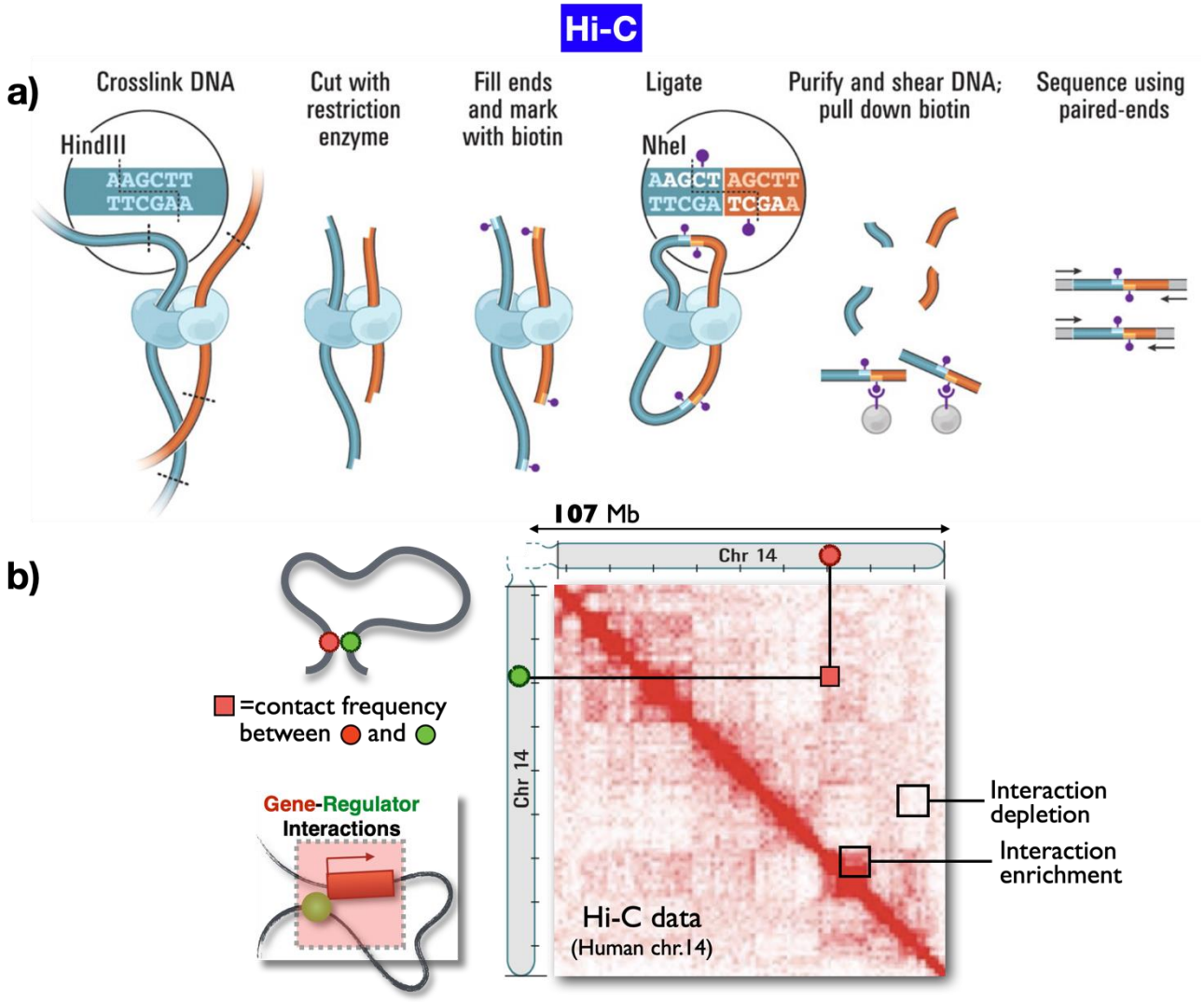

Figure 1.5: The Hi-C method provides genome-wide interaction maps. a) Overview of the protocol (see also the text). Cells are cross-linked with formaldehyde, producing covalent bonds between spatially close

chromatin segments. Chromatin is then digested into fragments by a restriction enzyme and the resulting ends are filled with biotin. Those ends are ligated, thus forming hybrid molecules made of pairs of chromatin segments in contact (i.e., spatially close) in the original nucleus. Finally, DNA is purified and ligation products sequenced along the genome. **b)** Hi-C data are organized into a symmetric pairwise contact matrix, whose axes represent the genomic coordinates of a considered chromosome (or of a smaller region, called "locus", along the chromosome). In this example, we report the Hi-C matrix of the entire human chromosome 14 (data from [13]) and the axis coordinates range through all the chromosome length (which comprises around 107Mb). Each entry (*i*, *j*) of the matrix is the number of contacts measured across the cell population (i.e., the contact frequency) between the DNA segments *i* and *j*. In the figure, for example, the highlighted red square is the contact frequency between the red and green circle sites, e.g., between a gene and its regulator. The matrix is typically plotted as an heatmap, where stronger colors represent enriched interactions. As contacts are inversely proportional to spatial distances, Hi-C provides a proxy of the spatial organization of chromosomes and, importantly, allows to detect the interactions between specific pairs of sites such as genes and regulators. Adapted from [13].

The analysis of the Hi-C matrices led to most of the discoveries in 3D genome organization in the last decades, revealing, e.g., the existence of structural units underlying chromosome folding and the emergence of complex patterns of interactions at multiple length scales. We will tackle those relevant findings in the Section 1.3. We stress that the Hi-C protocol summarized here is the one described in its original version [13], yet modifications and methodological improvements have been later implemented to increase the method efficiency. For instance, in the original protocol, chromatin is highly diluted after nuclei are lysed (i.e., their membranes disrupted) to hinder spurious ligations between different crosslinked fragments. However, such a dilution was shown to be not very effective [88,90,91], thus prompting to a new, more efficient protocol, called *in-situ* Hi-C, where ligation is performed "in-situ" inside the nucleus (i.e., a constrained space) rather than in solution (where DNA fragments are floating freely) [88]. These advancements dramatically changed the exploration of chromatin architecture, enabling the collection of an impressive amount of contact data for different cell types and across different organisms. Also, they further pushed new experimental developments and techniques, such as single-cell Hi-C [90], regarded as the first successful sequency-based single-cell chromatin technology [91], or Micro-C [92,93], which even more increased the resolution of local contacts (resolution of $\sim$200 bp) as chromatin fragmentation occurs without use of restriction enzymes. Fast and numerous are the technical advances built upon the first Hi-C protocol, as well as more refined variants where Hi-C is combined with other techniques to capture only interactions, for example, mediated by a specific protein of interest (e.g., HiChIP [94] and PLAC-seq [95]) or enriched within a specific genomic location (e.g, Capture-C [96–100]). The interested reader can find more details in the referenced works. In the following, we will refer to the basic, standard scheme of Hi-C reported above (**Fig. 1.5**).

1.2.3 Two novel ligation-free methods: SPRITE and GAM

Albeit the tremendous advancements that allowed to map genome-wide chromatin organization, Hi-C and the other 3C-based technologies have intrinsic technical limitations, which are mostly linked to the ligation process envisaged by their protocol. In fact, ligation is only partially efficient [91] and, importantly, provides information only on pairwise contacts, as multiple DNA

interactions, which occur simultaneously within the nucleus and are expected to exist, are not detected. To address those technical challenges, two main methods that do not rely on ligation of chromatin fragments have been invented: split-pool recognition of interactions by tag extension [14] (SPRITE, **Fig. 1.6a**) and genome architecture mapping [15] (GAM, **Fig. 1.6b**). Here, we review the key features of those ligation-independent technologies, which are extensively discussed with all details in the referenced papers.

In SPRITE, as also in Hi-C, nuclei are first crosslinked. Then, they are isolated and fragmented by sonication, and finally digested to obtain individual crosslinked chromatin fragments of approximately 150-1000bp in length. Next, those fragments are uniquely barcoded using multiple cycles of a split-and-pool strategy (**Fig. 1.6a**). In brief, all the complexes of crosslinked DNA are first randomly split across 96 wells of a plate; a specific ("tag") sequence of nucleotides is added to each DNA fragment within each single well; the complexes of DNA are then pooled into a single well ("pool") and again randomly shuffled inside the wells and tagged. After several rounds of split-pool tagging, all fragments have a unique series of ligated tags (each assigned at each split-pool round), forming a specific barcode. It follows that the fragments within the same crosslinked complex have the same barcode, as they are forced to stay together during the shuffling process into the wells because covalently linked; on the contrary, DNA fragments in separate complexes sort independently resulting in distinct barcodes. Indeed, the probability that fragments belonging to two independent clusters receive the same barcode decreases exponentially at each additional round of split-pool tagging (for instance, after six rounds, the number of possible unique barcode sequences is hugely greater than the typical number of unique DNA fragments in the initial sample, 10^{12} vs. 10^{9}). Finally, DNA fragments are sequenced and the fragments belonging to the same cluster (i.e., those forming a contact) are univocally identified via their identical barcode. In this way, pairwise, but also higher order n-wise interactions, can be counted at the considered experimental resolution. In the original paper [14], SPRITE was shown to accurately recapitulate in mouse genome previously published Hi-C data [16,88], but also to provide additional insights on chromosome structure, such as the identification of functionally related interactions, missed by Hi-C, occurring across large genomic distances. Moreover, a single-cell version of SPRITE has been recently developed, allowing the study of multiple contacts (e.g., between genes and their regulators) in individual cells [101].

GAM is the first genome-wide and ligation-free method for capturing three-dimensional proximities between any number of genomic loci. Albeit based on a completely different strategy compared to SPRITE to assay genome organization, it identifies multiway interactions and higher-order chromatin structures as well as more local contacts. In GAM, fixed nuclei are embedded in a sucrose solution and frozen. Then, they are cryo-sectioned (i.e., a ~220nm slice is laser-cut from each nucleus at random orientation) and the DNA from each slice (called also nuclear profile or NP) is extracted, amplified and sequenced (**Fig. 1.6b**). In this way, DNA sequences present in each nuclear slice can be identified (they are said to have "segregated" in the NP). The basic idea is that sites that are closer to each other in the nuclear space are detected in the same nuclear profile more frequently than distant, non-interacting sites. Hence, the co-segregation of all possible pairs of sites among a

large set of nuclear section profiles is used to generate a matrix of inferred locus proximities, from which the frequencies whereby pairs, triplets, or n-plets in general, segregate within the same slice can be derived. Finally, by using statistical tools and mathematical models, such as SLICE [15] (StatisticaL Inference of CosEgregation), the interactions most likely to be specific and significant (i.e., non-random with respect to genomic distance) are identified from GAM co-segregation data. For instance, the application of GAM to mouse embryonic stem cells revealed an exceptional abundance of triplet (3-way) contacts across the genome, particularly between functional regions subject to high levels of transcription, providing novel insights in genome architecture that were unattainable with previously available techniques (e.g., Hi-C and its derivatives).

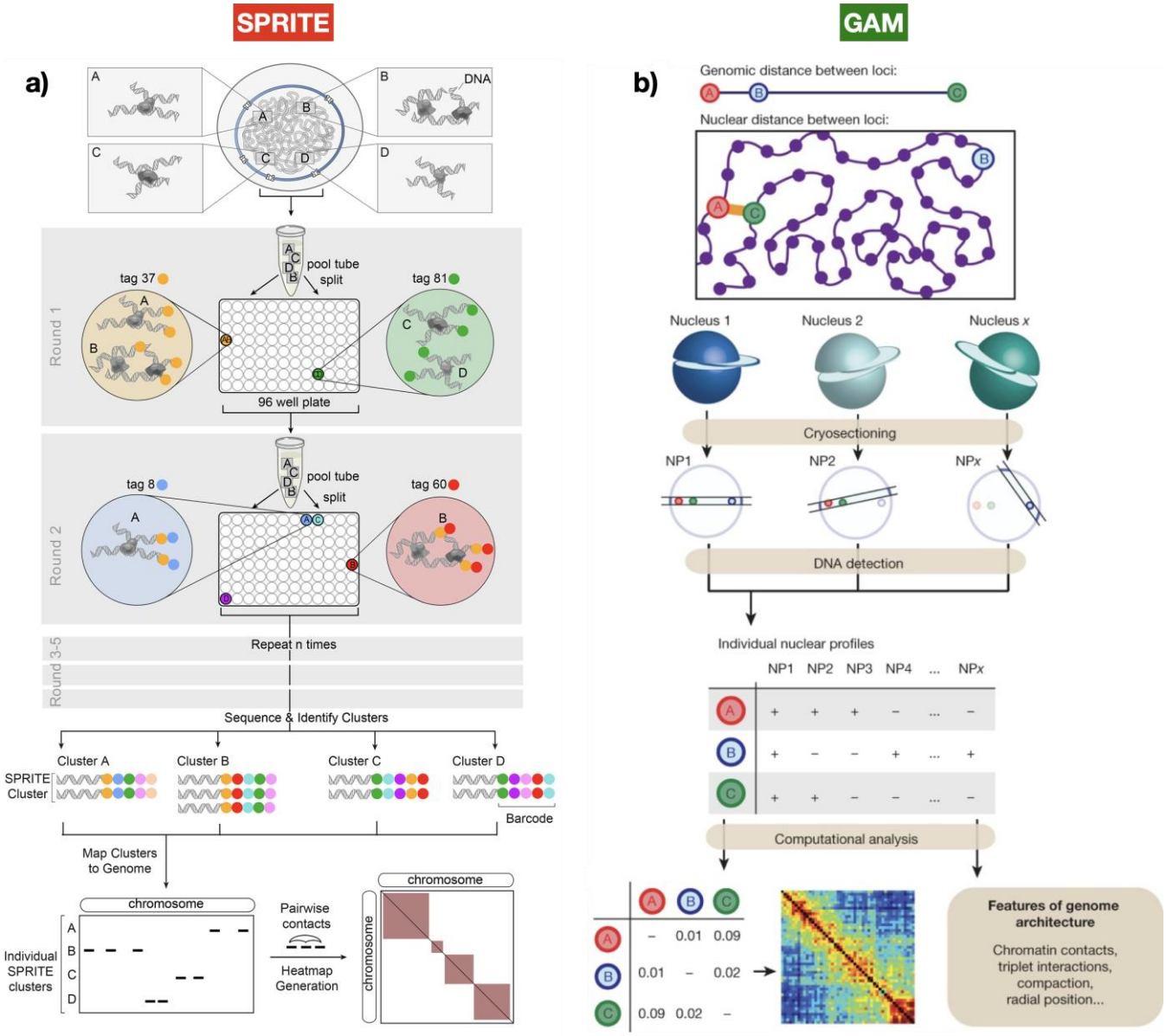

Figure 1.6: Overview of the SPRITE and GAM technologies. a) Outlook of SPRITE (see text): chromatin is crosslinked and digested into fragments, which are then uniquely barcoded using multiple cycles of a split-and-pool strategy. A specific ligated tag is assigned at each split-pool round to each fragment, so that after several rounds all fragments have a unique series of tags forming a specific barcode. Finally, DNA fragments are sequenced and those forming a contact (i.e., belonging to the same crosslinked cluster) can be easily detected because identically barcoded. Hence, the method naturally allows to detect interactions between

any number of genomic sites. Taken from [14]. **b)** Outlook of GAM (see text): fixed nuclei are cryo-sectioned into thin nuclear sections and the DNA from each slice (called also nuclear profile or NP) is extracted, amplified and sequenced. The basic concept is that physically proximal sites are found more frequently in the same NP than distant ones. Thus, by identifying the co-segregation of all possible pairs of sites among a large set of NPs, GAM allows to infer chromatin contacts and higher-order interactions. Taken from [15].

Overall, SPRITE and GAM both overcome the intrinsic limitations of ligation-based technologies (such as Hi-C), allowing, for instance, the detection of multiway contacts where multiple DNA sites simultaneously interact within the nucleus. By focusing on the pairwise level, Hi-C, SPRITE, and GAM all provide as final output a contact matrix, yet they are difficult to compare because contacts are measured differently in the different cases: Hi-C returns the number of ligation products, SPRITE the number of pairs belonging to the same crosslinked cluster, while GAM produces co-segregation data. A rigorous test of their performance would be indeed useful to assess the best technology to use under different experimental conditions. However, the considerable waste of resources required to implement the three methods and the absence of an independent benchmark make such a comparison extremely difficult to realize in real world. To overcome this challenge, we will discuss in **Chapter 4** a computational approach that we designed to compare in-silico, via models from polymer physics and computer simulations, the performance of those technologies [40].

1.3 Patterns in contact matrices

In this Section, we provide a basic, yet fundamental, bird's-eye view on the complex picture of chromosome 3D organization as emerging from the major experimental findings of the last decade. The analysis of Hi-C data, for instance, revealed the existence of specific patterns of contacts extending from the sub-Mb scale up to the whole chromosome range [2,13,88]. Hi-C maps typically appear as block-like matrices made of subsequent large domains within which chromatin sites tend to interact with each other more frequently than with sites from different domains. Those domains, which are hundreds of kilobases to several million bases in length, are called TADs (topologically associating domains) and they are regarded as the basic units of chromosome folding [3]. However, contacts at larger scales are also found, as TADs can interact with each other resulting into hierarchical higher-order structures ("meta-TADs") that span up to the entire chromosomal scale [18]. Above the mega-base scale, chromosomes are partitioned in two major spatial compartments, called A and B, which have been shown to be functionally related to active transcribed and repressed chromatin states, respectively [13]. Furthermore, all those large-scale structural features have been confirmed by other independent technologies, e.g., SPRITE [14] and GAM [15], and recently observed also at the single-cell level by high-resolution microscopy-based techniques (see Section 1.4). Finally, recent findings (see, for instance, [6–9,102]) have shown that the disruption of this complex higher-order chromatin organization, encompassing TADs and other structures, can result into spatial rearrangements altering, for example, the normal contacts between genes and their regulators, thereby affecting gene expression and causing diseases, such as congenital disorders [10] and cancers [11,12]. That proves the functional role of chromosome spatial organization and crucially highlights the importance of understanding the molecular and physical principles underlying its formation.

1.3.1 A/B compartments

The Hi-C matrix of a whole chromosome typically shows many large blocks of enriched and depleted interactions that result in a plaid-like pattern (**Fig. 1.7a**, the Hi-C matrix of the human chromosome 14 is shown). The emergence of those patterns implies that each chromosome can be partitioned, at the multi-Mb scale, into two classes of regions, called A and B compartments, so that contacts within each class are enriched, whereas those across classes are depleted [13]. In fact, the application of the principal component (PC) analysis to Hi-C data, as well as independent microscopy experiments, revealed a strong spatial compartmentalization of chromatin interactions in two distinct sets so that greater interaction occurs within, rather than across, each compartment [13,56,88] (**Fig. 1.7b**). Consistently, the correlation matrix *C*, i.e., a symmetric matrix whose entry c_{ij} is the Pearson correlation between the *i-t*h and *j*-th column of Hi-C, has even sharper plaid patterns (**Fig. 1.7b**): indeed, if two chromatin segments are near in space, then they will have same neighbors and correlated interaction profiles [13]. Those compartments are considerably large regions (with a typical size of 5-10Mb) and they alternate along each chromosome (**Fig. 1.7c**). Compartment A has been shown to strongly correlate with the presence of genes, higher expression levels and accessible ("open") chromatin; hence, it is mostly associated with actively transcribed chromatin regions. Conversely, site pairs at a given genomic distance within the compartment B have in general higher interaction frequencies than analogous pairs of sites in the compartment A, showing that B compartment is more densely packed and thus correlated with more "closed" and expression-inactive chromatin. That is consistent with the known presence in the nucleus of open and closed chromatin (euchromatin and heterochromatin, respectively). Finally, the increased resolution of in-situ Hi-C experiments has recently allowed to dissect the finer structure of the A/B compartments, revealing, e.g., the existence of additional sub-compartments with distinct genomic and epigenomic content [88].

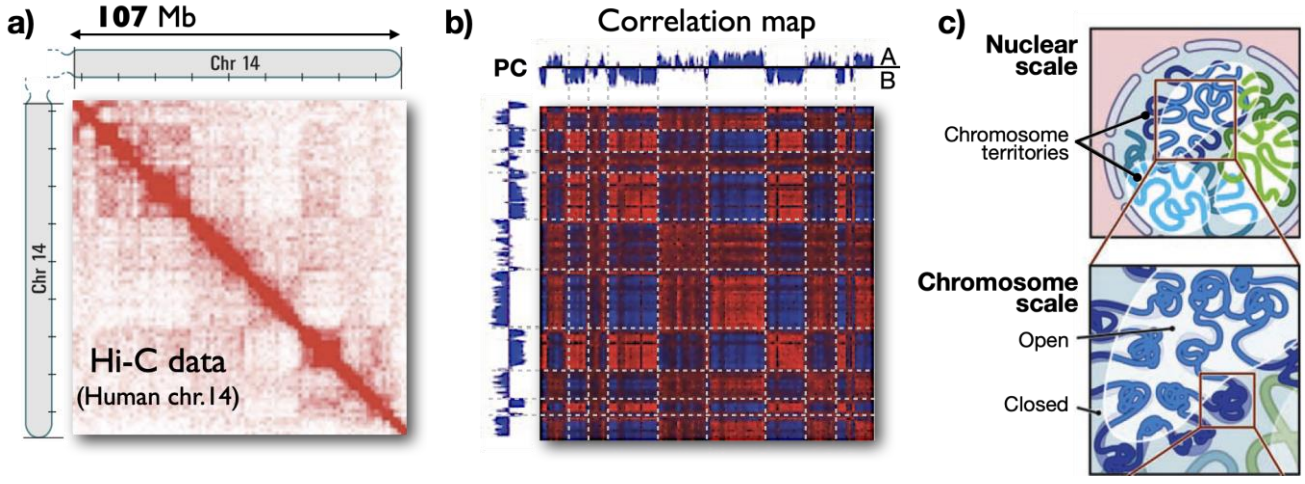

Figure 1.7: Chromatin is partitioned into A and B compartments above the megabase scale. a) Hi-C contact matrix of the human chromosome 14. As typical of Hi-C data, the matrix shows a peculiar plaid-like pattern, which indicates the existence of genomic regions that preferentially interact with each other. **b)** Pearson correlation matrix of the human chromosome 14 (see text; blue indicates low, while red high correlations) and its associated principal component (PC, top and left tracks). The PC profile correlates with the plaid pattern of the matrix, hence defining the compartment A (positive PC values) and B (negative PC values).

Contacts between A-A and B-B regions are enriched, while those across compartments are depleted. **c)** Schematic representation of chromatin organization at the nuclear scale where chromosome territories (hundreds of Mb) occupy distinct regions, and at the chromosome scale where more open and closer chromatin compartments (5-10Mb) alternately partition the genome. Adapted from [13].

1.3.2 Topologically Associating Domains (TADs)

A further analysis of the patters in Hi-C data revealed the existence of discrete, finer structures, much smaller than A and B compartments, where chromatin is marked by enriched levels of inner interactions. Those structural domains, which are largely conserved across species, cell types and tissue types and can be hundreds of kilobases in size (median size around 400-500kb), are called Topologically Associating Domains (TADs) [16,17,103]. They are visible in Hi-C (as well as in 5C) data as square, block-like, domains of high contact frequencies, signaling that DNA loci located within the same TADs tend to interact frequently with each other and much less frequently with loci located outside their domain (**Fig. 1.8a**). As Hi-C contact maps are symmetric, often only their upper or lower triangular part is represented: in this case, TADs appear as triangles, and not as squares, of enriched internal contacts (**Fig. 1.8b**). Those features enabled researchers to identify TADs throughout the human and mouse genomes by a quantitative inspection of Hi-C contact matrices. Several algorithms have been developed to this aim [16,18,88], which mostly rely on the detection of the TAD boundaries, defined as genomic positions where sharp changes in the average contact frequencies take place. Those analyses, also confirmed by microscopy experiments, showed that TADs are universal building blocks of chromosomes; the human and mouse genomes are each composed of more than 2000 TADs, covering over 90% of the genome [89]. Furthermore, increasing experimental evidence is pointing to important, functional roles for TADs: for instance, they are thought to act as insulating structures, spatially confining the activity of gene regulators to their proper targets [2,3,104]. Consistent with such a picture, disruption of TADs due to genomic rearrangements, such as deletions or inversions (the so-called "structural variants"), has been linked to a rewiring of gene-regulator contacts, resulting in gene misexpression and disease [6,7,102]. That highlights even more the importance of understanding the physical and molecular mechanisms, still largely unknown, underlying TAD formation. Indeed, many quantitative models [9,24,33–42,25,43–48,26–32] and computational approaches [49,50,59–61,51–58] have been proposed to understand the machinery that originates those contact patterns. In the next **Chapter 2**, we will discuss a first-principles approach based on polymer physics to tackle this challenging problem.

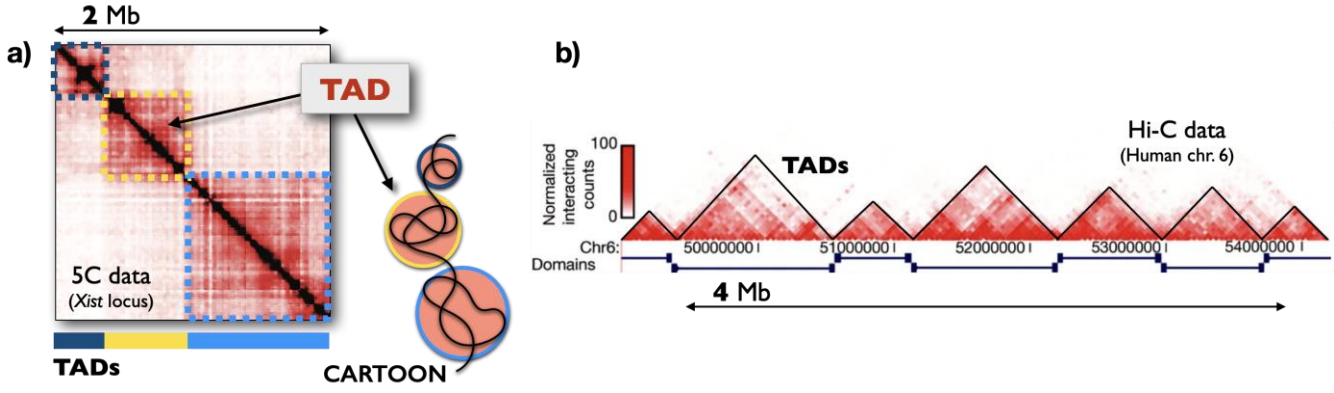

Figure 1.8: Chromatin folds into megabase sized self-interacting regions called TADs. a) Analyses performed on the high-resolution 5C interaction map of a 2Mb region encompassing the mouse X-chromosome inactivation centre (*Xist* locus, data from [17]) revealed a series of large structural domains, called TADs. Loci located within these TADs tend to interact frequently with each other, while they interact much less frequently with loci located outside their domain. TADs appear visually as subsequent square block-like domains located along the diagonal, which are enriched of inner contacts. In the example of the figure, three distinct TADs can be identified and they are highlighted by different colors (blue, yellow, cyan). **b)** Hi-C data of a 4Mb region along chromosome 6 in mouse embryonic stem cells. Here, only the upper triangular matrix is represented, as Hi-C is symmetric by construction. In this case, TADs clearly appear as subsequent triangles (and not squares) of increased inner contacts. Panel **b)** is adapted from [16].

1.3.3 The emerging complex picture of chromosome folding

Finally, we want to stress that the findings reviewed above represent only a limited and simplified, yet crucial, picture of the state-of-the-art of chromosome nuclear organization. In fact, we have seen before that mammalian genomes are organized into large (mega-base sized) domains, called TADs, that display a high degree of interaction [16,17]. However, interaction patterns can also form within and across TADs at lower as well as at larger scales. In the work [18], for instance, TAD higher-order interactions are investigated with Hi-C experiments through mouse neural differentiation and TADs are found to form a hierarchy of domains-within-domains (called "metaTADs", i.e., clusters of interacting TADs) extending across genomic scales up to the range of entire chromosomes. Also, long stretches of chromatin have been found to interact with the nuclear lamina (a network-like structure at the inner membrane of the nucleus), resulting in new, mostly transcriptionally repressed, contact domains called "LADs" (i.e., lamina-associated domains) [1]. At a lower scale (e.g., units of kb), the development of more and more refined technologies led to the discovery of new chromatin structural features. For example, high-resolution in-situ Hi-C data [88] allowed the identification of chromatin "loops", i.e., strongly enriched pointwise interactions that brings in close spatial proximity pairs of DNA sites (e.g., a gene and its regulator) that are distant along the linear sequence of the chromosome. Loops appear as particularly bright dot in Hi-C contact matrices and often they are found to occur at TAD boundaries (originating the so called "loop domains"); additionally, in many cases they coincide with pairs of properly oriented CTCF (CCCTC-binding factor) sites, indicating that CTCF protein may play a key role into TAD formation and genome spatial organization [88]. Indeed, CTCF sites and cohesin have been proposed to shape loops and TADs, for example via the loop extrusion polymer model [24,26] (see Section 1.5). At a finer scale (hundreds of bp), novel powerful techniques, such as Micro-C [93,105], revealed the existence of structures much smaller than TADs, like "micro-TADs", that encompass either single genes, multiple genes, or intergenic regions and may constraint specific transcriptional programs. Hence, in the emerging picture, chromosomes are folded into a complex 3D architecture [1–3] including a hierarchy of interactions, from loops [88] and TADs [16,17] to, above the mega-base scale, metaTADs [18] and A/B compartments [13] as revealed by population-averaged, e.g., Hi-C contact maps [13,88]. Such an organization serves important functional purposes as genes and their regulators are thought to establish specific physical contacts to regulate transcription. Furthermore, recent high-resolution microscopy experiments [19–23], which we will discuss with more detail in the next Section 1.4, are

allowing to probe genome organization at the single-cell level, revealing that contact patterns and complex functional structures, e.g., TADs and loops, also form in individual DNA single molecules.

1.4 Recent advancements of single-cell super-resolution microscopy technologies

Data from high-throughput sequencing-based technologies are becoming even more valuable nowadays thanks to the recent development of microscopy-related techniques, which enable direct measures of 3D spatial distances in the targeted genome and provide imaging-complementary information to molecular techniques (such as Hi-C or its derivatives), which, instead, are based on measuring contact frequencies. In DNA FISH (fluorescence in-situ hybridization) [106,107], for example, DNA probes are hybridized to cognate genomic regions of interest and visualized by fluorescence microscopy, hence allowing the measure of their localization, shape and inter-probe distances. However, the study of chromosome conformation under the microscope has been technically limited by the low number of DNA segments that can be probed simultaneously and by the limited spatial resolution of traditional light microscopes. These limitations have been removed thanks to major technological advancements in recent microscopy applications that overcame, for instance, the Abbe diffraction limit [91]. These methods, collectively called "super-resolution microscopy", markedly increased the achievable spatial resolutions and are currently among the most powerful technologies to interrogate genome 3D organization at the single-cell level [19–23]. For instance, TADs have been proposed as a fundamental structural unit of the genome organization (see Section 1.3), yet many of their basic properties, such as a clear physical understanding of TAD structure, remain unclear: do TADs exist in single-cells or are they a mere emergent property from cell population-averaging? Do they vary from cell-to-cell? What is a TAD in 3D space? Such questions, which have been unanswered in the last decade because of the lack of high-resolution and powerful enough technologies, can now be tackle thanks to the latest developments of microscopy-based methods. In this Section, we want to focus on a recent multiplex FISH imaging-based approach, reported with all details in the original paper [21], that provided for the first time a high-resolution visualization of the physical structure of chromatin in thousands of individual single-cells at the kilobase-to-megabase scale (i.e., the typical sizes of genes and regulatory domains). In the subsection 1.4.1 we provide some more technical details of the method, while in subsections 1.4.2 and 1.4.3 we discuss some of the major findings of the experiment. Overall, if, on the one hand, multiplex FISH methods have provided novel critical insights into 3D genome folding at the single-cell level, on the other hand they have disclosed new fundamental questions on the molecular mechanisms shaping chromatin organization in single-cells. In the **Chapter 3** we will discuss principled models from physics to tackle the challenging, microscopy-derived, findings described in this section.

1.4.1 Multiplex FISH microscopy for high-resolution chromatin tracing

In the paper [21], the authors reported a super-resolution chromatin tracing method to determine the structural features of the genome with nanometer-scale precision in single-cells. The rationale of the technique is essentially the following: if numerous consecutive chromatin segments can be identified and precisely localized in individual cells, then we can finely trace the 3D conformation of

the considered genomic region by simply connecting their measured spatial positions. However, a typical limitation of FISH microscopy methods is throughput, as they are limited to use two or three channels (i.e., "colors") that restrict the number of chromatin segments that can be simultaneously imaged. The authors in [21] overcame this challenge by partitioning the targeted genomic region into numerous segments, each 30 kb in length (which sets the resolution of the experiment), and by imaging individual segments using sequential optimized rounds of FISH (**Fig. 1.9**). The technical procedure, in brief, involves the following steps [21]: first, the considered region is labeled with a library of 10^3 probes, each specific for each 30kb segment to facilitate multiplexed FISH imaging; then, dye-labeled readout probes, complementary to the readout sequences along the genome, are added to allow the 3D imaging of individual 30kb segments; the sequential process of readout-probe labeling and imaging is repeated until all segments are imaged. This allows to generate, for each single-cell, a 3D super-resolution image of the chromatin region of interest, each reporting the position and structure of a contiguous 30kb segment with <50nm error in their localization (**Fig. 1.9**). In their work, the authors performed imaging of multiple Mb-sized chromatin regions, traversing different TADs and sub-TADs, of human chromosome 21 across different cell types. Some of the main results of their study are discussed in the next subsections.

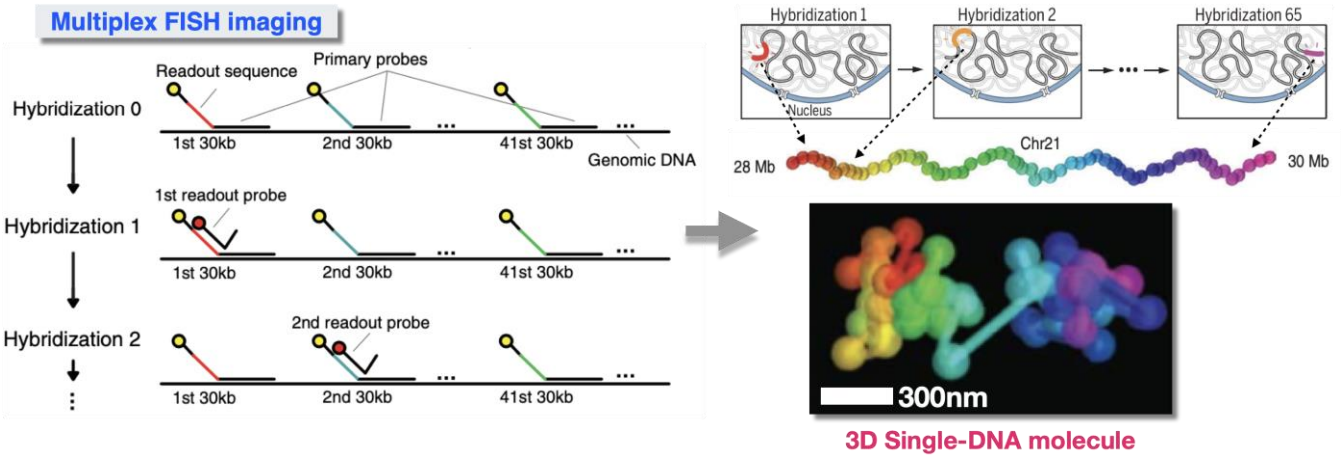

Figure 1.9: Multiplex FISH imaging allows to probe chromatin 3D organization with nanometer- and kilobase-scale resolution in single-cells. A scheme of the imaging approach from [21]. In brief, the genomic region of interest is divided into consecutive 30kb segments, each univocally labeled by roughly 300 probes that contain a readout sequence specific to each 30kb segment. Then, readout probes complementary to those readout sequences are added sequentially, allowing the imaging of individual 30kb segments and, hence, of the entire genomic region under investigation. The experiment allows to generate at the single-cell level (i.e., in a single DNA molecule) a 3D super-resolution image of the chromatin conformation of the studied genomic region with nanometer-scale precision. Adapted from [21].

1.4.2 TADs exist in single-cells and broadly vary from cell to cell

Super-resolution chromatin images allow to derive, for each cell in a population, a corresponding pairwise spatial distance matrix, i.e., a symmetric matrix containing the pairwise spatial distances between all chromatin segments of the imaged region in that cell. In other words, if D is the single-cell distance matrix, then the entry D_{ij} is the measured value of the physical (i.e., Euclidean) distance between the chromatin segments i and j in the considered cell. Averaging those single-cell distance

maps across the ensemble of imaged cells returns, by definition, the ensemble spatial distance matrix of the studied genomic region. In the paper [21], the authors performed imaging of multiple Mb-sized chromatin regions, traversing different TADs and sub-TADs, of human chromosome 21 across different cell types. For instance, the ensemble spatial distance matrix of a 2Mb genomic region (genomic coordinates chr21:28–30Mb, hg38) in IMR90 lung fibroblast cells is shown in **Fig. 1.10a** [21]. This matrix displays specific, block-like TAD structures and it is shown in the original work [21] to be in good agreement with the ensemble Hi-C contact matrix of the same region, thus revealing the consistency of the multiplex FISH approach with Hi-C at the population-average level. Furthermore, the single-cell spatial distance matrices (i.e., the distance maps of individual cells) also exhibit clearly visible TAD-like structures corresponding in 3D space to spatially segregated globular conformations (**Fig. 1.10b**). However, those TAD-like patterns are found to be broadly varying across the ensemble of imaged cells. Indeed, the genomic positions of the single-cell TAD boundaries are highly heterogenous, as they change from cell-to-cell (**Fig. 1.10b**).

