Università degli Studi di Napoli "Federico II"

DOTTORATO DI RICERCA IN FISICA

Ciclo XXXIII
Coordinatore: Prof. Salvatore Capozziello

Polymer-physics modeling to explore single-cell DNA architecture and characterize experimental methods

Settore Scientifico Disciplinare FIS/02

Dottorando:
Luca Fiorillo

Tutore:
Prof. Mario Nicodemi

Anni 2018/2021

CONTENTS

INTRODUCTION

Chromosomes in cell nuclei are structured in very complex architectures, spanning different spatial scales. Advanced microscopy technologies [1–3] and innovative sequencing methods [4–9] have revealed that chromosomes in mammal cells exhibit a complex hierarchy of structural features, from loops between distal genomic sites in the kilobase (kb) range [10] to globular domains of contacts at the megabase (Mb) scale (Topologically Associated Domains or TADs [11, 12]) and to higher-order interactions between neighboring TADs [13]. At the supermegabase scale, a checkerboard pattern of contacts emerges, identified as the organization of DNA in A/B compartments [7] and, zooming out at the genome wide level, the segregation of chromosomes in *territories* appear [14]. Even more, nuclear organelles as the nucleolus or the nuclear speckles or the nuclear lamina collect DNA filaments to form hubs of interactions [9, 15]. Notably, such complicated landscape of contacts and spatial proximities is crucially linked to the functionality of genome [16–22], because, for instance, DNA elements as genes and enhancers must be close in space to express their functions. Indeed, disruption of any of the mentioned spatial features can be associated to serious diseases [23–25]. That elucidates why the architecture of chromosomes and, especially, the molecular mechanisms driving their folding are object of intense research. In this context, polymer-physics models have been widely used to make sense of the experimental observations, providing useful insights into the processes which may generate the spatial arrangement of DNA, from the scale of few hundreds of base pairs (bp) up to whole nuclear scale. Many complementary or alternative models of chromosomes have so far been conceived [26–31], each proposing a possible driver of DNA tridimensional (3D) organization in the crowded nuclear environment. However, we are still far from an accepted omni comprehensive picture.

The present work of thesis is framed in such a dynamic context of research. Specifically, we focus on the role that polymer-physics plays in the field by discussing two relevant applications of polymer models.

First, we will show how key architectural features of DNA in mammal cells, the TADs [11, 12], can be explained by the phase-separation mechanism of classical polymers in the globule state [27]. Indeed, for a DNA region of human cells, we will illustrate that the stationary structures of phase-separated polymers can recapitulate not only its average architecture but also the variability of that architecture across cells, as observed by microscopy [32]. Additionally, the effects of cohesin depletion on DNA spatial conformations [32, 33] will be successfully described by a mechanism of phase reversal, i.e. by polymers switching from the globule phase-separated state to the coil thermodynamic state [27]. Overall, that will frame DNA architecture at the Mb scale in a steady-state scenario, where weak biochemical interactions (few k_BT) drive the folding of chromosomes through diffusion of molecules in a viscous bath. We will discuss how such scenario relates to the off-equilibrium model of chromosomes named *Loop Extrusion* [29, 34, 35], which has achieved much popularity in the field.

Next, we will present a different application of polymer models of DNA. We will show that known 3D conformations of model polymers can be used as simplified, yet fully controlled reference system to benchmark the performances of Hi-C [7], SPRITE [9] and GAM [8] technologies [36]. They are all powerful experimental methods designed to probe the architecture of DNA in nuclei. However, a clear assessment of their absolute and relative performances is missing, as they provide different measures of DNA spatial organization. Are all three technologies capturing faithfully the underlying conformations of chromosomes? Are they detecting different aspects of DNA architecture? Which

technology is less noisy when few cells are available? What is the impact of the experimental detection efficiency? What is the amount of noise affecting the outputs of those technologies for specific experimental conditions? We will show that these questions can be tackled by simulating Hi-C, SPRITE and GAM experiments on ensembles of fully known polymer structures, providing the first rigorous comparison of their characteristics, albeit simplified. The results and analyses we are going to illustrate provide insights on such powerful technologies and, eventually, may be of help in designing novel experiments.

All such studies and investigations were conducted in the last three years in the group of prof. Mario Nicodemi, in the Department of Physics of Università degli Studi di Napoli Federico II. Much of this work is published or under review process.

In **Chapter 1** we will provide basic knowledge of DNA biology, necessary for the comprehension of the present work of thesis. We will review the architectural features of chromosomes discovered by the latest experiments and briefly sketch the technologies that made possible those discoveries. We will focus, in particular, on the Topologically Associated Domains (TADs), as they are the main architectures investigated in **Chapter 2**. We will report how they were firstly observed and discuss important findings about their variability in cells. We will also offer an overview of the Loop Extrusion model, a polymer-physics model which can successfully account for the formation of TADs and other key experimental observations of DNA architecture. In **Chapter 2**, we will present the *Strings&Binders Switch* (SBS) model of DNA [37, 38] and the Molecular Dynamics approach used to extract the steady-state configurations of SBS polymers [39]. Then, we will show that such model can successfully explain the generation of TADs, their cell-to-cell variability and the effects of cohesin depletion based on the phase-separation mechanism and on the coil-globule phase transition of classical polymers, as said above. That is the content of a recently published work from our group [27]. Finally, **Chapter 3** is dedicated to the benchmark of the Hi-C, SPRITE and GAM technologies against known conformations of SBS polymers. We will review those experimental methods in detail and describe how they were modelled for implementation on polymer models. Then, we will perform Hi-C, SPRITE and GAM "computational experiments" and compare their detections with the known polymer 3D conformations. We will study the number of cells needed in an experiment to have statistically reproducible results and will assess the noise level at various experimental conditions. We will see that Hi-C, SPRITE and GAM have important differences and will thus clarify which method is the most suited for a given experimental target. The content of this chapter is deposited on BiorXiv [36] and is currently under review at *Nature Methods*.

1 THE ARCHITECTURE OF DNA IN CELL NUCLEI

The present chapter provides the reader with the background knowledge necessary for the comprehension of the thesis. All the topics will be reviewed and presented without wealth of details, which are referred to other specialized works.

In the first paragraph we will summarize basic concepts of DNA biology and illustrate why its 3D conformations in cell nuclei are object of intense research. In the second and third paragraphs we will review what is currently known of the spatial organization of chromosomes and will briefly describe the technologies allowing the investigation of such organization. In the fourth paragraph we will provide the definition of epigenetics and its relevance in the study of DNA spatial conformations. Next, we will focus on a specific architectural feature of mammal chromosomes, the *Topologically Associated Domains* (TADs)[11, 12], globular formations into which DNA is arranged all over the genome [11, 12]. We will illustrate their characteristics and describe a popular polymer-physics model (the *Loop Extrusion* model) used to explain their generation [29, 34, 35]. Finally, in the last paragraph, a recent microscopy experiment [32] which unveiled striking features of TADs in human cells will be reviewed, so paving the way to the content of **Chapter 2**, where a polymer-physics model alternative to Loop Extrusion will be shown to recapitulate the findings of such microscopy investigation [27].

1.1 DNA architecture in cell nuclei is functional and dynamic

In eukaryotic cells, the genetic information is encrypted in DNA (deoxyribonucleic acid) molecules. A DNA molecule consists in a pair of polynucleotide strands winded up in the shape of a double helix. The nucleotides making up each strand of the helix are composed of a nitrogen base and a phosphate group attached to the deoxyribose sugar, so that the sugars and phosphates form the backbone of a strand while the nitrogen bases project outwardly. Hence, the bases of the two polynucleotides bind to each other with interaction energy $\sim k_B T$ ($T \sim 300K$ in mammals), holding the double helix structure. There are four types of nitrogen bases in DNA, adenine (A), guanine (G), cytosine (C) and thymine (T), whose attractive interaction is highly specific, such that, across the two DNA strands, A can be paired only to C and G to T. That is commonly referred to as the *complementarity of the DNA strands.* The sequence of complementary base pairs along the DNA molecules of a cell constitutes its *genetic code*, that is the instructions to build all the compounds (proteins or RNA) necessary for the cell to live and act. A portion of the genetic sequence encoding for a specific protein or other compound is called *gene*.

Due to its vital role, in eukaryotic cells DNA is safely stored in the nucleus, where specialized proteins transcribe the genetic code so that the transcript may be used to construct functional molecules in specific sites outside the nucleus. That is known as the *transcriptional activity* of the genome. As we will describe in a moment, both (i) the confinement in the nucleus and (ii) the transcriptional activity are strongly linked to the way DNA molecules are spatially organized.

(i) The confinement in the nucleus compels DNA to an extraordinary folding process. To give a quantitative sense of that, the number of base pairs (bp) necessary to encrypt the human genetic code is 6.4×10^9, which corresponds to a 2m long DNA molecule constrained to pack in nuclei whose average length is 9-15μm. That requires a highly developed packing mechanism, which up to now is still unclear.

(ii) Albeit tightly folded in the nucleus, DNA strands must remain accessible for transcription factors to read the genetic code. Importantly, not all the genes of the genome need to be expressed (i.e. transcribed to produce a protein) altogether during the life of a cell, because some may be required only at specific moments or external conditions. Additionally, in multicellular organisms each cell is specialized in tissue-specific functions, hence only the subset of genes governing those functions must be expressed. Only in recent times it has become clear that such sharp regulation of cellular transcription is connected to the spatial organization of DNA in the nuclei [16–18]. A key mechanism illustrating that connection is the *promoter-enhancer* interaction [40–42], which works as follows. For a gene to be transcribed, the RNA-polymerase protein must anchor onto the gene *promoter*, i.e. a sequence of base pairs upstream of the gene which forms a stable binding site for the RNA-polymerase. Transcription factors (TFs) can, however, also attach to the promoter, facilitating or preventing the binding of the RNA-polymerase and so regulating the transcription. Hence, to enhance the expression of a gene, TFs facilitating the RNA-polymerase must get physically close to the promoter and that is accomplished by the *enhancers*. They are segments of DNA which bind to the TFs and carry them near the target promoter by looping. Since in mammal cells enhancers can be hundreds or thousands of bp away from their target promoters, the mere promoter-enhancer mechanism of regulation can orchestrate a complex network of loops and contacts all over the genome, shaping DNA 3D architecture. Similarly, other looping mechanisms exist, for instance between enhancers and *silencers*, which recruit the enhancers preventing them from activating the target promoters; or between *insulators*, DNA sequences where structural proteins anchor and bind to each other to insulate a loop of DNA from the rest. All these examples illustrate that DNA spatial organization in nuclei is highly functional and, also, far from static, as the expression of genes is enhanced or repressed according to the cellular needs.

From (i) and (ii) we get a picture where DNA molecules condense tightly in cell nuclei and nonetheless form complex patterns of contacts between DNA sites to regulate transcription. Such pattern is not rigid and can be rearranged in response to external or internal stimuli. In the next paragraph we will give an overview of the levels of DNA spatial organization in mammal cells, many of which are still openly debated. Specifically, we will discuss the architecture of DNA in the *interphase* period of the cells, i.e. the phase of full activity before duplication begins.

1.2 DNA architecture consists in several layers of organization

In nuclei, DNA is organized in many distinct molecules, called *chromosomes*. For example, human cells have 46 chromosomes, whereas murine cells 42. The folding of chromosomes in nuclei is orchestrated by numerous structural proteins binding to DNA. The collection of DNA filaments and bound structural proteins is known as *chromatin*.

The very first level of DNA folding involves the histone proteins. Along all the genome, eight histone molecules (two molecules each of histones H2A, H2B, H3 and H4) assemble and form a complex where DNA is wound up to the extent of about 150bp. The concentration of the histonic complexes along the genome is such to leave few loosen DNA, i.e. few bp to 80bp between each consecutive pair of histone blocks. The structure made of a histonic complex, the DNA rolled up on it and the residual loosen DNA is named *nucleosome* and constitutes the basic organizational unit of chromatin [43]. The thickness of nucleosomes is about 11nm and they provide a reduction of DNA linear length about 7-fold. At this stage, chromosomes are said to be in the *beads-on-a-string* structure, because nucleosomes appear as beads along the string of loosen linker DNA. The next step of folding is

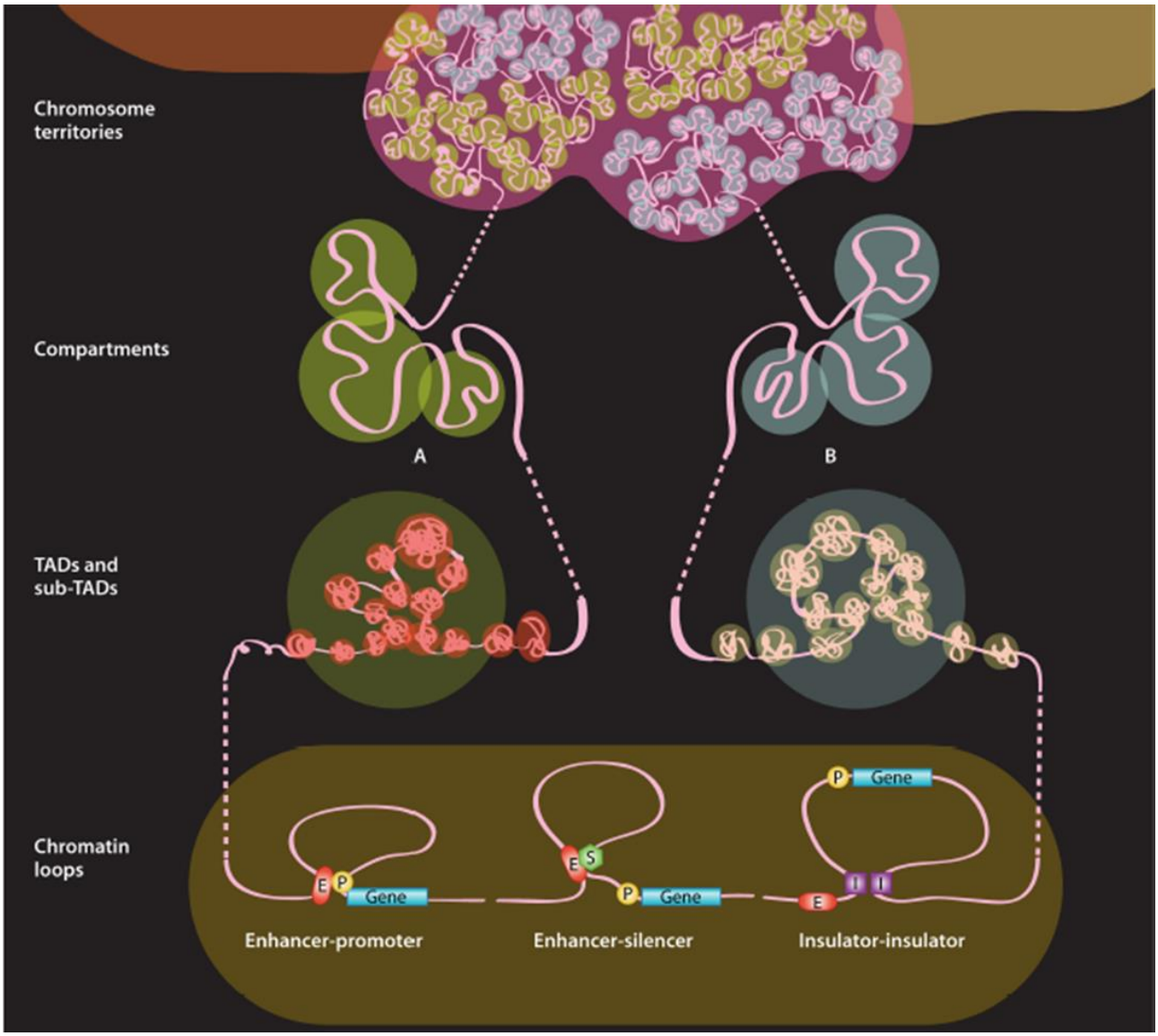


Figure 1.1*:* Cartoon of the main architectural features of chromatin at different scales, based on the latest experimental observations. From top to bottom, the length scale gets smaller. Chromosomes in nuclei are segregated in territories (top) [14], then, at the supermegabase scale, each chromosome is structured in insulated domains of interactions, known as A and B compartments [7]. At the megabase and submegabase scale, globular domains of contacts form, named Topologically Associated Domains (TADs) [11, 12]. Inside TADs, at the kb scale, functional contacts can take place in the form of loops, e.g. between enhancers and gene promoters, enhancers and insulators or between insulators (see **Main Text**). Taken from [44].

implemented by another type of histone, the H1, which collect the loosen DNA and wind it up to pack nucleosomes onto each other, making chromatin a fiber about 30nm thick. That is known as the *30nm fiber* structure. Importantly, the formation of such architecture has only been observed in *in-vitro* experiments [45], i.e. in laboratorial controlled conditions, and evidences have been collected that the 30nm fiber could be just a simplified model of what happens in real nuclei [46, 47]. In this sense, the existence of the 30nm fiber is not free of doubts. From there, uncertainties grow deeper, as the mechanisms enabling the successive steps of packing are still largely unclear.
As stated, it is known that chromatin engages an intricate nest of functional loops between distal DNA sites (e.g. promoter-enhancer), spanning kilobase (kb) ranges [10]. At the megabase (Mb) scale,

chromatin forms globular domains of enriched interactions named *Topologically Associated Domains* (TADs) [11, 12] and, additionally, domains of contacts with the nuclear lamina known as *Lamina Associated Domains* (LADs) [15]. Recent experiments indicate finer structures nested inside TADs, as *microTADs* or *subTADs* [48] and further interaction domains with elongated geometries, like *stripes* or *hairpins* [49, 50]. Above the Mb scale, TADs can interact with each other, spanning whole chromosomal ranges and creating huge clusters of interactions, the metaTADs [13]. On the same scale, chromosomes are organized in *A* and *B compartments* [7]. B compartments have been associated to domains of strongly condensed chromatin, which is overall inaccessible to transcription factors, while A compartments have shown high transcriptional activity [7]. Finally, chromosomes in nuclei segregate in almost spherical volumes (1-2μm effective radius) called *territories* [14]. Importantly, although the segregation in territories, contacts and loops across different neighboring chromosomes are also observed [51]. The picture emerging from those experimental findings (**Figure 1.1**) is that chromatin architecture presents several complex layers of organization, ranging from the few hundreds bp of nucleosomes to the megabase size of TADs and metaTADs.

Disruptions and alterations of any of those layers of organization result connected to diseases like congenital disorders or cancer [23–25], showing the strict bondage with genome functionality. That also illustrates why understanding the physical mechanisms driving each level of folding is a crucial task for contemporary biophysics. Indeed, many models of folding have been proposed for all scales [52–54]: quantum mechanics pictures are used to describe nucleotides interactions [55, 56], molecular mechanics approaches tackle the debate on the 30nm fiber [55, 57–59] and polymer-physics models are interrogated to explain the DNA architecture from the kb to the Mb scales [26–31, 35, 38, 60, 61], where torsional and confinement effects also become relevant [62]. Two of those polymer-physics models will be presented in much more details in the next chapter.

The wealth of observations summarized above on DNA 3D structure was obtained thanks to great technological advancements in experimental biology. In the following paragraph, we will briefly review the technologies that played and are still playing a key role in the dissection of chromatin architecture.

1.3 Technologies used to probe DNA architecture in nuclei

Data on DNA spatial organization are extracted by two broad categories of technologies: light microscopy methods and sequencing tools. The former directly observe physical distances between DNA sites in nuclei, while the latter detect the frequencies of contacts (or their analogues) between DNA segments and assume those are inversely proportional to their spatial distances in nuclei. In what follows, we are going to look closer both such categories of experiments.

1.3.1 Microscopy technologies

Microscopy methods were the first employed in the exploration of chromatin architecture and, specifically, the *fluorescence in situ hybridization* (FISH) technique played a pioneering role [63]. It relies upon probe sequences of DNA which bind to the target DNA sites in nuclei. The probe is endowed with a fluorochrome (or is made fluorescent by the action of cellular enzymes) and so, using a fluorescence microscope, the position of the target DNA can be detected. The size of the probe determines the bp resolution of FISH experiments, e.g. a 10kb probe size prevents from distinguishing DNA sites less than 10kb apart along chromosomes. In addition, the light diffraction

limit constrains the resolution of the Euclidean distances. FISH experiments have been realized at mega and kilobase resolutions, even if some were pushed to the scale of few nucleotides [64, 65]. As for the light diffraction limit, very recent advancements allowed resolutions at the nanometer scale (in this case, the expression *super-resolution microscopy* is commonly used) [2, 32].
FISH approaches have been and are still used to measure distances between DNA segments (e.g. promoter and enhancer, [66–68]); to investigate the location of a DNA site in relation to its chromosome [69]; to study the stability of TADs across cell populations [32] and to visualize the positions of chromosomes in the nucleus [14, 70]. Indeed, microscopy experiments led to the observation of chromosome territories (see **Figure 1.2**).