Such a variability is quantitatively assessed by the boundary probability function, which measures the probability to find in a single-cell a TAD boundary at a given genomic position (**Fig. 1.10c**, top). Consistently, as a reflex of the substantial single-cell variation of the domain boundaries, this function is found to be nonzero throughout the imaged region (i.e., each genomic position has a finite probability to be a TAD boundary in single-cells). The local peaks of the boundary function correspond to the most likely TAD boundaries, i.e., those visible in the ensemble distance map of the locus at the population-average level. Those peaks, interestingly, show a preference to reside at genomic positions associated with binding sites of CTCF and cohesin (**Fig. 1.10c**, bottom bar. Cohesin is marked by RAD21, one of its core subunits), that is consistent with the known role of those proteins in shaping chromatin structure.

Additionally, to measure the level of spatial segregation between chromatin segments along the imaged genomic region, the authors introduced a separation score function [21] (see **Fig. 1.11** for the technical definition): a score equal to 1 means a complete spatial separation between chromatin regions on the right and left side of the considered genomic position, while values <1 indicate a degree of chromatin intermingling around that position. The high values of the separation score (in the range 0.7-0.9) indicate a nearly complete spatial segregation at many of the identified single-cell TAD boundaries (**Fig. 1.10c**, bottom), confirming that the imaged region tends to fold into separated globule-like domains in single-cells. Similar experimental results, not discussed here for brevity, are also found for other human cell lines (e.g., K562 erythroleukemia and A549 lung epithelial carcinoma cells) [21].

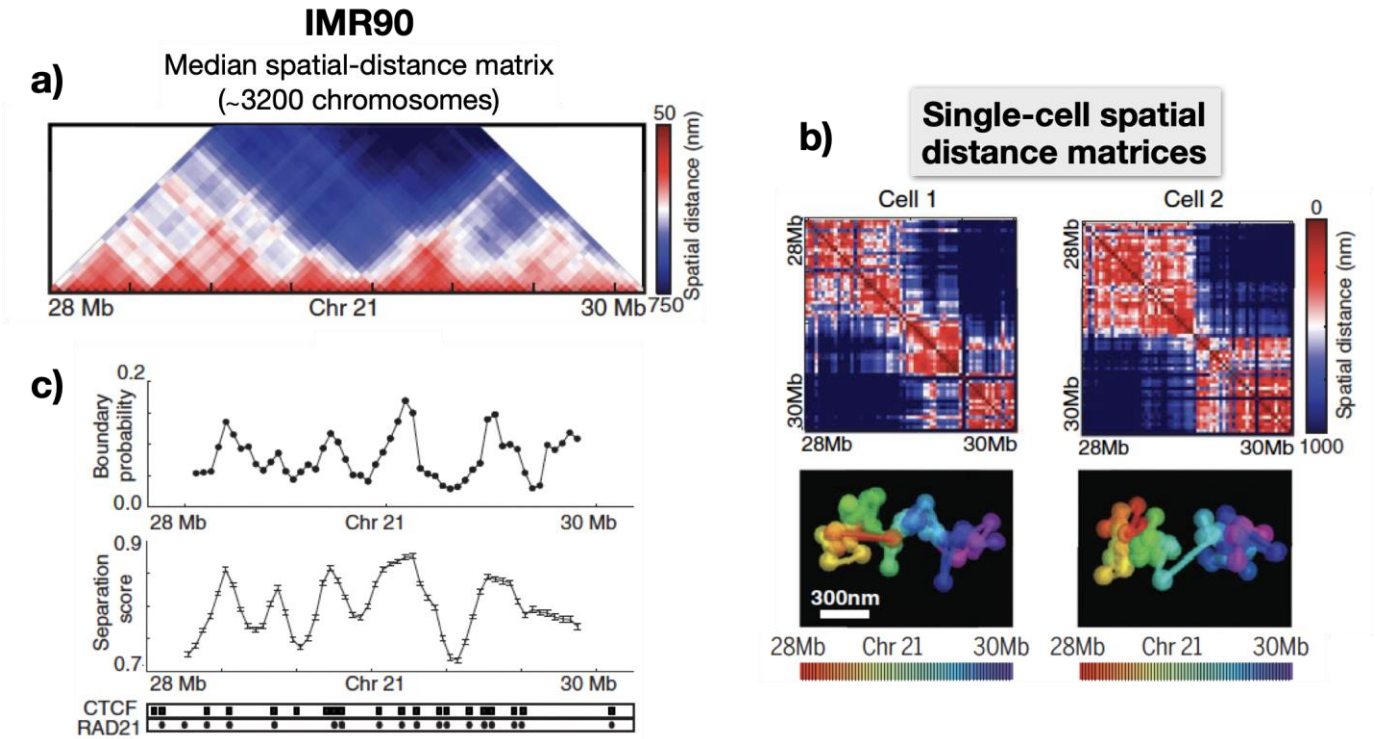

Figure 1.10: Multiplex FISH microscopy reveals the existence of TADs in single-cells. a) Median spatial distal matrix of the locus chr21:28–30Mb in human IMR90 cells. **b)** Examples of single-cell imaged distance matrices of the locus and corresponding 3D structures. TAD-like patterns broadly vary from cell-to-cell. **c)** Top: Boundary probability of the locus, i.e., the probability for each genomic position across the locus to be the boundary of a single-cell TAD-like domain. This function, consistent with the high variability of single-cell TAD boundaries, is nonzero at all genomic positions and is particularly enriched at CTCF and cohesin sites (bottom tracks). Bottom: The separation score function measures the level of spatial segregation around each genomic position across the locus (see **Fig. 1.11** for its technical definition) and has a behavior similar to the boundary probability. The bottom bar shows the CTCF and RAD21 (a cohesin subunit) binding sites of the locus. Adapted from [21].

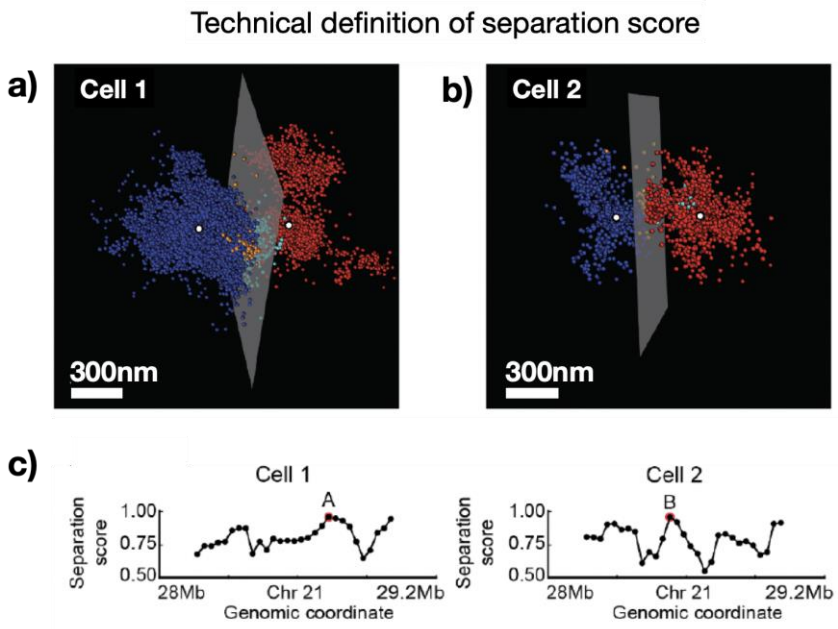

Figure 1.11: Technical definition of separation score in single-cells. The separation score of a genomic coordinate "P" is defined by considering: *i)* a 180kb region (i.e., 6 imaged segments) downstream P (i.e., on the left side of P); *ii)* a 180kb region upstream P (i.e., on the right side of P); *iii)* the midplane between the mass centers of the downstream and upstream regions of P. Then, the separation score is the ratio $(n_{left}+n_{right})/n_{tot}$, where n_{left} is the number of 3D imaging localizations on the proper side of the midplane (i.e., the number of 3D localizations of the downstream 180kb region on the left side of the midplane), n_{right} is the number of localizations of the upstream 180kb region on the right side of the midplane, and n_{tot} the total number of localizations of the whole 360kb region. A separation score equal to 1 indicates that the downstream and upstream regions are completely separated. The panels **a)** and **b)** help to visualize such technical definition. **a)** 3D imaging localizations of six chromatin segments (i.e., 180kb) located downstream (colored with blue and cyan) and upstream (colored with red and orange) of the genomic position indicated as "A" in cell 1 of panel **c)**. The centers of mass of the downstream (blue+cyan) and upstream (red+orange) chromatin regions are indicated as white dots; the midplane between the two mass centers is also shown as a grey sheet. **b)** Same as in **a)** but for the genomic position "B" in cell 2 of panel **c)**. **c)** Separation score across a 1.2Mb region (Chr21:28-29.2Mb) for two example single-cells. Definitions and figures taken from [21].

Summarizing, multiplex FISH microscopy experiments [21] have revealed that chromatin in single-cells forms TAD-like structures with sharp domain boundaries that correspond to 3D globular conformations with strong physical segregation between neighboring domains. Those single-cell TAD-like domains, however, show substantial cell-to-cell variability, as TAD boundaries occur with nonzero probabilities at all genomic positions, albeit preferentially at CTCF and cohesin sites. This shows, in particular, that TADs are physical structures present in single-cells and not an emergent property of population-averaging, resolving an important, long-standing debate on the nature of those domains.

1.4.3 Cohesin depletion erases patterns at the population-average level, but not in single-cells

Another major finding discussed in [21] is the investigation of how single-cell chromatin structures change upon the removal of the cohesin complex, which is a key known architectural protein. To this aim, the authors imaged a 2.5Mb region (chr21:34.6–37.1Mb) in human HCT116 cells (colon cancer cells) and compared the chromatin structures in normal - "wild type" (WT) - conditions (**Fig. 1.12**, left) and after inducing an auxin-based treatment for cohesin depletion (**Fig. 1.12**, right). While the ensemble spatial-distance matrix derived from imaging data has several pronounced TAD structures in the WT case (**Fig. 1.12a**, left), the matrix derived after auxin treatment is mostly featureless (**Fig. 1.12a**, right), as the population-averaged TADs and sub-TADs of the locus are largely eliminated. This result is consistent with previous Hi-C experiments where cohesin loss is found to erase contact patters at the ensemble level [88]. However, well beyond the average picture, chromatin TAD-like domains in single-cells persisted in both the WT (**Fig. 1.12b**, left) and cohesin depleted case (**Fig. 1.12b**, right). Conversely, a major difference in the absence of the cohesin complex is related to the genomic location and probability of the single-cell domain boundaries, which are uniformly spread across all genomic coordinates in HCT116+Auxin cells (**Fig. 1.12c**, right) and no longer show a preferential positioning at CTCF and cohesin sites as in the WT case (**Fig. 1.12c**, left).

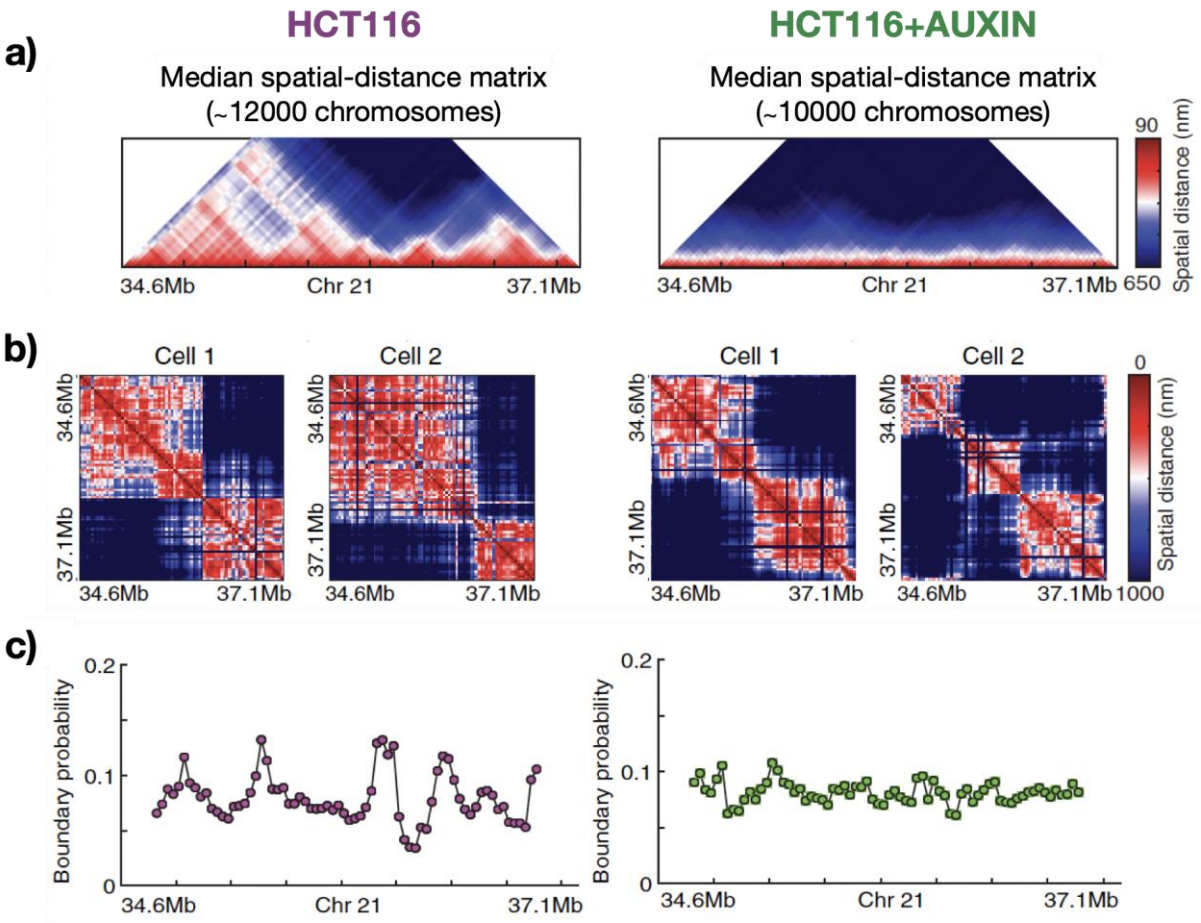

Figure 1.12: Effects of cohesin depletion on chromatin structure at the single-cell level. a) Median spatial distance matrix of the locus chr21:34.6–37.1Mb in human HCT116 cells in normal "wild-type" (WT) conditions (left) and upon cohesin depletion (HCT116+Auxin, right). In the WT case, the median distance map has specific TADs and sub-TADs, which appear to be erased after cohesin removal. **b)** Examples of single-cell distance matrices of the locus in WT (left) and HCT116+Auxin (right) cells. Albeit cohesin depletion implies loss of structure at the population-averaged level, contact patterns still persist in single-cells. **c)** Boundary probability of the locus in the normal case (left) and in absence of cohesin (right). Adapted from [21].

To summarize, cohesin depletion leaves contact patterns at the TAD-scale intact in single-cells, but domain boundaries become equally likely to locate at any genomic position, therefore abolishing TADs at the population-average level. In the emerging picture, cohesin is not required for the maintenance of TAD-like structures in single cells and its role in the formation of ensemble TADs is to establish preferred genomic boundaries for the single-cell domains.

1.5 Physical mechanisms of chromosome spatial organization

The increasing level of details provided by experimental advances has triggered the need to further develop quantitative models from physics to understand the molecular mechanisms shaping chromosome spatial organization. To this aim, first-principled models have been developed that try to explain the complexity of chromatin contact patterns in a coherent, mechanistic framework (see, e.g., [9,24,33–42,25,43–48,26–32]). All those models rely on the basic idea that chromatin has a polymeric nature and that its folding properties can then be quantitatively investigated via first-

principled approaches based on the laws of polymer physics. Hence, chromosome conformations are built based on molecular folding processes, which are postulated a-priori, and experimental data are used to test and validate model hypotheses and predictions. Here, we aim to quickly introduce the reader with two recently discussed classes of models that reflect two classical, yet distinct biological scenarios to explain the formation of chromatin contacts: loop-extrusion and phase-separation based polymer models (**Fig. 1.13**). The content of this Section is adapted from our recent, more extensive review articles on the topic (see, e.g., [39,108,109]).

The loop-extrusion (LE) model [24,26,35,42,110] (**Fig. 1.13a**) envisages a picture where physical proximity between distal sites is attained by a protein complex that binds to DNA and actively extrudes a chromatin loop until encountering specific anchor (i.e., extrusion-halting) sites. While the polymer compacts into a linear array of loops, specific contacts form between the motor anchor sites where extrusion halts, thus defining boundaries between subsequent chromatin regions. The technical details of the model and its implementation can be found in the referenced paper. The LE describes an active, off-equilibrium physical process, as the molecular motor requires energy, e.g., ATP, consumption to perform extrusion. However, also variants of the model have been developed where extrusion is driven by thermal diffusion [35] or transcription-induced supercoiling [42] rather than by an energy-burning complex. The peculiar LE motor activity has been primarily ascribed to a widely conserved class of proteins known as structural maintenance of chromosomes (SMC) complexes, which particularly include cohesin and condensin, while properly oriented CTCF sites can act as blocking sites [110]. Computer simulations have shown that the LE model can explain with good accuracy loops and TADs visible in bulk Hi-C contact matrices as well as, for example, experiments on mitotic chromosome compaction and segregation [24,26,35,42,110,111]. Furthermore, recent in-vitro single-molecule experiments have provided direct observation of the extrusion activity driven by factors such as condensin and cohesin, albeit in simplified conditions [112–114]. Also, the model has been shown to recapitulate the structural rearrangements upon disruption of specific CTCF binding sites [24,115,116]. However, other important experiments have provided evidence that chromatin 3D architecture is only partially explained by the LE model. For instance, Hi-C data upon cohesin and CTCF depletion show that interactions persist at the A/B compartment level and within former loops or TADs [117–119]. In addition, super-resolution multiplex FISH microscopy experiments [19–22] have shown that TAD domains are broadly varying in single-cells and that TAD boundaries occur with nonzero probability at any genomic location, not only at a subset of CTCF sites [21] (see also Section 1.4). Those findings hint that there are other folding physical mechanisms beyond LE.

The second class of polymer models [9,25,36–41,43–46,27,48,28–34] considers another classical scenario of folding, where specific attractive interactions exist between corresponding cognate types of distal DNA binding sites (**Fig. 1.13b**), either established by direct DNA-DNA contact (e.g., by DNA-bound histone molecules) or established by diffusing molecules (such as Transcription Factors) that can bridge those sites. Statistical mechanics dictates that thermodynamic phases are independent of the specific origin of the interactions (i.e., direct or mediated by diffusing molecular factors), hence those different models can belong to the same universality class [62]. Since

interactions are homotypic, cognate DNA binding sites spontaneously self-assembles along the sequence into specific globular blocks via a thermodynamic mechanism called “polymer phase-separation”. This mechanism is a typical phenomenon observed in block-copolymers made of incompatible chemical components [120,121], in which all the different blocks, due to homotypic interactions, fold separately one from each other thus crumpling the polymer. Importantly, while loop extrusion requires energy consumption, phase separations do not require energy input beyond the thermal bath. Growing experimental evidence is supporting physical mechanisms of phase-separation as a robust paradigm of cell nuclear organization and of transcriptional regulation [78,79,81,82,122]. For instance, cooperative multi-molecular protein assemblies, such as combinations of RNA polymerase II with transcription factors and co-factors (as the Mediator complex), have been shown to control gene activity by the formation of functionally relevant, phase-separated, chromatin hubs [81].

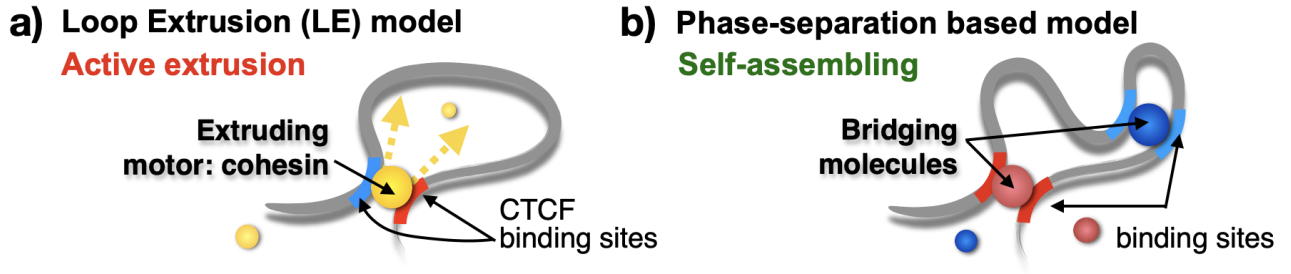

Figure 1.13: Two different physical mechanisms to explain the formation of chromatin contacts. a). Cartoon of the loop-extrusion (LE) model. In the model, physical proximity between distal sites is established by an active molecular motor (i.e., the cohesin complex), which extrudes chromatin loops until encountering specific anchor sites (envisaged as pairs of CTCF binding sites with opposite orientation), in an out-of-equilibrium process that requires energy (e.g., ATP) consumption. **b)** In phase-separation based polymer models, distal sites come into close spatial proximity either by diffusing cognate molecular factors (e.g., Transcription Factors and other proteins) that can bridge those sites (as shown in the cartoon) or via direct polymer-to-polymer homotypic attractive interactions. In both cases, driven by those long-range homotypic interactions, the system spontaneously self-assembles into specific globular domains via a thermodynamic mechanism of polymer phase-separation (see text), which involves no energy consumption as it is sustained by the thermal bath. Adapted from [108].

In the next **Chapter 2**, we will discuss with all details a well-known example in the class of phase-separation based polymer physics models, i.e., the Strings and Binders (SBS) polymer model [43,45], which has been broadly applied, for example, to investigate chromosome large-scale 3D organization [25,123], to understand the impact of disease-associated mutations [9] and also to predict in-silico chromatin 3D structure at the level of single DNA molecules [38].

2. Polymer phase-separation-based physics models of chromosomes

As discussed in **Chapter 1**, chromosomes fold into a complex 3D architecture within the cell nucleus, including a hierarchy of multiple interactions serving important functional purposes (e.g., gene regulation), yet the organizational mechanisms shaping chromosome structure remain poorly understood. In this Chapter, we describe how the large-scale features of chromosome 3D organization, such as TADs and population-averaged contact patterns, can be explained by a molecular mechanism based on thermodynamic phase-separation. To this aim, we focus here on the Strings and Binders (SBS) polymer model [9,38,43,45], which envisages a textbook scenario where physical contacts between distal DNA sites are established via diffusing molecular factors, such as Transcription Factors, that can bridge those sites. The model has been shown to accurately describe Hi-C, GAM and FISH data across different genomic loci and cell types [9,18,25,32,36,45,124–126], also at the single-molecule level [38,40]. In Section 2.1, we discuss the polymer physics implementation of the SBS model, where a chromosome filament is represented as a polymer chain along which different types of binding sites are located for cognate diffusing binders. Depending on the number or energy affinity of the binders, the model undergoes a thermodynamic phase-transition from a coil state, where the polymer is randomly folded due to prevailing entropic effects, to an equilibrium globule phase, where instead attractive interactions between binders and cognate polymer sites result in more condensed, compact structures. As known in polymer physics [62], the system steady-state 3D conformations fall in those two main structural classes (i.e., coil and globule) corresponding to its thermodynamics phases. We show that those phase transitions result in conformational changes of the polymer chain, which establish, via a phase-separation based mechanism, contact patterns (e.g., TADs and other structures) similar to those observed in real data. To identify the binding sites of the SBS model, we developed a recursive statistical computational algorithm, called PRISMR [9,38] and detailed in Section 2.2, that combines machine learning and polymer physics. In brief, by taking as input only bulk pairwise contact data (such as Hi-C or GAM contact maps) available in a considered genomic region of interest, PRISMR infers the minimal set of binding sites sufficient to recapitulate, through only physics, the input contact patterns. Then, in our strategy, we perform massive Molecular Dynamics (MD) simulations of the inferred polymer model to derive a thermodynamic ensemble of in-silico, single-molecule 3D conformations of the studied genomic region. The details of the MD implementation of the model are all discussed in Section 2.3. Finally, as an application of our approach to real data, we discuss in Section 2.4 the SBS model of a genomic region crucially involved in limb development, i.e., the *HoxD* gene region [36]; we show that the basic ingredients of the model can explain with high accuracy the complex and highly specific patterns of interactions observed in the experiments.

2.1 The Strings and Binders (SBS) polymer model

In the Strings and Binders (SBS) model a chromatin region is represented as a self-avoiding walk (SAW) chain of beads, along which are located specific binding sites for diffusing, cognate molecular binders [25,43,45] (**Fig. 2.1a**). The binders have a molar concentration, *c*, and can form loops by bridging pairs of distal polymer sites, hence driving the folding of the chain. The particle interaction

potentials of the model are based on classical studies of polymer physics [127] and include overall three distinct terms (see Section 2.3 for a detailed discussion): a finite extensible (FENE) spring between consecutive beads on the chain; a repulsive Lennard-Jones (LJ) potential to account for exclude volume effects; a short-range attractive LJ between beads and binders, with an energy scale, E_{int}, in the weak biochemical energy range (few K_BT). The binders, as well as the polymer beads, move under the Langevin equation, investigated by Molecular Dynamics computer simulations (Section 2.3).

As the number of binders (or their affinity strength) grows above a threshold point, the polymer model undergoes a thermodynamics phase transition from a coil, i.e., randomly folded, to a globule state (**Fig. 2.1b**). For the explored weak biochemical affinities, the threshold concentration typically falls in the fractions of µmol/l range [25,38], values compatible, e.g., with transcription factor concentrations. As known from block-copolymer theory [120,128], the coil-globule transition triggers sharp conformational rearrangements of the chain, as in the coil state entropic forces produce more fleeting and transient contacts whereby the polymer folds into random open conformations that fall in the universality class of the free SAW [62,129], while in the globule state attractive interactions thermodynamically prevail and shrink the polymer into a compact, globular structure [38]. As dictated by polymer physics [62], the equilibrium states of the model fold just in a few conformational classes, which correspond to the system thermodynamics phases, i.e., the coil or globule state. The model with its basic ingredients envisages an organizational mechanism of folding based on reversible phase transitions, which occur spontaneously and are sustained by the thermal bath. Polymer conformational changes can be sharply regulated in a switch-like manner, as the system only needs, e.g., to establish an above threshold concentration of binders (or affinity), with no need of fine tuning their number or strength.

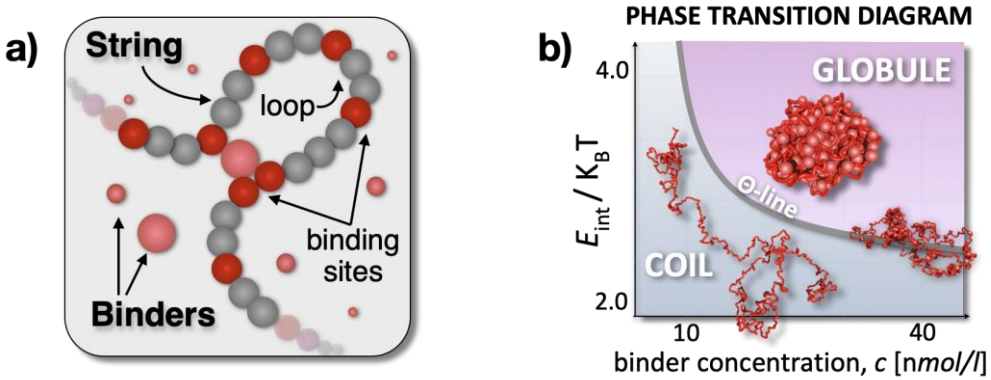

Figure 2.1: The Strings and Binders (SBS) polymer model and its phase diagram. a) Cartoon of the Strings and Binders (SBS) polymer model. A chromatin filament is represented as a self-avoiding polymer chain made of beads that cannot overlap. Along the chain specific binding sites (in red) are located for the molecular binders of the model. The binders can bridge distal, cognate sites on the chain, thus looping the polymer. **b)** Upon increasing the binder concentration or affinity, the model has a phase transition from a coil, i.e., randomly folded, to a more condensed globule state. As dictated by polymer physics [62], the equilibrium conformations of the model fold in those two main conformational classes (i.e., coil and globule), which correspond to its thermodynamics phases. The system can switch from one state to another simply by crossing the phase boundary, with no need of parameter fine tuning. Adapted from [25].

2.1.1 Polymer phase-separation as a mechanism of TAD formation

Here, we aim to discuss a very simplified example to make sense of how contact patterns (e.g., TADs) emerge in the model picture. To this aim, we consider a basic variant of the SBS model discussed previously, in which now two distinct types of binding sites, visually represented by different colors (i.e., red and green), are located on the polymer chain (**Fig. 2.2a**). In particular, the red and green sites are placed in two separated halves of the polymer and each binding domain is provided with a specific affinity to its cognate binders [45]. Inert, non-interacting sites are also included and visualized in grey. In these conditions, the physical interactions between specific cognate binding sites along the sequence promote a polymer-to-polymer phase-separation whereby the chain self-assembles, at thermodynamic equilibrium, in two distinct, spatially segregated globules (**Fig. 2.2b**, left). Consistently, the steady-state average pairwise contact matrix of the model shows two squared blocks of enhanced interactions along the diagonal, reflecting the partitioning of the chain in two separated globular domains (**Fig. 2.2b**, right). Notably, those domains closely resemble the ones, commonly referred to as TADs (see Section 1.3), observed, for example, in 3C-based interaction frequency data [16,17] (**Fig. 2.2c**). Polymer physics dictates that the thermodynamic phases do not depend on the specific origin of the interactions, e.g., whether they are direct or binder-mediated, thus different physics models can belong to the same universality class [62]. For that reason, a model based, for instance, on homotypic direct interactions between polymer sites, rather than mediated by diffusing factors, has same behaviors, as it provides the same thermodynamic phases of the SBS model. Those phase transitions result in structural changes of the chain that spontaneously establish, via a phase-separation based mechanism, contact or segregation of specific distal sites, such as genes and their regulators [38].

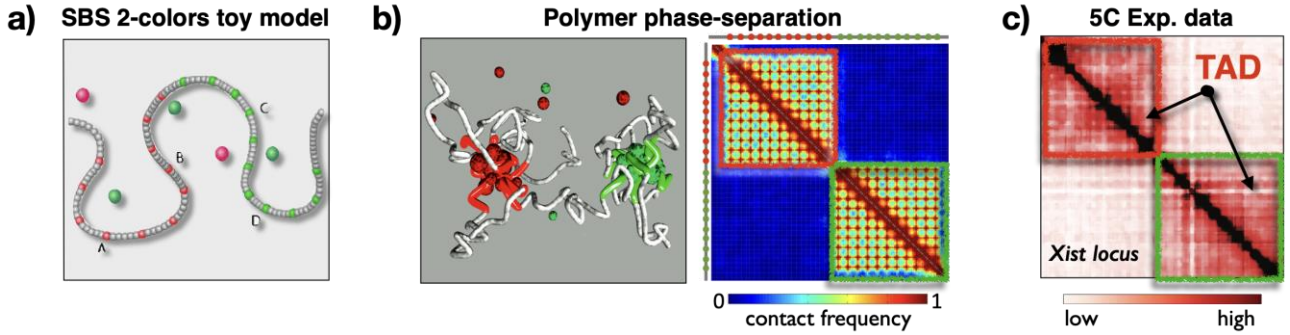

Figure 2.2: Contact patterns result from polymer phase-separation in the SBS model. a) SBS polymer chain with two distinct types of binding sites, visually represented by different colors (i.e., red and green). **b)** Left: the homotypic attractive interactions between polymer binding sites and cognate binders drive a polymer-to-polymer phase separation whereby the chain spontaneously folds, at equilibrium, in two distinct, segregated lumps (respectively, the red and the green globules). Right: the pairwise contact matrix of the phase-separated polymer chain has enriched interaction squares along the diagonal, which reflect the spatial segregation of the model in two distinct globular domains. **c)** 5C interaction data of a 2Mb wide region encompassing the *Xist* gene in mouse embryonic stem cells [17]. The contact patters observed in the experiments (e.g., TADs) result in the SBS model from a physics mechanism of polymer phase-separation [38]. Adapted from [45].

2.2 Inference of the model binding sites via a polymer-based recursive statistical computational method

We can use the basic ingredients of the SBS model discussed before to investigate the mechanisms whereby real genomic regions fold in 3D space [9,25,36,38]. However, a naïve SBS block-copolymer, as the one discussed in Section 2.1, only provides a poor description of folding, as chromatin loci involving, for instance, genes and their regulators, typically have complex, specific interaction patterns, as observed in Hi-C or GAM experiments. For that reason, we developed a more refined SBS model, in which different types of binding sites, visually represented by many different colors, are non-trivially arranged along the chain to capture such a contact specificity [9,25]. To infer the genomic location and the types of the putative binding sites of the SBS model of a given genomic region of interest, we designed a polymer-based recursive statistical computational approach, called PRISMR [9], which provides the minimal, best polymer model by taking as input only contact data (e.g., Hi-C or GAM). Here, we summarize the main features of the PRISMR method [9] (subsection 2.2.1), and also discuss its improved, more recent version implemented in [38] (subsection 2.2.2). Most of the material discussed in this Section is adapted from the original papers [9,38].

2.2.1 The PRISMR algorithm

Based only on physics, our PRISMR algorithm infers the minimal number of distinct types of specific binding sites (i.e., distinct colors) and their positioning along the chain that best reproduce the input contact matrix of a given genomic locus of interest (**Fig. 2.3**). Formally, an SBS polymer model is identified by the arrangement of binding sites of different types along the chain of beads, i.e., by the set $\{c_i\}$ of its color variables, where c_i=0, 1,…, *n* labels the distinct colors on the chain (*n* is the total number of colors, while 0 corresponds to gray, inert sites), and the index i=1,…,*N* labels the *i*-th bead. Hence, the output of PRISMR is the best, minimal arrangement $\{c_i\}_{min}$ of beads along the chain to describe the input contact matrix. The sequence of the genomic region to model is divided into *M* windows (i.e., bins), according to the genomic resolution of the considered experiment, e.g., Hi-C or GAM. As a single DNA window could include many binding sites, we suppose that our model can accommodate up to *r* binding sites (beads) in each DNA window. Thus, the total number of beads in the SBS model chain of the considered locus is the *M***r*. The procedure also estimates the optimal value of *r*, *r**. For instance, for the genomic locus in human HCT116 cells investigated in [38], we used 30 kb resolution Hi-C data with *M*=83 windows and found *r**=10 (see Section 3.1), thus returning polymer models made of 830 beads. The method is general to account for any type of input contact data, e.g., Hi-C, GAM, SPRITE, or microscopy-based average distances, albeit in the following, for simplicity, we focus on the case of Hi-C data.

PRISMR is based on a standard simulated annealing Monte Carlo (SA) optimization procedure [130], which minimizes the distance between the predicted polymer model and the input contact matrix, under a Bayesian weighting term to avoid overfitting. To this aim, the algorithm employs a cost function, $H=H_0+H_\lambda$, which accounts for two distinct contributions: H_0 is the standard mean squared error function, i.e., the average squared distance between the input and model-derived contact matrix, while H_λ is the Bayesian term (a chemical potential), weighted by a regularization parameter

$\lambda \geq 0$ that penalizes the addition of new interacting beads on the polymer chain (the larger λ is, the more the addition of a new binding site is penalized). PRISMR seeks the minimum of the total cost function, H, in the space of all SBS polymers with n allowed colors by the SA iterative procedure (**Fig. 2.3**): starting from a random initial assignment of the binding sites along the chain, the type (i.e., the color) of a randomly chosen bead is changed at random, the average contact matrix of the new polymer is computed out of physics, and the cost function evaluated until convergence. Such a procedure is repeated many times using different initial conditions to scan the space of the parameters (n, λ, r), in order to find their optimal values, (n^*, λ^*, r^*), that explain the input Hi-C contact matrix within a given accuracy.

A computationally demanding step of PRISMR is the calculation of the equilibrium thermodynamics average contact frequency, $C(i,j)$, during each iteration of the SA procedure. That can be achieved, for instance, by molecular dynamics (MD) computer simulations (see Section 2.3), which, however, may require huge computational efforts. To speed up the computation and make our procedure feasible, e.g., over genomic scales, we implement, as typical of statistical mechanics [131], a mean-field approximation, in which $C(i,j)$ is estimated based on the average contact frequency between two sites at the same genomic separation in a homopolymer SBS model with the same number of beads [9]. In particular, two sites, i and j, of the same color have as average contact frequency the value in the corresponding SBS homopolymer of interacting beads [45], whereas in case they belong to different color types, $C(i,j)$ is equal to the interaction frequencies between beads belonging to different types in different domains in the SBS model [9,45] (which is typically negligible compared to the other case). We tested by extensive MD simulations that such an approximation performs comparatively well and provide similar results of the full-scale MD simulations [9].

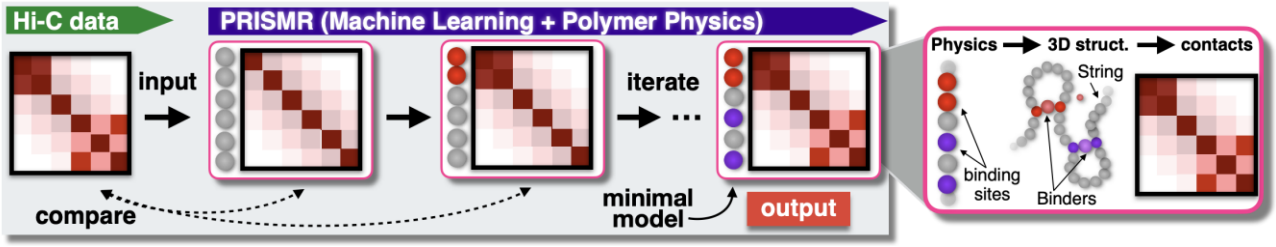

Figure 2.3: Scheme of the PRISMR iterative procedure. PRISMR is a polymer physics and machine learning based procedure that infers the genomic location and types of the putative binding sites of the SBS polymer model of a given genomic region. The approach takes as input only bulk (e.g., Hi-C) contact data, with no use of epigenetic tracks or additional parameters. Left: The method involves five main iterative steps: *(i)* consider a polymer model with a given arrangement of binding sites; *(ii)* derive a thermodynamics ensemble of its 3D equilibrium conformations out of physics; *(iii)* compute its contact matrix and *(iv)* compare it with the input contact data; *(v)* change the polymer model accordingly and repeat until convergence. The output of the procedure is the minimal polymer model of the considered genomic region. Right: By massive MD simulations of the optimal polymer model inferred by PRISMR we derive a thermodynamic ensemble of single-molecule 3D conformations of the considered genomic region. Different types of binding sites are visually represented by distinct colors. Molecular binders can bridge their cognate sites on the chain, thus folding the chain via a mechanism of polymer phase-separation. Adapted from [9].