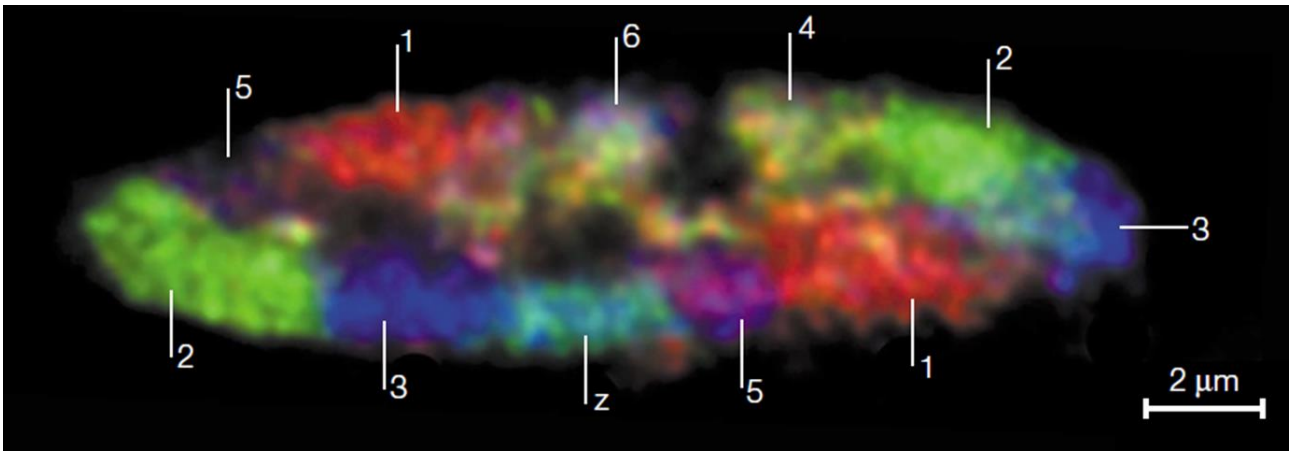


Figure 1.2: Experimental evidence of chromosome territories. Light optical section of a chicken fibroblast nucleus extracted from a FISH experiment. Different chromosomes, marked with diverse colors, clearly segregate in mutually exclusive territories. Taken from [14].

Besides the resolution limits discussed before, the main limitation of microscopy technologies is that, at kb resolutions, they permit the investigation only of a restricted region of chromatin (few Mb). Indeed, probing all the genome at the kb (or lower) scale and visualizing it at microscope is still too cumbersome, albeit promising advancements have been made [1, 3]. That prompted the invention of sequencing-based methods, as the *3C-technologies*.

1.3.2 Sequencing-based technologies

Sequencing a DNA segment means reading the sequence of base pairs that makes it up. Human and murine genomes have been completely sequenced, so the sequence of bases composing the genetic code in humans and mice is known. Hence, for instance, when a human DNA segment is sequenced, its sequence of bases is compared against that of the whole genome and the identity of the DNA region is discovered.
The first technology devised to study DNA 3D organization and based on sequencing was called *3C* (Capturing Chromosome Conformations, [4]). From that, several variants were conceived (4C [5], 5C [6], Hi-C [7] and others) which are all referred to as *3C-technologies*. Here, we will only describe the 3C-technology named Hi-C, since it has gained great popularity in the field [7]. Further details on the Hi-C method will also be given in **Chapter 3**.

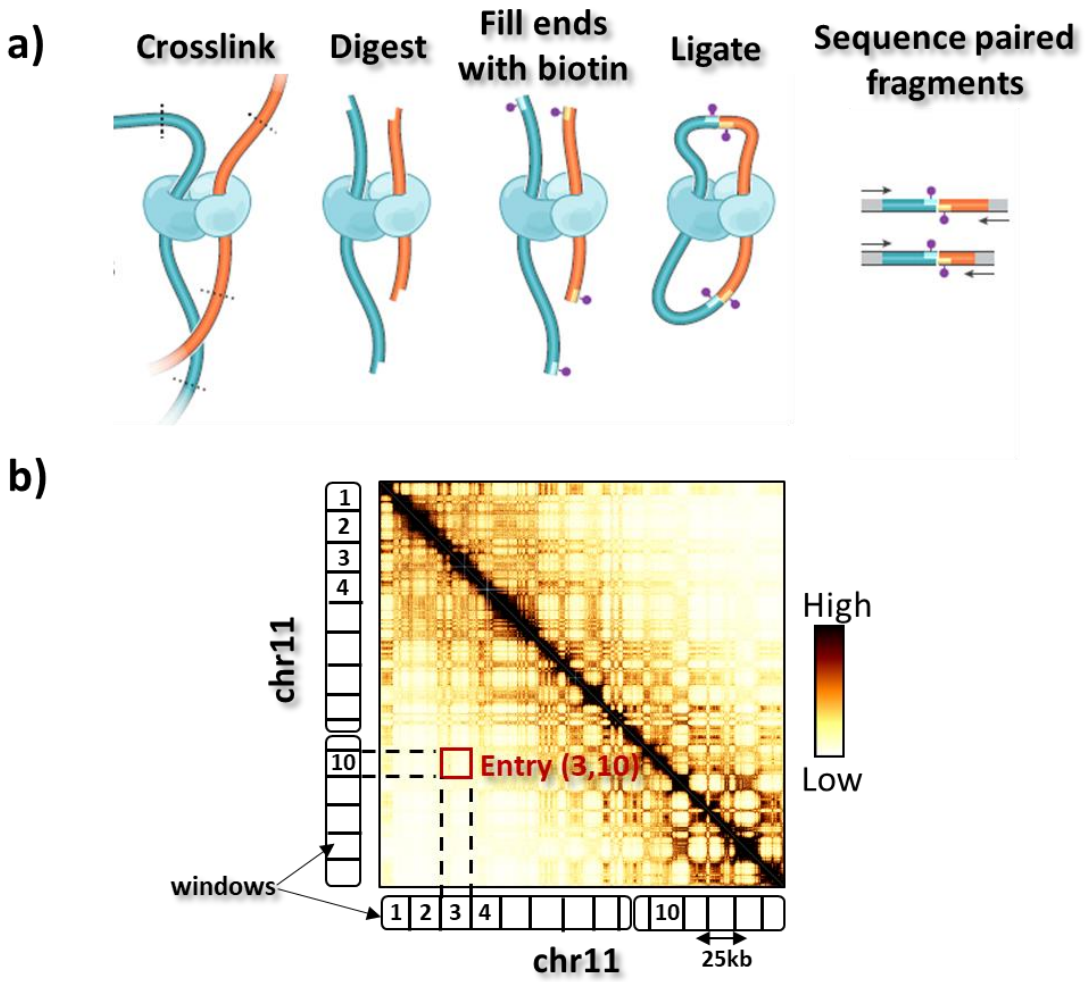


Figure 1.3: a) Scheme of the Hi-C protocol [7]. Chromatin from a population of nuclei is crosslinked with formaldehyde, so to freeze the pattern of contacts. Then DNA is cut into fragments by restriction enzymes and the loose ends of those fragments are filled with biotin. Pairs of biotinylated fragments are ligated, forming hybrid molecules (ligation products) of DNA fragments which were in contact in their nucleus. Fragments of each ligation product are sequenced and called as in contact, so that in the end the number of contacts across all the nuclei population is returned. Adapted from [10].
b) The contacts detected in a Hi-C experiment are typically arranged in a 2-dimensional matrix, where each entry (i, j) displays the measured number of contacts between the DNA windows i and j in a population of cells. Such contact matrix is usually visualized as a heatmap. Here, as example, it is shown the heatmap of the Hi-C contact matrix for the chromosome 11, from murine staminal cells (data from [71]). The resolution is 25kb, i.e. chromosome 11 is organized into windows 25kb long.

In Hi-C experiments (**Figure 1.3a**), a population of cell nuclei is treated with formaldehyde so to cross-link DNA segments which are spatially close (10-100nm distance range). Specifically, formaldehyde molecules bind together DNA sites which are close enough to be in physical interaction, i.e. are *in contact* with each other. The bond is covalent, thus formaldehyde freezes the pattern of contacts present in the nuclei population at a certain time. Then, nuclei membranes are disrupted and chromatin is cut into fragments (i.e. *digested*) by specific proteic machineries called restriction enzymes. In this way, chromatin is reduced to a set of crosslinked DNA fragments. The loose ends of the fragments are filled with biotinylated nucleotides; hence, the DNA-ligase enzyme is deployed to ligate the biotinylated ends of pairs of crosslinked fragments. That generates hybrid

molecules (*ligation products*) made of ligated pairs of DNA segments which were in contact in the original nucleus. Finally, formaldehyde is washed away, and all the ligated pairs of fragments are sequenced. Since each ligation product identifies a contact, the output of a Hi-C experiment is the number of contacts detected for all pairs of fragments across the population of nuclei and, thereby, the frequencies of contacts. Such data are typically organized in a *contact matrix* C, whose entries (i, j) indicate the number of contacts between the DNA segments i and j (**Figure 1.3b**). To this aim, the genomic sequence is divided in *windows* of equal length and contacts are assigned to each possible pairs of DNA windows. The bp size of the windows defines the resolution of the Hi-C dataset. In principle, the contact data can be arranged at any resolution above the average length of the digestion fragments, which ranges from few hundreds of bp to units of kb depending on the restriction enzyme used [72]. Up to now, the highest resolution reached in a standard Hi-C experiment is the order of units of kb [71], while the typical resolutions are in the range of tens of kb. Anyway, to facilitate their visualization, contact matrices are usually plotted as heatmaps (**Figure 1.3b**).

The pattern of Hi-C contacts across the whole genome gives indirect information on the spatial organization of DNA because contacting windows are necessarily close in space. That revolutionized the exploration of chromatin architecture, allowing for the collection of an impressive amount of contact data for different cell types and organisms. The analysis of Hi-C contact matrices led to the observation of TADs [11], A/B compartments [7], loops and stripes [34, 49]. Hi-C is currently widely employed in the field and further variants or improvements of the protocol are continuously proposed [10, 72–75]. Of these, we only mention the *single-cell Hi-C*, i.e. variants of the Hi-C protocol which are applied to a single cell nucleus rather than on an entire population [72, 75–77]. Interestingly, these methods return the pattern of contacts for a single specific cell, avoiding the averaging over a population of nuclei intrinsic to the original Hi-C design [7]. In the following, when we use the term Hi-C we will generally refer to an experiment conducted over a population of cells, while if we need marking the difference with the single-cell variant, we will explicitly state *ensemble*/*average/bulk* Hi-C as opposed to *single-cell* Hi-C.

As every technology, Hi-C has limitations. Among those, we stress that ligation can only detect pairwise contacts, while multiple contacts between many DNA windows could also occur. To overcome such limitation, along with an alternative 3C method [78], two ligation-free technologies were recently invented, the Genome Architecture Mapping (GAM) method [8] and the Split-Pool Recognition of Interactions by Tag Extension procedure (SPRITE) [9]. Based on sequencing as Hi-C, nonetheless they do not use ligation as strategy to detect contacts and can, in principle, yield 3-wise, 4-wise etc contact patterns genome wide. In **Chapter 3** we will describe the details of the GAM and SPRITE procedures, in relation to a study we conducted on their performances [36].

When new chromatin structural features are discovered by any of the reported technologies, their functionality is interrogated. In the next paragraph, we will give few details on how the functionality of DNA structures can be assessed, introducing the concept of epigenetic.