2.2.2 An improved version of the algorithm

In [38], we changed and redesigned different aspects of the PRISMR algorithm. First, to better control for genomic distance effects, we implemented a new cost function, in which each term of the mean squared error function H_0 is normalized by the average Hi-C contact frequency at that genomic distance. That prevents the Hi-C matrix elements close to the diagonal to dominate the calculations, as they have much higher values than those corresponding to larger genomic distances.

Next, we developed a more refined and quantitative procedure to estimate the optimal parameters (n^*, λ^*, r^*). Based on a standard approach of supervised learning, we divide the input Hi-C data in two complementary sub-sets: a training set and a test set (**Fig. 2.4a**). During the SA procedure the cost function is evaluated only on the matrix elements of the training set, while the test dataset is used to test the model predictions. In [38], for instance, we randomly split the Hi-C dataset into 70% training and 30% test set, yet we verified that the estimated parameters are robust upon changing the size of the training set, e.g., from 50% to 80% of the Hi-C data. To assess, e.g., n^*, i.e., the optimal number of different types (colors) of binding sites of the putative polymer model of a considered genomic region, we iterate the SA procedure for different values of n and evaluate the cost function of the corresponding output model on both the training and test sets. While the function $H_0(n)$ over the training set decreases with n until plateauing [9], the one evaluated over the test set first decreases up to reach a minimum and then increases, signaling overfitting (**Fig. 2.4b, c**). The minimum of $H_0(n)$ over the test set provides the optimal n^*, for which the model has its best predictive power. For each n, we typically run at least 20 independent SA simulations with distinct random selections of the training set and varying initial conditions (different random initializations of the polymer model). For example, in the case of the loci studied in Conte20, we found n^*=4 in the HCT116 model, n^*=3 in HCT116+ Auxin and n^*=7 in IMR90 (see Sections 3.1 and 3.4).

The optimal values of λ and r, respectively λ^* and r^*, are estimated by using a similar procedure. Specifically, to find λ^*, we fix n^* and search the minimum of the extended cost function $H=H_0+H_\lambda$ at varying λ values. As described before, the input Hi-C dataset is split in two complementary training and test sets (e.g., 70% and 30% respectively) and PRISMR is only performed on the training set. Then, we run a battery of independent SA simulations at varying initial conditions and identify λ^* as the minimum of the H_0 function over the test set. The values returned by the procedure in [38] are: $\lambda^* = 10^{-5}$ in HCT116 and HCT116 + Auxin, and $\lambda^* = 10^{-4}$ in IMR90. Next, by fixing n^* and λ^*, we proceed in the same way to estimate r^*, i.e., the optimal number of polymer beads corresponding to a given window (bin) of the input Hi-C contact matrix. For instance, we found r^*=10 in the models discussed in [38]. As a final step, to identify the model corresponding to the absolute minimum of the cost function, we perform additional 10^2 SA simulations from different initial conditions by using the estimated optimal parameters (n^*, r^*, λ^*). The models corresponding to the lower 10% minima are found to be statistically similar to each other, hence proving the robustness of the procedure [9].

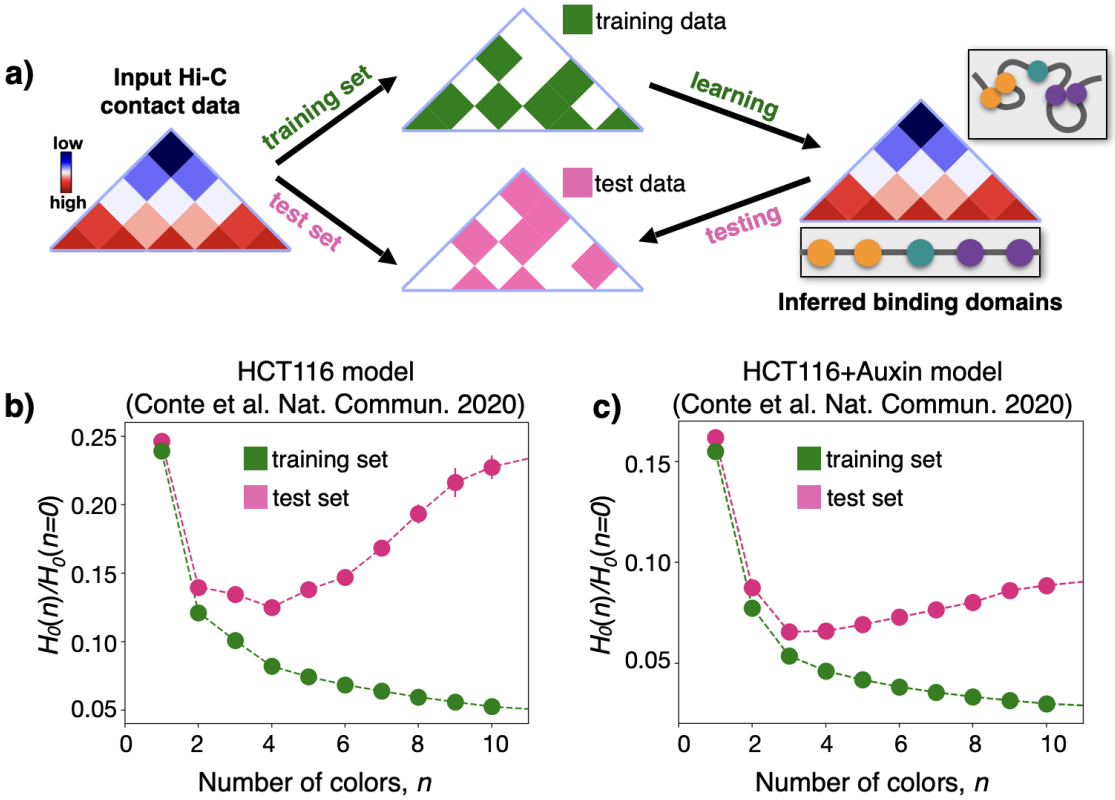

Figure 2.4: The improved PRISMR procedure. a) In its improved version [38], PRISMR takes as input bulk, e.g., Hi-C, contact data, which are split in two complementary training (green) and test (pink) sets. To infer the minimal SBS polymer model that best explains the input contact data, the machine learning procedure is run on the training dataset, while the test set is used to evaluate the contact frequencies predicted by the model. **b)** The cost function H_0, i.e., the normalized mean squared difference between model and experimental contact maps, is plotted against the number of distinct types of binding sites (colors), n, over the training (green) and the test (pink) sets in the case of the SBS model of a 2Mb wide locus in human HCT116 cells [38]. Data shown are mean values +/- s.e.m. by taking 20 independent simulations for each point. The value of n corresponding to the minimum of H_0 in the test set is where the model has its best predictive power. In the plot, the function $H_0(n)$ is normalized by $H_0(n=0)$, i.e., the corresponding value of a polymer model with no binding sites. **c)** Same as **b)**, for the SBS model of a 2Mb wide locus in human HCT116+Auxin cells. Adapted from [38].

2.3 Molecular Dynamics (MD) simulations of the SBS model

To derive a thermodynamic ensemble of single-molecule 3D structures of specific genomic loci, we perform massive Molecular Dynamics (MD) simulations of the optimal polymer models inferred by PRISMR [9,38]. We employ the publicly available LAMMPS software [132] to integrate the system motion equations, which are numerically solved via the velocity-Verlet algorithm until thermodynamic equilibrium is reached. MD simulations are performed as detailed in classical studies of polymer physics simulations [127] and all technical details reported below are summarized from published works [25,38,133,134].

2.3.1 Motion equation and physical interaction potentials

Polymer beads and binders are modelled as hard spheres having, for simplicity, the same mass, m, and diameter, σ. They are Brownian particles embedded in the nuclear medium and subject to the Langevin equation:

$$m\frac{d^2}{dt^2}\vec{x}(t) = -\zeta\frac{d}{dt}\vec{x}(t) - \vec{\nabla}U\big(\vec{x}(t)\big) + \vec{\xi}(t)\,, \tag{1}$$

where $\vec{x}(t)$ is the vector of the particle spatial coordinates, U is the total potential acting on the particle (see below), $\vec{\xi}(t)$ the white noise term of the randomly fluctuating force produced by the viscous nuclear environment. The friction coefficient ζ is linked to the cell nucleus viscosity, η, via the Stokes relation $\zeta=3\pi\eta\sigma$. As usual in MD implementations [127], the simulations are performed in dimensionless units (reduced or Lennard-Jones units): m, σ, K_BT (the thermal energy) are set equal to 1. A typical value in LJ units for the friction coefficient is ζ=0.5 [46,47,127]. To avoid boundary effects, the system moves in a simulation cubic box with periodic boundary conditions, whose linear size is proportional to the gyration radius of a SAW chain with the same number, N, of beads ($\propto N^{0.59}$) [25].

We use in our simulations standard interaction potentials developed in classical polymer physics studies [127] (**Fig. 2.5**). The physics potential, U, in the Langevin equation (1) is the sum of three distinct contributions: *i)* a finitely extensible non-linear elastic (FENE) potential with standard parameters [127] between consecutive monomers of the chain:

$$V_{FENE}(r) = -\frac{1}{2}k\left(\frac{R_0}{\sigma}\right)^2\ln\left[1-\left(\frac{r}{\sigma}\right)^2\right], \tag{2}$$

where k is the FENE spring strength, set to $30K_BT/\sigma^2$, and R_0 is its maximal extension, i.e., the maximal length of the bond ($V_{FENE}=\infty$ if $r \geq R_0$; $R_0=1.6\sigma$); *ii)* the Weeks-Chandler-Andersen (WCA) potential, which models the hard-core repulsion between two adjacent polymer sites:

$$V_{WCA}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 + \frac{1}{4}\right], \tag{3}$$

where $\varepsilon=K_BT$ is the energy unit and r the distance between two particle centres. Such a potential is null if r is greater than $2^{1/6}\sigma\approx1.12\sigma$ [25,127]; *iii)* a truncated short-range, attractive Lennard-Jones (LJ) potential [46] between polymer beads and cognate binders:

$$V_{LJ}(r) = 4\varepsilon_{int}\left[\left(\frac{\sigma_{bb}}{r}\right)^{12} - \left(\frac{\sigma_{bb}}{r}\right)^6 - \left(\frac{\sigma_{bb}}{r_{int}}\right)^{12} + \left(\frac{\sigma_{bb}}{r_{int}}\right)^6\right], \tag{4}$$

where ε_{int} is the parameter controlling the strength of the interaction, σ_{bb} is the sum of bead and binder radii, r the particle centre-to-centre distance and r_{int} the cut-off distance value that sets the interaction range (i.e., $V_{LJ}=0$ if $r \geq r_{int}$). Finally, the absolute value of the minimum of the potential V_{LJ}, E_{int}, defines the energy scale of the specific interaction between beads and cognate binders:

$$E_{int} = \left| 4\varepsilon_{int} \left[\left(\frac{\sigma_{bb}}{r_{int}} \right)^6 - \left(\frac{\sigma_{bb}}{r_{int}} \right)^{12} - \frac{1}{4} \right] \right| . \quad (5)$$

We also considered the case where along the chain there are unspecific binding sites for binders, characterized by a lower energy affinity [38]. In our computer simulations, typical values of the specific, as well as unspecific, binding energy affinities are in the weak biochemical energy range [9,25,38] (e.g., few units of K_BT). For instance, in [38] we explored a spectrum of specific and unspecific affinities between binders and binding sites, respectively, from 3.1 to 8.0K_BT (for simplicity equal across the different types) and from 0 to 2.7K_BT (see, e.g., Section 3.1). The minute details of those parameters, however, do not affect the general, equilibrium properties of the model, because of the Statistical Mechanics concept of universality in phase transitions [62].

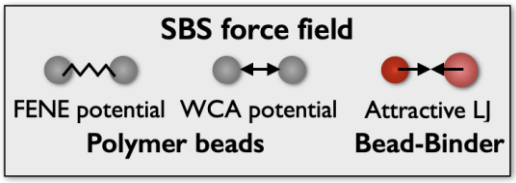

Figure 2.5: Summary scheme of the interaction potentials of the SBS model. An elastic FENE spring connects consecutive beads on the chain, yet their overlap is prevented by a repulsive Weeks-Chandler-Andersen (WCA) potential. A truncated, short-range Lennard-Jones (LJ) potential produces the attractive interaction between beads and cognate binders.

2.3.2 Conversion of MD parameters into physical units

We use standard MD procedures to map the dimensionless parameters of our simulations into physical units [47,135]. For instance, the model length scale, i.e., the bead diameter σ, can be estimated by assuming that the local density of chromatin loci equals the expected average density of DNA in the entire nucleus. As a result, $\sigma = (s_0/G)^{1/3}D_0$, where G is the total genomic content of DNA in the cell, D_0 the average nuclear diameter of the considered cell type and s_0 the genomic content of each chain polymer bead [45]. Alternatively, if available, experimental, e.g., microscopy-based, distance data can be used to calibrate the length scale of the models. In the paper [38], for example, σ is estimated by using single cell multiplexed FISH imaging data at 30kb resolution [21] available in several human genomic loci (see, e.g., subsection 3.3.1). Similarly, the molar binder concentration, c, is converted into μmol/l via the relation c=P/(VN_A) [25], where P is the absolute number of binders, V the box volume, and N_A the Avogadro number. The MD time scale, τ, is given by $\tau = 6\pi\eta\sigma^3/(K_BT)$, where η is the solvent viscosity. Changes to the viscosity, whose reference values range around 0.01-0.03P [38,46], only proportionally change the time scale.

2.3.3 Steady-state 3D conformations from initial states

The initial states of our MD simulations are distinct SAW conformations, prepared as described in [127]: first, we generate a random walk chain in which the distance between two consecutive beads is 0.97σ, i.e., the average length of an equilibrium SAW conformation subject to the FENE (2) and WCA (3) potentials described above; next, to remove overlaps between polymer beads, we let the system equilibrate, for 10^7 MD iteration timesteps [25], under the soft potential:

$$V_{soft}(r) = A\left[1 + cos\left(\frac{\pi r}{L}\right)\right], \tag{6}$$

where the factor A linearly increases in time and $L=2^{1/6}\sigma$ ($V_{soft}=0$ if $r \geq L$). To check that the polymers are in the SAW universality class, we perform a scaling analysis by computing the average pairwise contact probability $P_c(s)$, as well as the mean squared Euclidean spatial distance $R^2(s)$, as a function of the genomic distance (i.e., the polymer contour length). Indeed, textbook results provide, respectively, $P_c(s) \sim s^{-\alpha}$ and $R^2(s) \sim s^{2\nu}$, where α=-2.1 and $\nu \simeq 0.59$ [62,129]. As an example, we show in **Fig. 2.6** the scaling exponents of the SAW conformations used in [38] as initial states of the MD simulations of a 2Mb wide locus in human HCT116 cells.

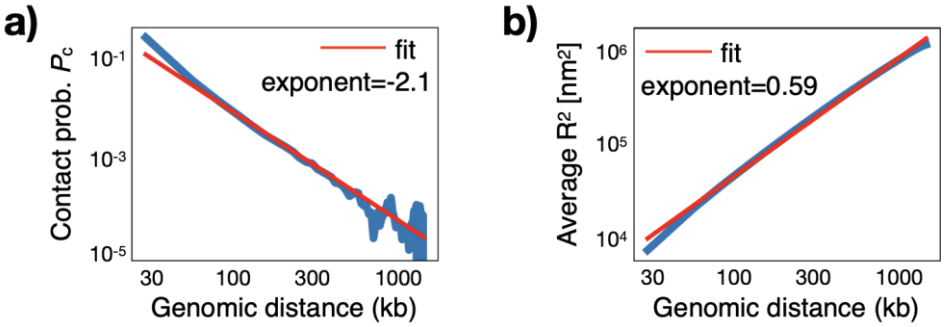

Figure 2.6: Scaling laws of the SAW universality class. To check that the initial states of our MD simulations are folded in the SAW universality class, we compute the scaling exponents of **a)** the pairwise contact probability, P_c, and **b)** the average squared distance, R^2, as a function of the genomic distance along the chain (i.e., the polymer contour length). As expected from textbook results [62,129], those functions scale as, respectively, $P_c(s) \sim s^{-2.1}$ and $R^2(s) \sim s^{2*0.59}$.

Then, we introduce the binders in the simulation box and let them interact, as described above, with specific and unspecific binding sites of the polymer chain, so driving its folding. The system evolves up to when stationarity is reached, e.g., typically up to 10^8 MD time iterations steps [25,38]. To monitor system dynamics, we use as control parameter, for instance, the gyration radius of the chain, defined as [129]:

$$R_g \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r_i} - \vec{r}_G)^2}\,, \tag{5}$$

where N is the number of beads of the polymer and $\vec{r}_G$ the coordinates of its mass center. Upon choosing high enough binder concentrations or affinities (see **Fig. 2.1b**), the gyration radius has a sharp drop in time, which signals the coil-to-globule phase transition of the system. At large times, its asymptotic plateauing marks the reached stationarity (**Fig. 2.7**). A similar drop is also found for other order parameters of the chain, such as its binding energy or separation score (Section 3.1). Finally, to produce a robust statistical ensemble of conformations, we run thousands of independent polymer simulations.

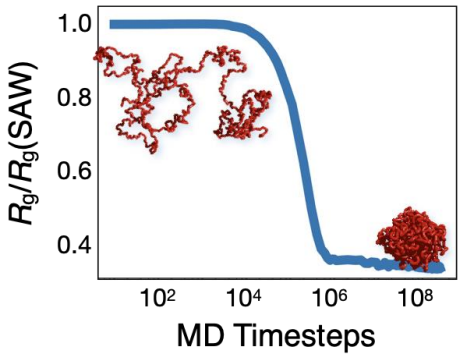

Figure 2.7: The coil-to-globule phase transition of the system is monitored by the time course of its gyration radius. For high enough binder concentrations or affinities, the model undergoes a thermodynamics phase transition from an initial coil state, described by the statistics of the SAW chain, to a globule phase. The model transition is marked by a sharp decrease of the system order parameters, such as its gyration radius as a function of the MD time. In this example, the gyration radius of the chain is normalized by its initial SAW value [38].

2.4 Application to real data: SBS model of the *HoxD* gene region

As an application of the above-described methods to real data, we briefly discuss here the SBS model of a 7Mb region (chr2:71160000-78160000, mm10) around the *HoxD* genes in mouse embryonic stem cells (ESCs) and cortical neuronal cells (CNCs). Those genes are particularly relevant because they are critically involved in limb development and genomic mutations within their region, such as duplications or inversions, have been functionally linked to malformations and diseases [136–138]. The material discussed in this Section is adapted from our published paper [36].

To infer the SBS models of the *HoxD* locus, i.e., the minimal number of distinct types of specific binding sites and their positioning along the polymer chain, we employed our PRISMR approach (see Section 2.2) by taking as input Hi-C data at 5kb resolution [139] available in the studied cell types (ESCs and CNCs). Next, to derive an ensemble of polymer 3D conformations, we performed MD simulations of the models inferred by PRISMR as described above (Section 2.3). To test the accuracy of our models, we compared the experimental Hi-C data in ESCs and in CNCs against the corresponding model-derived pairwise contact matrices (**Fig. 2.8**). The Pearson's correlation between model and Hi-C is r=0.92 and r=0.93 in ESCs and CNCs, respectively. To account for genomic distance effects (i.e., the average decay of Hi-C interactions at increasing genomic distances), we computed the genomic distance-corrected Pearson correlation coefficient, r', between model and experimental maps, that is the Pearson correlation evaluated on contact matrices whose diagonals are subtracted (in both model and experiment) by their average value at that genomic distance. We found this finer correlation to provide high values too (r'=0.48 and r'=0.59, respectively), thus showing that the models well recapitulate the observed Hi-C contact patterns in both cell types (compare, e.g., top panels of **Fig. 2.8a, b** for ESCs and top panels of **Fig. 2.8d, e** for CNCs).

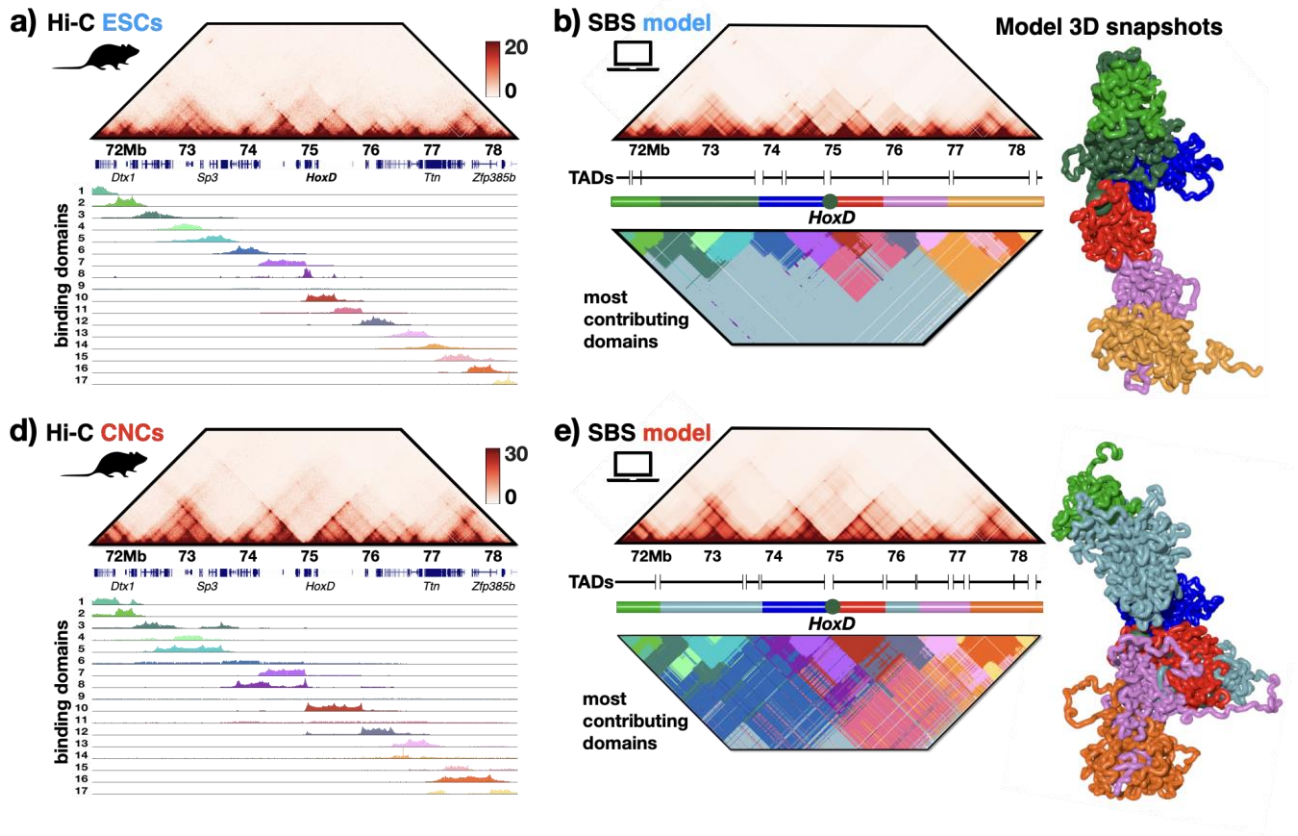

Figure 2.8: The SBS model describes with good accuracy Hi-C contact patterns of the *HoxD* region. a) Top: Hi-C data [139] of the studied *HoxD* region in mouse embryonic stem cells (ESCs). A track with the genes of the locus is also shown. Bottom: The SBS model inferred main binding domains. **b)** Top: SBS model-derived average pairwise contact matrix of the *HoxD* locus. Black segments show the TADs of the region [139]. Bottom: Representation of the model binding domains most contributing to the contacts observed in Hi-C. A single-molecule model 3D conformation of the HoxD region is also shown on the right with its color scheme. **c)** As in panel **a)** but for cortical neuronal cells (CNCs). **d)** Same as in **b)** but for CNCs.

To provide a principled interpretation of the patterns visible in the contact maps of the locus, we investigated how those patterns originate from polymer physics by the interactions of the model binding sites. In ESCs, the model identified 17 main binding domains (visually represented by different colors), each having a significant overlap with a single TAD or sub-TAD of the locus (**Fig. 2.9a**). In fact, by visually plotting the most contributing domain to each pairwise contact, we found that the TADs visible in the Hi-C data roughly correspond to DNA regions particularly enriched by contacts linked to one of the binding sites of the model (see **Fig. 2.8b**, bottom). Yet, binding domains overlap with each other along the considered DNA sequence; that allows to faithfully capture the observed finer TAD internal structures, as interactions within a TAD can be associated with more than a single binding domain. The model also identifies more spread binding domains (**Fig. 2.9a**), not directly associated to a single TAD, that originate the weaker, but not negligible, long-range interactions of the locus (e.g., the interactions across the different TADs). Similar results are also found in CNCs, where, interestingly, the binding domains exhibit a stronger genomic overlap with each other, resulting in much more marked higher level of inter-TAD interactions as observed in Hi-C data (**Fig. 2.8e** and **Fig. 2.9b**). Finally, to guess the molecular nature of the inferred binding domains of the models, we computed the Pearson correlation between their genomic positions and

the profiles of available epigenetic and histone marks [139,140] (**Fig. 2.9a, b**, rightmost panels; all details of the calculations are in the original work [36]). Intriguingly, we found that each type of binding sites corresponds to a different combination of epigenetic marks (e.g., CTCF/H3K4me3, H3K4me1/H3K27ac, and so on), rather than a single factor.

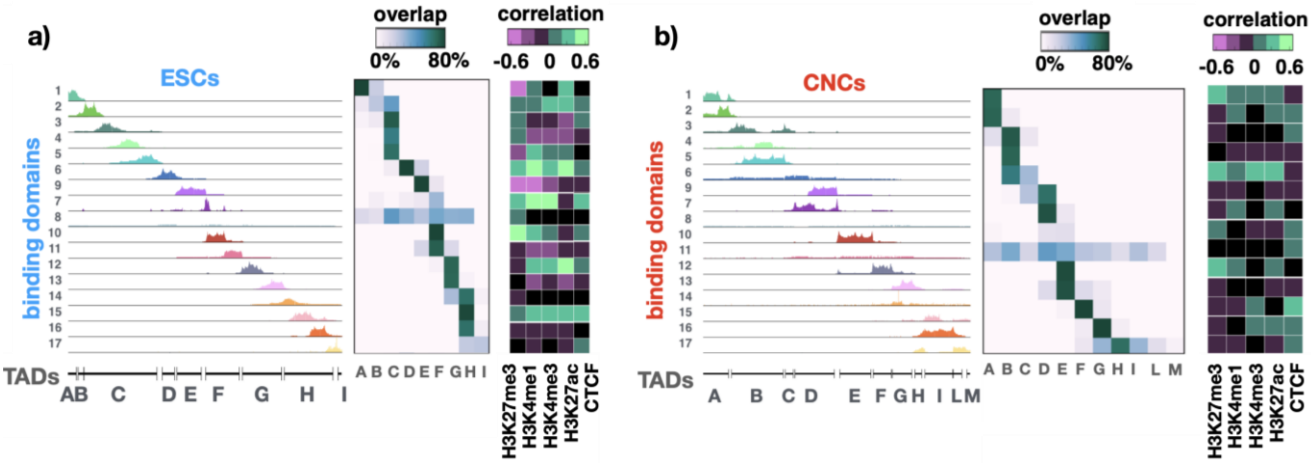

Figure 2.9: Correlation of binding domains with TADs and available epigenetic marks. Genomic location of the model different binding domains (left) inferred from Hi-C data [139], their overlaps with TADs as defined in [36,139] (middle) and their significant Pearson correlations with available chromatin marks [139,140] (right) in **a)** ESCs and **b)** CNCs.

Summarizing, the high correlations between models and experiments support a principled interpretation of chromatin contact patterns through polymer physics, as envisaged by the SBS model. In the emerging scenario, structures such as TADs and metaTADs can be explained as regions enriched for contacts between specific types of binding sites, which self-assemble along the polymer chain via a thermodynamic process of phase-separation and are correlated each with a specific combination of known epigenetic factors (e.g., including, but not limited to, CTCF).

3. A thermodynamic mechanism of polymer phase-separation unveils the origin of contact patterns in single-cells

In this Chapter, which is entirely based on the paper [38], we aim to investigate the molecular mechanisms shaping chromatin folding at the level of single DNA molecules by using polymer-physics-based approaches, such as the SBS model discussed before. To this aim, we focus on two 2Mb wide loci in human HCT116 and IMR90 cells, where both Hi-C [88,117] and independent single-cell imaging data are available [21]. In our strategy, the SBS polymer models are inferred from bulk Hi-C, whereas model predictions on single-molecules are validated against single-cell microscopy conformations. Specifically, in Section 3.1 we report with all details the polymer models of the considered HCT116 and IMR90 loci and discuss their phase transition from a coil to a globule phase-separated state. In Section 3.2, we describe a method to directly compare pairs of experimental-model 3D structures and show that all single-cell imaged conformations of the studied loci statistically map onto conformations of the model belonging to its thermodynamics globule phase-separated state. To further validate our model, in Section 3.3 we compare the features of its predicted 3D structures against those observed in single-cell experiments, such as the TAD boundary probability function or the average separation score, and find a significant agreement in all those comparisons. Next, we quantitatively investigate the cell-to-cell structural variability of chromatin single-cell conformations and show that it naturally results from the inherent folding degeneracy of the phase-separated conformations of the model. We also test, in Section 3.4, the predictions of our polymer model upon removal of the cohesin complex, which is a known key chromatin architecture organizing factor. Our results, consistent with single-cell imaging experiments [21], indicate that cohesin depletion tends to reverse chromatin globule phase separation to the coil thermodynamics state in most single cells, resulting in much more variable and transient contact patterns in single molecules. Finally, in Section 3.5, we explore the steady-state time dynamics of chromatin structure at the single-molecule level and discuss how stochasticity of DNA interactions can coexist with contact specificity. The overall agreement between single-cell imaged and model-derived conformations supports the view whereby, in the studied HCT116 and IMR90 loci, chromatin folding is explained at the single-cell level by a thermodynamics physics mechanism of polymer phase-separation. All the material reported in this Chapter is taken or adapted from [38].

3.1 SBS polymer models of two 2Mb wide genomic regions in human HCT116 and IMR90 cells

We describe here the SBS models of two 2Mb wide loci in human HCT116 and IMR90 cells (genomic coordinates: chr21:34.6–37.1Mb and chr21:28–30Mb, hg38), where bulk Hi-C [88,117] and independent single-cell multiplex FISH data [21] are available. In the subsection 3.1.1, we report the experimental Hi-C contact data of those loci that we use in our approach as input to infer the binding domains of the SBS models. The details of the models and their thermodynamic coil-to-globule phase transition are discussed in subsections 3.1.2 (for HCT116) and 3.1.3 (for IMR90). Finally, in subsection 3.1.4 we analyze the epigenetic signature of the inferred SBS binding domains. The content of this Section is adapted from the paper [38].

3.1.1 Hi-C contact data of the studied loci

To infer the putative binding sites of the SBS models of the considered loci, we employed our PRISMR routine (Section 2.2) by taking as input bulk Hi-C data from [117] and [88], respectively, in HCT116 (**Fig. 3.1a**, top) and IMR90 (**Fig. 3.1b**, top) cells. We used 5kb resolution Hi-C data, which we re-binned at 30kb in our modeling to match the resolution of multiplexed FISH data [21]. The Hi-C maps of both cases have specific patterns of TAD and sub-TAD structures, showing that in both loci chromatin is partitioned into preferential self-interacting domain structures at the population-averaged level. The boundary probability function derived from imaging data [21], i.e., the probability for each genomic position to be the boundary of a single-cell domain, is enriched in correspondence of the main TAD boundaries of the contact matrices, yet it is nonzero across all genomic positions because of the high cell-to-cell variability of TAD-like structures (**Fig. 3.1**, middle; see also Section 1.4). Consistently, the separation score [21], which measures the level of spatial segregation at each genomic position, has peaks matching those of the boundary probability function and, again, is nonzero along the studied loci, revealing that chromatin folds into spatially segregated globular domains that, however, broadly vary in single cells (**Fig. 3.1**, bottom). As detailed in the next subsections, by machine learning from only Hi-C data (**Fig. 3.1**), we infer the SBS models of the considered HCT116 and IMR90 loci and, next, by MD simulations we derive a corresponding ensemble of in-silico 3D structures.

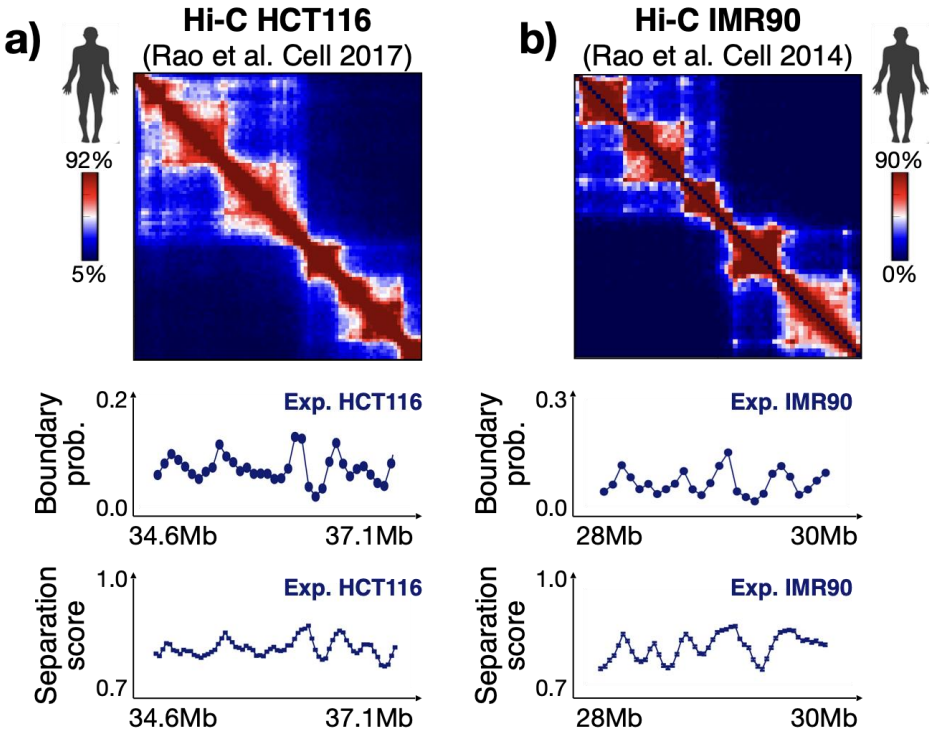

Figure 3.1: Bulk Hi-C contact data of the studied HCT116 and IMR90 loci. a) Top: Hi-C contact data [117] of the studied chr21:34.6-37.1Mb locus in human HCT116 cells. Experimental resolution is 30kb. Middle: Genomic boundary probability of the locus in HCT116 as measured in single-cell imaging experiments [21]. It is enriched in correspondence of the genomic locations of the main TAD structures of the locus, yet it is found to be spread across all genomic positions as a reflex of the high cell-to-cell structural variability. Bottom: The spatial segregation [21] quantifies the level of spatial segregation along the locus and behaves similarly to

the boundary probability function. **b)** Top: Hi-C contact data [88] of the studied chr21:28-30Mb locus in human IMR90 cells. Experimental resolution is 30kb. Middle: Boundary probability of the IMR90 locus derived from imaging data [21]. Bottom: Separation score [21] of the imaged IMR90 locus.

3.1.2 SBS model of the HCT116 locus and its coil-to-globule phase transition

In the HCT116 locus, by taking as input Hi-C data from [117] (**Fig. 3.2a**), PRISMR returns four distinct types of specific binding sites (n^*=4), visually represented by different colors and each defining a binding domain (**Fig. 3.2**). Similarly, the inferred optimal number, r^*, of polymer beads within a single window (bin) of the experimental Hi-C matrix is r^*=10 (see Section 2.2 for definitions). As the considered locus is 2.5Mb wide and the Hi-C resolution is equal to 30kb, the experimental contact matrix is divided into M=2500/30=83 bins, hence providing polymer chains made of $M \times r^*$=830 beads. The polymer system is investigated at different binder concentrations and affinities (for simplicity equal for all types) by massive parallel MD simulations to derive, for each different concentration or binder affinity, a thermodynamic ensemble of single-molecule 3D conformations of the model of the locus. In our study we explored a spectrum of specific and unspecific affinities between binders and binding sites in the weak biochemical energy range, respectively from 3.1 to $8.0 K_B T$ and from 0 to $2.7 K_B T$. After setting the affinities, we sampled almost three orders of magnitude in binder concentrations, from 0 to 0.5μmol/l.

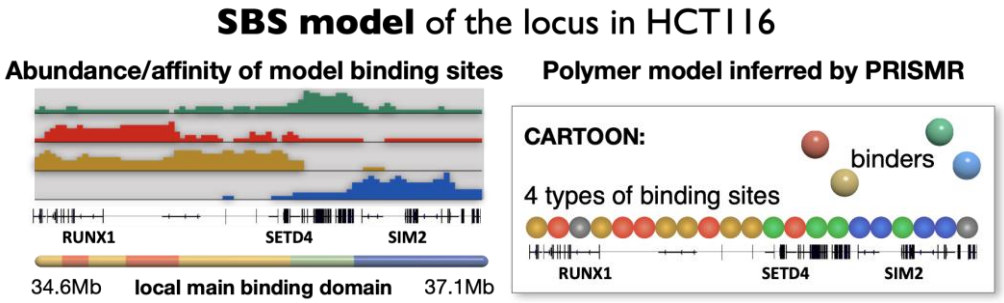

Figure 3.2: SBS model of the HCT116 locus. Left: The inferred SBS model of the studied chr21:34.6-37.1Mb locus in human HCT116 cells has four distinct types of binding sites, forming four binding domains that are visually represented by a different color. Their genomic location and abundance are shown. Right: Cartoon of the SBS model polymer chain with the track of genes (UCSC Genome Browser) highlighted. Adapted from [38].

Upon increasing the concentration of binders, at a characteristic threshold (**Fig. 3.3a**) the polymer undergoes a thermodynamics phase transition from a coil to a globule phase separated state [62], a more condensed structure where the different binding domains self-assemble by action of (and along with) their cognate binders to form more compact and partially separated globules. The conformational rearrangement of the system is signaled by a sharp decrease of its order parameters, such as the equilibrium gyration radius R_g (**Fig. 3.3a**, top) or the system binding energy (**Fig. 3.3a**, middle), i.e., its total potential energy. Similarly, the sharp decrease of the separation score, which measures the level of spatial separation between chromatin segments on either side of a given genomic position, signals that, when the number of binders (or their affinity) increases above threshold, distinct spatially segregated globules self-assemble along the polymer chain (**Fig. 3.3a**, bottom). In the HCT116 model, we find the transition threshold concentration to be about

50nmol/l (**Fig. 3.3a**). More generally, for the explored weak biochemical affinities, such a threshold lies within the fractions of µmol/l range [25], which are values consistent with typical transcription factor concentrations.

To check whether the SBS polymer conformations recapitulate the Hi-C data [117] of the HCT116 locus used to infer its binding sites, we computed the average pairwise contact matrix of the model in the two thermodynamic phases of the theory, i.e., in the coil and the globule phase-separated. To this aim, we performed extensive MD simulations to produce in both states an ensemble of model conformations at thermodynamic equilibrium (**Fig. 3.3b**). Specifically, in the simulations we set the specific and unspecific binding energy equal to $3.1K_BT$ and $2.7K_BT$, respectively, and used as reference a binder concentration c=0.03µmol/l for the coil and c=0.11µmol/l for the phase-separated state (**Fig. 3.3a**). Different choices of binder concentrations in the thermodynamic phases of the system do provide the same results because of the concept of universality in Statistical Mechanics [62]. To compute the average contact matrix of the model we used a standard literature approach [25,46]. In brief, we first compute the pairwise contact matrix of a single-molecule conformation, i.e., a square symmetric matrix C where the entry C_{ij} is equal to one or zero depending on whether the polymer beads, i and j, are in contact or not. Two any polymer sites are in contact if their relative spatial distance is less than a fixed distance threshold. For instance, in the HCT116 model we varied such a threshold from 3σ up to 10σ and checked that our results are robust [38]. Then, the average pairwise contact map is simply the ensemble average of the above-described single-molecule contact matrices. In the coil phase, where entropic forces keep the polymer in randomly folded states, contacts within single-molecules are fleeting and variable, hence producing a structureless contact matrix at the population-averaged level (**Fig. 3.3c**). Conversely, in the globular state, the system exhibits a pattern of TADs and sub-TADs similar those in Hi-C data (**Fig. 3.3c**). To quantify the similarity between model-derived and experimental contact maps, we first computed a standard Pearson correlation coefficient, r, which returned r=0.60 in the case of the coil state and a higher correlation, r=0.88, in the globule phase-separated state (**Fig. 3.3c**). However, since such a measure does not account for genomic distance effects, we also computed the genomic distance-corrected Pearson correlation, r'. We found a low r' correlation value between the coil state of the model and Hi-C contact data, r'=0.13, and a comparatively higher correlation, r'=0.68, in the phase-separated state. In Section 3.2, by using independent single-cell multiplexed FISH data [21] available in the studied HCT116 locus, we will demonstrate that all imaged conformations statistically map onto model single molecules falling in the phase-separated globule state, consistently with the findings discussed here.

Summarizing, these results indicate that the basic physics ingredients of our polymer model are sufficient to recapitulate the bulk Hi-C input contact data used by PRISMR to infer the SBS binding domains of the HCT116 locus. Specifically, based on a systematic MD investigation of the model in its two main thermodynamic phases, we found that the model conformations in the globule phase-separated state provide an accurate fit of the experimental Hi-C contact matrix (r=0.88, r'=0.68), which is, instead, only poorly captured by the coil state (**Fig. 3.3c**).

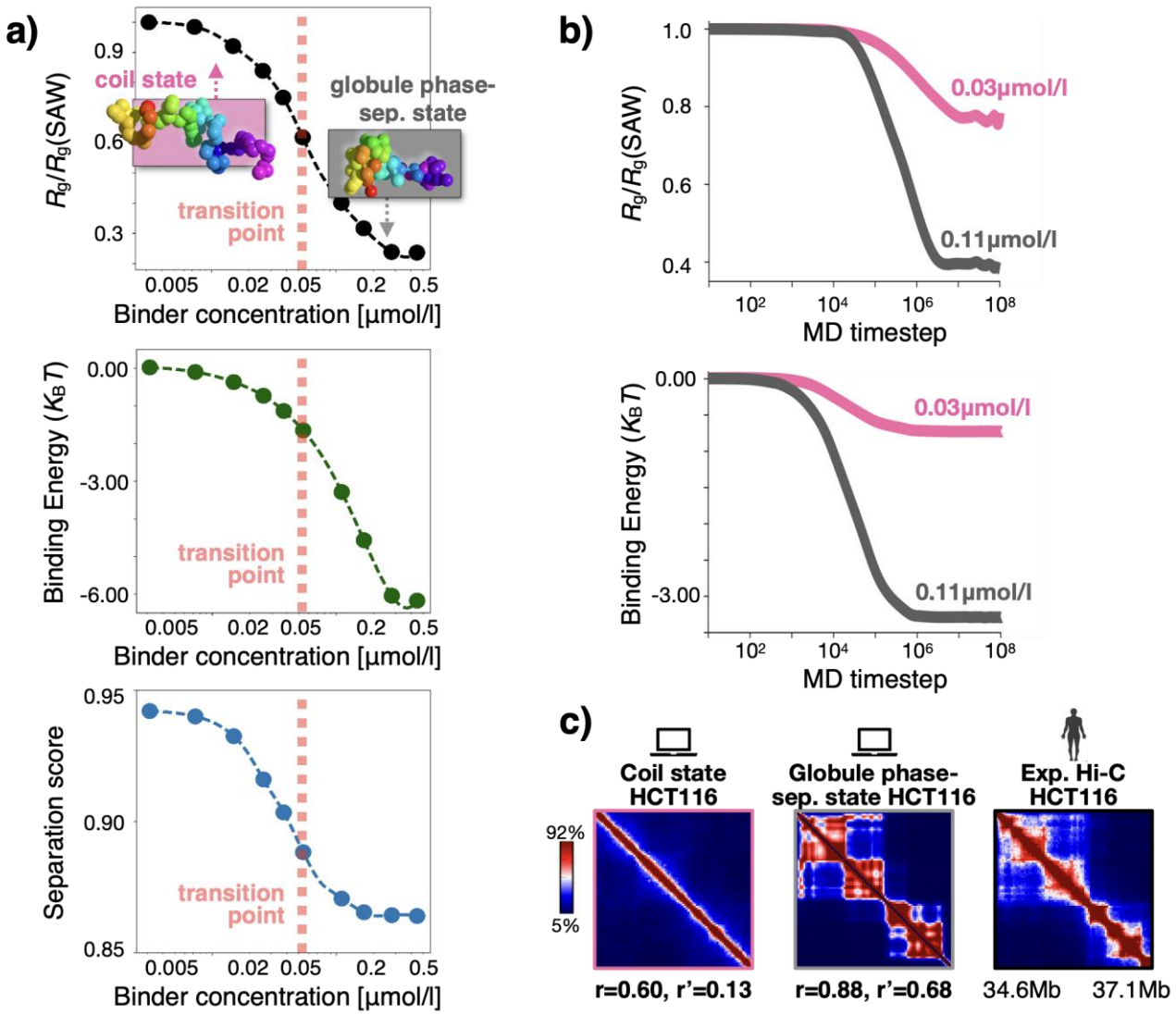

Figure 3.3: The SBS model of the HCT116 locus has a phase transition from a coil to a globule phase-separated state. a) The model phase transition from a coil to a more compact, phase-separated state is signaled by a sharp drop of its order parameters, such as the gyration radius (top), the system binding energy (middle) and the degree of spatial segregation along the chain as measured by the separation score (bottom). In the case of the HCT116 model, the transition threshold concentration is around 50nmol/l. **b)** Extensive MD simulations of the HCT116 model were performed in both the coil (e.g., c=0.03μmol/l) and globule phase-separated state (e.g., c=0.11μmol/l) to derive in each of those phases a statistical ensemble of single-molecule polymer conformations at thermodynamic equilibrium. **c)** The average pairwise contact matrix is computed in both the coil (left) and globular (middle) states of the model and then compared against bulk Hi-C data [117] (right) of the studied HCT116 locus. While the population-averaged contact matrix is featureless in the coil phase, the one from the phase-separated state returns interaction patterns similar to the experiment, as quantified by its comparatively higher correlations. Adapted from [38].

3.1.3 SBS model of the IMR90 locus and its coil-to-globule phase transition

In the IMR90 locus, based on Hi-C data from [88] (**Fig. 3.1b**), PRISMR returns a polymer model made of 630 beads with seven distinct binding domains (**Fig. 3.4a**). As in the case of the HCT116 locus, we systematically explored by MD simulations different binder affinities (in the range 0-8K_BT) and concentrations (from 0 to 0.02μmol/l). The simultaneous sharp drop of the system order

parameters, such as the chain gyration radius (or, similarly, its binding energy) and the average separation score, signals the phase transition of the system from the coil random state to a phase-separated state in which the polymer, due to attractive interactions, forms more compact, separated globules (**Fig. 3.4b**). For the considered weak energy affinities, the collapse of the order parameters occurs at a characteristic threshold concentration, which is around 20nmol/l in the IMR90 case (**Fig. 3.4b**). To check whether the SBS structures recapitulate the bulk Hi-C data [88] of the locus, we derived by MD simulations an ensemble of equilibrium single-molecule conformations of the model in both the coil and globule states and computed the corresponding average pairwise contact matrices that we compared with the experiment (**Fig. 3.4c**). While in the coil phase, as discussed before, average contact patterns are overall erased, in the phase-separated state the SBS contact matrix mimics with high accuracy the specific contacts visible in Hi-C, as highlighted by its high correlations with the data (r=0.94 and r'=0.74, **Fig. 3.4c**). In the next Section 3.2, by performing an all-against-all comparison between model and microscopy [21] 3D structures of the studied IMR90 locus, we will show that the SBS conformations of the globule phase-separated state are indeed a statistical bona-fide representation of the imaged chromatin single-cells.

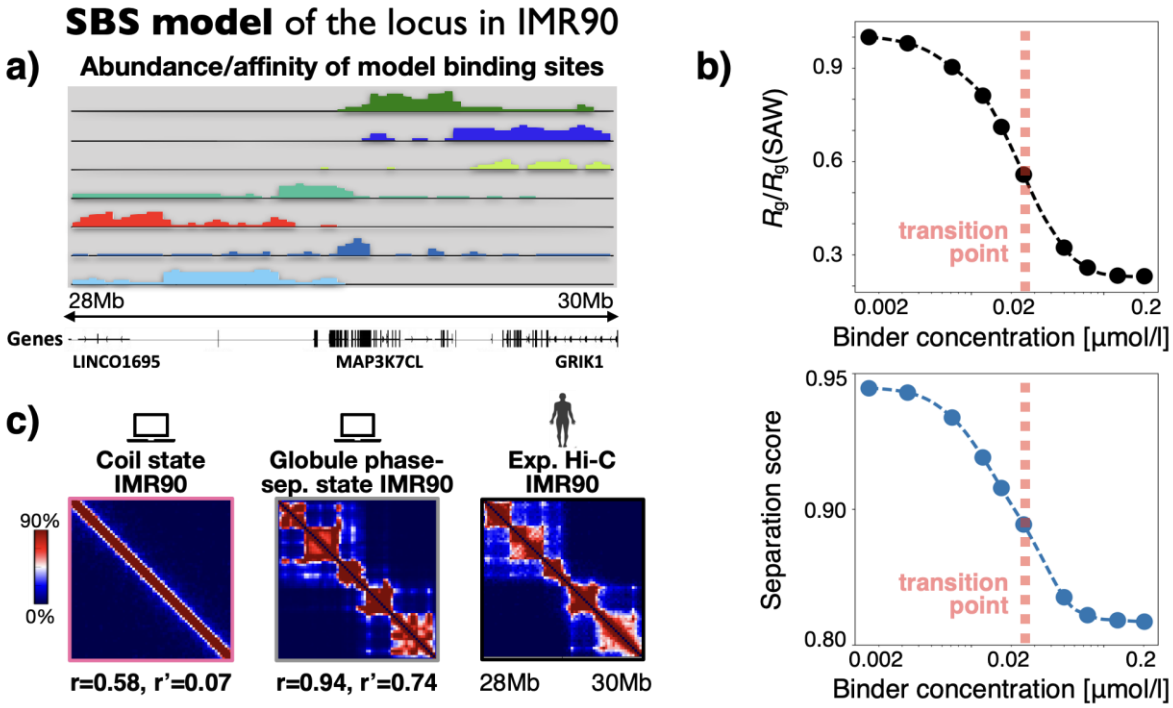

Figure 3.4: The SBS model of the HCT116 locus has a phase transition from a coil to a globule phase-separated state. a) Genomic location and abundance of the inferred SBS model of the studied chr21:28-30Mb locus in human IMR90 cells. The model has seven distinct types of binding sites, forming seven binding domains that are visually represented by a different color. **b)** The model of the IMR90 locus undergoes a thermodynamic phase transition from a coil to a phase-separated state as the number of binders (or their energy affinity) grows above a threshold value. In correspondence of such a threshold, which is around 20nmol/l in the IMR90 case, the order parameters of the system, such as the gyration radius or the separation score of the chain, have a sharp drop. **c)** The average pairwise contact matrix is computed in both the coil (left) and globular phase-separated (middle) states of the model and then compared against bulk Hi-C data [88] (right) of the studied IMR90 locus. We find that only the thermodynamic globular state of the theory recapitulates experimental contact data, as quantified by its higher correlations. Adapted from [38].

3.1.4 Epigenetics signatures of the binding domains of the SBS models of the studied loci

To gain insights into the molecular nature of the inferred model binding domains (**Fig.s 3.2a, 3.4a**), which are responsible of folding, we correlated their genomic positions with a set of epigenetic tracks available in the studied loci (**Fig. 3.5**). To this aim, we used Chip-seq data from [117] and from the ENCODE database [141] and proceeded as detailed in [38]. Specifically, for each of the considered loci, we first binned the epigenetic tracks at 30kb resolution (i.e., the resolution of the employed Hi-C data) and then we computed a Pearson correlation coefficient between each pair of binding domain-epigenetic mark. To test the statistical significance of the association, we compared such correlations against a random control model [9] where the Pearson coefficients are computed between chromatin marks and randomized binding domains derived by bootstrapping the positions of their binding sites (we considered more than 100 different realizations for each case). Hence, we considered significant positive correlations those above the 90^{th} percentile of the control distribution, negative correlations those below the 10^{th} percentile. Small changes in those significance thresholds do not affect our results [9,38]. Finally, significant correlations are visually represented as heatmaps (**Fig. 3.5**).

Interestingly, we found that the different model binding types (visually represented by different colors) are each associated to a specific combination of known architecture organizing factors. For instance, in HCT116 the first binding domain of the model (green, in **Fig. 3.5a**) has statistically significant Pearson correlations with the CTCF/Smc1 (Cohesin) system, the second one (red) manly correlates with active marks (e.g., H3K27ac and transcription factors) and less with Smc1, the third (brown) with repressive marks (e.g., H3K27me3), while the fourth (blue) with H4K16ac and specific transcription factors. Similar results are also found in the IMR90 locus (**Fig. 3.5b**), as each of the model inferred binding domains has significant correlations with different, specific combinations of chromatin architectural factors (rather than a single one), such as CTCF/Cohesin, H3K4me3 or H3K4me1. Overall, those results support a picture where the 3D architecture of the genome is spontaneously shaped by the combinatorial action of different molecules, modulating each other activity [38].

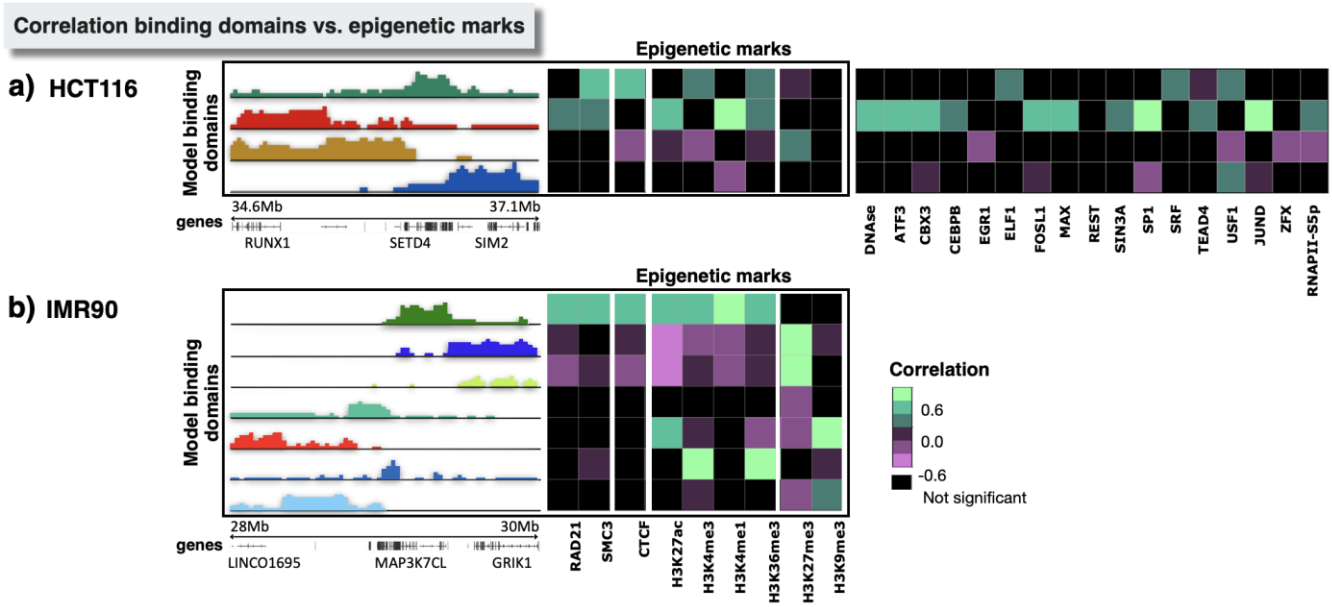

Figure 3.5: Epigenetic signature of the inferred SBS model binding domains. a) The binding domains of the SBS model of the locus chr21:34.6-37.1Mb in human HCT116 cells (left) have statistically significant correlations each with a specific combination of available chromatin epigenetic marks and architecture factors [117] (right). The rightmost panel includes additional factors from the ENCODE database [141] available in HCT116, but not in IMR90 cells. **b)** The binding domains of the SBS model of the locus chr21:28-30Mb in human IMR90 cells are shown with their significant correlations with epigenetics factors [141] (right). Adapted from [38].

3.2 All-against-all comparison of single-cell imaged and single-molecule model 3D structures

We first asked whether the single-molecule 3D structures predicted by the SBS model are a bona-fide representation of chromatin conformations in HCT116 and IMR90 loci. To that aim, we performed an all-against-all structural comparison between single-cell imaged [21] and model predicted 3D structures. We used a computational method [142] that performs a roto-translational alignment between two centered 3D structures (derived, e.g., from experiments and models) by minimizing the root mean square deviation (RMSD) of their coordinate positions (**Fig. 3.6**). Hence, each microscopy conformation is univocally associated to a corresponding, best-matching model 3D structure by searching for the least RMSD. To efficiently implement the structural comparison, we used the object-oriented MDAnalysis Python library, which ensures a rapid calculation of RMSDs and least-square rotation matrices using a quaternion-based characteristic polynomial [143].

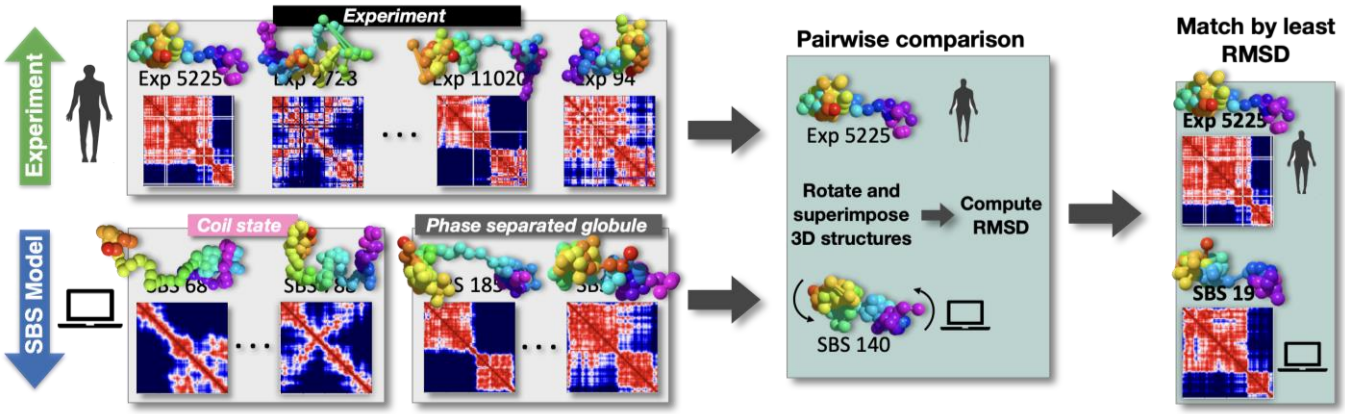

Figure 3.6: All-against-all structural comparison of single-cell imaged and model 3D structures via the RMSD criterion. By using the least RMSD criterion, each 3D structure from imaging data is univocally associated to a corresponding, best-matching 3D structure from the model. Adapted from [38].

In the HCT116 case, we found that all the best-matching model 3D structures of the experimental conformations fall in the globule phase-separated state of the theory, i.e., no imaged structures are optimally mapped onto model conformations of the coil state (**Fig. 3.7a**). That is consistent with our previous results on average contact matrices, where only globule conformations of the SBS model, instead of those in the coil phase, recapitulate bulk Hi-C data (see subsections 3.1.2). To test the association is statistically significant, we compared the RMSD distribution of the experiment-model optimal matches to a control distribution where RMSD is computed between random pairs of experimental structures (null model): the two distributions are statistically different (Mann–Whitney test p value = 0) with only 2% of entries of the former falling above the first quartile of the latter (**Fig. 3.7b**, RMSD is in dimensionless units as a standard z-score is applied on both experimental and model coordinates to have a fair comparison). Additionally, we also checked that each model globule conformation is significantly associated to at least one structure from the experiment, showing that the model well represents the ensemble of microscopy conformations [38]. Finally, the experiment-model best-matching pairs identified via the RMSD criterion indeed return similar 3D structures and interaction patterns, as shown in **Fig. 3.7c.**

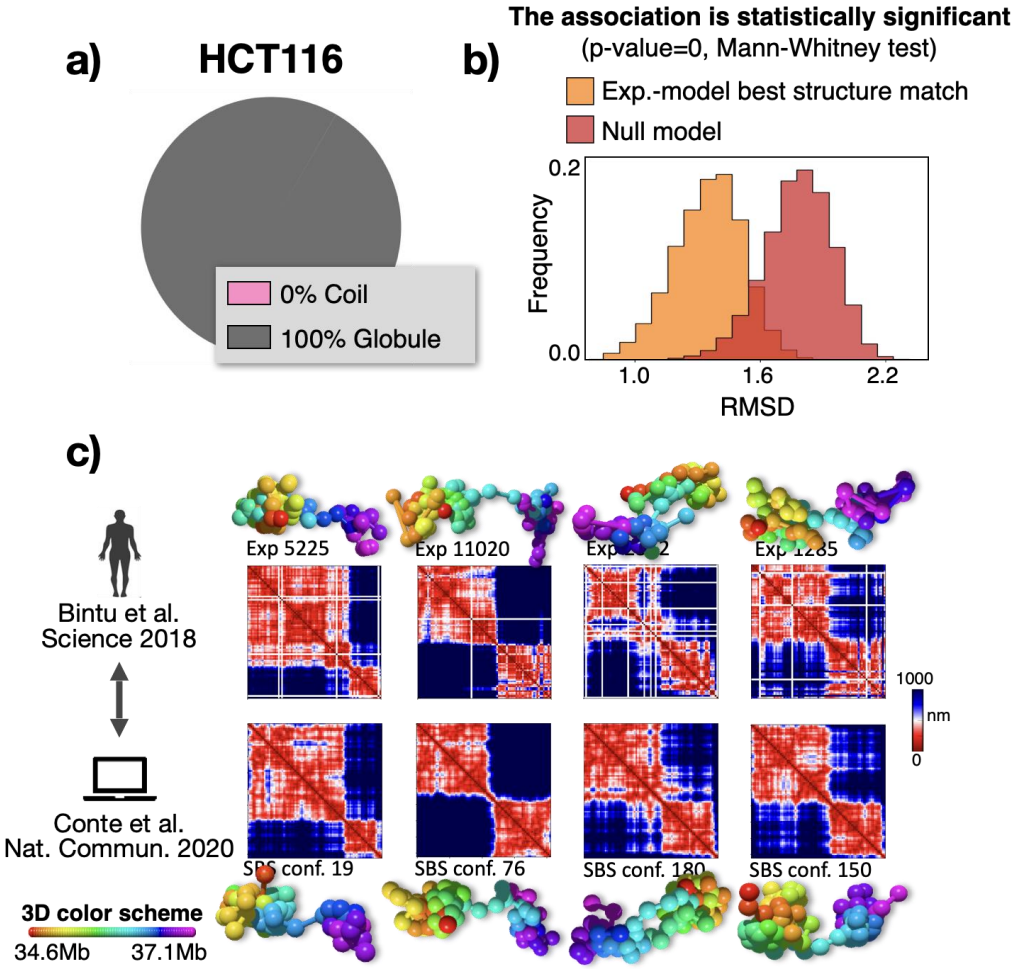

Figure 3.7: Experiment-model best RMSD structure match in the HCT116 locus. a) In the HCT116 case, all imaged 3D structures from the experiment are optimally mapped onto model structures in the globule phase-separated state. **b)** The mapping is statistically significant because the distribution of the optimal RMSD experiment-model best matches is different from the null control model (p value=0, two-sided Mann-Whitney test). **c)** Examples of least RMSD best matches between single-cell 3D structures from imaging data (top) and model globule phase-separated 3D structures (bottom), along with their corresponding distance matrices, in the HCT116 locus. Adapted from [38].

The analysis of the IMR90 locus provided same results (**Fig. 3.8**). Consistent with our other findings on average contact maps (subsections 3.1.3), we found that 99% of experimental structures are mapped onto SBS conformations of the globule state (**Fig. 3.8a**). The association is statistically significant, as the RMSD distribution of the experiment-model best matches and the control RMSD distribution of pairwise comparisons between experimental structures are statistically different (Mann–Whitney test p value = 0) with only 2% of entries of the former falling above the first quartile of the latter (**3.8b**). Also, each model conformation is significantly similar to at least one experimental structure (examples of least RMSD best matches in **Fig. 3.8c**).

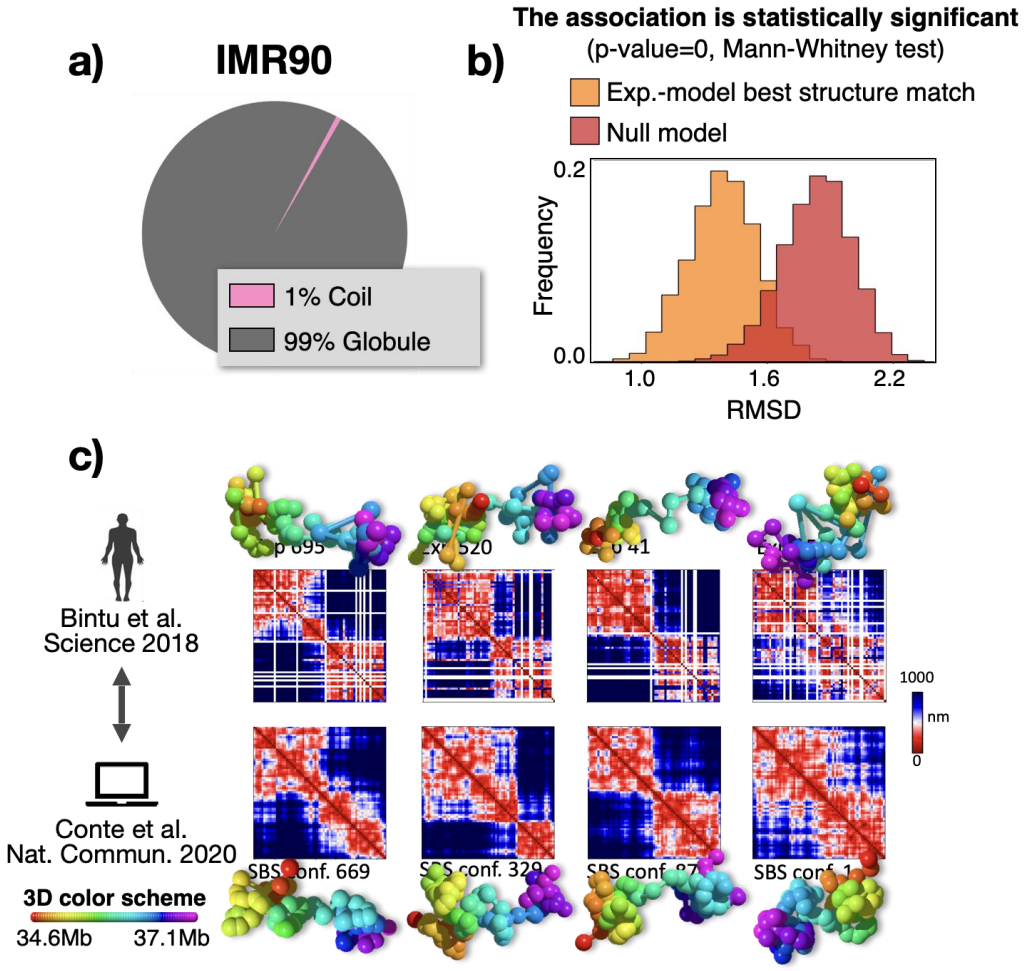

Figure 3.8: Experiment-model best RMSD structure match in the IMR90 locus. a) In the IMR90 case, 99% of imaged 3D structures is optimally mapped onto SBS model structures in the globule phase-separated state. **b)** The association is significant because the distribution of the optimal RMSD experiment-model best matches is statistically distinguishable from the null model (p value=0, two-sided Mann-Whitney test). **c)** Examples of least RMSD best matches between single-cell 3D structures from imaging data (top) and globule phase-separated model 3D structures (bottom), along with their corresponding distance matrices, in the IMR90 locus. Adapted from [38].

Summarizing, based on a quantitative all-against-all structural comparison between experimental and model single-molecules, we found that all 3D structures from imaging data are mapped onto model conformations in the globule phase-separated state. Hence, the SBS model of the studied loci provides a statistical bona-fide representation of chromatin structure in single-cells, as its polymer conformations in the globule state well represent the experimental ensemble.

3.3 Model predicted 3D structures are validated against independent single-cell imaging data

Once assessed that the phase-separated state of the model returns a statistically significant description of chromatin single-cells, we systematically compared the structural features of its predicted 3D conformations against those from imaging experiments [21]. Specifically, in this Section, we show that the SBS polymer structures can recapitulate population-averaged distance

data of the studied loci as well as finer, local properties of chromatin folding, such as the single-molecule TAD boundary probability and strength or the degree of spatial segregation along the locus as measured by the average separation score (subsection 3.3.1). Next, we demonstrate that our ensemble of in-silico structures explores the same conformational space of the imaged conformations, as they have the same degree of single-molecule structural variability (subsection 3.3.2). Finally, to further assess the significance of model predictions, we consider a control block-copolymer model and show that, differently from the SBS, it returns only a poor description of chromatin structure at the single-molecule level (subsection 3.3.3).

3.3.1 The model ensemble of phase-separated single-molecule conformations has features similar to those found in single-cell imaging experiments

To perform a rigorous comparison against multiplexed FISH data [21], we first had to calibrate the dimensionless length scale of the model, i.e., the bead diameter σ, into physical units (e.g., nanometres). To this aim, we computed in both the imaging and SBS datasets the ensemble distribution of gyration radius and estimated σ by equating the median values of the model and experimental distributions (**Fig. 3.9**). We found $\sigma = 45$ nm in the HCT116 and $\sigma = 60$ nm in IMR90 model [38].

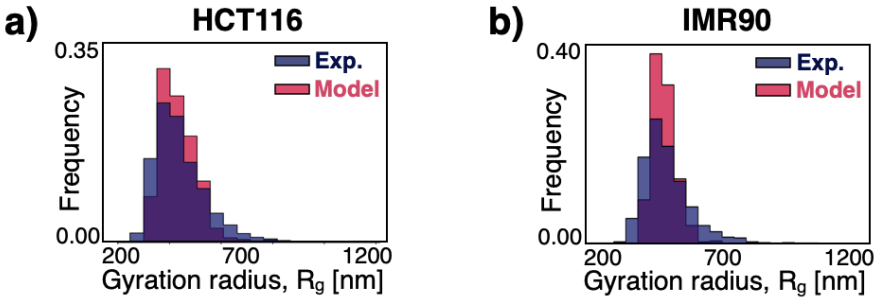

Figure 3.9: Calibration of the model length scale into physical units. The model length scale, i.e., the bead diameter, is converted into physical units by equating the medians of model and experimental gyration radius distributions. **a)** Comparison between model and imaged gyration radius distribution in HCT116 (two-sided Mann–Whitney p value = 0.40). **b)** Same as in **a)** for IMR90 (two-sided Mann–Whitney p value = 0.68). Adapted from [38].

As an initial validation of our model and of its Hi-C-inferred putative binding sites, we compared its predicted median distance matrix in the globular state against imaging data of the studied loci (**Fig. 3.10a, b**). In the HCT116 case, we found that the model and experimental median distance maps are very similar (**Fig. 3.10a**), as quantified by the Pearson, r=0.95, and distance-corrected correlation, r'=0.84, which are, interestingly, even higher than correlations with Hi-C data (see **Fig. 3.3c**). Same results are also found in the IMR90 locus (**Fig. 3.10b**), where, again, the correlations between model and experiment, r=0.96 and r'=0.77, are higher than those with the corresponding bulk Hi-C data used by PRISMR to infer the binding sites of the model (see **Fig. 3.4c**). Next, we derived in our model a set of single-molecule local properties of chromatin folding that we compared against the corresponding quantities from the experiment. For instance, we computed the TAD-like boundary probability function, i.e., the probability for each genomic position to be the

boundary of a single-cell TAD domain (**Fig. 3.10c, d**). As discussed above (subsection 1.4.2), that function is nonzero across each of the considered genomic loci, revealing that domain boundaries broadly vary from cell to cell as they have a finite probability of being located at any genomic position, albeit preferentially at CTCF and cohesin binding sites [21] (**Fig. 3.10c, d**). We found that our model accurately captures the experimental boundary probability function of both the HCT116 and IMR90 loci, as highlighted by the corresponding high Pearson correlation values (r=0.79 and r=0.60, respectively). We also computed the separation score, which locally measures the level of spatial segregation around each genomic position (see **Fig. 1.11** for its technical definition), and found again the predictions of the model to be very close to the experiments (**Fig. 3.10e, f**. Pearson correlations r=0.85 and r=0.79, respectively, in HCT116 and IMR90).