1.4 Epigenetics and the functionality of DNA structures

As stated at the beginning of this thesis, the promoter-enhancer interactions can inform the generation of DNA loops to enable transcription. So, studying the enrichment of promoter-enhancer pairs is a possible approach to identify chromatin structures with a transcriptionally active function.

More generally, the functional role of chromatin organizational units can be investigated studying their *epigenetic signature*.

Epigenetics is the study of phenotype alterations which are not caused by modifications of the genetic code and, nonetheless, are heritable. Indeed, given the genomic sequence of bases, the expression of the genes can be altered (and thus the phenotype affected) by chemical modifications of the DNA molecules which, importantly, can be transmitted from parental to children cells. Methylation, i.e. the addition of a methyl group, is a common example of epigenetic modification of DNA. Specifically, methylations of cytosine at CpG sites (regions of DNA where cytosine is followed by guanine) are often observed at gene promoters with repressive function. Also, the aminoacids composing histone tails can undergo chemical alterations, as methylation, acetylation, phosphorylation, sumoylation, ADP-ribosylation or ubiquitinylation [79–81]. All these modifications exhibit an active or repressive function for gene transcription, as they affect the DNA chemical accessibility or recruit proteic factors which, in turn, favor or block transcriptional processes.

The chemical state of chromatin thus defines its *epigenetic signature* and may provide strong indication that a region of DNA is actively expressed or repressed. In this sense, studying the epigenetic signature of a chromatin architecture can help elucidating its functional meaning. For instance, the A compartments are abundant of genes and rich of histone modifications associated to active chromatin [7], which brought the idea of A compartments as hubs of active chromatin. Conversely, B compartments were found to correlate with repressive histone modifications, suggesting they are collectors of inactive DNA [7].

All in all, epigenetic modifications and the spatial organization of DNA are connected to each other because they both contribute to regulate transcription. Epigenetic signatures can help understanding the function of chromatin structural units, albeit the exact interplay between epigenome, architecture and transcription, i.e. which one informs the other, is still unclear.

1.5 Topologically Associated Domains (TADs)

In this section we focus on the DNA structures named Topologically Associated Domains, or, shortly, TADs. We will recapitulate their characteristics and review a model recently proposed to explain their formation, so to ensure the comprehension of **Chapter 2**.

1.5.1 Experimental findings

In mammals, TADs represent the main organizational unit of chromatin at the submega and megabase scales. They were first observed through Hi-C and 5C experiments [11, 12], but their existence was also confirmed by FISH observations [12, 32]. Examining Hi-C contact matrices, TADs are visible as square domains of high contact frequencies, i.e. regions with much more abundant contacts than the surroundings (**Figure 1.4**). They indicate groups of DNA windows which tend to contact more among each other than with external sites. From **Figure 1.4**, it can be seen that TADs typically follow each other closely along the main diagonal of the Hi-C matrices, suggesting they are a widespread organizational feature of the genome. Moreover, it was found that TADs can exhibit further subdomains inside them, which are generally called subTADs [82] (**Figure 1.4**). At microscopy observations [32], TADs appear in space as globular insulated domains.

Several algorithms have been elaborated to identify TADs in Hi-C datasets [10, 11, 13], and all of them rely on the detection of the boundaries of TADs, defined as genomic positions where sharp changes in the average contact frequencies take place. Regardless of the specific routine used, TAD

calling showed that about 2000 domains characterize murine and human genomes, covering all the chromosomes. That confirmed TADs as the dominant form of DNA organization at the submega and megabase scale.

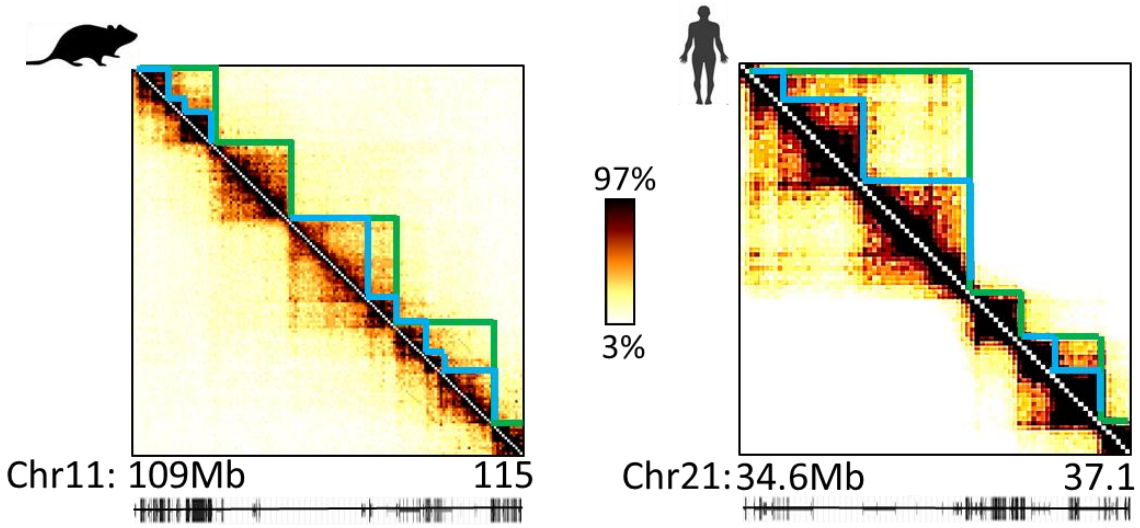


Figure 1.4: Two examples of Hi-C contact matrices exhibiting TADs and subTADs structures. On the left, the contact matrix for a 6Mb-long murine locus of chromosome 11, from staminal cells at 40kb resolution [11]. Below, the list of genes populating the locus are shown (data from UCSC Genome Browser). On the right, the heatmap for a 2.5Mb-long human locus of chromosome 21, from the HCT116 cell line at 30kb resolution [33]. Below, the list of genes is displayed (UCSC Genome Browser). In both matrices, squares of enriched contact numbers are clearly visible along the diagonal (TADs), with inner smaller squares of further enrichments (subTADs), signaling that chromatin orchestrates globules of contacts in space. In the heatmaps, the boundaries of TADs and subTADs are qualitatively marked with green and light blue, respectively. Colorbar indicates the percentiles of the heatmaps.

TADs are enriched both for active and repressive histone modifications and genes present inside them are typically expressed together [12], that suggests TADs can collect DNA sites in need to be proximal both for repressive and active transcriptional scopes. Interestingly, TAD boundaries tend to be rich of binding sites for CTCF and cohesin, two proteins known to exert relevant architectural functions [44]. Pairs of CTCF proteins, for instance, can mediate the formation of contacts between distal DNA sites and, importantly, that seems to occur preferentially when the two CTCF binding sites are in convergent orientation along the genome, i.e. when one binding base sequence is forward and the other is reversed [10]. Strikingly, removal of CTCF binding sites at the boundary between two TADs causes their fusion and so, generally, rewires the pattern of contacts with dramatic consequences on gene expression [83]. Similarly, the depletion of cohesin over regions of chromatin was observed to wipe out the TADs [32, 33]. Overall, such findings strongly suggest that CTCF and cohesin play a decisive role in the formation of domains and, in the next paragraph, we will see how this is accounted for in the Loop Extrusion model of chromatin.

1.5.2 The *Loop Extrusion* as model of TAD formation

The mechanisms which could generate TADs are widely discussed in the scientific community, because of their importance. Recently, the molecular mechanism proposed in the *Loop Extrusion*

model (a polymer-physics model of DNA) gained great popularity and, as such, will be briefly described here.

Originally thought to explain mitotic compaction of chromosomes [84, 85], the *Loop extrusion* (LE) model of chromatin is conceived to tackle DNA organization from the kb scale or above [29, 34, 35, 86]. In that range, a DNA-tracking mechanism is assumed to oversee the architecture: molecular complexes called *loop extruding factors* (LEFs) are supposed to bind to chromatin and form progressively expanding loops by extruding the DNA sequence (**Fig 1.5**). A LEF is pictured as two connected molecular motors which slide along chromatin in opposite directions, gradually extruding DNA (**Fig 1.5**). The process is thought as driven by active energy consumption, i.e. LEFs are pumped by ATP combustion. Formed loops disappear when LEFs detach from chromatin due to thermic fluctuations in the nucleoplasm. Hence, births and deaths of loops drive chromatin into a stationary off equilibrium state, where nested loops or neighboring loops should determine the pattern of contacts observed, for instance, in Hi-C matrices.

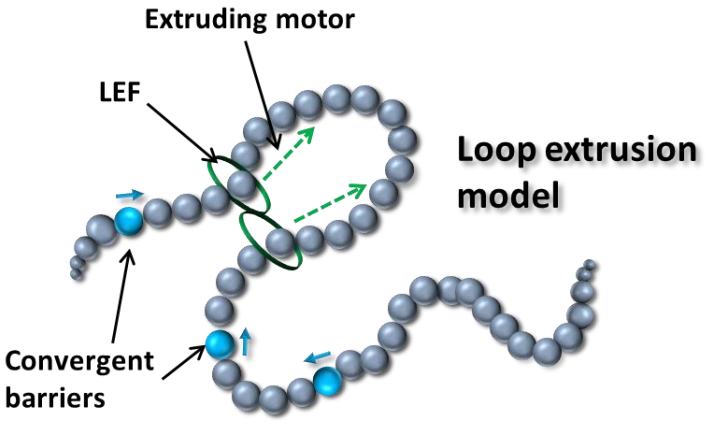


Figure 1.5: Cartoon of the Loop Extrusion (LE) model of chromatin [29, 34, 35]. Two-ringed proteic machineries, the Loop Extruding Factors (LEFs), are assumed to progressively extrude chromatin, pumped by ATP-consuming motors. That shapes DNA filaments in loops. LEFs stop their activity upon encountering specific convergent barriers, i.e. DNA sites made of reversed base sequences. That limits the expansion of loops along chromatin, together with the spontaneous detachment of LEFs due to thermic fluctuations. Adapted from [52].

Specifically, in computer simulations chromatin is depicted as a self-avoiding polymer chain with a weak aspecific self-interaction (few $k_B T$) to account for generic attractive forces between DNA segments [53]. Consecutive beads of the polymer are held together by a harmonic potential, self-avoidance is realized with a truncated, shifted Lennard-Jones force and the self-interaction is also implemented with a truncated, attractive Lennard-Jones potential. LEFs are modelled as diffusive particles able to anchor on the polymer and move along it for a fixed average permanence time, τ. Beads extruded by a LEF are bound by harmonic or Lennard-Jones potentials so to generate stable loops.