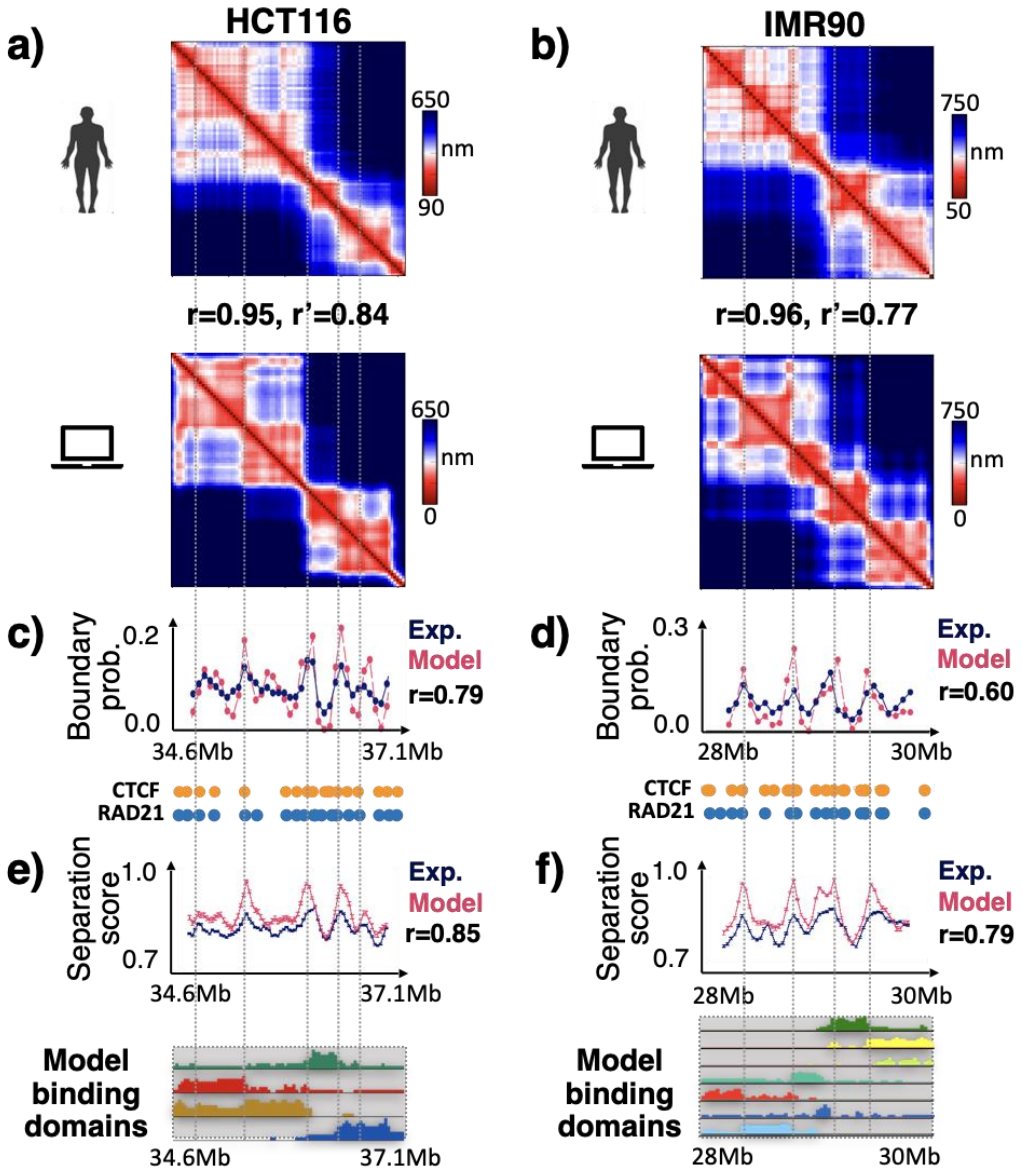

Figure 3.10: Model validation against independent single-cell imaging data. We perform a systematic investigation of the structural features of the single-molecule conformations predicted by the model [38] to test whether they match or not those observed in super-resolution single-cell experiments [21]. **a)** In the considered 2.5Mb wide locus chr21:34.6-37.1Mb of human HCT116 cells, the model median distance matrix in the globule separated state compares well against imaging data (Pearson and genomic-distance corrected

correlations are, respectively, r = 0.95 and r' = 0.84). **b)** The model derived median distance matrix in the globular state of the IMR90 model (chr21:28-30Mb) recapitulates the experiment as well (r=0.96, r'=0.77). **c)** The probability of a domain boundary in 3D conformations along the HCT116 and **d)** IMR90 loci also match (model-experiment correlation r = 0.79 and r=0.60, respectively). **e)** The separation score function, which quantifies the degree of spatial segregation around each genomic position (see **Fig. 1.11** for its definition), is similar in both model and imaged single-molecule conformations in the HCT116 locus (model-experiment correlation r=0.85). **f)** Same as **e)** for the IMR90 locus (r=0.79). The location of ChIP-seq CTCF (orange circles) and cohesin (RAD21, blue circles) sites [141], and the abundance of the model binding sites along the considered loci (bottom, as in **Fig.s 3.2a, 3.4a**) are also shown. The vertical dotted lines help to better compare the panels. Adapted from [38].

To further test the TAD boundary features predicted by the model, we averaged the boundary probability function over all genomic positions of the studied loci. The experimental average boundary probability is found to be close to 8% in both the HCT116 and IMR90 loci and, interestingly, such a value is consistent, within the statistical errors, with the SBS predictions (**Fig. 3.11a, b**). Additionally, for each identified TAD boundary we computed the boundary strength [21], which describes how steeply is the change of the spatial distance across the considered boundary position. The experimental and model distributions of boundary strengths turned out to be similar in both the HCT116 (**Fig. 3.11c**) and IMR90 loci (**Fig. 3.11d**), as also their corresponding average values (**Fig. 3.11e, f**).

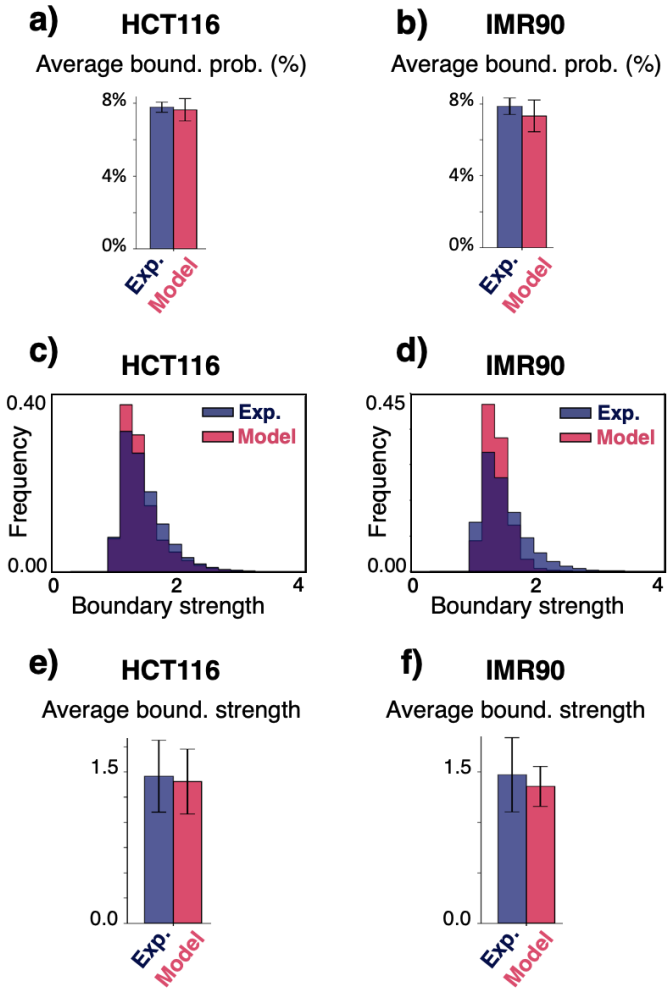

Figure 3.11: Model and imaged single-molecule 3D structures have similar TAD boundary probabilities and strengths. The average boundary probabilities match in super-resolution imaging experiments [21] and in the models [38] of the studied loci in HCT116 (panel **a)**) and IMR90 cells (panel **b)**). Also, the boundary strength distributions of the models in HCT116 (panel **c)**) and IMR90 (panel **d)**) are similar to the experiments, as well as their corresponding average values (panel **e)** and **f)**, respectively). Error bars shown here are the standard errors. Adapted from [38].

Our SBS conformations recapitulate with high accuracy the many different boundary properties of the imaged loci, yet no free parameters are available in all those comparisons. Consistent with the experimentally reported chromatin segregation in globules at the single-cell level [21], the in-silico structures have distance matrices with sharp, and highly varying, TAD boundaries that correspond, as in the imaged structures, to spatially segregated globular 3D conformations in single-molecules (**Fig. 3.12**). Taken together, our results show that the polymer globule phase-separated state of the model returns single molecule structures with features consistent with those found in super-resolution single-cell imaging experiments.

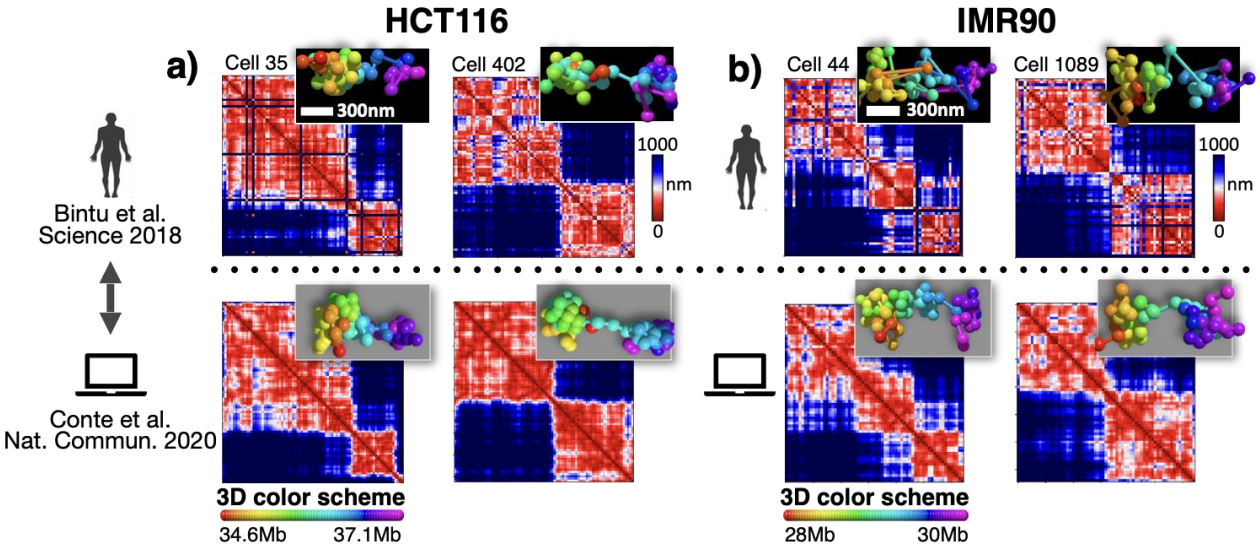

Figure 3.12: The structures predicted by the model reflect the microscopy observed spatial segregation in globules of the studied chromatin loci in single-cells. The model single molecules of the studied loci have distance matrices with sharp, yet highly varying, TAD boundaries that correspond, consistent with microscopy data [21], to spatially segregated globular structures in 3D space. Examples of best-matching experiment-model structure pairs in **a)** HCT116 and **b)** IMR90. The matching is performed via the RMSD criterion (see Section 3.2). Adapted from [38].

3.3.2 Thermodynamic degeneracy in polymer phase-separation explains cell-to-cell structural variability

As a next step, we tested whether the variability of the model single-molecule conformations [38] reflects the experimental structural variability of chromatin single-cells [21]. To this aim, the variability of the imaged single-molecule structures is measured by the distribution of r' correlations between pairs of distance matrices and is compared to the variability of in-silico structures. Specifically, we computed three distinct distributions (see summary scheme in **Fig. 3.13a**): *(i)* the

distribution of r' correlations between pairs of experimental and model single-molecule distance matrices (referred in the following as exp.-model r' distribution); *(ii)* the r' correlation distribution between all pairs of experimental single-cell distance matrices; *(iii)* the distribution of r' correlations between all pairs of model single-molecule distance matrices.

In the HCT116 locus, we found that the experiment-experiment r' distribution is broad with a nonzero average correlation r' = 0.27 (**Fig. 3.13b**, blue histogram), signaling that the imaged single-cell conformations have a significant degree of structural similarity, albeit they are broadly varying. Notably, our model-model r' distance correlation has a similar distribution (red histogram) and, additionally, the distribution of correlations between microscopy and model single-molecule distance matrices (average r' = 0.22) is not statistically distinguishable from the one between experiments (dark grey, two-sided Mann–Whitney p-value = 0.19). As a control case (grey), we computed the r' correlation between pairs of experimental single-molecule distance matrices with bootstrapped diagonals and found, as expected in randomized samples, a zero-peaked distribution statistically different from the others (p-value =0). We also checked that similar results are found by using the simple Pearson correlation, r [38].

We repeated all the above analyses in the IMR90 locus and found that the r' correlation between pairs of single-molecule distance matrices from imaging data has a broad distribution with an average of 0.23 (**Fig. 3.13c**, blue histogram). It is similar to the distribution of r' correlations between model distance matrices (red histogram) and is statistically not distinguishable from the distribution of correlations between single-molecule imaged and model distance matrices (dark grey, two-sided Mann-Whitney p-value=0.02). As before, the r' distribution of the control case (grey) is statistically distinguishable from the distributions of correlations from both models and experiments (p-value = 0).

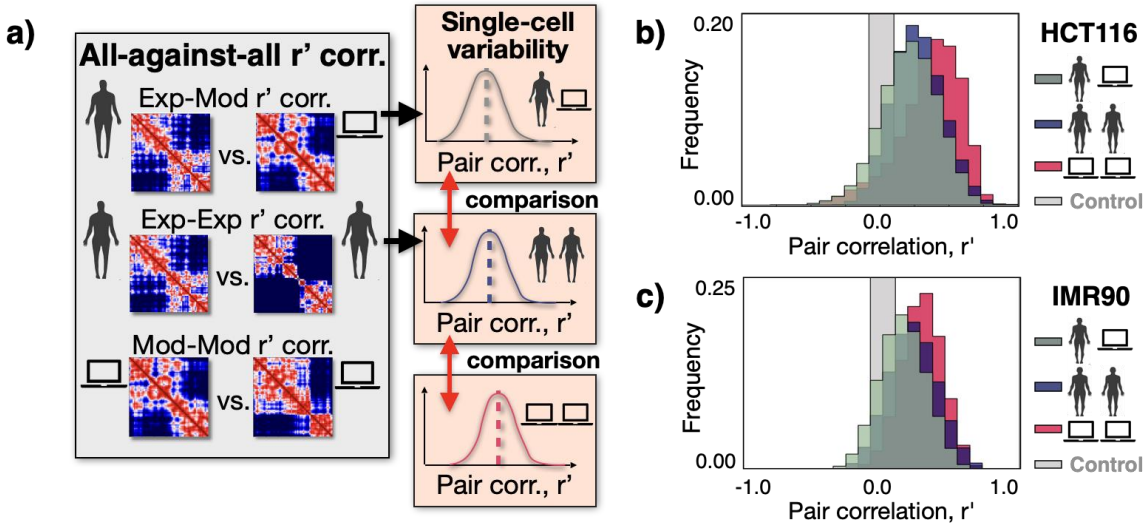

Figure 3.13: Structural variability of single-cell imaged and model-predicted 3D structures. a) To have a quantitative measure of structural variability, we computed the genomic-distance corrected Pearson correlation, r', between all pairs of experimental distance matrices (Exp-Exp r' corr.), all pairs of experimental and model distance matrices (Exp-Mod) and all pairs of model distance matrices (Mod-Mod). **b)** In the HCT116 locus, the r' distribution between pairs of imaged structures (blue, average r' = 0.27) is broad and

statistically not distinguishable from the r' distribution between imaged and model distance matrices (dark gray, two-sided Mann–Whitney p value = 0.19). The model-model r' distribution is in red and in gray a control where r' is computed between pairs of randomized distance matrices. **c)** In IMR90, the Exp-Exp. r' distribution (blue) is broad too and has an average of 0.23. It is statistically not distinguishable from the distribution of correlations between single-molecule imaged and model distance matrices (dark grey, two-sided Mann-Whitney p-value=0.02). The model-model correlation is in red and in grey a random control. Adapted from [38].

Those results show that the 3D structures predicted by our model have the same degree of variability as measured in single-cell experiments, to the point that single-molecules from the model are statistically indistinguishable from experimental single-cell structures. Based on the systematic, quantitative agreement between model predictions and experimental evidence, we can try to explain from a theoretical point of view the origin of the microscopy reported single-cell variability. To this aim, we can consider, as a simplified example, two different single-molecule conformations predicted by the model in the HCT116 case (**Fig. 3.14**): the first structure has a distance matrix with two asymmetrical TAD-like domains (the upstream domain is larger than the downstream one), which correspond to two asymmetrical globules in 3D space; the second structure forms, instead, two more symmetrical domains and globules. We can thus speculate on why those polymer structures are different. In the considered example, the structural diversity mainly results from the green binding domain of the model (**Fig. 3.14**), as in the first structure this domain is proximal to the upstream red and yellow ones, while in the other case it is bridged to the downstream blue domain. This occurs because the green binding domain of the model, although locally enriched, is also spread along regions of the polymer sequence that overlap with the other domains (e.g., the yellow or the blue). For that reason, the model does not fold in a single, naïve state, as in protein folding, but in a much broad set of possible structures.

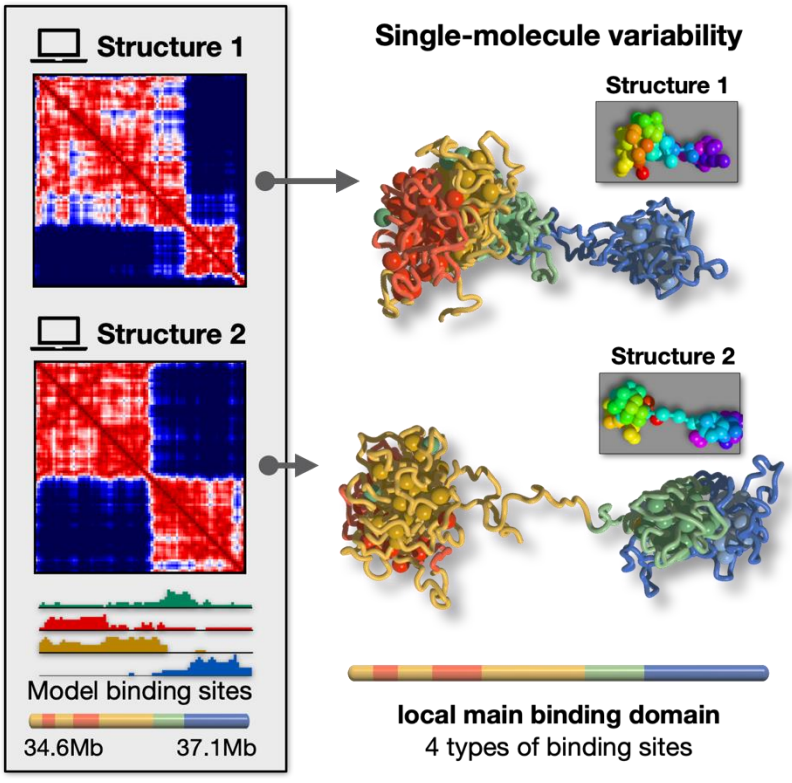

Figure 3.14: Origin of the single-molecule structural variability. Within the SBS framework, the broad variability of single-molecule 3D globular structures, reflected in distance matrices with varying locations of TAD-like domain boundaries, naturally results from the intrinsic folding degeneracy of the phase-separated conformations, enhanced by the overlapping genomic organization of the different binding domains. Adapted from [38].

The overlapping genomic distribution of the model binding sites increases the intrinsic degeneracy of the system microstates, which can fold in a multiplicity of phase-separated 3D conformations. The overall agreement between single-cell imaging data and the independently derived model conformations supports the view whereby, in the studied HCT116 and IMR90 loci, chromatin folding is explained at the single-cell level by a thermodynamics mechanism of globule phase-separation, driven by the interactions of a few different types of binding sites, non-trivially arranged along the genome and each associated to specific combinations of chromatin organizing factors, including, but not limited to, CTCF and cohesin (see, e.g., **Fig. 3.5**). The intrinsic thermodynamic degeneracy of the globule phase-separated state, enhanced by the overlapping genomic organization of the model binding domains, manifests in a broad variety of single-molecule conformations, reflected in the variability of TAD-like contact patterns, consistent with single-cell imaging data.

3.3.3 A control block-copolymer model poorly reflects the complexity of single-cell imaged structures.

As a comparison with our SBS model, we considered a control block-copolymer model designed specifically to reproduce the four main TAD-like structures visible in bulk Hi-C data of the HCT116 cell locus (**Fig. 3.15a**). The block-copolymer, by construction, has the same number of degrees of freedom of our SBS model, i.e., the same number of binding site types (colors) and beads, yet with no intertwining (i.e., overlap) between them. By running massive MD simulations to derive an ensemble of equilibrium polymer conformations in the phase-separated state, we used such a model as a control where to repeat all our single-molecule analyses.

First, as expected by construction, we found that the block-copolymer distance matrix returns the main TAD structures of the considered locus, albeit inter- and intra-TAD signals, which are confirmed by both Hi-C and super-resolution microscopy, are not captured by the model (**Fig. 3.15b**). In fact, its correlation with the experimental median distance map is r'=0.54, which is lower than the SBS value (r'=0.84). Thus, the overlapping nature of the binding sites in our SBS model is crucial to explain important experimental evidence, such as inter-TAD interactions and higher-order contacts, missed by the control model. Similarly, the boundary probability of the block-copolymer recapitulates the genomic location of the main TAD boundary peaks (**Fig. 3.15c**), yet the data are overall less well described by the control (correlation r = 0.47) than by our model (r = 0.79). For instance, the peaks of the control model are four times higher than those from imaging and from our model, showing that the separation of the globules in the block-copolymer is much stronger than in real data and in the SBS model. Finally, we investigated the structural variability of the control by computing the all-against-all r' correlation distribution between pairs of model single-molecule distance matrices that we compared against the corresponding distribution from the

experiment (**Fig. 3.15d**). We found, again, that the control model provides a poorer fit of the experimental distribution than our SBS model. In particular, the average value of r' in the block-copolymer model is 33% higher than in the experiment, showing that in the former there is a lower degree of conformational variability.

As a result, a control block-copolymer model with four non-intertwining binding domains, which are specifically designed to mimic the main TAD-like structures visible at the population-averaged level, only poorly reflects the complexity of the observed imaged single-cell conformations and their cell-to-cell structural variability.

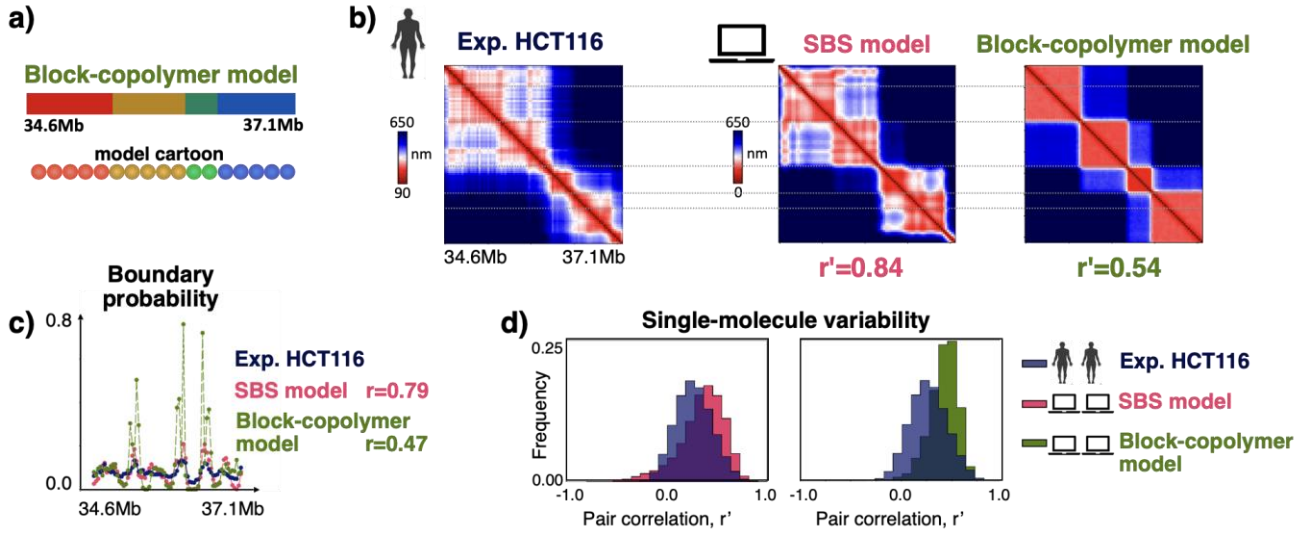

Figure 3.15: A control block-copolymer model provides a poorer description of imaging data. a) Cartoon representing the linear arrangements of the different binding site types (colors) of the control block-copolymer model of the HCT116 locus. **b)** The median distance matrix is shown (left) of the considered locus in HCT116 cells [21]. The SBS model (middle) well explains those data, as quantified by its high distance-corrected Pearson correlation (r'=0.84). Conversely, the control block-copolymer model (right) less well captures the complexity of the observed contact patterns (r'=0.54). **c)** The boundary probability of the control block-copolymer (green) also returns a worse fit of real data (blue) than the SBS model (red). For instance, the control model peaks are four times higher than those from microscopy and from the SBS model, showing that globule separation is sharper in the control than in the SBS model. **d)** The distribution of r' correlations between pairs of distance matrices from the SBS model (red) is closer to the experimental one (blue) than the block-copolymer model (green). The average value of r' in the experimental data (r'=0.3) is approximately equal to the SBS value, whereas in the control model it is 33% higher. Adapted from [38].

3.4 Cohesin depletion reverses phase separation into the coil state in most single-cells

In this Section, we investigate how acute cohesin depletion impacts single-molecule chromatin conformations. To this aim, we considered the same locus chr21:34.6–37.1Mb in HCT116 Auxin treated (HCT116+Auxin) cells, where again both Hi-C [117] and independent imaging data [21] are available. The Hi-C map of the cohesin depleted cells lacks the TAD-like structures of the wild type (WT) locus and retains only a faint pattern of interactions [117] (**Fig. 3.16a**). Similarly, the flat

domain boundary probability and separation score in HCT116+Auxin cells reflect the absence of those contact domains at the population-averaged level [21] (**Fig. 3.16a**). Albeit TADs and sub-TADs are abolished in bulk data, recent super-resolution microscopy data [21] revealed that contact patterns persist in single cells even after cohesin depletion and they are highly variable from cell-to-cell (**Fig. 3.11b**, see also Section 1.4.3). As cohesin has a key role in establishing chromatin architecture [117–119], those diverse results raise questions on the nature of those observed single-cell contact patterns, as well as on the molecular mechanisms underlying their formation. Here, based on the results published in [38], we show that a physical mechanism of polymer phase separation, as envisaged by the SBS model, is consistent with both Hi-C and independent single-cell imaging data in HCT116+Auxin cells. In particular, as fully detailed below (subsections 3.4.1 and 3.4.2), our results depict a scenario where acute cohesin depletion tends to reverse globule phase separation into the coil (i.e., randomly folded) state in most cells, hence explaining why TAD-like domains are erased at the population-averaged level, whereas much more variable and transient contact patterns are found in single molecules.

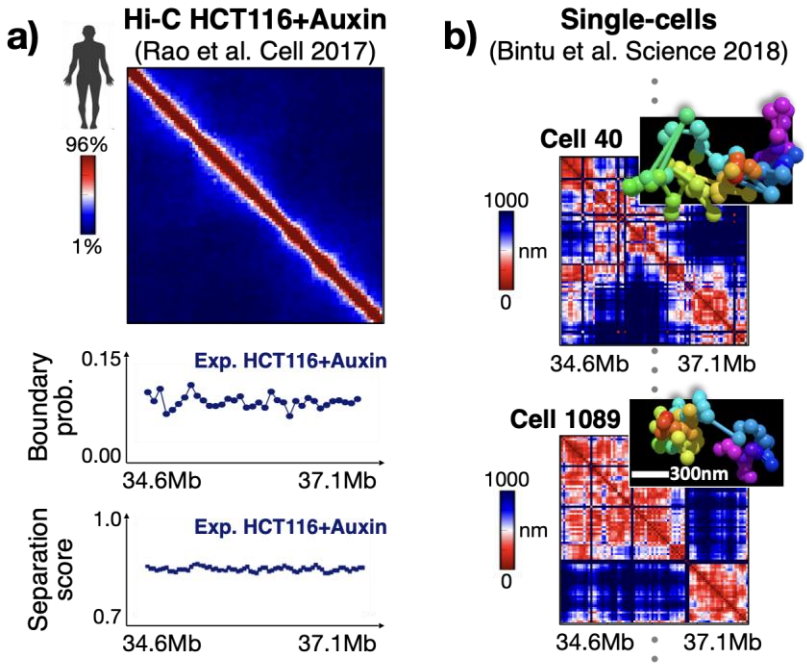

Figure 3.16: Cohesin depletion erases TAD-like structures at the ensemble-averaged level, yet contact patterns remain in single-cells. a) Top: Bulk Hi-C contact data [117] of the locus chr21:34.6–37.1Mb in HCT116 Auxin treated (HCT116+Auxin) cells. Experimental resolution is 30kb. Middle: The boundary probability function [21] of the HCT116+Auxin locus is uniformly spread across the genomic coordinates, consistently with the absence of relevant TAD-like features in Hi-C data. Bottom: The separation score is flat as well in HCT116+Auxin cells. **b)** Differently from bulk data, single-cell imaged distance matrices [21] have specific, highly variable contact patterns. Adapted from [38].

3.4.1 SBS model of the HCT116+Auxin locus

We inferred, as before (see Section 3.1), the new SBS polymer binding sites from Hi-C data in HCT116+Auxin cells (**Fig. 3.16a**) [117] and then derived by MD simulations an equilibrium ensemble of model 3D conformations to compare with imaging data available in the same cell type upon

cohesin depletion [21]. Interestingly, in this case our procedure returned only three types of specific binding sites in the locus (**Fig. 3.17**). In fact, the domain strongly correlated with cohesin in WT HCT116 cells (green domain in **Fig. 3.2a**) now disappears, whereas the other WT domains, although weakened and shrunk, are overall maintained at their genomic locations and their epigenetic signatures partially preserved (see **Fig.s 3.2a**, **3.17a**). On the one hand, such finding further supports the epigenetic significance of the SBS binding sites, as the WT and cohesin depleted polymer models mainly differ for one specific binding domain, which is, consistently, the one mostly associated to cohesin signals (e.g., to RAD21 and SMC); on the other hand, it ensures that similar results would have been obtained by following a different strategy, that is removing the green binding domain in WT HCT116 cells and then run MD simulations of the new three-color model without prior PRISMR training on HCT116+Auxin bulk data, hence highlighting the generality of our approach.

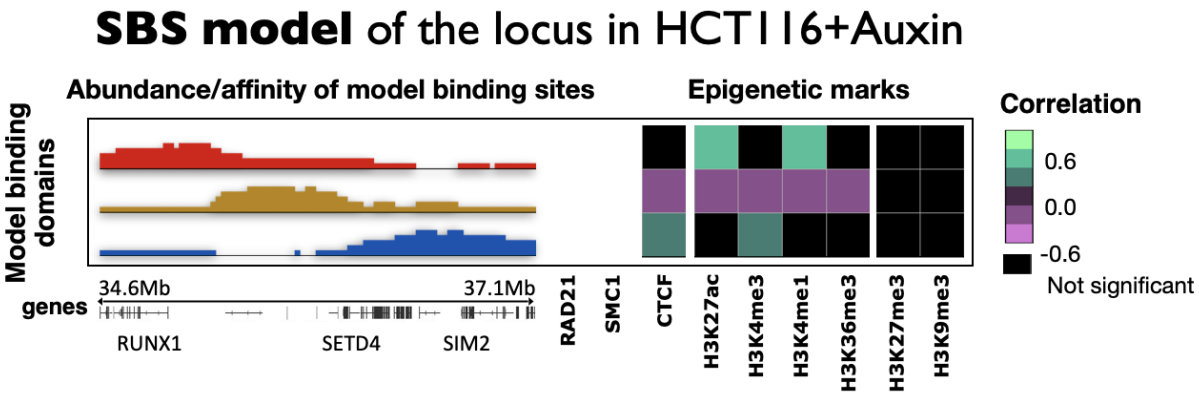

Figure 3.17: SBS model of the locus in HCT116+Auxin cells and epigenetic signature of its inferred binding domains. The model of the HCT116 locus in cohesin depleted cells (left) has three distinct binding domains, each correlated with a specific combination of epigenetic marks (right). The inferred binding sites are similar to those of the HCT116 case (see **Fig. 3.2a**), except for the WT green domain, which is strongly correlated with the cohesin complex and consistently disappears in the HCT116+Auxin model. Adapted from [38].

As in the WT case, the HCT116+Auxin polymer model also undergoes a thermodynamic phase transition from a coil to a globule phase-separated state, yet at higher binder concentrations, e.g., around 500 nmol/l, if the same energy affinities of the WT HCT116 case study model are used (**Fig. 3.18a**). By running MD simulations, we derived a population of equilibrium 3D conformations in the coil and globule states and in each of them we computed the average pairwise contact matrix of the model to be compared with Hi-C data. However, the contact maps computed in the pure states of the theory (e.g., in the pure coil or phase-separated globule) do not return the optimal correlations with the experiment, as we found that a mixture of coil and globule states is required to best explain Hi-C data [38]. For instance, the Pearson, r, and genomic-distance corrected correlation, r', between model and experimental contact maps are, respectively, r=0.59 and r'=0.16 in the pure globule state (**Fig. 3.18**, coil fraction=0%); in the coil state we observe a markedly higher Pearson value, r=0.95, but a poorer r' correlation, r'=0.02 (**Fig. 3.18,** coil fraction=100%). We thus considered a population mixture of polymer structures, composed of a fraction, f, of coil states and of 1-f globule conformations, and repeated for each case the computation of the contact matrix. We found that an ensemble composed 80% of single-molecule conformations in the coil and 20% in the globule state best recapitulates Hi-C data, as it maximizes the r' correlation between model and experiment

(r'=0.33) by keeping, additionally, a high Pearson coefficient value (r=0.96, see the grey dotted line in **Fig. 3.18b**). As the mixture is mainly composed of coil, i.e., randomly folded, states, the average contact map of the model is featureless and lacks the WT relevant TAD-like structure as in real Hi-C data (**Fig. 3.18c**). In the next subsection, we will validate the model at the single-molecule level against independent imaging data and will derive more formally the 80-20% coil-globule mixture, identified here based on empiric considerations (i.e., the comparison between model and Hi-C contact matrices), by using the RMSD structural criterion as in Section 3.2.

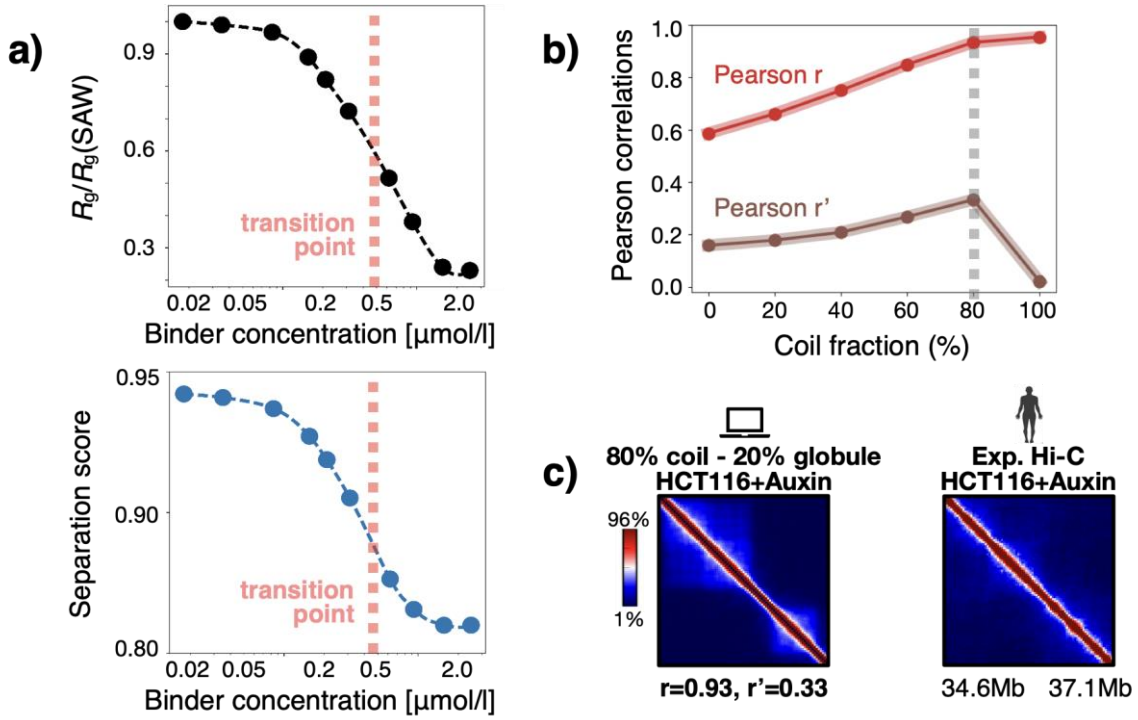

Figure 3.18: A mixture of coil and globule conformations is required to best explains bulk Hi-C data in cohesin depleted cells. a) The SBS model of the locus chr21:34.6Mb-37.1Mb in HCT116+Auxin cells is thermalized to equilibrium at different concentrations of the binders. The same energy affinities for specific and unspecific binding sites of the WT HCT116 model are used (see Section 3.1). The model undergoes a phase transition from a coil to a globule phase separated state, marked by a sharp decrease of its order parameters, such as the chain gyration radius (top) and separation score (bottom). **b)** The best agreement between model and experimental average contact maps is found by taking a population mixture of polymer structures composed 80% of single-molecule conformations in the coil and 20% in the globule state. **c)** The contact matrix of the model mixture compares well against real data [117], as quantified by its high correlation values. Adapted from [38].

3.4.2 Model validation against independent single-cell imaging data in cohesin depleted cells

To further assess the significance of the previous model mixture, we performed an independent all-against-all structural comparison whereby each 3D structure from imaging data [21] is univocally associated by the least RSMD criterion (Section 3.2) to a corresponding, best-matching model 3D structure in the coil or phase-separated globule state. Consistent with our results on average contact maps, we found that 80% experimental structures from microscopy data optimally map onto model conformations in the coil and 20% in the globule state (**Fig. 3.19a**) in a statistically significant association (**Fig. 3.19b**). The distance matrix of single molecules has non-trivial patterns in both

states, but in the coil state (**Fig. 3.19b**, left) contacts originate from random collisions rather than stable phase separated globule domains (**Fig. 3.19b**, right). Overall, the identified best match pairs have very similar distance matrices and 3D conformations (**Fig. 3.19b**).