The dynamics and stationarity of such system is controlled by two parameters, the processivity of a LEF (λ) and the average separation of LEFs along the polymer (d). The processivity indicates the mean size of a loop extruded by a LEF during the average permanence time τ, if the LEF sliding is not obstructed. Calling v the velocity whereby LEFs slide over the polymer, $\lambda = v\tau$. For N LEFs attached to the polymer, and if L is the polymer length, then $d = L/N$. The λ/d ratio determines the type of

steady state of the LE model. If $\lambda/d \gg 1$, loops are tightly packed along the polymer, they block each other from expanding and nested structures are very common, where multiple LEFs reinforce the same loop. Stationarity is easily reached as deaths of loops are rapidly compensated by the creation of new ones and, since loops constrain each other, the final average size of a loop is less than λ. This is named the *dense regime* of the LE model. If $\lambda/d \ll 1$, then loops are quite far apart from each other and stationarity is reached with average loop size equal to λ. This is the *sparse regime* of the LE model. Clearly, intermediate values of λ/d imply polymer states in the middle of those two extreme cases.

While the dense regime has been employed to explain mitotic compaction [87], the LE model, as described so far, is unable to reproduce Hi-C experimental observations of interphase chromatin. To this aim, interestingly, it is necessary to add extruding barriers on the polymer chain, i.e. polymer sites preventing LEFs to move further (**Figure 1.5**). Simulations of LE systems with extruding barriers and, intriguingly, for $\lambda/d \sim 1$, have been shown to return the patterns observed in Hi-C experiments and, in particular, TADs [35]. The extruding barriers work as boundaries of TADs, restraining loops to form across them. Since experimental investigations found TAD boundaries enriched with CTCF and, also, CTCF removal is associated to TAD melting (see previous paragraph), the extruding barriers were associated to CTCF binding sites of chromatin. Similarly, the abundance of cohesin at TAD ends, its known architectural functions and, especially, its chemical structure made of two connected rings led to the hypothesis of cohesin as LEF.

The popularity of the Loop Extrusion model derives from important experimental findings supporting its predictions. Experiments where cohesin or CTCF proteins were artificially depleted have shown the complete loss of domains and reduction of local compaction [83]. Conversely, experiments where cohesin concentration was increased resulted in stronger TADs and compaction levels [88, 89]. Additionally, the LE model can easily explain why CTCF pairs tend to mediate contacts only if convergent [10]: the extruding barriers stopping the LEFs must have convergent orientations, as the two components of a LEF move in opposite directions (**Figure 1.5**).

An objection posed to the LE model is that cohesin may not be able to slide over chromatin at enough speed to shape its architecture [53, 60]. Indeed, given that the permanence time of cohesin on chromatin is measured to be about 20min [90, 91], to form loops of 100kb cohesin should move at $\sim$5kb/min and that should increase 10 times for loops of 1Mb. No evidences have been so far collected that cohesin move at such speed. The only indication comes from *in-vitro* experiments where a protein very similar to cohesin, the *condensin*, was found to proceed on pure DNA at $\sim$3.6kb/min [92], but what happens *in-vitro* is arguably faithful to what takes place in real nuclei on chromatinized DNA. On this basis, a variant of Loop Extrusion was proposed where LEFs move by simple diffusion rather than ATP-driven (the *Slip Link* model [60]), so that the speed of cohesin is determined by diffusivity and not by an ATP pumped motor. However, that diffusion can meet the velocity required for cohesin to orchestrate loop extrusion is still to be proven.

Notably, among all the features of DNA spatial organization, the LE model was proposed as mechanism to explain the generation of TADs, loops and stripes, while it cannot reproduce the A/B compartments organization [86, 93]. Indeed, while cohesin or CTCF deletions cancel the pattern of TADs, they leave almost unaltered the compartments structure [33]. That indicates other and different mechanisms can play equally important roles in shaping chromatin architecture at the Mb scale [86] and maybe contribute also to TAD generation.

Very recently, a super-resolution microscopy experiment posed a serious challenge to the idea that loop extrusion based on CTCF and cohesin is the master mechanism for the formation of TADs [32]. Given its relevance and because it will be the starting point for the analyses of **Chapter 2**, we will dedicate the next paragraph in summarizing its findings.

1.5.3 Microscopy investigation of TADs brings new perspectives

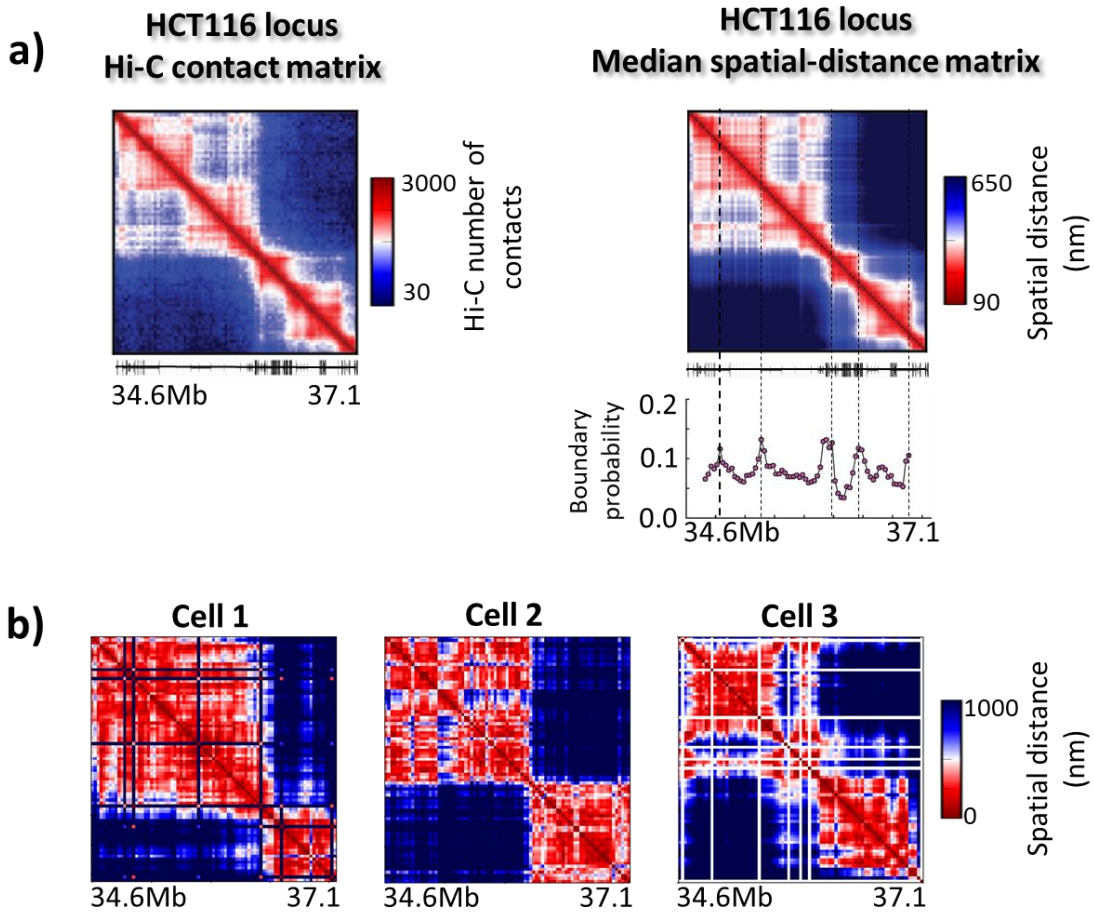


Figure 1.6: a) For a 2.5Mb-long locus of the HCT116 human cells, the ensemble Hi-C matrix [33] is very similar to the corresponding median distance matrix from an independent microscopy experiment over a population of cells [32]. Both maps are at 30kb resolution. Below the distance matrix, the list of genes (UCSC Genome Browser) is shown, and the measured boundary probability is reported (**Main Text** and [32]). Notably, no DNA window has null probability, indicating that TAD boundaries strongly vary from cell to cell (see panel b). The probability peaks correspond to the boundaries observed in the median distance matrix (dashed vertical lines) and, importantly, were found to correlate with CTCF+cohesin binding sites [32]. Adapted from [32].
b) For the same human locus, the distance matrices imaged from an allelic copy in single cells (*single-molecule matrices*) can be very different from each other, with TADs changing their arrangement [32]. Here, three single-molecule distance matrices from different cells are shown as example. Notably, each of them also differs from the median distance matrix (panel a). Data from [32].

TADs and subTADs of different DNA regions in human cell types were studied by multiplexed super-resolution FISH imaging [32]. That is a microscopy technique allowing for the observation of chromatin in single cells at the kb scale. Specifically, several DNA regions (*loci)* 2-3Mb long were investigated at 30kb resolution across thousands of human cells, with less than 50nm uncertainty

on the positioning of each 30kb DNA window. For definiteness, we will only review the findings about a 2.5Mb long locus in the human HCT116 cell line (chr21:34.6-37.1Mb), because analogous observations were found for all the other loci explored.

The 3D conformation of the HCT116 locus was imaged across thousands of cells at the *single-molecule* level, i.e. conformations were snapshotted from each of the allelic chromosomes in every cell. Hence, the map of the median distances was extracted over the cell population, that is a 2D matrix whose entries contain the median physical distances between all possible pairs of DNA windows. That was found well correlated with the Hi-C matrix of the same locus [33], and, in particular, the TADs detected by Hi-C are also seen in the median distance matrix (**Figure 1.6a**).

On the other hand, looking at the distance maps for each single chromosome rather than the median matrix, the pattern of TADs was found surprisingly variable across cells, with notable differences from the median arrangement (**Figure 1.6b**). This shows, importantly, that the single cell 3D configurations are not well represented by their average, as significant deviations can be present. Indeed, the frequency whereby a 30kb window of the locus acts as a TAD boundary across cells (*the boundary probability*) was found greater than zero for almost every window composing the HCT116 locus (**Figure 1.6a**), with the highest probabilities for windows with CTCF and cohesin binding sites [32]. That quantitatively illustrates the arrangement of TADs remarkably varies from cell to cell, but also indicates that CTCF and cohesin somehow determine the preferential locations of TAD boundaries. Hence, when investigating the average architecture over cell populations (as in ensemble Hi-C experiments), only the more frequent TADs emerge, i.e. those with CTCF and cohesin at their boundaries: CTCF-cohesin boundaries are organizational features emergent from ensemble averages, while TADs per se are not.

Next, the HCT116 locus was treated with auxin to deplete cohesin. As expected, the median distance matrix shows loss of domains, in agreement with the Hi-C matrix extracted for the same locus after the same treatment [33] (**Figure 1.7a**). However, the single-molecule distance maps still exhibit clearly visible TAD-like structures, variable from cell to cell as in the untreated case (**Figure 1.7b**). The boundary probability is almost uniform across all the 30kb windows of the locus, i.e. the peaks of probability found in the untreated case in correspondence of CTCF or cohesin binding sites vanished (**Figure 1.7a**). These findings are consistent with the hypothesis that CTCF and cohesin together define preferential locations of TAD boundaries but do not originate TADs per se: depleting one of the two makes the positioning of boundaries equally possible everywhere along the DNA locus, so, on average, no TADs emerge; yet, in each cell, TAD-like domains still form.

The observation that TADs form in single chromosomes despite the lack of cohesin challenges the Loop Extrusion mechanism as basic and exclusive principle of TAD generation. Other processes not based on cohesin and CTCF may explain the segregation in domains. However, cohesin and CTCF do play a fundamental role as they determine preferred locations of TAD boundaries, which are ultimately responsible of the average chromatin architecture. The mechanism whereby CTCF+cohesin realize that could be loop extrusion but further investigations need to be conducted. For instance, the observations above are also compatible with the physical scenario whereby depletion of CTCF or cohesin amounts at providing the chromatin fiber with translational symmetry, resulting in featureless average contact patterns. In this perspective, CTCFs and cohesin interactions may impose preferred domains thanks to a symmetry breaking process [94]. Additionally, while LE simulations can reproduce faithfully the TADs of ensemble Hi-C matrices, they fail to return the TADs

observed at single-cell level [32, 35]. Again, that allows for a picture of DNA 3D organization where loop extrusion co-exists with other essential mechanisms, rather than being the master process.
In the next chapter, we will propose a polymer-physics model alternative to Loop Extrusion whereby all the described findings (e.g. single-cell variability of TADs, boundary probabilities) are reproduced and explained by simple equilibrium thermodynamics [27].

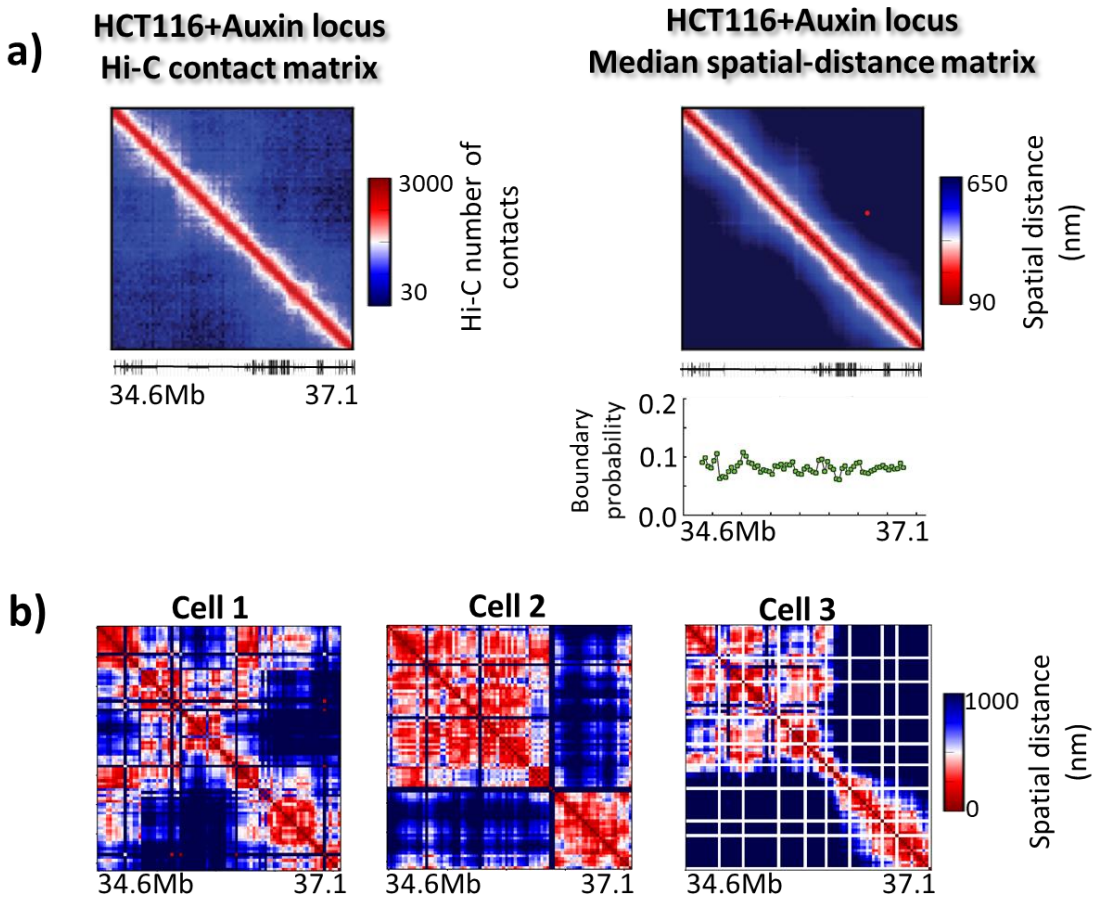


Figure 1.7: a) Hi-C (left) and microscopy data (right) are shown for the locus of the HCT116 human cells when treated with Auxin, so to remove cohesin (see **Main Text**). Specifically, the ensemble Hi-C matrix from [33] appears as featureless as the corresponding median distance matrix from an independent microscopy experiment [32]. Under the distance map, the measured boundary probability [32] lacks the peaks observed in the wild type locus (**Figure 1.6a**), showing uniform values across the 30kb windows. That means TADs still form in single cells, yet without preferential boundary locations (panel b)). Adapted from [32].
b) The single-molecule distance maps from [32] exhibit TAD-like structures, which can vary wildly from cell to cell. That is despite no TADs are seen in median distance matrix (panel a). Here, three examples of single-molecule distance matrices are shown from three different cells. Data from [32].

2 CHROMATIN FOLDING PROPERTIES IN SINGLE CELLS ARE EXPLAINED BY THE PHASE-SEPARATION MECHANISM OF POLYMERS

Given the experimental findings about TADs formation and variability in single cells [32] (see **paragraph 1.5.3**), here we describe how they can be explained by a polymer-physics model based on the mechanism of equilibrium phase separation [27]. The model is the *Strings&Binders Switch* (SBS) model of chromatin [37, 38]. Based on simple equilibrium thermodynamics of polymers, the SBS model has been used successfully to reproduce ensemble features of chromatin architecture, like TADs, A/B compartments or contact probability decay with genomic distance [38, 95–97]. Yet, as illustrated before, the average picture of chromatin organization is far from exhaustive. In the next pages, we interrogate the SBS model to explain the cell-to-cell variability of TADs detected by microscopy [32] in a 2.5Mb-long DNA human locus of the HCT116 cell line (chr21:34.6-37.1Mb), in two conditions: the wild-type (WT) natural condition and after treatment with Auxin to deplete cohesin. The WT and Auxin-treated loci will be thereafter called the *HCT116 locus* and the *HCT116+Auxin locus*. In brief, we will report how the optimal SBS polymers were designed for modeling the two loci and how their possible stationary 3D configurations in space were extracted by Molecular Dynamics simulations and then examined thoroughly to explain the experimental findings [27]. We will show that the classical physics process of phase separation successfully describes TADs formation and fluctuations across cells, challenging the picture of loop extrusion as master mechanism to generate TADs and suggesting it may be rather involved in determining their preferential boundaries. The effect of cohesin removal will also be explained as a thermodynamic switch from the phase-separated state to the coil state of classical polymers.

In paragraph 2.1 we provide general details on the SBS model of chromatin, illustrating the physical forces at play and the parameters to fix. Then, in paragraph 2.2, we describe the machine learning procedure enabling the construction of the best SBS model to represent a given genomic region. Next, we explain the Molecular Dynamics simulations carried to obtain the putative 3D conformations of the loci, illustrating the key dynamical and equilibrium properties of the SBS systems. In paragraphs 2.4 and 2.5, we show the predictions of the models match and rationalize the experimental observations for, respectively, the HCT116 and the HCT116+Auxin loci [32]. Finally, in paragraph 2.6, we exploit the Molecular Dynamics approach to study the temporal dynamics of the loci model conformations. All the contents presented below have been recently published [27].

2.1 The Strings&Binders Switch model of chromatin

The *Strings&Binders Switch* (SBS) model describes a chromatin filament as a self-avoiding polymer chain made of beads. Diffusive particles, named binders, can attractively interact with specific beads along the chain, acting as binding sites. Different kinds of binders and cognate binding sites are typically employed in applications [27, 95–97] (**Figure 2.1**) and the set of all binding sites of the same type constitutes a *binding domain* of the polymer. Importantly, a single binder can simultaneously attract many cognate binding sites, driving them close in space. Thus, binders act as mediators of interaction between distal beads of the polymer chain and determine its folding. In general, beads not acting as binding sites (*inert beads*) are also envisaged (**Figure 2.1**).

Such picture of chromatin relies on the textbook description of the nuclear cell environment: an aqueous solution where thousands of proteic molecules swarm around DNA filaments to shape their conformations, regulate transcription or repair damages.

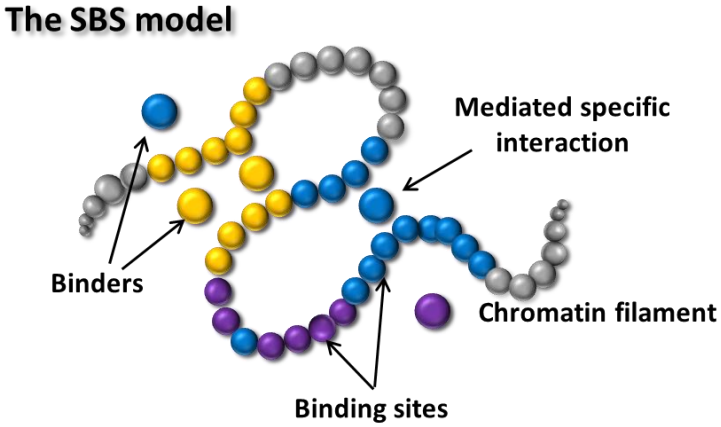


Figure 2.1: Scheme of the *Strings&Binders Switch* (SBS) model of chromatin [37, 38]**.** A chromatin filament is represented by a self-avoiding polymer chain made of beads. Diffusive particles named binders move all around in a viscous solution and can interact with cognate binding sites (visualized as same-colored beads). By interacting with multiple beads at the same time, binders generate mediated attractions between distal binding sites of the same kind, driving the folding of the polymer. Inert not-interacting beads are also included and are visualized in grey.

The 3D equilibrium configurations of an SBS polymer are obtained through Molecular Dynamics (MD) simulations: the polymer chain and the binders are put in certain initial conditions inside a simulation volume, then the equations of motion of the system are solved to return the spatial coordinates of beads and binders until thermodynamic equilibrium is reached.
The number of beads making up the polymer chain, the number of types of binding sites and their distribution along the chain must be set accurately for the SBS 3D configurations to model nicely the conformations of a specific real DNA locus, as they discriminate the folding properties of the polymer. In the next paragraph we will describe the complex procedure whereby the features of a SBS polymer are selected according to the DNA region to model.