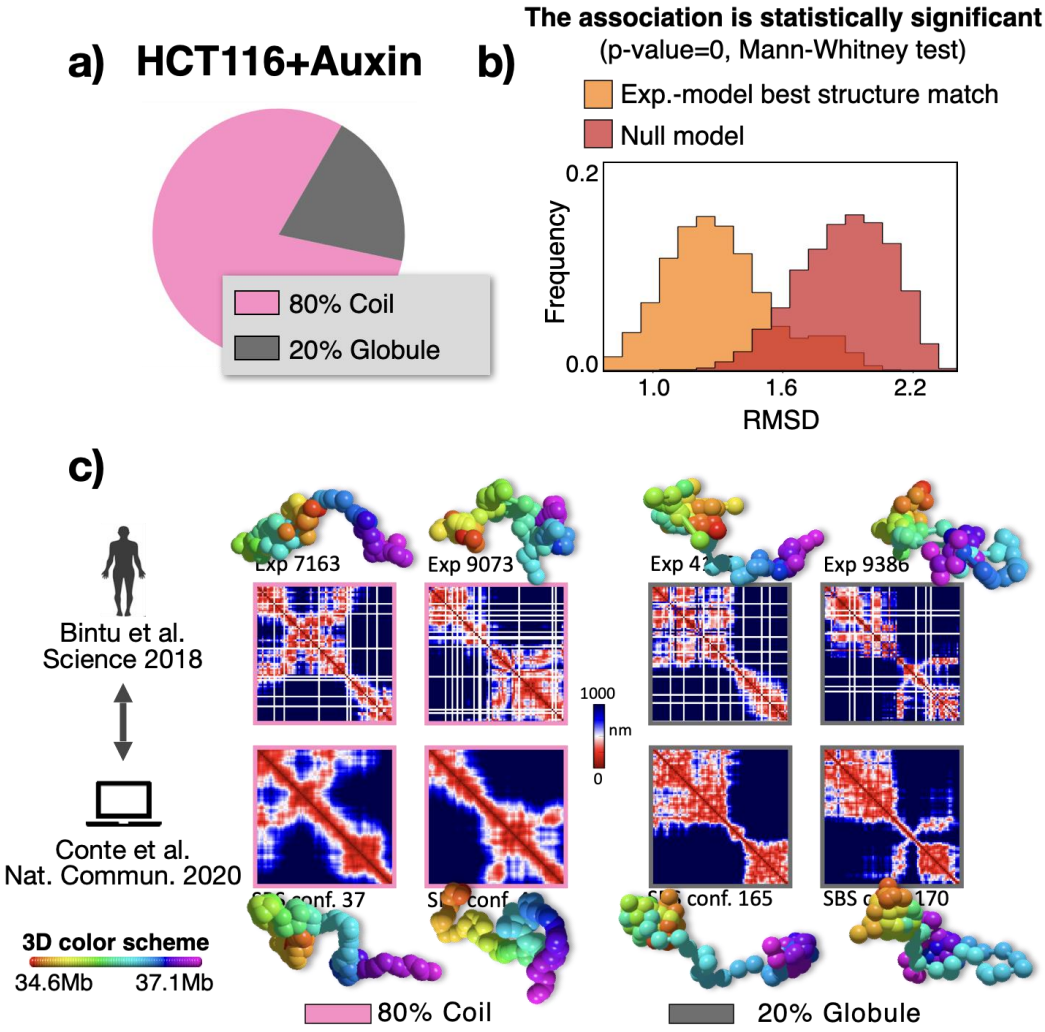

Figure 3.19: The model mixture in HCT116+Auxin cells is confirmed by the RMSD all-against-all single-molecule structural comparison. a) In the HCT116+Auxin case, 80% of imaged 3D structures is optimally mapped via the RMSD criterion onto SBS model structures in the thermodynamic coil phase and 20% in the globule phase-separated state. **b)** The association is significant because the distribution of the optimal RMSD experiment-model best matches is statistically distinguishable from a null model made of random pairs of imaged structures (p value=0, two-sided Mann-Whitney test). **c)** Examples of best-matching experiment-model pairs. The 3D conformations of the model mixture include globular states as in WT (right), but 80% of single-molecules are in the coil state (left) whose contact patterns reflect more transient, random chromatin collisions rather than stably folded contacts as in WT. Adapted from [38].

Next, we compared the single-molecule predictions of our model mixture against independent microscopy data (**Fig. 3.20**). As a first validation, we focused on the experimental median distance map of the locus (**Fig. 3.20a**, top), which is as featureless as the corresponding Hi-C data. We found that the median distance matrix predicted by the model (**Fig. 3.20a**, bottom) compares well against imaging data: the Pearson and genomic-distance corrected correlations are, respectively, $r = 0.96$ and $r' = 0.57$, even higher than those with Hi-C data (**Fig. 3.18c**). Consistent with a scenario where

cohesin depletion promotes more random and fleeting interactions, the boundary probability is flat along the locus in both model and experiment (**Fig. 3.20b**, r=0.19), as well as the average separation score (**Fig. 3.20c**, r=0.41). Conversely, as consistently predict by our model, the average boundary probability and strength are similar to their corresponding WT values (compare, e.g., **Fig. 3.20d** with **Fig. 3.11a, e**). We also checked that the boundary strength distribution of the model is similar to the experimental one (**Fig. 3.20e**), as much as the ensemble gyration radius distribution (**Fig. 3.20f**, two-sided Mann–Whitney p value = 0.10), whose average value is 23% larger than in the WT case (540 nm vs. 440 nm), showing that the locus is indeed more open upon cohesin depletion. Finally, as done in subsection 3.3.2, we measured the structural variability of single-molecule conformations by computing in both model and experiment the all-against-all r' correlation distribution between pairs of distance matrices (**Fig. 3.20g**). Notably, we found that the imaged single-cell 3D conformations have a higher variability than WT ones, as the average r' value between pairs of distance matrices is r' = 0.0 (whereas the WT value is r'=0.27) and its distribution is broader (blue histogram in **Fig. 3.20g**). The model single-molecule conformations, which mostly fold in random states, have also a high variability and they have an r' correlation distribution with imaged distance matrices (dark grey, average r'=0.0), and with each other (red, average r'=0.0), statistically similar to the one between experiment pairs (two-sided Mann-Whitney p value = 0.48).

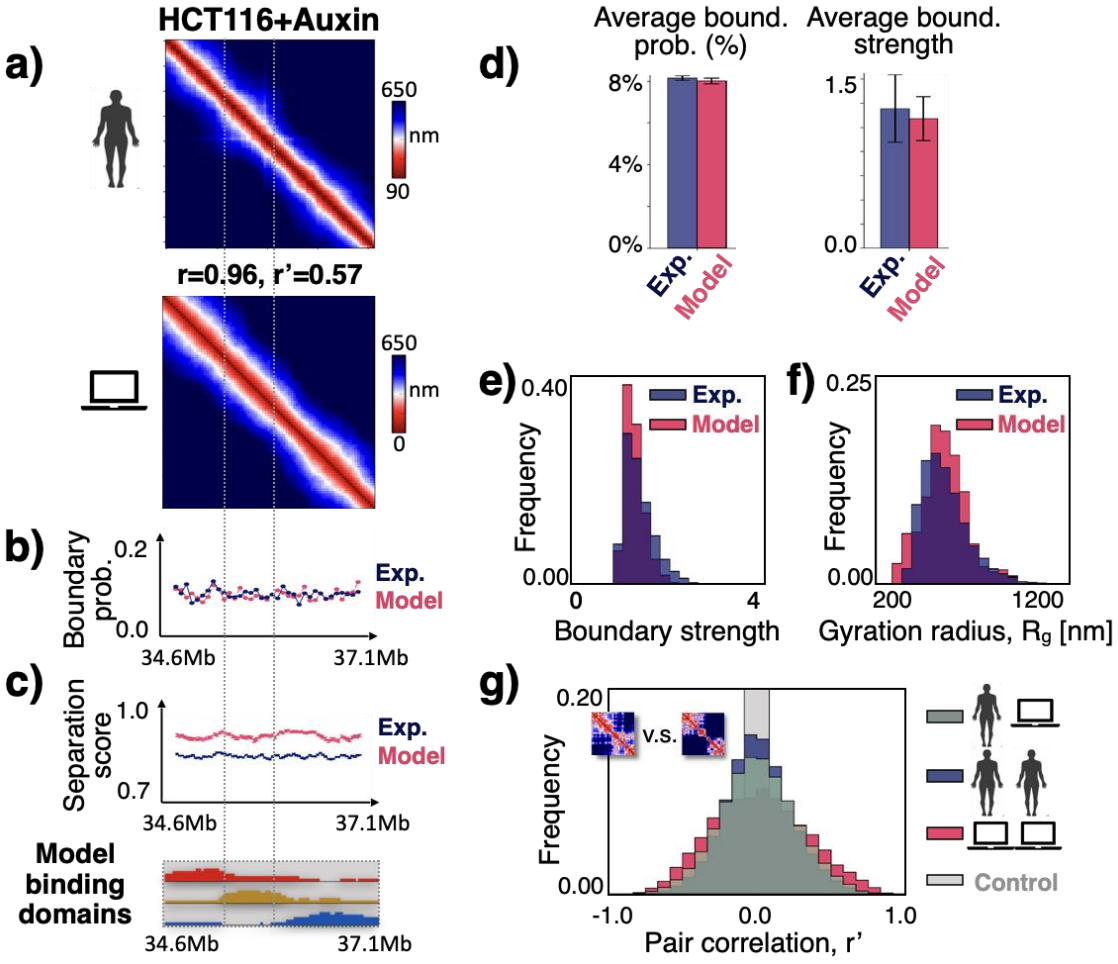

Figure 3.20: The SBS model of the HCT116+Auxin locus is consistently validated against independent super-resolution single-cell imaging data. a) The comparison of our model prediction on the median distance matrix against the independent imaging data [21] gives high correlations (r = 0.96, r' = 0.57). A mixture of

model 3D conformation is required, however, 80% in the coil and 20% in the globule state. **b)** Consistent with the data, the boundary probability of the model (r=0.19) and **c)** its separation score (r=0.41) are flat along the locus, reflecting the absence of TAD-like domains in the median matrices. Bottom: the model has three distinct binding domains, as the cohesin-correlated WT domain (green domain in Fig. **3.2a**) now disappears. **d)** The average boundary probability (left) and the average boundary strength (right) of the model return values that are consistent with the experiment (error bars shown are standard errors). **e)** Also, the distribution of boundary strengths in HCT116+Auxin cells is similar in both model and experiment. **f)** The gyration radius distributions are not distinguishable too (two-sided Mann–Whitney p value = 0.10) and have a higher average value than wild-type (540 nm vs. 440 nm of **Fig. 3.9a**), consistent with our interpretation that cohesin depletion tends to form in single-cells more fleeting, and thus more open, contact structures. **g)** The distribution of genomic-distance corrected, r', correlations between single-cell imaged distance matrices (blue) is broader than in WT (compare with **Fig. 3.13b**) and its average is r' = 0.0 (significantly lower than WT average r'=0.27), revealing a higher cell-to-cell structural variability upon cohesin depletion. It is similar to the correlations between model distance matrices (red) and statistically not distinguishable from the distribution of r' correlations between experimental and model distance matrices (dark gray, two-sided Mann–Whitney p value = 0.48). Adapted from [38].

The systematic agreement between model and independent microscopy data in the HCT116+Auxin case supports a scenario where, consistent with the known role of cohesin as a key architecture organizing factor, cohesin depletion reverses phase separation in most cells as their corresponding model single-molecule structures are mainly in the coil rather than in the globule state, contrary to the WT HCT116 case. We also note that those findings are consistent with important, recent experimental evidence in the yeast genome, where the cohesin complex is observed to phase-separate with DNA into liquid droplets by ATP-independent DNA bridging [144]. In this sense, the depletion of cohesin hinders bridging-induced chromatin phase-separation, hence producing, in agreement with microscopy data [21] and model predictions, more open and variable structures that tend to form mostly random and transient interactions.

Summarizing, the emerging scenario shows that in WT cells (e.g, the HCT116 and IMR90 loci) chromatin folds mostly in the globule phase-separated state, whose inherent thermodynamics structural degeneracy is manifested in the varying genomic positions of TAD-like patterns across single-molecules (**Fig. 3.21a**). Conversely, in cohesin depleted cells, globule phase separation is reversed into the coil state in most cells, producing much more variable and transient contact patterns in single-molecules (**Fig. 3.21b**), abolishing therefore population-averaged TAD-like domains. Taken together, our results indicate that, in the studied loci, chromatin folding is driven at the single-molecule level by such a mechanism of polymer phase separation. However, other molecular physical processes (such as active loop-extrusion) are likely to play a role and they could compete, contribute, or co-exist with thermodynamic phase-separation in shaping chromatin architecture.

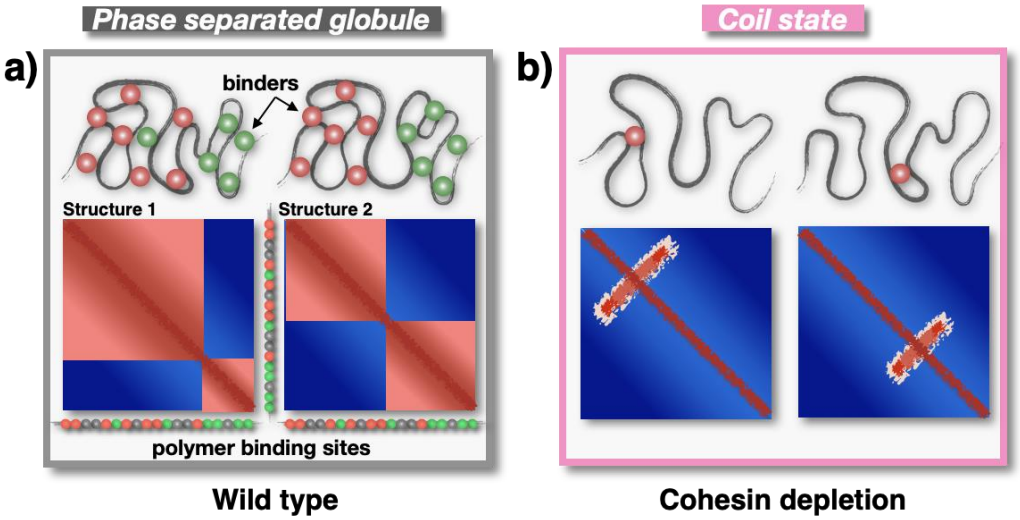

Figure 3.21: Thermodynamic polymer phase-separation explains chromatin structure across single-cells. In the SBS model, according to the abundance or affinity of binding molecules, the polymer folds in conformational classes corresponding to the thermodynamics phases of the system, e.g., the coil and the phase separated globule state. Those thermodynamics states recapitulate single-cell architecture data of the studied loci. **a)** The globule phase-separated state, for example, produces local, spatially segregated compact environments where specific contacts are highly enhanced between polymer regions enriched for cognate binding sites. The intrinsic thermodynamic degeneracy of conformations results in a variety of single-molecule 3D structures, reflected in the variability of the genomic position of TAD-like patterns, consistent with imaging data [21] of loci in human HCT116 and IMR90 cells. **b)** Cohesin depletion, as in HCT116+Auxin cells, tends to reverse globule phase-separation into the coil state in most cells, abolishing patterns in bulk Hi-C data and resulting in much more variable and fleeting contacts in single-molecules. Adapted from [38].

3.5 Steady-state time dynamics of single-molecule polymer conformations

Next, we explored the steady state time dynamics of our in-silico structures, particularly how the spatial conformations of single DNA molecules change in time and how specific patterns of contact or insulation are established. Although a rigorous, full test of our model results would require experiments following in time entire chromatin loci (e.g., super-resolution live-cell microscopy), which, however, are not available at least in the studied loci, we performed in [38] an initial validation of the model time behavior by comparing its predictions against single-cell imaging data [21]. We discuss in this Section the major findings of our single-molecule time dynamics analysis, fully reported in the paper [38], showing how globule phase-separation can either establish contact specificity or spatial insulation between chromatin regions in the stochastic nuclear environment.

First, we investigated how the equilibrium single-molecule conformations of the model behave in time in its two main thermodynamic phases. To this aim, we studied at different time-points the steady-state distance maps and corresponding 3D structures of single polymer molecules, respectively, in the coil state of the HCT116+Auxin model and in the globule phase-separated state of the HCT116 model. Albeit the model 3D structures vary in time due to thermal fluctuations in both states of the theory, important differences mark the two phases (**Fig. 3.22a-d**). In the coil state,

the contact patterns visible in the distance matrix of a single molecule are highly transient and fleeting, as they mainly result from random polymer-to-polymer collisions, and suddenly change in time (**Fig. 3.22a**, top). Consistently, the average r' self-correlation function (i.e., the population-averaged r' correlation computed between steady-state single-molecule distance matrices at different lag times) approaches zero at long times (**Fig. 3.22b**), in agreement with the ensemble distribution of r' correlation between different replicates (see **Fig. 3.20g**). In the phase-separated state, the single-molecule distance matrix also varies in time and the 3D structure breathes and rearranges due to thermal effects (**Fig. 3.22c**, top); however, the long-time average r' self-correlation remains well above zero (**Fig. 3.22d**, in the HCT116 model it plateaus to 0.39), again consistent with the average non-zero correlation between replicates (**Fig. 3.13b**), showing that the folded globules change, yet they persist in time (**Fig. 3.22c**, top). Consistent with such a picture, the conformation average de-correlation time, defined as the lag time where the average r' self-correlation has spanned 95% of its total variation range, is almost one order of magnitude larger in the globule state than in the coil state; its scale can be roughly guessed by using estimates of the viscosity of the nuclear medium reported in the literature (which are around 0.03P) [46,145]: for example, it results to be 9s and 60s respectively in the coil state of the HCT116+Auxin and in the phase-separated state of the HCT116 model (**Fig. 3.22b**, **d**).

Then, we asked how globule formation establishes, in the face of a varying environment, domain boundaries and specific contact loops at the single-molecule level. To that aim, we investigated the relative spatial distances of specific pairs of sites: (i) a pair of sites (orange in **Fig. 3.22**), located 1.2 Mb apart from each other in different sub-TADs, having in HCT116 cells a strong point-wise (loop) interaction in the median distance matrix; (ii) a pair of 0.6 Mb distant sites (green) separated by a strong TAD boundary in between; (iii) a control pair of sites (brown), almost 0.6 Mb apart, enclosed within the same sub-TAD. The coordinates of the considered pairs are: 34.69–35.80 Mb (orange), 35.59–36.25 Mb (green), 36.43–36.91 Mb (brown). In the HCT116+Auxin model, where molecules are mostly folded in the coil state, the average physical distances of the green and brown pair are comparable to each other (around 620 nm; **Fig. 3.22a**, bottom), while the orange sites are more separated (average distance 660 nm; **Fig. 3.22a**, bottom) simply because of their larger genomic separation. The HCT116+Auxin model distance distributions of, e.g., the orange and green pairs (**Fig. 3.22e**, in pink) are comparatively broad and similar to their corresponding experimental distributions (**Fig. 3.22f**, pink). Conversely, in the phase-separated state of the HCT116 model the average distance of the orange and brown pairs is reduced of factor 2.5 down to around 280 nm (**Fig. 3.22c**, bottom). In fact, the orange pair (or, similarly, the brown) is located within polymer regions enriched with cognate binding sites, which are highly likely to be bridged in the globule compact environment, thus resulting in a strong loop visible in bulk data. The green sites, by contrast, tend to be trapped each in a different globule and for that reason their relative distance broadly fluctuates in time around the coil-state distance values (more than 600nm; **Fig. 3.22c**, bottom). Those examples demonstrate how globules can form, respectively, specific contacts between chromatin sites (see, e.g, the orange and brown pairs) or insulating boundaries between them (see, e.g., the green pair) in the highly stochastic nuclear environment. As expected, the ensemble distance distribution of the orange (as well as the brown) pair is much narrower in the

HCT116 than in HCT116+Auxin case in both model (**Fig. 3.22e**, left, grey) and experiment (**Fig. 3.22f**, left, grey), whereas the model and microscopy distributions of the green pair are similar in both cell types (compare right panels of **Fig. 3.22e** and **Fig. 3.22f**). Considering the basic character of the model, its predicted distributions are comparatively close to the experimental ones, albeit no free parameters are available in all the above comparisons (**Fig. 3.22e**, **f**).

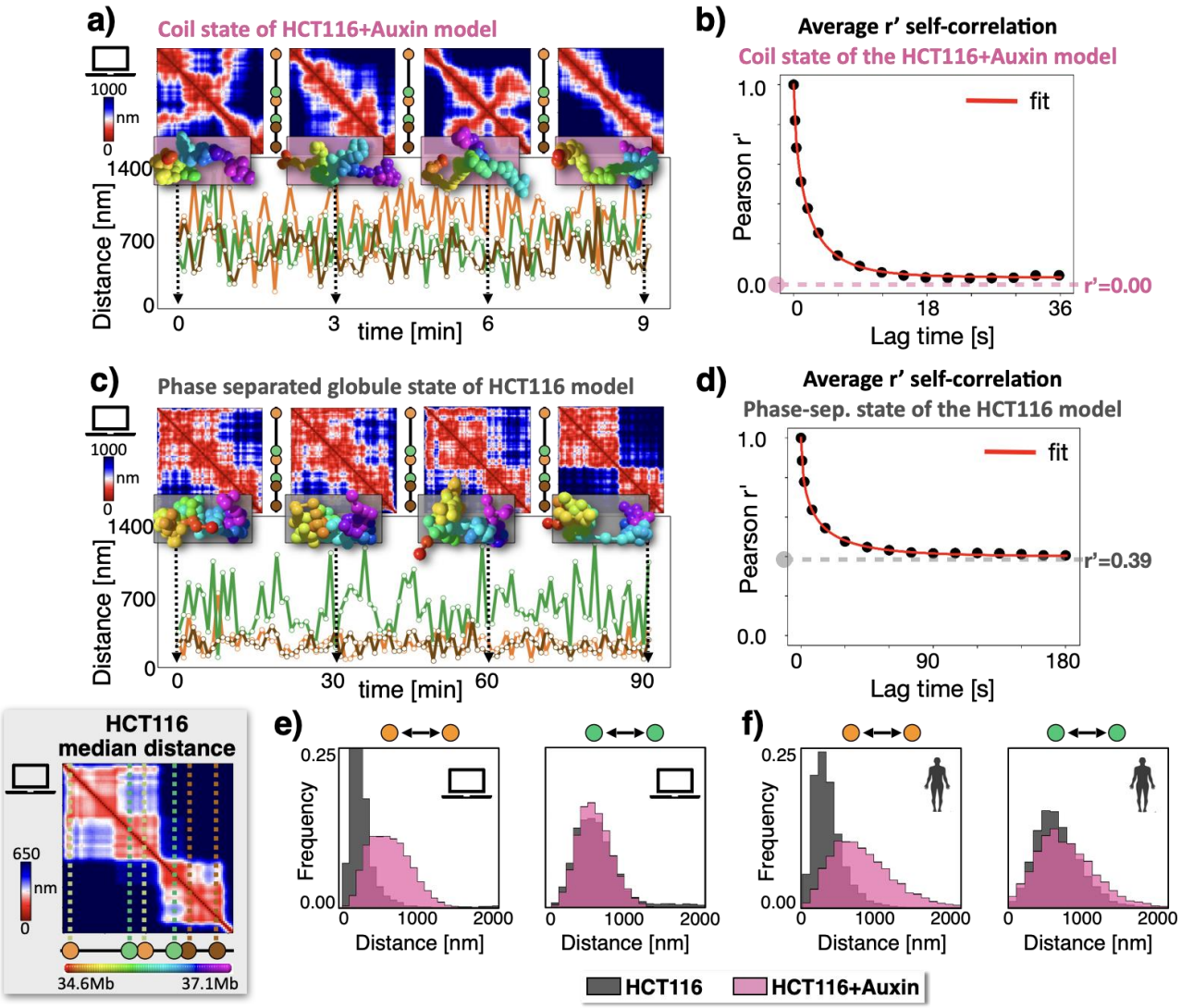

Figure 3.22: Single-molecule time dynamics. a) Steady state time behaviour of a single-molecule coil conformation in the model of the chr21:34.6–37.1Mb locus in cohesin depleted (HCT116+Auxin) cells. Top: Time course of a single-molecule distance matrix in the coil state of the theory: contact patterns are random and transient, and they fully change in time. Bottom: The relative distances of specific pairs of sites (see text) have all wide fluctuations in the coil state, as contacts are fleeting. **b)** Average r' self-correlation in the coil state of the HCT116+Auxin model. The long-time self-correlation value approaches zero, consistent with the ensemble correlation analysis (see **Fig. 3.20g**). Superimposed is a stretched exponential fit. **c)** Steady state time behaviour of a single-molecule phase-separated conformation in the model of the chr21:34.6–37.1Mb locus in wild type (HCT116) cells. Top: The interaction patterns visible in the distance matrix show that globules vary, yet they persist in time. Bottom: Time tracks of the relative distances of specific site pairs (see text). Specific contacts are enhanced between pairs sharing abundant cognate binding sites (orange and brown), whereas pairs in different globules remain insulated, hence forming a boundary (green). **d)** Average r' self-correlation in the phase-separated state of the HCT116 model. Its nonzero long-time value (close to

0.39) is consistent with our previous correlation analysis between replicates (see **Fig. 3.13b**). Superimposed is a stretched exponential fit. **e)** Left: Model derived distance distributions of the considered orange pair in the HCT116 model (grey) and in the HCT116+Auxin model (pink). The orange pair is on average much closer and its distance distribution much narrower in the HCT116 than in the HCT116+ Auxin model. Right: The model distance distributions of the green pair are similar in both HCT116 and HCT116+Auxin. **e)** Experimental counterpart of the distance distributions shown in panel **e).** The agreement between model and microscopy distance data further supports the view that chromatin folds in different thermodynamics states in WT and cohesin depleted cells. Adapted from [38].

To summarize, single-molecule contacts in the coil state are random and transient, as binders establish only fleeting interactions that are promptly overwhelmed by thermal fluctuations (**Fig. 3.23a**). That results in broad distance distributions between site pairs and average time self-correlations rapidly decaying to zero. Conversely, in the globule phase-separated state, because of the abundance of cognate binding sites, self-interacting globules vary, yet they persist in time (**Fig. 3.23b**). That enhances specific local contacts between site pairs within the globule compact environment and produces persistent, yet stochastically varying, "memory" conformations reflected in nonzero long-time average self-correlations.

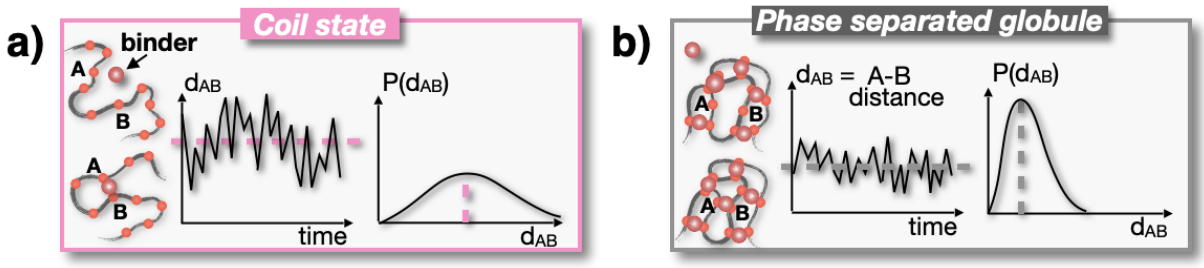

Figure 3.23: Summary cartoon of the model time structural variability. a) In the coil state of the model, the polymer is subject to random and highly transient interactions that result in open randomly folded conformations where the relative distances between sites pairs broadly fluctuate in time around high average values. **b)** In the phase-separated state of the model, different binding sites self-assemble by action of their cognate binders to form more compact and stable globules, where specific contacts are highly favored over stochastic encounters. Adapted from [38].

Overall, our analysis of the time dynamics of single molecules illustrates the diverse modes of action of globules in shaping specific spatial interactions or insulation between distal chromatin sites. While segregating neighboring regions along the sequence (see, e.g., the green pair), they can create stable, local compact environments where specific contacts (e.g., the loops of the brown and orange pairs) are boosted between regions sharing abundant cognate binding sites. That exemplifies how stochasticity of DNA interactions, typical of weak biochemical affinities, can coexist with contact specificity in the varying nuclear environment.

4. An in-silico experiment: benchmarking sequencing-based technologies via polymer physics modeling

In this final Chapter, we discuss how models from polymer physics can be used to assess advantages and limitations of benchmarked experimental methods for determination of chromosomal structure. We focus, in particular, on three powerful and popular sequencing-based approaches, i.e., Hi-C [13,88], SPRITE [14] and GAM [15], which have been used to probe chromatin interactions genome wide, unveiling the complex 3D organization of mammalian genomes that includes, for example, DNA loops, TADs, higher-order structures (such as metaTADs), and A/B compartments (see Section 1.3). Differently, e.g., from multiplexed FISH experiments, which allow directly visualizing genomes under the microscope, the three technologies, albeit technically different, detect in a population of cells, or in single cells, the frequency at which pairs of DNA loci (or also multiple loci) interact with each other. Those measured frequencies result in genome interaction maps, which can be taken as a proxy of the studied chromatin conformations [89]. However, it remains unclear to what extent those technologies are faithful to the underlying genome 3D structure and, importantly, how they perform relative to each other in different applications, as they provide distinct measures of interactions and no benchmark exists. For instance, is GAM or SPRITE as faithful to genome structure as Hi-C? How many cells are required to ensure the statistical reproducibility of experimental outcomes? How is data quality affected by the detection efficiency? Which method is more suited to capture interactions at large genomic distances? To answer those questions, in our published work [40] we devised a computational experiment where we implemented in-silico those distinct three methods on an ensemble of known 3D polymer structures from validated SBS models of real chromosomal loci, thus building a simplified, yet fully controlled, framework where to analyze their outputs. Here, we summarize the main results of the study, which is fully reported and detailed in the original publication [40]. Specifically, in the Section 4.1 of this Chapter, we validate our approach by showing that the population-averaged in-silico Hi-C, SPRITE and GAM data significantly match their corresponding experiments. Such a consistent agreement ensures that our polymer models can be successfully employed to compare in-silico the performance of the three technologies. In Section 4.2, we investigate the behavior of the simulated Hi-C, GAM and SPRITE with respect to key experimental parameters, such as the detection efficiency, genomic separation and cell numbers. For instance, while all three methods comparatively well reproduce the reference model 3D structures, we found that the minimal number of cells required to return statistically similar contacts is different across the technologies, being lowest in SPRITE and highest in GAM under the same conditions. Similarly, noise-to-signal level in contact matrices increases as a power law by decreasing the efficiency and it varies with genomic distance differently in the different cases, with GAM being the least affected by noise at larger genomic separations (>1Mb). Overall, by combining polymer physics and computer simulations, our analyses provide a quantitative benchmark to assess how well different experimental methods represent the 3D structure of the genome and to test their relative performance in different conditions.

The reader can find an outline of the experimental Hi-C, SPRITE and GAM protocols in the introductory **Chapter 1** (see Section 1.2). The in-silico versions of those technologies are designed

to mimic each step performed in the corresponding, real experiments. Their technical implementation is extensively discussed in the paper [40] and, for brevity, is not reported also here. The algorithms that we developed for simulations of in-silico Hi-C, SPRITE and GAM are published in [40] and deposited at the link: https://github.com/fmusella/In-silico_Hi-C_GAM_SPRITE. All the material of this Chapter is taken or adapted from [40].

4.1 In-silico Hi-C, SPRITE and GAM significantly reproduce independent experimental data

In the paper [40], we considered the SBS models of a set of different loci across different cell types, such as the 6Mb wide genomic regions around the *Sox9* (chr11:109-115Mb, mm9) and *HoxD* (chr2:71-78Mb) genes in mouse embryonic stem cells (mESCs) [25,36] and around the *Epha4* gene in mouse CHLX-12 cells (chr1:73-79Mb) [9], as well as the 2.5Mb locus in human HCT116 cells [38] investigated in recent microscopy experiments [21] (the same discussed in **Chapter 3**). Those loci are particularly interesting because, for example, related to pathogenic structural variants inducing gene misexpression (*Sox9* and *Epha4*) [6,9,10] or establishing tissue-specific regulatory 3D architectures during differentiation (e.g., *HoxD*) [137,146]. In this Section, we first describe the computational approach developed to compare in-silico Hi-C, SPRITE and GAM (subsection 4.1.1). Then, we validate our in-silico contact data against known polymer 3D structures at both the population-averaged (subsection 4.1.2) and single-cell level (subsection 4.1.3). For brevity, we take in the following Sox*9* as main case study, yet our conclusions remain unchanged also for the other mentioned loci, as extensively shown in [40].

4.1.1. Our approach for comparing in-silico the Hi-C, SPRITE and GAM technologies

The model of the *Sox9* locus [25], i.e., the optimal genomic locations of its binding sites, is inferred via the PRISMR algorithm [9] (Section 2.2) based on available 40kb resolution Hi-C data from [16]. Next, by extensive MD simulations, an ensemble of single-molecule 3D polymer structures is derived in the thermodynamic steady state of the system [25]. On those 3D structures we implemented the Hi-C, GAM, and SPRITE methods by simulating their corresponding steps in in-silico contact data (**Fig. 4.1a**) [40]. In brief, in the case of in-silico Hi-C two polymer chains represent the two *Sox9* alleles in each cell; they are fragmented in equal segments, then crosslinked fragments are ligated and ligation products are counted to derive an in-silico analog of Hi-C contact frequencies (see the original work [40] for all details of the simulations). The overall efficiency of the process is the product of the in-silico crosslinking, digestion, biotinylation, ligation and sequencing efficiencies. Similarly, in-silico SPRITE is performed by counting chain fragments tagged with the same barcode. Finally, in-silico GAM is implemented by cutting randomly oriented slices from a sphere (representing the nucleus) where two single-molecule 3D structures, i.e., the two alleles, are randomly positioned; thus, by listing the polymer sites falling within each slice, a co-segregation matrix of the studied locus is derived. The overall efficiency comprises the detection and sequencing efficiency of such sites. By using that procedure, we derived in-silico contact maps of the known polymer 3D structures and investigated how the different technologies are affected, for instance, by the detection efficiency, by the number of pairs, *N*, of 3D single-molecule structures included in

the analysis (for simplicity, we refer hereafter to *N* as the number of in silico cells) and how they depend on the genomic separation.

4.1.2. In-silico contact data are derived from known SBS 3D structures

As the SBS polymer model of *Sox9* is inferred from Hi-C data [16], we first checked that its in-silico Hi-C contact map (i.e., the contact data averaged over the ensemble of 3D structures) reproduces the corresponding real bulk Hi-C data [16] of the locus (**Fig. 4.1b**, left). To this aim, we computed the correlation between model and experimental contact maps and found that different coefficients, such as Spearman (r_s), Pearson (r), and HiCRep (stratum adjusted correlation coefficient, scc [147]), consistently returned all high values: r_s=0.83, r=0.83 and scc=0.80, respectively, as also reported in [25]. Based on this preliminary check, we next validated our approach by comparing the in-silico SPRITE and GAM contact matrices derived from the same ensemble of *Sox9* model 3D structures against the corresponding, independent SPRITE [14] and GAM [148] bulk experimental matrices. We found, again, high correlations between model and experiments, respectively, r_s=0.92 and r_s=0.79, r=0.75 and r=0.80 (**Fig. 4.1b**, middle and right). The HiCRep score, which is mostly designed for comparison of Hi-C data, is also statistically significantly high—respectively, scc=0.57 and scc=0.40. Similar results were found for the other considered loci, such as *HoxD* [40].

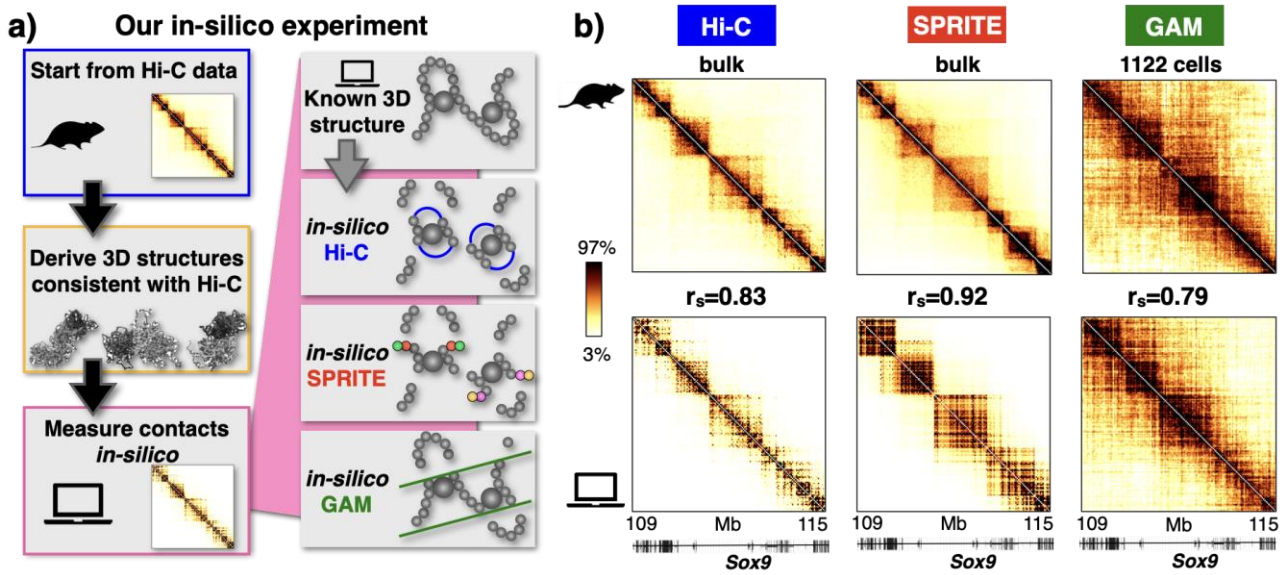

Figure 4.1: In silico Hi-C, SPRITE and GAM bulk contact maps recapitulate experimental data. a) Schematic cartoon illustration of our computational experiment: a thermodynamic ensemble of single-molecule 3D conformations of the polymer model of the DNA locus of interest is derived from bulk Hi-C data using the PRISMR [9] procedure and polymer physics simulations. Then, the experimental protocols of Hi-C, SPRITE and GAM are implemented in-silico on the ensemble of model 3D structures. **b)** In our *Sox9* case study (chr11:109–115Mb, mm9), the SBS model conformations, which are inferred from only Hi-C data [16], return average contact maps (bottom) that significantly match Hi-C and independent SPRITE [14] and GAM data (top), as quantified, e.g., by the high Spearman correlation (r_s) values. GAM data are from a new dataset made of 1122 F123 cells. Adapted from [40].