2.2 The machine learning procedure to infer the best SBS polymer model

The SBS model of a DNA region is, ultimately, characterized by the number of beads of the polymer (N), the number of types of binding sites (n) and their distribution along the polymer chain. Clearly, their choice cannot be random as they are required to effectively model the architecture of the region under study. As said, the average 3D organization of a chromatin locus can be detected, for instance, by Hi-C experiments. In this sense, the best SBS model of a chromatin region can be identified as that polymer yielding the pattern of contacts most similar to that observed in a Hi-C experiment: this is the basic principle of the PRISMR algorithm (*Polymer-based Recursive Statistical Inference Method*) [96]. Indeed, PRISMR is a computational procedure which infers the optimal SBS polymer model for a given DNA locus, starting from its Hi-C contact matrix. Specifically, in applications, the Hi-C contact matrix is first smoothed and normalized by standard algorithms [98]. Here, we are going to present the latest version of PRISMR [27], as significant improvements have been made since the original implementation [96].
At the core of PRISMR there is a Simulated Annealing (SA) routine, whereby many different SBS polymers are scanned to find the one yielding the contact map most alike to the Hi-C input matrix.

To this aim, a cost function is used, H, measuring the similarity between the SBS and the Hi-C contact maps so that the best polymer returns the minimum of H. Specifically, the cost function is made of two terms: $H = H_0 + H_\lambda$. H_0 is the average squared difference between the two matrices, where each term is normalized by the average value of the Hi-C map at the corresponding genomic distance. This is done to prevent values at short genomic distances from dominating H_0, as contacts are generally much more abundant between DNA windows close along the genome. H_λ is a Bayesian term penalizing the adding of further kinds of binding sites, to reduce overfitting. It is weighted by a factor, λ, controlling the strength of the penalization. For a polymer with N beads, n allowed kinds of binding sites and for fixed λ, at each SA iteration a randomly chosen bead is turned into a binding site of a certain type, the cost function is evaluated and the change is accepted or rejected according to the standard Metropolis algorithm [99]. That is done until H converges, i.e. plateaus to a minimum. Then, the whole procedure is repeated for many initial casual conditions of the polymer to check the robustness of the plateau. That eventually provides the best arrangement of binding sites along the polymer chain, given N, n and λ. Notably, to compute the contact matrix of the SBS polymer at each SA iteration MD simulations should in principle be conducted to extract its 3D configurations and then calculate the model contact frequencies. Since that would be too demanding, the contact frequencies are computed recurring to a mean-field approximation, whereby the probability of contact between two cognate or different binding sites at a given distance along the chain is approximated by that valid for a toy SBS model. In the case of cognate sites, the toy model consists in a polymer whose beads are all binding sites of the same type and which is completely folded; in the eventuality of different binding sites, the toy polymer is made only of inert not-interacting beads. The reader who may be interested in further details is referred to [96].

The procedure described so far finds the optimal arrangement of binding sites for given N, n and λ. To set N, the resolution of the Hi-C input matrix is considered, i.e. the length of the DNA windows into which the locus is divided (**paragraph 1.3.2**). If L is the genomic size of the locus and res the resolution of the Hi-C data, then the number of DNA windows is L/res. To account for the presence of various binding sites inside each window, PRISMR poses $N = r * L/res$, where r is an integer number. The first guess is $r = n$, so to collapse two parameters (N and n) into one. Then, the best n is searched with the following standard procedure of supervised learning.

The Hi-C contact data are split into two complementary sets, the training and the test sets. The SA routine is run over the training set for several values of n, using H_0 as cost function (so ignoring the Bayesian term, see **Figure 2.2a**). Hence, the best polymers found for each n are benchmarked against the test set, evaluating again the cost function. As dictated by Machine Learning theory, the minimal H_0 values found on the training set must decrease indefinitely for growing n, while the values obtained on the test set reach a minimum, then increase with increasing n, signaling overfitting. As example, in **Figure 2.2b**, we report the plots of H_0 against n obtained for the modeling of the HCT116 and HCT116+Auxin human loci. In these cases, the training set was the 70% of the used Hi-C data [33] and the remaining 30% the test set, but other proportions were checked to return analogous results [27]. The value of n returning the minimum of H_0 on the test set defines the best estimate n^*.

Further on, the optimal Bayesian parameter λ^* is analogously determined, working with the complete cost function H and with $n = n^*$. Given n^* and λ^*, at last the learning is performed to seek the best r value (r^*), removing the initial constraint $r = n$.

Once the key parameters are fixed, a final battery of SA runs is launched for different random initial conditions of the polymer, to get the optimal definitive arrangement of binding sites. Of all the outputs obtained from different initial conditions, the arrangements corresponding to the 10% lower minima of H are generally highly similar to each other, proving the robustness of the approach [96]. To give a sense of the computational effort involved, up to 500 different SA runs were performed for modelling each of the HCT116 loci [27].

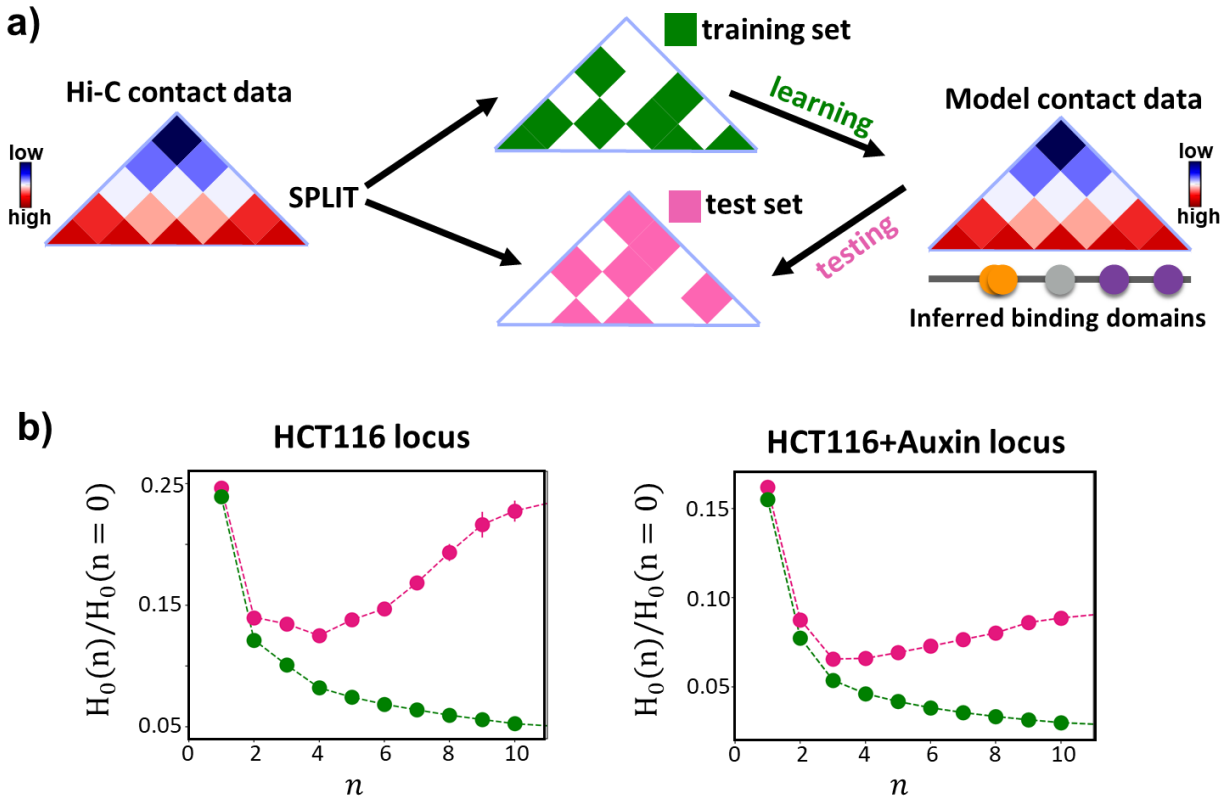


Figure 2.2: a) Outline of the supervised learning procedure used by PRISMR to define the optimal number of binding site types, n. The input experimental Hi-C data are split into two parts, the training set and the test set. For a given n, the Simulated Annealing routine of PRISMR (**Main Text**) is applied on the training set to return the model contact matrix yielding the minimal cost function $H_0(n)$. Then H_0 is estimated over the test data for the same model contact matrix. The procedure is repeated for several values of n. As dictated by Machine Learning theory, the values of H_0 computed on the test set will have a minimum, which defines the optimal n. Adapted from [27].
b) The cost function H_0 is plotted against n in the case of the modelling made for the HCT116 locus (left) and the HCT116+Auxin locus (right) [27]. The values of H_0 are normalized dividing by the value at $n = 0$, i.e. when the polymer model is made only of inert beads. The green curves show the minima of H_0 found on the training set, while the pink curves display the corresponding values of H_0 extracted from the test set. As expected, the green curves decrease with growing n for both the loci, while the pink curves exhibit a minimum, after which overfitting begins. The n values giving the minimal H_0 on the test data are those selected for the best SBS model of the considered loci ($n = 4$ and $n = 3$ for, respectively, the HCT116 and the HCT116+Auxin locus). Adapted from [27].

For sake of completeness, we mention that PRISMR has been accommodated to run also on GAM matrices [100] or distance maps derived by microscopy. Indeed, PRISMR only needs experimental information on the architecture of the target genomic locus, so to benchmark the SBS polymer

model. Thus, one can made the algorithm run on any kind of architectural data upon adequate modifications of the cost function. Also, the mechanism of PRISMR is general and could be similarly applied to other polymer models, changing accordingly the optimal parameters to seek.

2.3 The Molecular Dynamics simulations

To model the possible architectures of a real chromatin region, the dynamics of the best SBS polymer and relative binders is computed through Molecular Dynamics (MD). The system is placed into many random initial conditions and for each of them a MD simulation is performed until stationarity. Molecular Dynamics is computed using the open-source LAMMPS software (*Large Atomic Molecular Massive Parallel Simulator*), which integrates the equations of motion via the Verlet algorithm [101]. In the end, an ensemble of steady-state configurations of the SBS system is obtained.

We will now present the equations of motion and potentials characterizing the dynamics of SBS systems and describe how the MD runs are typically conducted [27, 95–97]. Such topic was systematically reviewed in [39].

2.3.1 Equations of motion of the SBS system

For simplicity, beads and binders are implemented as spheres with same diameter, σ, and mass, m. They move in a viscous solution, representing the nuclear environment. Hence, they all follow the Langevin equation of motion:

$$m\frac{d\vec{v}}{dt} = -\vec{\nabla}V - m\zeta\vec{v} + \vec{f} \tag{1}$$

V is the total energy potential of the particle, ζ is the friction coefficient and $\vec{f}$ is the stochastic force caused by the viscous solution. The friction coefficient is connected to the viscosity of the embedding solution (η) by the Stokes relation for a sphere: $m\zeta = 6\pi\eta\sigma$.