4.1.3. Validation of the ensemble of single-molecule polymer 3D conformations

As broadly discussed in **Chapter 3**, the SBS model 3D structures represent chromatin structure well beyond the population-averaged data, as they provide a bona-fide representation of chromatin conformations in single-cells [38]. To this aim, we took advantage of published multiplex FISH super-resolution microscopy data [21] for a 2.5Mb region in human HCT116 cells (chr21:34.6–37.1 Mb, see **Chapter 3**), because here we can compare experimental and model-derived single-molecule 3D structures [21,38] as well as Hi-C data [117] (GAM and SPRITE data are not available for that cell type). By using the minimum RMSD criterion (Section 3.2), we performed an all-against-all structural comparison whereby each SBS model single-molecule conformation was univocally associated to a corresponding imaged 3D structure (**Fig. 4.2a, b**). To set a control, we generated self-avoiding random-walk (SAW) polymer chains having the same number of beads and the same average gyration radius (i.e., same linear size) as the real images of the locus. Thus, we compared the RMSD distribution of the model-experiment best matches against those from the SAW control model and found the two distributions to be statistically different (two-sided Mann–Whitney test p=0) with 93% of the former falling below the first tertile of the latter (**Fig. 4.2a**). Similar results hold if we consider the RMSD distribution of the experiment-model best matches against the control case (**Fig. 4.2b**, two-sided Mann–Whitney test p=0, with 70% of the entries of the former below the first tertile of the latter). As a further validation, we verified that the SBS-predicted [38] and microscopy [21] mean distance matrices have a high correlation value (r_s=0.96), as well as the model [38] and experimental Hi-C [117] contact matrices (r_s=0.94, **Fig. 4.2c**). Importantly, the in-silico SPRITE and GAM average matrices also represent with high accuracy the mean distance data (correlation, r_s=−0.98 and r_s=−0.99, respectively, **Fig. 4.2c**) and our results are robust across different correlation metrics (see **Table I**).

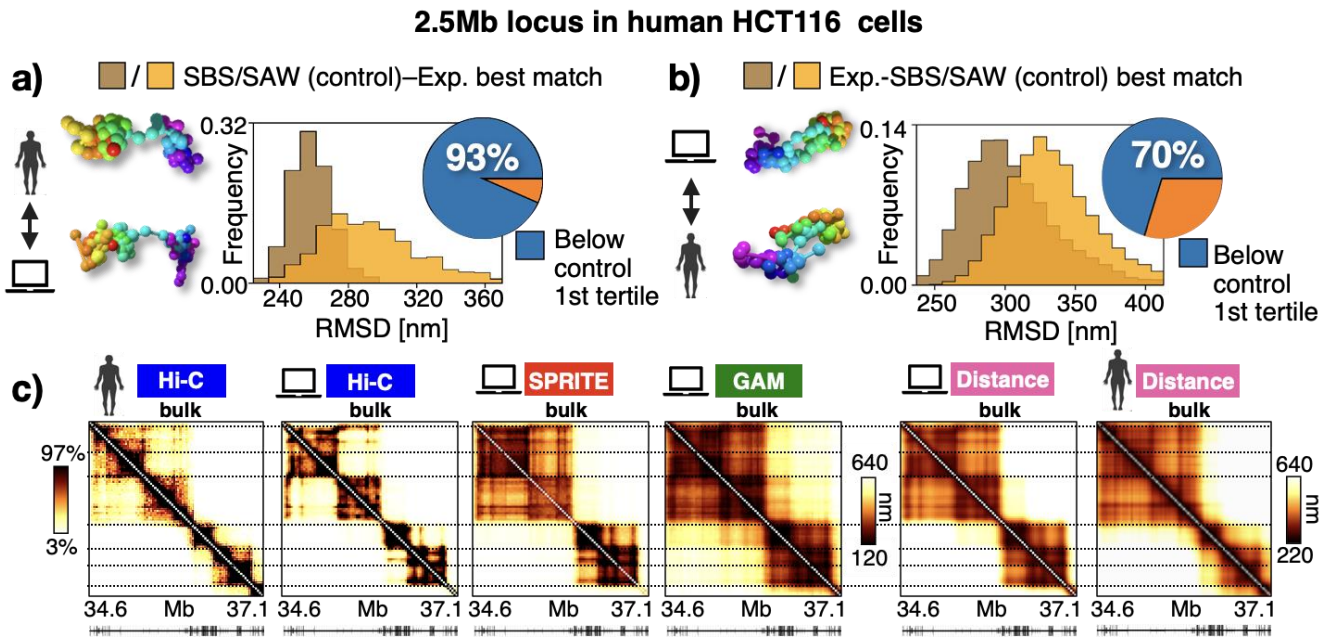

Figure 4.2: Model 3D structures are consistently validated at the single-molecule level. a) The distribution of RMSD between the 3D conformations of the SBS model and their best-matching experimental structures [21] (brown) is compared to that between control SAW structures and their best-matching

microscopy conformations (orange) in the imaged 2.5Mb locus in human HCT116 cells (chr21:34.6–37.1Mb). The two distributions are statistically different (two-sided Mann–Whitney test p = 0) with 93% of the former below the first tertile of the latter. **b)** The distributions of RMSDs is computed between the experimental structures and their best-matching structures from the SBS (brown) and the control SAW model (orange). The distributions are statistically different (p=0) and well separated (70% of the former is below the first tertile of the latter). **c)** In the HCT116 locus, the experimental Hi-C [117] and the in-silico Hi-C, SPRITE, and GAM contact matrices (left) are compared to the model and experimental average distance maps (right). Adapted from [40].

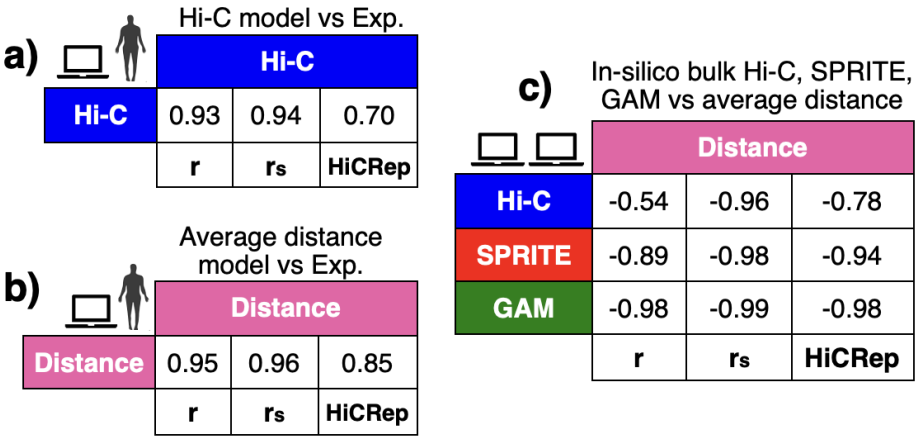

Table I: Pearson (r), Spearman (r_s), and HiCRep correlation values in the HCT116 locus for: **a)** in-silico and experimental Hi-C contact maps, **b)** in-silico and microscopy average distance matrices, **c)** in-silico Hi-C, SPRITE, and GAM against average distance data. Adapted from [40].

As a further control, we compared the experimental imaged average distance matrix [21] of the locus against those derived from the SBS [38] and SAW models (**Fig. 4.3a-c**). While the SAW matrix has no TADs or patterns and poorly compares against the experiment (genomic-distance-corrected Pearson r'=0.32), the SBS model well captures the features of the data and returns a much higher correlation (r'=0.84, see also **Chapter 3**). We also computed the distribution of r' correlations between the pairs experiment–experiment, experiment–SBS and experiment–SAW single-molecule distance matrices and found that while the first and second distributions are not statistically distinguishable (two-sided Mann–Whitney test p=0.19), the experiment–SAW distribution is clearly different (p= 0; **Fig. 4.3d**).

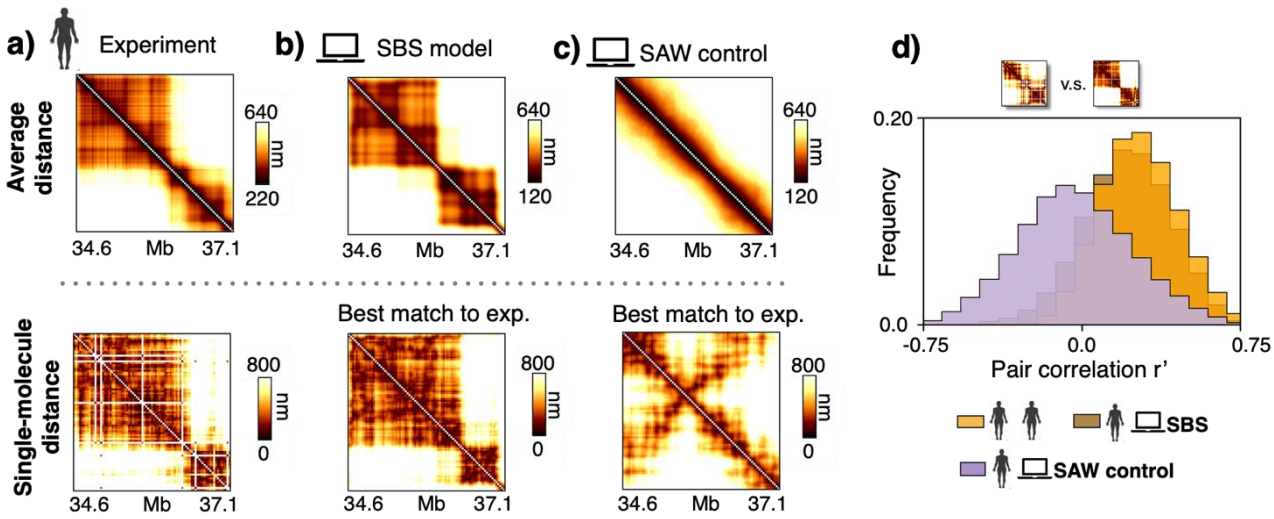

Figure 4.3: Single-molecule distance maps of the SBS model significantly correlate with single-cell imaging data in HCT116 cells. The average distance matrix (top) and a representative example of single-molecule distance map are shown for: **a)** multiplexed FISH imaging data [21] of the studied HCT116 locus; **b)** the SBS polymer model of the locus [38]; **c)** a control SAW model with same linear size as real microscopy conformations. While the SAW model tends to form fleeting and random interactions that result in featureless bulk data, the SBS model recapitulates the TAD-like structures visible in the experiment and also captures the highly specific contacts established in single-molecules. **d)** Distribution of r' correlation values between pairs of experiment-experiment (orange), of experiment-SBS (brown) and of experiment-SAW (violet) single-molecule distance matrices. The first and second distributions are statistically not distinguishable (two-sided Mann-Whitney test p=0.19) to each other and statistically different from the control (p=0). Taken from [40].

Overall, the agreement between model and experiments provides a validation of our polymer model as its 3D structures, which are inferred from Hi-C data, faithfully recapitulate independent SPRITE, GAM and microscopy data, also at the single-molecule level, consistently across different experiments, loci and cell types. Additionally, our in-silico approach does not reveal biases favoring Hi-C, SPRITE, or GAM, hence allowing a fair significant comparison of their performance.

4.2 Results of our in-silico experiment

In this Section, we discuss the results of the comparison of our in-silico Hi-C, SPRITE and GAM. In brief, we show that all three methods faithfully recapitulate bulk 3D distances (subsection 4.2.1), yet they are less accurate at the single-cell level because of the intrinsic structural variability of single-molecule conformations (subsection 4.2.2). Then, while SPRITE is less sensitive to the number of cells employed in the experiment, GAM is found to be the most affected (subsection 4.2.3). Finally, we show that, differently from Hi-C and SPRITE, GAM has the best and lower varying noise-to-signal level for long-distance interactions (subsection 4.2.4), thus highlighting the different experimental conditions where each technology is most effective.

4.2.1. Bulk data from in-silico Hi-C, SPRITE, and GAM are faithful to average 3D distances

First, we aimed to investigate whether our in-silico technologies do reflect the spatial structure of the underlying ensemble of model single-molecule 3D conformations. Hence, in our *Sox9* case study we computed the average distance matrix of the known model 3D structures, which represents their typical folding, and we compared it against the in-silico Hi-C, SPRITE and GAM bulk contact data (**Fig. 4.4a**). Albeit bulk interaction patterns of those technologies are similar to each other, GAM is found to better capture longer-range contacts between TADs (**Fig. 4.4a**). That is consistent with the result that the noise-to-signal ratio is the lowest in GAM for larger genomic separation, as will be explained in the subsection 4.2.4. All three methods well reproduce the known TAD-like structures of the average distance matrix (**Fig. 4.4b**, different colors in the bottom bar), as quantified by their high correlation values (Spearman correlations are negative as contacts and distances are inversely related). Different correlation measures provide the same scenario, as well as the analyses performed in the other loci [40]. Taken together, our results show that bulk data from in-silico Hi-C,

SPRITE and GAM faithfully reflect the spatial structure of the underlying 3D conformations of the studied loci, as they provide overall comparable information on average distances.

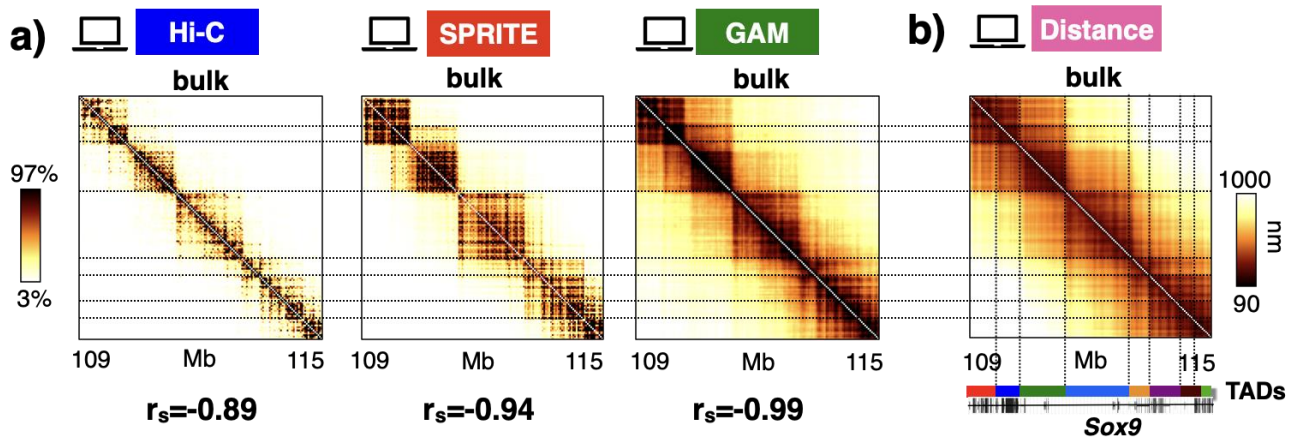

Fig. 4.4: Bulk Hi-C, SPRITE and GAM data faithfully recapitulate average 3D distances. a) In-silico bulk Hi-C, SPRITE and GAM maps of the *Sox9* locus are compared to the average 3D distance matrix of the known single-molecule 3D conformations of the locus model (panel **b)**). On the bottom, spearman correlation coefficients are listed between each contact map and the average 3D model distance matrix (similar values are also found for the Pearson and HiCRep coefficients [40]). **b)** Average distance matrix of the ensemble of in-silico model single-molecule 3D conformations. A colored bar on the bottom highlights the main TAD structures of the *Sox9* locus. Taken from [40].

4.2.2. The intrinsic variability of single-molecule 3D conformations affects in-silico single-cell contact data

We also investigated how the structural variability of single molecules impacts on contact maps beyond the averaged-bulk level. Consistent with single-cell imaging data [21–23], single-molecule *Sox9* conformations exhibit strong variability in the ensemble of model 3D structures, as their single-cell distance matrices have, for instance, broadly varying Spearman correlations with the average distance matrix of the locus (**Fig. 4.5a**; mean r_s=0.88). Also, the correlation of an in silico single-cell Hi-C, SPRITE or GAM contact map with its corresponding single-cell distance matrix is found to be much lower than in the case of the bulk data. For example, in the ideal case of in-silico experiments with efficiency set to 100%, the distribution of correlations between in-silico single-cell Hi-C contact and corresponding distance maps has an average r_s=-0.37, while r_s=-0.46 is the average measured for SPRITE (**Fig. 4.5b**). For GAM the correlation between single-cells maps is even lower, average r_s=-0.15, and its distribution much broader (in the range -0.4< r_s<0). At lower, realistic values of detection efficiency, in-silico single-cell contacts are further worsened and the correlations with corresponding distance maps decrease (**Fig. 4.5c**). By taking, e.g., an efficiency of 0.5, the average correlation between in-silico single-cell maps is around r_s=0.2, 0.4, and 0.1 for Hi-C, SPRITE and GAM, which, importantly, are all values consistent with real single-cell experiments [40,90]. Hence, the stochasticity of in-silico Hi-C, SPRITE and GAM data reflects the intrinsic structural variability of single-molecule chromatin conformations. For that reason, in-silico single-cell data are less faithful than bulk data to the corresponding single-cell distances, even in the ideal case of 100% efficiency.

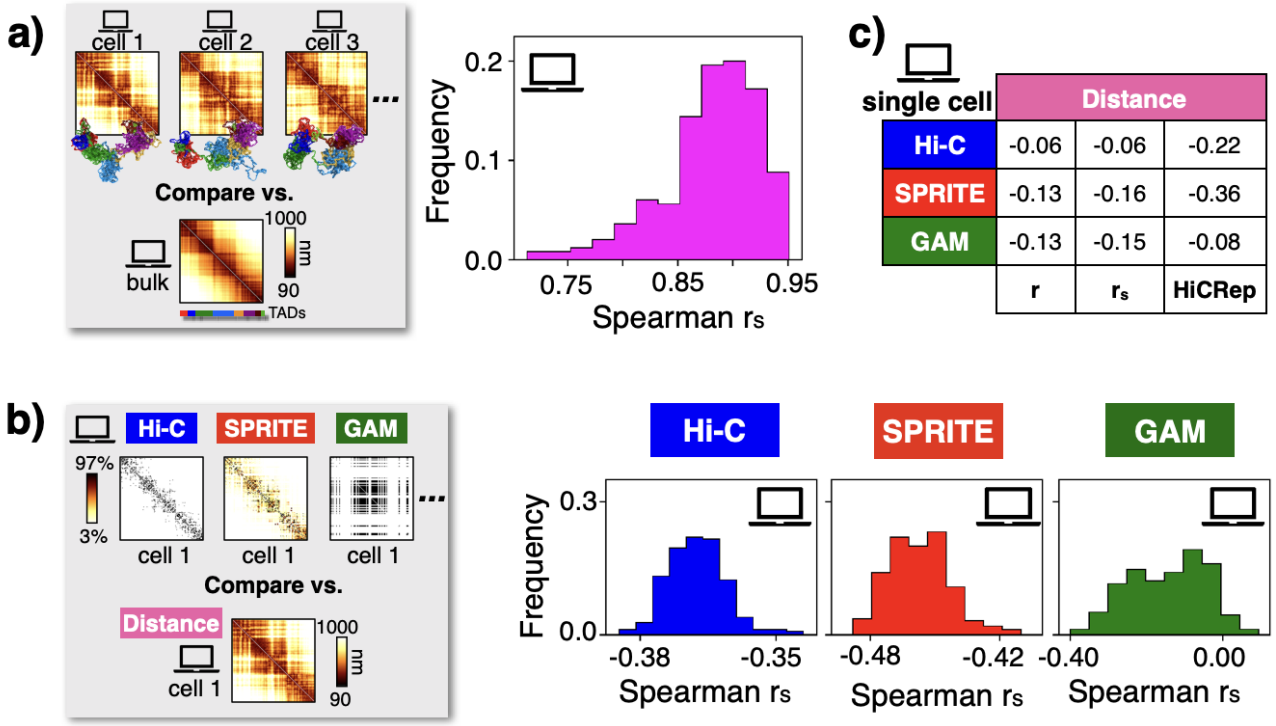

Fig. 4.5: The variability of single-molecule 3D conformations is reflected in the highly stochasticity of single-cell contact maps. a) Left: In silico single-cell distance maps of the *Sox9* locus are compared against the average distance matrix. Right: The distribution of Spearman correlations between in-silico single-cell and average distance maps is broad and has an average value r_s=0.88. Pearson and HiCRep correlations provide similar results [40]. **b)** The distributions of Spearman correlations between in-silico single-cell contact matrices at efficiency=1 and their corresponding in-silico single-cell distance matrices return correlations much lower than in the case of bulk data (see **Fig. 4.4**). **c)** Mean correlation values between in-silico single-cell contact maps and corresponding single-cell distance maps at efficiencies similar to those of the real experiments (e.g., we used here 0.05 for Hi-C and SPRITE and 0.5 for GAM). Adapted from [40].

4.2.3. The minimal number of cells for replicate in-silico experiments is different for Hi-C, SPRITE and GAM

Next, we studied the effect of the number of in-silico cells (*N*) on the quality of in-silico Hi-C, SPRITE, and GAM contact maps. By setting, for instance, the efficiency to 100%, we found that those matrices become sharper and stabilize as *N* increases, e.g., from 10 to 10000 cells (**Fig. 4.6a**). This is also confirmed in the case of efficiencies comparable to typical experimental values (**Fig. 4.6b**), i.e., 0.05 in the case of Hi-C (taken as an upper limit of the values reported in the literature [90,142,149,150]) and SPRITE, and 0.5 for GAM (whose efficiency is roughly one order of magnitude larger than the Hi-C and SPRITE). However, our simulations suggest that the critical, threshold value of *N* to reach saturation is strongly dependent on both the efficiency level and the considered technology (**Fig. 4.6a, b**). For that reason, we aimed to quantitatively identify the minimal number of cells that, at a given efficiency, is required for replicate in-silico experiments to approach the bulk limit (i.e., to return similar contact maps). To this aim, we measured in our *Sox9* case study the similarity between pairs of identical experiments (i.e., with same *N* and efficiency) by computing the average Pearson correlation between their corresponding contact maps (**Fig. 4.6c**; other correlations, e.g., Spearman and HiCRep, returned similar results [40]). Such a replicate

correlation increases as a function of *N* and reaches a plateau to 1 in the large *N* limit (**Fig. 4.6d, e**); importantly, this behaviour is independent of the experimental efficiency as the average over a large number of cells compensates for reduced efficiency values (in **Fig. 4.6d** efficiency is set to 0.1, while in **Fig. 4.6e** it is close to the real experimental values). Then, we heuristically defined the minimal number of cells, *M*, required for statistically reproducible results across different replicates as the value of *N* where the correlation is greater than a given threshold, r_t=0.9. In the paper [40], we prove that such heuristic criterion, as well as the above definitions and features of *M*, can be fully grounded on the Central Limit Theorem (CLT). Interestingly, we found that *M* varies with the different technologies: for instance, if the efficiency is 0.1, we have *M*=200, 100, and 2000 for Hi-C, SPRITE and GAM, respectively (**Fig. 4.6d**). Additionally, our data show that *M* also depends on the considered detection efficiency, as for real efficiency values (e.g., 0.05 for Hi-C and SPRITE, and 0.5 for GAM) the corresponding values are 650, 250 and 800, respectively (**Fig. 4.6e**). Importantly, GAM data include both random and non-random cosegregation events (i.e., specific contacts), which can be dissect by using statistical and math tools, such as SLICE [15]. When GAM is combined with SLICE, which is specifically designed to capture significant interactions, we find that noise effects are drastically reduced and *M* becomes nearly half than the value required for GAM alone under same conditions (hence, by using SLICE, *M*=400 instead of 800 for real efficiencies) [40]. For that reason, SLICE can reliably enhance the performance of GAM in real applications, for instance on in-vivo sample tissues where the number of available cells is small.

Then, we aimed to rigorously investigate how the experimental efficiency affects the quality of in-silico data, i.e., the minimal number, *M*, of cells required for replicate similarity. We found that this number strongly depends on efficiency, as it increases approximately as an inverse squared power law as the efficiency decreases (**Fig. 4.6f**). In general, we observe that the estimated *M* for SPRITE is two times smaller than the one for Hi-C, and roughly one order of magnitude smaller than for GAM. In the paper [40], we show that those findings have general validity, as they hold also for other genomic loci (e.g., the murine *HoxD*, *Epha4* and the human HCT116 loci). Moreover, they do not depend on the specific approach of our SBS model, as similar results are found if the polymer model is inferred, e.g., from GAM rather from Hi-C or if a toy block-copolymer, unrelated to real chromatin loci, is considered [40]. Overall, our analyses illustrate how the statistical reproducibility of in-silico contact matrices is affected by both the number of cells, *N*, and experimental efficiency, and how the different technologies under different situations perform relative to each other.

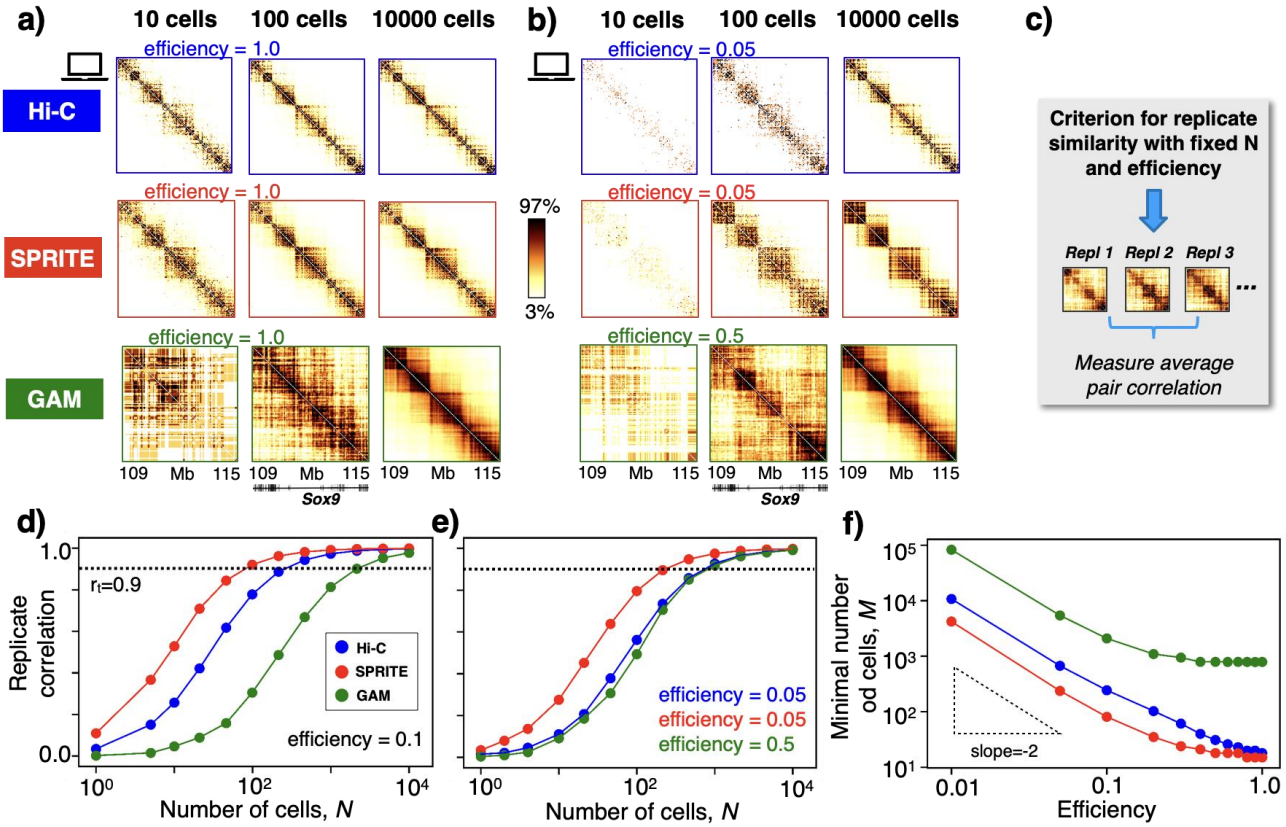

Fig. 4.6: Hi-C, SPRITE and GAM require a different number of cells to ensure replicate reproducibility. a) In-silico Hi-C, SPRITE and GAM contact maps of the *Sox9* locus are shown for different numbers of in-silico cells (*N*). Efficiency is set to 1 (ideal case). **b)** As in panel **a)**, but efficiency is similar to that of real experiments (0.05 for Hi-C and SPRITE, 0.5 for GAM). **c)** Scheme of the heuristic criterion to assess similarity between replicates: the minimal number of cells, *M*, providing reproducible (i.e., statistically similar) contact maps is defined as the value of *N* whereby the average Pearson correlation between replicates is higher that the threshold r_t=0.9. **d)** Pearson correlation between replicate experiments as a function of *N* for Hi-C, SPRITE and GAM at a given efficiency (0.1 in the case shown). The dashed line is the threshold correlation value r_t=0.9. **e)** As in panel **d)** for efficiencies corresponding to typical experimental values. **f)** The minimal number of cells, *M*, strongly depends on the detection efficiency and follows approximately an inverse squared power law. This behaviour, as fully discussed in [40], is consistent with the Central Limit Theorem (CLT). Adapted from [40].

4.2.4. Noise-to-signal ratio levels vary differently in Hi-C, SPRITE and GAM

Finally, we considered the noise-to-signal level of the entries of contact matrices and investigated how it varies with genomic separation, with the number of cells *N* and with the efficiency of the in-silico experiments. The noise-to-signal ratio of each entry, C_{ij}, of a contact matrix is defined as the ratio between the standard deviation, σ_{ij}, and the mean, μ_{ij}, of that entry across replicate experiments under the same conditions. By fixing *N* and the in-silico efficiency, we found that the average noise-to-signal ratio, σ/μ (average over the entries with same genomic separation), strongly depends on the genomic distance (**Fig. 4.7a**). For instance, in our *Sox9* locus, while SPRITE has the lowest σ/μ value at genomic separations below 1Mb, this ratio is the lowest for GAM at large genomic scales (>1Mb), as it is almost one order of magnitude lower than Hi-C and SPRITE. Also, for both Hi-C and SPRITE, σ/μ increases by more than one order of magnitude in the genomic distance

range 0.5-5Mb and steeply boosts above 1Mb, whereas GAM has an overall lower varying (i.e., more stable) noise-to-signal level. Conversely, by keeping fixed genomic distance and efficiency, σ/μ decreases as a function of N (**Fig. 4.7b**). That behaviour is expected as in the large N limit the contact matrices overall are not sensitive to the considered efficiency value, consistent with our previous observations (see, e.g., **Fig. 4.6a, b**). In particular, σ/μ exhibits an inverse squared power law in N (i.e., $\sigma/\mu \sim N^{-0.5}$), which we show in the paper [40] to be consistent again with the CLT. Next, for a fixed genomic separation and N, the noise-to-signal ratio strongly depends, as expected, on the efficiency (**Fig. 4.7c**, σ/μ decreases roughly as an inverse power law of the experimental efficiency). As shown in [40], all the above results also hold for the other investigated loci (*HoxD*, *Epha4*, and the human HCT116 loci) as well as for polymer toy models, supporting the view that quantitative comparison of the performance of in-silico Hi-C, SPRITE and GAM has a more general validity.

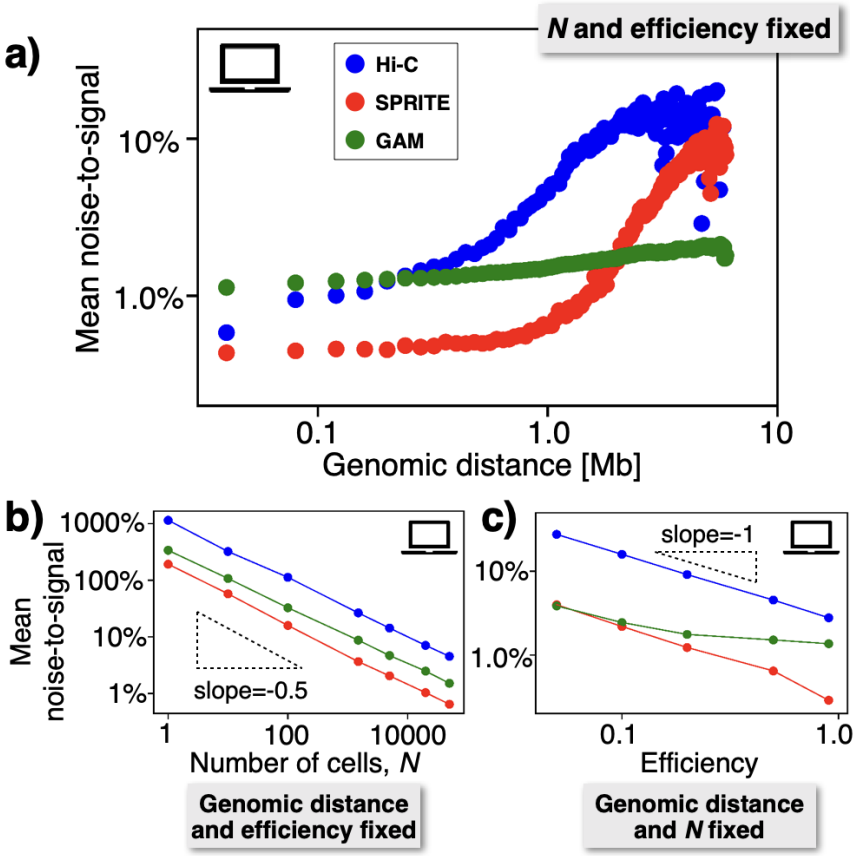

Fig. 4.7: Noise-to-signal ratio levels vary differently with genomic distance, number of cells and efficiency in Hi-C, SPRITE and GAM. a) The mean noise-to-signal ratio (σ/μ) of a contact map depends on the genomic distance at fixed number of cells (N) and efficiency. In this example, N=50000 and efficiency=0.5. **b)** The σ/μ ratio at fixed genomic distance and efficiency decreases with N as an inverse square root, consistent with the CLT [40]. In this plot, genomic separation is 1Mb and efficiency set to 0.5. **c)** At fixed genomic distance and number of cells, σ/μ scales approximately as an inverse power law of the efficiency. In this case, genomic distance=1Mb and N=50000. Adapted from [40].

Overall, based on computer simulations and polymer physics modeling, we discussed an in-silico experiment to test how well Hi-C, SPRITE and GAM, i.e., three different powerful technologies currently employed to probe DNA contacts genome wide, represent the 3D structure of the genome. By using a validated ensemble of known polymer 3D structures as benchmark, we performed in this

simplified, yet fully controlled, framework the first quantitative comparison of their performance. We systematically analyzed in silico Hi-C, SPRITE and GAM under different experimental conditions, including key parameters such as the number of cells considered in the experiment, detection efficiency and genomic separation scales. The results of our study are summarized in **Table II**.

	Faithfulness to 3D structure		Replicate similarity vs. cell number	Noise vs. detection efficiency	
	Bulk data	Single-cell		< 1Mb	> 1Mb
Hi-C	***	**	**	**	*
SPRITE	***	***	***	**	**
GAM	***	*	*	*	***

Table II: Summary table of the performance of in-silico Hi-C, SPRITE and GAM. Adapted from [40].

CONCLUSIONS AND PERSPECTIVES

Recent technological advancements from molecular biology have allowed to measure chromosome physical interactions at increasing resolution and scale [13–15,88]. They have revealed that our genome is folded into a complex 3D architecture [1–3,10,104] within the cell nucleus involving a hierarchy of structural patterns, such topologically associating domains (TADs) [16,17], i.e., large genomic regions that display a high degree of interaction, as well as higher-order structures extending up to the whole chromosome scale [13,18]. Importantly, such an organization serves important functional purposes: TADs, for instance, are thought to act as insulating structures, spatially confining the long-distance interactions between genes and their regulators [2,3,104]. At the same time, novel approaches from FISH microscopy [19–23] are allowing to resolve single-molecule DNA structures at very fine scales, e.g., few nanometers, showing that those organizational features (e.g., TADs and other structures) physically exist in individual cell nuclei [21]. Yet, the self-organizational principles of the system remain unclear. How are chromosome contacts established? What is the physical mechanism shaping genome 3D structure in single-cells? How to explain the experimentally reported cell-to-cell structural variability of genome conformations?

In this thesis, we aimed to tackle those questions by using principled models from polymer physics. In particular, we focused on the Strings and Binders (SBS) polymer model [38,43,45] whereby a chromosome region is represented as a self-avoiding chain of beads, in a thermal bath, with specific binding sites for cognate diffusing molecular binders, such as Transcription Factors, that can bridge those sites. The equilibrium 3D conformations of the model fall in structural classes corresponding to its thermodynamics phases: upon increasing binder concentration or affinity above a threshold value, the system undergoes a phase transition from a coil (i.e., randomly folded) to a phase-separated state where the attraction between polymer binding sites and their associated cognate binding molecules drives a micro-phase-separation of the chain into distinct globules. By performing extensive computer simulations of the model, we derived predictions about DNA single-molecule 3D structures in real human genomic regions that we compared against super-resolution imaging data in single-cells [21,38]. We showed that the model conformations in the globule phase-separated state of the theory display structural features (e.g., 3D spatial distances, contact patterns, TAD boundary probabilities and strengths, separation scores, gyration radius distributions) consistent with those measured in microscopy experiments and they also reflect the same degree of observed structural variability. In the SBS picture, such a variability naturally results from the intrinsic thermodynamic folding degeneracy of the model, as polymer conformations do not fold in a single, naïve structure, as in protein folding, but in a much broad set of possible microstates. We also tested the predictions of the model upon removal of the cohesin functional complex, which is a known genome organizing factor. Our results, consistently validated against microscopy data, indicate that cohesin depletion tends to reverse phase-separation from the globule to the coil (randomly folded) thermodynamics state in most single-cells, resulting in much more variable and transient contacts in single-molecules. Finally, we explored the steady-state time dynamics of the polymer conformations of the theory and showed that thermodynamic globule phase-separation can either establish contact specificity or spatial insulation between different genomic regions, thus

providing a quantitative picture on how contacts, e.g., between genes and distal regulators, can be controlled at the molecular level.

To summarize, the consistent agreement between single-cell imaged and model-derived conformations supports the view whereby, in the studied genomic regions, chromosome folding is driven at the single-cell level by a thermodynamics physics mechanism of polymer phase-separation [38]; the observed cell-to-cell structural variability spontaneously results from the intrinsic thermodynamic conformational degeneracy of polymer folding; the depletion of cohesin, consistent with the key role of this molecular factor in shaping chromosome organization, promotes a reversal of phase-separation into more open randomly folded structures. Overall, globule phase separation is shown to be a robust yet reversible mechanism of chromosome organization where stochasticity of DNA interactions, which is typical of weak biochemical affinities, can coexist with specificity.

Next, we also discussed a different, original application of models from polymer physics, which can be used to evaluate advantages and limitations of benchmarked experimental methods for determination of genome structure [40]. We considered three important methods, i.e., Hi-C, SPRITE and GAM, which are currently used to probe genome-wide chromosomal physical interactions. However, it is unclear to what extent those technologies are faithful to the underlying 3D structure of the genome and how they perform relative to each other in different applications, because they return distinct measures of interactions and no benchmark exists. For instance, is GAM or SPRITE as faithful to chromosome structure as Hi-C? Which method requires the minimal number of cells to achieve statistically significant results? And which is more suited in detecting interactions at large genomic distances? To answer such questions, we designed a computational experiment where we simulated “in-silico” those distinct three methods on ensembles of fully known and validated SBS polymer structures. That allowed us to systematically investigate in-silico Hi-C, SPRITE and GAM under different experimental conditions, examining their behavior against crucial experimental parameters such as detection efficiency, cell numbers and genomic separation scales. We found, for instance, that all three technologies faithfully describe the average chromosome structure from a population of cells, whereas they are less reliable at the single-cell level because dominated by noise. Then, under equal conditions, SPRITE is the technique requiring the lowest number of cells to ensure the statistical reproducibility of the measures, while GAM the highest. The noise-to-signal ratio follows an inverse power law with detection efficiency and grows with genomic distance differently among the three methods, with GAM having an overall lower varying noise-to-signal level, especially at larger genomic separations. Overall, although simplified, those studies can help in designing real-world applications of those technologies for specific purposes, as well as in guiding experimentalists on the best approach to use to interrogate genome structure in different contexts.

Lastly, we are currently involved in many other ongoing research projects. As an example, we are investigating further mechanisms of genome folding in single-cells beyond polymer phase-separation. For instance, loop-extrusion (LE) has been proposed as another major physical process shaping chromosome large-scale spatial organization. In the LE picture, physical proximity between distal sites is established by molecular motors that bind to DNA and extrude a loop in an out-of-

equilibrium, active physical process involving external energy (e.g., ATP molecule) consumption. However, does loop-extrusion explain single-cell data? How does it compare to thermodynamic polymer phase-separation? Do those distinct physical mechanisms compete or coexist in establishing genome architectures? Those are some of the questions we are trying to answer to. On another side, we are working in collaboration with the Institute of Human Genetics in Montpellier (France, Giacomo Cavalli's Lab) to investigate the mechanisms of gene regulation at the single-molecule level. Understanding the 3D genome has indeed important implications for real life, as for instance aberrant chromosome organizations are increasingly recognized as a hallmark of various diseases ranging from common cancers to rarer genetic disorders [4,10]. Hence, in such an aspiring yet challenging diagnostic perspective, quantitative models from physics can be essential to make predictions on the effects of pathogenic mutations on genome architecture based on the comprehension of the underlying fundamental molecular mechanisms. That is the golden path we aim to push our models through.

REFERENCES

[1] W. A. Bickmore and B. Van Steensel, *Genome Architecture: Domain Organization of Interphase Chromosomes*, Cell.

[2] J. Dekker and L. Mirny, *The 3D Genome as Moderator of Chromosomal Communication*, Cell.

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

[4] J. Dekker and T. Misteli, *Long-Range Chromatin Interactions*, Cold Spring Harb. Perspect. Biol. **7**, (2015).

[5] A. Pombo and N. Dillon, *Three-Dimensional Genome Architecture: Players and Mechanisms*, Nature Reviews Molecular Cell Biology.

[6] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions*, Cell (2015).

[7] D. Hnisz, A. S. Weintrau, D. S. Day, A. L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, *Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods*, Science (80-.). (2016).

[8] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, *Breaking TADs: How Alterations of Chromatin Domains Result in Disease*, Trends in Genetics.

[9] S. Bianco, D. G. Lupiáñez, A. M. Chiariello, C. Annunziatella, K. Kraft, R. Schöpflin, L. Wittler, G. Andrey, M. Vingron, A. Pombo, S. Mundlos, and M. Nicodemi, *Polymer Physics Predicts the Effects of Structural Variants on Chromatin Architecture*, Nat. Genet. (2018).

[10] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, *Structural Variation in the 3D Genome*, Nature Reviews Genetics.

[11] A. L. Valton and J. Dekker, *TAD Disruption as Oncogenic Driver*, Current Opinion in Genetics and Development.

[12] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stütz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm, and J. O. Korbel, *Pan-Cancer Analysis of Somatic Copy-Number Alterations Implicates IRS4 and IGF2 in Enhancer Hijacking*, Nat. Genet. (2017).

[13] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker,

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Science (80-.). (2009).

[14] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman, *Higher-Order Inter-Chromosomal Hubs Shape 3D Genome Organization in the Nucleus*, Cell **174**, 744 (2018).

[15] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. De Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. W. Edwards, M. Nicodemi, and A. Pombo, *Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping*, Nature **543**, 519 (2017).

[16] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions*, Nature (2012).

[17] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. Van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, *Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre*, Nature (2012).

[18] J. Fraser, C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken, S. Q. Xie, K. J. Morris, M. Itoh, H. Kawaji, I. Jaeger, Y. Hayashizaki, P. Carninci, A. R. Forrest, C. A. Semple, J. Dostie, A. Pombo, and M. Nicodemi, *Hierarchical Folding and Reorganization of Chromosomes Are Linked to Transcriptional Changes in Cellular Differentiation*, Mol. Syst. Biol. **11**, 852 (2015).

[19] A. N. Boettiger, B. Bintu, J. R. Moffitt, S. Wang, B. J. Beliveau, G. Fudenberg, M. Imakaev, L. A. Mirny, C. T. Wu, and X. Zhuang, *Super-Resolution Imaging Reveals Distinct Chromatin Folding for Different Epigenetic States*, Nature **529**, 418 (2016).

[20] D. I. Cattoni, A. M. C. Gizzi, M. Georgieva, M. Di Stefano, A. Valeri, D. Chamousset, C. Houbron, S. Déjardin, J. B. Fiche, I. González, J. M. Chang, T. Sexton, M. A. Marti-Renom, F. Bantignies, G. Cavalli, and M. Nollmann, *Single-Cell Absolute Contact Probability Detection Reveals Chromosomes Are Organized by Multiple Low-Frequency yet Specific Interactions*, Nat. Commun. **8**, 1753 (2017).

[21] B. Bintu, L. J. Mateo, J.-H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A. N. Boettiger, and X. Zhuang, *Super-Resolution Chromatin Tracing Reveals Domains and Cooperative Interactions in Single Cells*, Science (80-.). (2018).

[22] A. M. Cardozo Gizzi, D. I. Cattoni, J. B. Fiche, S. M. Espinola, J. Gurgo, O. Messina, C. Houbron, Y. Ogiyama, G. L. Papadopoulos, G. Cavalli, M. Lagha, and M. Nollmann, *Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms*, Mol. Cell **74**, 212 (2019).

[23] E. H. Finn, G. Pegoraro, H. B. Brandão, A. L. Valton, M. E. Oomen, J. Dekker, L. Mirny, and T. Misteli, *Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization*, Cell **176**, P1502 (2019).

[24] A. L. Sanborn, S. S. P. P. Rao, S.-C. C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, *Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes*, Proc. Natl. Acad. Sci. **112**, E6456 (2015).

[25] A. M. Chiariello, C. Annunziatella, S. Bianco, A. Esposito, and M. Nicodemi, *Polymer Physics of Chromosome Large-Scale 3D Organisation*, Sci. Rep. **6**, (2016).

[26] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, *Formation of Chromosomal Domains by Loop Extrusion*, Cell Rep. (2016).

[27] D. Jost, P. Carrivain, G. Cavalli, and C. Vaillant, *Modeling Epigenome Folding: Formation and Dynamics of Topologically Associated Chromatin Domains*, Nucleic Acids Res. (2014).

[28] B. Zhang and P. G. Wolynes, *Topology, Structures, and Energy Landscapes of Human Chromosomes*, Proc. Natl. Acad. Sci. U. S. A. **112**, (2015).

[29] C. A. Brackley, J. M. Brown, D. Waithe, C. Babbs, J. Davies, J. R. Hughes, V. J. Buckle, and D. Marenduzzo, *Predicting the Three-Dimensional Folding of Cis-Regulatory Regions in Mammalian Genomes Using Bioinformatic Data and Polymer Models*, Genome Biol. **17**, (2016).

[30] M. Di Stefano, J. Paulsen, T. G. Lien, E. Hovig, and C. Micheletti, *Hi-C-Constrained Physical Models of Human Chromosomes Recover Functionally-Related Properties of Genome Organization*, Sci. Rep. **6**, (2016).

[31] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, *Transferable Model for Chromosome Architecture*, Proc. Natl. Acad. Sci. **113**, 12168 (2016).

[32] M. Barbieri, S. Q. S. Q. Xie, E. Torlai Triglia, A. M. A. M. Chiariello, S. Bianco, I. De Santiago, M. R. M. R. Branco, D. Rueda, M. Nicodemi, and A. Pombo, *Active and Poised Promoter States Drive Folding of the Extended HoxB Locus in Mouse Embryonic Stem Cells*, Nat. Struct. Mol. Biol. **24**, 515 (2017).

[33] A. Buckle, C. A. Brackley, S. Boyle, D. Marenduzzo, and N. Gilbert, *Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci*, Mol. Cell **72**, (2018).

[34] G. Shi, L. Liu, C. Hyeon, and D. Thirumalai, *Interphase Human Chromosome Exhibits out of Equilibrium Glassy Dynamics*, Nat. Commun. (2018).

[35] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo, *Nonequilibrium Chromosome Looping via Molecular Slip Links*, Phys. Rev. Lett. **119**, 138101 (2017).

[36] S. Bianco, C. Annunziatella, G. Andrey, A. M. Chiariello, A. Esposito, L. Fiorillo, A. Prisco, M. Conte, R. Campanile, and M. Nicodemi, *Modeling Single-Molecule Conformations of the HoxD Region in Mouse Embryonic Stem and Cortical Neuronal Cells*, Cell Rep. (2019).

[37] A. M. Chiariello, S. Bianco, A. M. Oudelaar, A. Esposito, C. Annunziatella, L. Fiorillo, M. Conte,

A. Corrado, A. Prisco, M. S. C. Larke, J. M. Telenius, R. Sciarretta, F. Musella, V. J. Buckle, D. R. Higgs, J. R. Hughes, and M. Nicodemi, *A Dynamic Folded Hairpin Conformation Is Associated with α-Globin Activation in Erythroid Cells*, Cell Rep. (2020).

[38] M. Conte, L. Fiorillo, S. Bianco, A. M. Chiariello, A. Esposito, and M. Nicodemi, *Polymer Physics Indicates Chromatin Folding Variability across Single-Cells Results from State Degeneracy in Phase Separation*, Nat. Commun. (2020).

[39] S. Bianco, A. M. Chiariello, M. Conte, A. Esposito, L. Fiorillo, F. Musella, and M. Nicodemi, *Computational Approaches from Polymer Physics to Investigate Chromatin Folding*, Current Opinion in Cell Biology.

[40] L. Fiorillo, F. Musella, M. Conte, R. Kempfer, A. M. Chiariello, S. Bianco, A. Kukalev, I. Irastorza-Azcarate, A. Esposito, A. Abraham, A. Prisco, A. Pombo, and M. Nicodemi, *Comparison of the Hi-C, GAM and SPRITE Methods Using Polymer Models of Chromatin*, Nat. Methods **18**, (2021).

[41] D. Plewczynski and M. Kadlof, *Computational Modelling of Three-Dimensional Genome Structure*, Methods.

[42] D. Racko, F. Benedetti, J. Dorier, and A. Stasiak, *Transcription-Induced Supercoiling as the Driving Force of Chromatin Loop Extrusion during Formation of TADs in Interphase Chromosomes*, Nucleic Acids Res. **46**, (2018).

[43] M. Nicodemi and A. Prisco, *Thermodynamic Pathways to Genome Spatial Organization in the Cell Nucleus*, Biophys. J. (2009).

[44] M. Bohn and D. W. Heermann, *Diffusion-Driven Looping Provides a Consistent Provides a Consistent Framework for Chromatin Organization*, PLoS One (2010).

[45] M. Barbieri, M. Chotalia, J. Fraser, L.-M. L. M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, *Complexity of Chromatin Folding Is Captured by the Strings and Binders Switch Model*, Proc. Natl. Acad. Sci. U. S. A. **109**, 16173 (2012).

[46] C. A. Brackley, S. Taylor, A. Papantonis, P. R. Cook, and D. Marenduzzo, *Nonspecific Bridging-Induced Attraction Drives Clustering of DNA-Binding Proteins and Genome Organization*, Proc. Natl. Acad. Sci. U. S. A. (2013).

[47] A. Rosa and R. Everaers, *Structure and Dynamics of Interphase Chromosomes*, PLoS Comput. Biol. (2008).

[48] J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur, and L. A. Mirny, *Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation*, Proc. Natl. Acad. Sci. U. S. A. **115**, (2018).

[49] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci, *3D Genome Reconstruction from Chromosomal Contacts*, Nat. Methods **11**, (2014).

[50] H. Tjong, W. Li, R. Kalhor, C. Dai, S. Hao, K. Gong, Y. Zhou, H. Li, X. J. Zhou, M. A. Le Gros, C. A. Larabell, L. Chen, and F. Alber, *Population-Based 3D Genome Structure Analysis Reveals Driving Forces in Spatial Genome Organization*, Proc. Natl. Acad. Sci. U. S. A. (2016).

[51] S. Zhang, D. Chasman, S. Knaack, and S. Roy, *In Silico Prediction of High-Resolution Hi-C Interaction Matrices*, Nat. Commun. **10**, (2019).

[52] G. Fudenberg, D. R. Kelley, and K. S. Pollard, *Predicting 3D Genome Folding from DNA Sequence with Akita*, Nat. Methods **17**, (2020).

[53] R. Schwessinger, M. Gosden, D. Downes, R. C. Brown, A. M. Oudelaar, J. Telenius, Y. W. Teh, G. Lunter, and J. R. Hughes, *DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning*, Nat. Methods **17**, (2020).

[54] Y. Wang, Y. Zhang, R. Zhang, T. van Schaik, L. Zhang, T. Sasaki, D. Peric-Hupkes, Y. Chen, D. M. Gilbert, B. van Steensel, A. S. Belmont, and J. Ma, *SPIN Reveals Genome-Wide Landscape of Nuclear Compartmentalization*, Genome Biol. **22**, (2021).

[55] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. Filion, and M. A. Marti-Renom, *Automatic Analysis and 3D-Modelling of Hi-C Data Using TADbit Reveals Structural Features of the Fly Chromatin Colors*, PLoS Comput. Biol. **13**, (2017).

[56] G. Nir, I. Farabella, C. Pérez Estrada, C. G. Ebeling, B. J. Beliveau, H. M. Sasaki, S. H. Lee, S. C. Nguyen, R. B. McCole, S. Chattoraj, J. Erceg, J. AlHaj Abed, N. M. C. Martins, H. Q. Nguyen, M. A. Hannan, S. Russell, N. C. Durand, S. S. P. Rao, J. Y. Kishi, P. Soler-Vila, M. Di Pierro, J. N. Onuchic, S. P. Callahan, J. M. Schreiner, J. A. Stuckey, P. Yin, E. L. Aiden, M. A. Marti-Renom, and C. T. Wu, *Walking along Chromosomes with Super-Resolution Imaging, Contact Maps, and Integrative Modeling*, PLoS Genet. **14**, e1007872 (2018).

[57] D. Lin, G. Bonora, G. G. Yardimci, and W. S. Noble, *Computational Methods for Analyzing and Modeling Genome Structure and Organization*, Wiley Interdiscip. Rev. Syst. Biol. Med. **11**, e1435 (2018).

[58] M. Di Stefano, J. Paulsen, D. Jost, and M. A. Marti-Renom, *4D Nucleome Modeling*, Current Opinion in Genetics and Development.

[59] Q. Li, H. Tjong, X. Li, K. Gong, X. J. Zhou, I. Chiolo, and F. Alber, *The Three-Dimensional Genome Organization of Drosophila Melanogaster through Data Integration*, Genome Biol. **18**, 145 (2017).

[60] H. J. Kim, G. G. Yardımcı, G. Bonora, V. Ramani, J. Liu, R. Qiu, C. Lee, J. Hesson, C. B. Ware, J. Shendure, Z. Duan, and W. S. Noble, *Capturing Cell Type-Specific Chromatin Compartment Patterns by Applying Topic Modeling to Single-Cell Hi-C Data*, PLoS Comput. Biol. **16**, (2020).

[61] Y. Qi and B. Zhang, *Predicting Three-Dimensional Genome Organization with Chromatin States*, PLoS Comput. Biol. **15**, (2019).

[62] P. G. De Gennes, *Scaling Concepts in Polymer Physics. Cornell University Press.*, Ithaca N.Y., (1979).

[63] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, *Capturing Chromosome Conformation*, Science (80-.). **295**, (2002).

[64] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat, *Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by*

Chromosome Conformation Capture-on-Chip (4C), Nat. Genet. **38**, (2006).

[65] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, *Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements*, Genome Res. **16**, (2006).

[66] T. Misteli, *The Inner Life of the Genome*, Sci. Am. **304**, (2011).

[67] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (2007).

[68] et al. Lodish H, Berk A, Zipursky SL, *Molecular Cell Biology 5th Ed* (2000).

[69] S. A. Grigoryev and C. L. Woodcock, *Chromatin Organization - The 30nm Fiber*, Experimental Cell Research.

[70] S. V. Razin and A. A. Gavrilov, *Chromatin without the 30-Nm Fiber Constrained Disorder Instead of Hierarchical Folding*, Epigenetics.

[71] C. L. Woodcock, *A Milestone in the Odyssey of Higher-Order Chromatin Structure*, Nature Structural and Molecular Biology.

[72] G. Felsenfeld and M. Groudine, *Controlling the Double Helix*, Nature.

[73] A. Jansen and K. J. Verstrepen, *Nucleosome Positioning in Saccharomyces Cerevisiae*, Microbiol. Mol. Biol. Rev. **75**, (2011).

[74] D. Carter, L. Chakalova, C. S. Osborne, Y. feng Dai, and P. Fraser, *Long-Range Chromatin Regulatory Interactions in Vivo*, Nat. Genet. **32**, (2002).

[75] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan, *Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation*, Cell **148**, (2012).

[76] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, *The Long-Range Interaction Landscape of Gene Promoters*, Nature **489**, (2012).

[77] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. De Laat, *Looping and Interaction between Hypersensitive Sites in the Active β-Globin Locus*, Mol. Cell **10**, (2002).

[78] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, *A Phase Separation Model for Transcriptional Control*, Cell.

[79] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, and R. A. Young, *Coactivator Condensation at Super-Enhancers Links Phase Separation and Gene*

Control, Science (80-.). **361**, eaar3958 (2018).

[80] R. Galupa and E. Heard, *X-Chromosome Inactivation: New Insights into Cis and Trans Regulation*, Current Opinion in Genetics and Development.

[81] W. K. Cho, J. H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, *Mediator and RNA Polymerase II Clusters Associate in Transcription-Dependent Condensates*, Science (80-.). **361**, 412 (2018).

[82] Y. Shin and C. P. Brangwynne, *Liquid Phase Condensation in Cell Physiology and Disease*, Science (80-.). **357**, eaaf4382 (2017).

[83] T. Jenuwein and C. D. Allis, *Translating the Histone Code*, Science.

[84] T. Misteli, *Chromosome Territories: The Arrangement of Chromosomes in the Nucleus.*, Nat. Educ. **1**, (2008).

[85] T. Cremer, C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn, *Rabl's Model of the Interphase Chromosome Arrangement Tested in Chinise Hamster Cells by Premature Chromosome Condensation and Laser-UV-Microbeam Experiments*, Hum. Genet. **60**, (1982).

[86] T. Cremer and C. Cremer, *Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells*, Nature Reviews Genetics.

[87] C. Lanctôt, T. Cheutin, M. Cremer, G. Cavalli, and T. Cremer, *Dynamic Genome Architecture in the Nuclear Space: Regulation of Gene Expression in Three Dimensions*, Nature Reviews Genetics.

[88] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping*, Cell (2014).

[89] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, *Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data*, Nature Reviews Genetics.

[90] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser, *Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure*, Nature **502**, 59 (2013).

[91] I. Jerkovic´ and G. Cavalli, *Understanding 3D Genome Organization by Multidisciplinary Methods*, Nature Reviews Molecular Cell Biology.

[92] T. H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, *Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C*, Cell **162**, (2015).

[93] T. H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, and X. Darzacq, *Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding*, Mol. Cell **78**, (2020).

[94] M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang,

HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture, Nat. Methods **13**, (2016).

[95] R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, and B. Ren, *Mapping of Long-Range Chromatin Interactions by Proximity Ligation-Assisted ChIP-Seq*, Cell Research.

[96] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe, and P. Fraser, *The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements*, Genome Res. **25**, (2015).

[97] S. Schoenfelder, R. Sugar, A. Dimond, B. M. Javierre, H. Armstrong, B. Mifsud, E. Dimitrova, L. Matheson, F. Tavares-Cadete, M. Furlan-Magaril, A. Segonds-Pichon, W. Jurkowski, S. W. Wingett, K. Tabbada, S. Andrews, B. Herman, E. Leproust, C. S. Osborne, H. Koseki, P. Fraser, N. M. Luscombe, and S. Elderkin, *Polycomb Repressive Complex PRC1 Spatially Constrains the Mouse Embryonic Stem Cell Genome*, Nat. Genet. **47**, (2015).

[98] B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. Leproust, G. A. Follows, P. Fraser, N. M. Luscombe, and C. S. Osborne, *Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C*, Nat. Genet. **47**, (2015).

[99] J. R. Hughes, N. Roberts, S. Mcgowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs, *Analysis of Hundreds of Cis-Regulatory Landscapes at High Resolution in a Single, High-Throughput Experiment*, Nat. Genet. **46**, (2014).

[100] J. O. J. Davies, J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs, and J. R. Hughes, *Multiplexed Analysis of Chromosome Conformation at Vastly Improved Sensitivity*, Nat. Methods **13**, (2015).

[101] M. V. Arrastia, J. W. Jachowicz, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, D. A. Selck, R. F. Ismagilov, and M. Guttman, *Single-Cell Measurement of Higher-Order 3D Genome Organization with ScSPRITE*, Nat. Biotechnol. (2021).

[102] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W. L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, *Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications*, Nature **538**, (2016).

[103] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, *Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome*, Cell.

[104] E. H. Finn and T. Misteli, *Molecular Basis and Biological Function of Variability in Spatial Genome Organization*, Science (80-.). **365**, eaaw9498 (2019).

[105] N. Krietenstein, S. Abraham, S. V. Venev, N. Abdennur, J. Gibcus, T. H. S. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. A. Mirny, J. Dekker, and O. J. Rando, *Ultrastructural Details of Mammalian*

Chromosome Architecture, Mol. Cell **78**, (2020).

[106] M. Schardin, T. Cremer, H. D. Hager, and M. Lang, *Specific Staining of Human Chromosomes in Chinese Hamster x Man Hybrid Cell Lines Demonstrates Interphase Chromosome Territories*, Hum. Genet. **71**, (1985).

[107] T. Cremer, C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders, *Analysis of Chromosome Positions in the Interphase Nucleus of Chinese Hamster Cells by Laser-UV-Microirradiation Experiments*, Hum. Genet. **62**, (1982).

[108] A. M. Chiariello, S. Bianco, A. Esposito, L. Fiorillo, M. Conte, E. Irani, F. Musella, A. Abraham, A. Prisco, and M. Nicodemi, *Physical Mechanisms of Chromatin Spatial Organization*, FEBS Journal.

[109] L. Fiorillo, S. Bianco, A. Esposito, M. Conte, R. Sciarretta, F. Musella, and A. M. Chiariello, *A Modern Challenge of Polymer Physics: Novel Ways to Study, Interpret, and Reconstruct Chromatin Structure*, Wiley Interdisciplinary Reviews: Computational Molecular Science.

[110] E. J. Banigan and L. A. Mirny, *Loop Extrusion: Theory Meets Single-Molecule Experiments*, Current Opinion in Cell Biology.

[111] J. H. Gibcus, K. Samejima, A. Goloborodko, I. Samejima, N. Naumova, J. Nuebler, M. T. Kanemaki, L. Xie, J. R. Paulson, W. C. Earnshaw, L. A. Mirny, and J. Dekker, *A Pathway for Mitotic Chromosome Formation*, Science (80-.). **359**, (2018).

[112] M. Ganji, I. A. Shaltiel, S. Bisht, E. Kim, A. Kalichava, C. H. Haering, and C. Dekker, *Real-Time Imaging of DNA Loop Extrusion by Condensin*, Science (80-.). **360**, (2018).

[113] Y. Kim, Z. Shi, H. Zhang, I. J. Finkelstein, and H. Yu, *Human Cohesin Compacts DNA by Loop Extrusion*, Science (80-.). **366**, (2019).

[114] M. Kong, E. E. Cutts, D. Pan, F. Beuron, T. Kaliyappan, C. Xue, E. P. Morris, A. Musacchio, A. Vannini, and E. C. Greene, *Human Condensin I and II Drive Extensive ATP-Dependent Compaction of Nucleosome-Bound DNA*, Mol. Cell **79**, (2020).

[115] Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, and Q. Wu, *CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function*, Cell **162**, (2015).

[116] E. de Wit, E. S. M. Vos, S. J. B. Holwerda, C. Valdes-Quezada, M. J. A. M. Verstegen, H. Teunissen, E. Splinter, P. J. Wijchers, P. H. L. Krijger, and W. de Laat, *CTCF Binding Polarity Determines Chromatin Looping*, Mol. Cell **60**, (2015).

[117] S. S. P. Rao, S. C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K. R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden, *Cohesin Loss Eliminates All Loop Domains*, Cell **171**, 305 (2017).

[118] W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny, F. Spitz, C. H Haering, L. Mirny, and F. Spitz, *Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal.*, Nature **551**,

51 (2017).

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

[120] F. S. Bates and G. H. Fredrickson, *Block Copolymer Thermodynamics: Theory and Experiment*, Annu. Rev. Phys. Chem. **41**, (1990).

[121] A. Y. Grosberg, A. R. Khokhlov, H. E. Stanley, A. J. Mallinckrodt, and S. McKay, *Statistical Physics of Macromolecules*, Comput. Phys. **9**, (1995).

[122] A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes, and R. A. Young, *Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains*, Cell **175**, 1842 (2018).

[123] M. Conte, L. Fiorillo, C. Annunziatella, A. Esposito, F. Musella, A. Abraham, S. Bianco, and A. M. Chiariello, *Dynamic and Equilibrium Properties of Finite-Size Polymer Models of Chromosome Folding*, Phys. Rev. E **104**, 054402 (2021).

[124] B. K. Kragesteen, M. Spielmann, C. Paliou, V. Heinrich, R. Schöpflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerković, I. Harabula, P. Guckelberger, M. Pechstein, L. Wittler, W. L. Chan, M. Franke, D. G. Lupiáñez, K. Kraft, B. Timmermann, M. Vingron, A. Visel, M. Nicodemi, S. Mundlos, and G. Andrey, *Dynamic 3D Chromatin Architecture Contributes to Enhancer Specificity and Limb Morphogenesis*, Nat. Genet. (2018).

[125] L. Fiorillo, S. Bianco, A. M. Chiariello, M. Barbieri, A. Esposito, C. Annunziatella, M. Conte, A. Corrado, A. Prisco, A. Pombo, and M. Nicodemi, *Inference of Chromosome 3D Structures from GAM Data by a Physics Computational Approach*, Methods (2019).

[126] A. Esposito, A. M. Chiariello, M. Conte, L. Fiorillo, F. Musella, R. Sciarretta, and S. Bianco, *Higher-Order Chromosome Structures Investigated by Polymer Physics in Cellular Morphogenesis and Differentiation*, Journal of Molecular Biology.

[127] K. Kremer and G. S. Grest, *Dynamics of Entangled Linear Polymer Melts: A Molecular-Dynamics Simulation*, J. Chem. Phys. (1990).

[128] Hamley I.W., *The Physics of Block Copolymers* (Oxford University Press, 1999).

[129] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamic* (Oxford Science Publications, Clarendon Press, Oxford, 1986).

[130] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by Simulated Annealing*, Science (80-.). **220**, (1983).

[131] G. Parisi, *Statistical Field Theory* (Westview Press, New York, 1998).

[132] S. Plimpton, *Fast Parallel Algorithms for Short-Range Molecular Dynamics*, J. Comput. Phys. (1995).

[133] M. Conte, A. Esposito, L. Fiorillo, R. Campanile, C. Annunziatella, A. Corrado, M. G. Chiariello, S. Bianco, and A. M. Chiariello, *Efficient Computational Implementation of Polymer Physics Models to Explore Chromatin Structure*, Int. J. Parallel, Emergent Distrib. Syst. (2019).

[134] C. Annunziatella, A. M. Chiariello, A. Esposito, S. Bianco, L. Fiorillo, and M. Nicodemi, *Molecular Dynamics Simulations of the Strings and Binders Switch Model of Chromatin*, Methods.

[135] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids (Oxford Science Publications) SE - Oxford Science Publications*, Oxford Univ. Press (1989).

[136] D. Noordermeer, M. Leleu, P. Schorderet, E. Joye, F. Chabaud, and D. Duboule, *Temporal Dynamics and Developmental Memory of 3D Chromatin Architecture at Hox Gene Loci*, Elife **2014**, (2014).

[137] G. Andrey, T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz, and D. Duboule, *A Switch between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs*, Science (80-.). **340**, (2013).

[138] T. Montavon, L. Thevenet, and D. Duboule, *Impact of Copy Number Variations (CNVs) on Long-Range Gene Regulation at the HoxD Locus*, Proc. Natl. Acad. Sci. U. S. A. **109**, (2012).

[139] B. Bonev, N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J. P. Hugnot, A. Tanay, and G. Cavalli, *Multiscale 3D Genome Rewiring during Mouse Neural Development*, Cell (2017).

[140] E. A. Feingold, P. J. Good, M. S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F. S. Collins, T. R. Gingeras, D. Kampa, E. A. Sekinger, J. Cheng, H. Hirsch, S. Ghosh, Z. Zhu, S. Patel, A. Piccolboni, A. Yang, H. Tammana, S. Bekiranov, P. Kapranov, R. Harrison, G. Church, K. Struhl, B. Ren, T. H. Kim, L. O. Barrera, C. Qu, S. van Calcar, R. Luna, C. K. Glass, M. G. Rosenfeld, R. Guigo, S. E. Antonarakis, E. Birney, M. Brent, L. Pachter, A. Reymond, E. T. Dermitzakis, C. Dewey, D. Keefe, F. Denoeud, J. Lagarde, J. Ashurst, T. Hubbard, J. J. Wesselink, R. Castelo, E. Eyras, R. M. Myers, A. Sidow, S. Batzoglou, N. D. Trinklein, S. J. Hartman, S. F. Aldred, E. Anton, D. I. Schroeder, S. S. Marticke, L. Nguyen, J. Schmutz, J. Grimwood, M. Dickson, G. M. Cooper, E. A. Stone, G. Asimenos, M. Brudno, A. Dutta, N. Karnani, C. M. Taylor, H. K. Kim, G. Robins, G. Stamatoyannopoulos, J. A. Stamatoyannopoulos, M. Dorschner, P. Sabo, M. Hawrylycz, R. Humbert, J. Wallace, M. Yu, P. A. Navas, M. McArthur, W. S. Noble, I. Dunham, C. M. Koch, R. M. Andrews, G. K. Clelland, S. Wilcox, J. C. Fowler, K. D. James, P. Groth, O. M. Dovey, P. D. Ellis, V. L. Wraight, A. J. Mungall, P. Dhami, H. Fiegler, C. F. Langford, N. P. Carter, D. Vetrie, M. Snyder, G. Euskirchen, A. E. Urban, U. Nagalakshmi, J. Rinn, G. Popescu, P. Bertone, S. Hartman, J. Rozowsky, O. Emanuelsson, T. Royce, S. Chung, M. Gerstein, Z. Lian, J. Lian, Y. Nakayama, S. Weissman, V. Stolc, W. Tongprasit, H. Sethi, S. Jones, M. Marra, H. Shin, J. Schein, M. Clamp, K. Lindblad-Toh, J. Chang, D. B. Jaffe, M. Kamal, E. S. Lander, T. S. Mikkelsen, J. Vinson, M. C. Zody, P. J. de Jong, K. Osoegawa, M. Nefedov, B. Zhu, A. D. Baxevanis, T. G. Wolfsberg, G. E. Crawford, J. Whittle, I. E. Holt, T. J. Vasicek, D. Zhou, S. Luo, E. D. Green, G. G. Bouffard, E. H. Margulies, M. E. Portnoy, N. F. Hansen, P. J. Thomas, J. C. McDowell, B. Maskeri, A. C. Young, J. R. Idol, R. W. Blakesley, G. Schuler, W. Miller, R. Hardison, L. Elnitski, P. Shah, S. L. Salzberg, M. Pertea, W. H. Majoros, D. Haussler, D. Thomas, K. R. Rosenbloom, H. Clawson, A. Siepel, W. J. Kent, Z. Weng, S. Jin, A. Halees, H. Burden, U. Karaoz, Y. Fu, Y. Yu, C. Ding, C. R. Cantor, R. E. Kingston, J. Dennis, R. D. Green, M. A. Singer, T. A. Richmond, J. E.

Norton, P. J. Farnham, M. J. Oberley, D. R. Inman, M. R. McCormick, H. Kim, C. L. Middle, M. C. Pirrung, X. D. Fu, Y. S. Kwon, Z. Ye, J. Dekker, T. M. Tabuchi, N. Gheldof, J. Dostie, and S. C. Harvey, *The ENCODE (ENCyclopedia of DNA Elements) Project*, Science.

[141] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, P. Kheradpour, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. A. Feingold, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. C. Giddings, T. R. Gingeras, R. Guigó, T. J. Hubbard, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, M. L. Eaton, A. Dobin, A. Tanzer, J. Lagarde, W. Lin, C. Xue, B. A. Williams, C. Zaleski, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, Y. Hayashizaki, A. Reymond, S. E. Antonarakis, G. J. Hannon, Y. Ruan, P. Carninci, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, L. L. Grasfeder, P. G. Giresi, A. Battenhouse, N. C. Sheffield, K. A. Showers, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, V. R. Iyer, M. Zheng, P. Wang, J. Gertz, J. Vielmetter, E. C. Partridge, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, M. A. Muratet, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, J. S. Newberry, S. E. Levy, D. M. Absher, W. H. Wong, M. J. Blow, A. Visel, L. A. Pennachio, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, C. Davidson, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, T. Hunt, I. Jungreis, M. Kay, E. Khurana, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, E. Tapanari, M. L. Tress, M. J. Van Baren, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, A. Valencia, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, D. Raha, L. Ramirez, B. Reed, M. Shi, T. Slifer, H. Witt, L. Wu,

X. Xu, K. K. Yan, X. Yang, K. Struhl, S. M. Weissman, L. O. Penalva, S. Karmakar, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, A. Victorsen, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, G. Jain, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, M. E. Sanchez, R. S. Sandstrom, A. O. Shafer, A. B. Stergachis, S. Thomas, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, K. Beal, A. Brazma, P. Flicek, N. Johnson, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, W. Miller, P. J. Bickel, B. Banfai, N. P. Boley, H. Huang, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, and L. Lochovsky, *An Integrated Encyclopedia of DNA Elements in the Human Genome*, Nature **489**, 57 (2012).

[142] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue, *3D Structures of Individual Mammalian Genomes Studied by Single-Cell Hi-C*, Nature **544**, 59 (2017).

[143] D. L. Theobald, *Rapid Calculation of RMSDs Using a Quaternion-Based Characteristic Polynomial*, Acta Crystallogr. Sect. A Found. Crystallogr. **61**, (2005).

[144] J. K. Ryu, C. Bouchoux, H. W. Liu, E. Kim, M. Minamino, R. de Groot, A. J. Katan, A. Bonato, D. Marenduzzo, D. Michieletto, F. Uhlmann, and C. Dekker, *Bridging-Induced Phase Separation Induced by Cohesin SMC Protein Complexes*, Sci. Adv. **7**, (2021).

[145] M. Baum, F. Erdel, M. Wachsmuth, and K. Rippe, *Retrieving the Intracellular Topology from Multi-Scale Protein Mobility Mapping in Living Cells*, Nat. Commun. **5**, 4494 (2014).

[146] D. Noordermeer, E. De Wit, P. Klous, H. Van De Werken, M. Simonis, M. Lopez-Jones, B. Eussen, A. De Klein, R. H. Singer, and W. De Laat, *Variegated Gene Expression Caused by Cell-Specific Long-Range DNA Interactions*, Nat. Cell Biol. **13**, (2011).

[147] T. Yang, F. Zhang, G. G. Yardımci, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li, *HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient*, Genome Res. **27**, (2017).

[148] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. Ritland Politz, J. Shendure, and S. Zhong, *The 4D Nucleome Project*, Nature.

[149] T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. Mendelson Cohen, S. Wingett, P. Fraser, and A. Tanay, *Cell-Cycle Dynamics of Chromosomal Organization at Single-Cell Resolution*, Nature **547**, 61 (2017).

[150] D. Lando, T. J. Stevens, S. Basu, and E. D. Laue, *Calculation of 3D Genome Structures for*

Comparison of Chromosome Conformation Capture Experiments with Microscopy: An Evaluation of Single-Cell Hi-C Protocols, Nucleus **9**, (2018).