To treat beads or binders as hard spheres, a repulsive shifted Lennard-Jones interaction is turned on between any pair of particles, commonly known as the Weeks-Chandler-Anderson potential [102]:

$$V_{WCA}(r_{ij}) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} + \frac{1}{4}\right] & if\ r_{ij} < 2^{\frac{1}{6}}\sigma \\ 0 & otherwise, \end{cases} \tag{2}$$

where r_{ij} is the distance between the particles i and j and ϵ is an energy parameter set equal to k_BT, i.e. the typical scale of biochemical energies. The presence of σ in the equation ensures a rapidly divergent repulsion takes place if the particles come closer than the sum of their radii, enforcing the feature of hard spheres. The truncation at $r_{ij} = 2^{1/6}\sigma$ removes the attractive part of the Lennard-Jones potential.

Consecutive beads along the chain are held together by a standard finitely extensible non-linear elastic potential (FENE) [103]:

$$V_{FENE}(r) = -\frac{kR_0^2}{2}\ln\left(1-\frac{r}{R_0}\right), \quad (3)$$

where k is the strength of the elastic potential, r the distance between the consecutive beads and R_0 sets the maximum possible elongation. k is posed equal to $30k_BT/\sigma^2$ and R_0 to 1.6σ [39]. Such values of the parameters guarantee that the combined action of (2) and (3) makes the preferred length between two consecutive beads approximately equal to σ, i.e. adjacent beads are made tangent to each other. This is important to avoid cross-interactions and nots during MD simulations. The specific interaction between binders and cognate binding sites at the basis of the SBS model is again implemented with a Lennard-Jones shifted and truncated potential:

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

r_{int} is the range of action of the interaction, ϵ_{LJ} is the energy parameter and σ_{LJ} defines the distance below which the Lennard-Jones potential becomes divergent. Consistent with the picture of hard spheres, σ_{LJ} is posed equal to σ. Finally, a weaker aspecific Lennard-Jones attraction can also be turned on between any pair of beads and binders, regardless of their respective type [27]. That accounts for generic London forces which can establish between DNA and proteic molecules.
The minimum of a Lennard-Jones potential (4), in absolute value, is given by the following:

$$E_{int} \equiv |\min(V_{LJ})| = \left|4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r_{int}}\right)^{6} - \left(\frac{\sigma_{LJ}}{r_{int}}\right)^{12} - \frac{1}{4}\right]\right|. \quad (5)$$

E_{int} defines the energy scale of the Lennard-Jones interaction. In units of k_BT and σ, ϵ_{LJ} and r_{int} are generally set so that E_{int} is in the weak biochemical range of energies (units of k_BT) for both the aspecific and specific interactions. For sake of simplicity, the specific potentials are generally assumed identical across the various possible types of binding sites. For instance, in the modeling of the HCT116 loci, ϵ_{LJ} and r_{int} were fixed for the specific and aspecific potentials such that E_{int} resulted, respectively, $3.1k_BT$ and $2.7k_BT$ [27].

2.3.2 Physical value of the MD parameters

The equations of motion (1) are integrated by LAMMPS in adimensional units. That is, the mass of the particles (m), their diameters (σ), and the thermal energy k_BT are put to 1. So, to make sense of the quantities computed, a conversion in physical units is required.
The physical value of σ can be estimated in different ways. A possible approach is to assume that the local genomic density equals that of the whole cellular nucleus [95]. Assuming the nucleus is spherically shaped, that returns $\sigma = (g/G)^{1/3}D$, where g is the genomic content of a bead (genomic length of the modelled locus divided by the number of beads), G is the size in bp of the entire genome in the nucleus and D is the nuclear diameter. Another method is based on experimental estimates of the average compaction of chromatin, i.e. the mean number of bp per nm [36, 96]. As the genomic content of a bead is known, the value of σ is simply computed from the

compaction data. Finally, σ can be estimated by comparison whenever a length scale of the model can be matched against an experimental equivalent [27].
Temperature is the typical one for the organism of the investigated locus (around 37°C for mice and human beings), which settles the energy unit.
The concentration of binders (c) in the simulation box in mol/l is obtained from the relation $c = P/(VN_A)$, with P is the absolute number of binders, V the simulation volume and N_A the Avogadro number.
The MD time scale, τ, is given by the relation $\tau = 6\pi\eta\sigma^3/k_BT$, where η is the viscosity of the embedding solution. Experimental estimates of nucleoplasm viscosity are around 0.03P [26, 28], which is the value used in the SBS models [39].
Finally, the mass of beads and binders can be calculated from $m = \tau\sigma/\sqrt{k_BT}$.

2.3.3 Running the MD simulations

As initial conditions, the SBS polymer is placed in several random self-avoiding walk (SAW) configurations. This is achieved locating all consecutive beads at the fixed distance of σ but at random relative orientations (random walk of step σ) and then running preliminary MD runs with the following soft potential to enforce self-avoidance [39, 103]:

$$V_{soft}(r) = A\left[1 + \cos\left(\frac{\pi r}{2^{\frac{1}{6}}\sigma}\right)\right].$$

A is an energy parameter and r is the distance between a pair of particles.

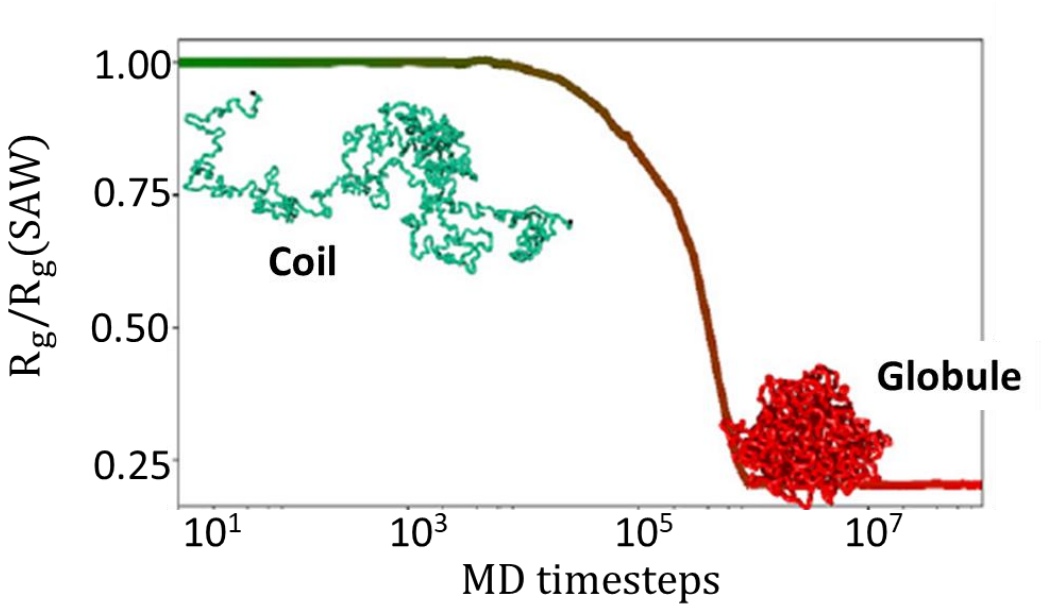


Figure 2.3: The Molecular Dynamics simulations for an ensemble of SBS polymers are tracked by the temporal evolution of the average gyration radius. For high enough concentrations of binders, the polymers switch from the initial SAW coil state to the folded globule phase. The transition is marked by a sharp drop of the average gyration radius, which finally plateaus, signaling that the polymers have fallen in stationary globule conformations. In this example, the gyration radius is normalized by the initial SAW value. Adapted from [39].

To check that the polymer configurations turn from the random to the self-avoiding walk regime, their gyration radii are tracked. Indeed, the gyration radius, on average, is expected to increase from the random walk value up to a plateau, marking the SAW state. The quality of the self-avoiding configurations can be further verified by checking for the scaling properties of a SAW polymer, e.g. controlling that the gyration radius scales with the polymer length as $R_g \sim N^{0.588}$ [104], where N is the number of beads.

Each SBS polymer in a SAW random configuration is inserted in a simulation cubic box with length scale approximately equal to $N^{0.588}$. Binders are added randomly in the box with given concentration c. Then, for every polymer, the MD runs are started with the potentials and equations described above (**paragraph 2.3.1**). As we will illustrate in detail for the HCT116 loci, for high enough concentrations of binders the polymers undergo a phase transition from the coil SAW state to a globule phase-separated state [39]. As known from block-copolymer theory [105, 106], that happens because binders drive distal cognate binding sites in close proximity, forming separated globules for each binding domain. Correspondingly, the average gyration radius exhibits a sharp drop, ultimately reaching a plateau (**Figure 2.3**): the gyration radius acts as order parameter of the transition. Hence, the MD simulations are typically tracked following the gyration radius evolution across the ensemble of polymers. When, on average, the gyration radius plateaus, then the globular stationary state is attained and the simulations are arrested.

The compact globular configurations can be used for modeling real DNA loci structures, given their confinement in the crowded nuclear environment. However, in agreement with previous studies [95], we will see that the coil conformations also play a role in the description of chromatin architecture.

2.4 The variability of TADs across cells can be explained by the thermodynamic degeneracy of phase-separated polymers

2.4.1 Transition from the coil state to the phase-separated globule state

We now focus on the model developed for the HCT116 locus. As said, that is a 2.5Mb long locus (chr21:34.6–37.1Mb) of human HCT116 cell line, deeply investigated by microscopy experiments [32]. All the following content is summarized from published work [27].

The PRISMR algorithm was applied on the Hi-C data of the locus at 30kb resolution [33] to infer the best SBS model (**paragraph 2.2**). We found the optimal polymer consists of 830 beads (10 beads per each 30kb window) with 4 types of binding sites distributed as shown in **Figure 2.4a**. The MD simulations were carried out setting the energy scale of the specific potential to $3.1 k_B T$ for all types of binding sites, while $2.7 k_B T$ was used as energy scale of the aspecific interaction. The runs were performed exploring different values of binder concentrations (equal for all the binder types), ranging from 0 to 0.5 μmol/l and, for each concentration, an ensemble of 1000 3D configurations was derived. For every investigated concentration the MD simulations were tracked analyzing the gyration radius evolution with time, averaged over the ensemble of polymers. **Figure 2.4b** shows the mean gyration radius for two different binder concentrations along all the duration of the MD simulations. In both cases the gyration radius decreases to a plateau as result of the folding activity of binders, marking that stationarity has been achieved. However, the entity of the decrease depends on the considered concentration. To seek the concentration of binders determining an

actual coil-globule phase transition, we studied the average gyration radius at stationarity for all the values of concentration explored. We also analyzed other two order parameters of the transition, the mean separation score and the total binding energy of the system. The former measures the average level of spatial separation between segments on either side of a given polymer position (see [27, 32] for the mathematical details); the latter is the total potential energy of the SBS system.

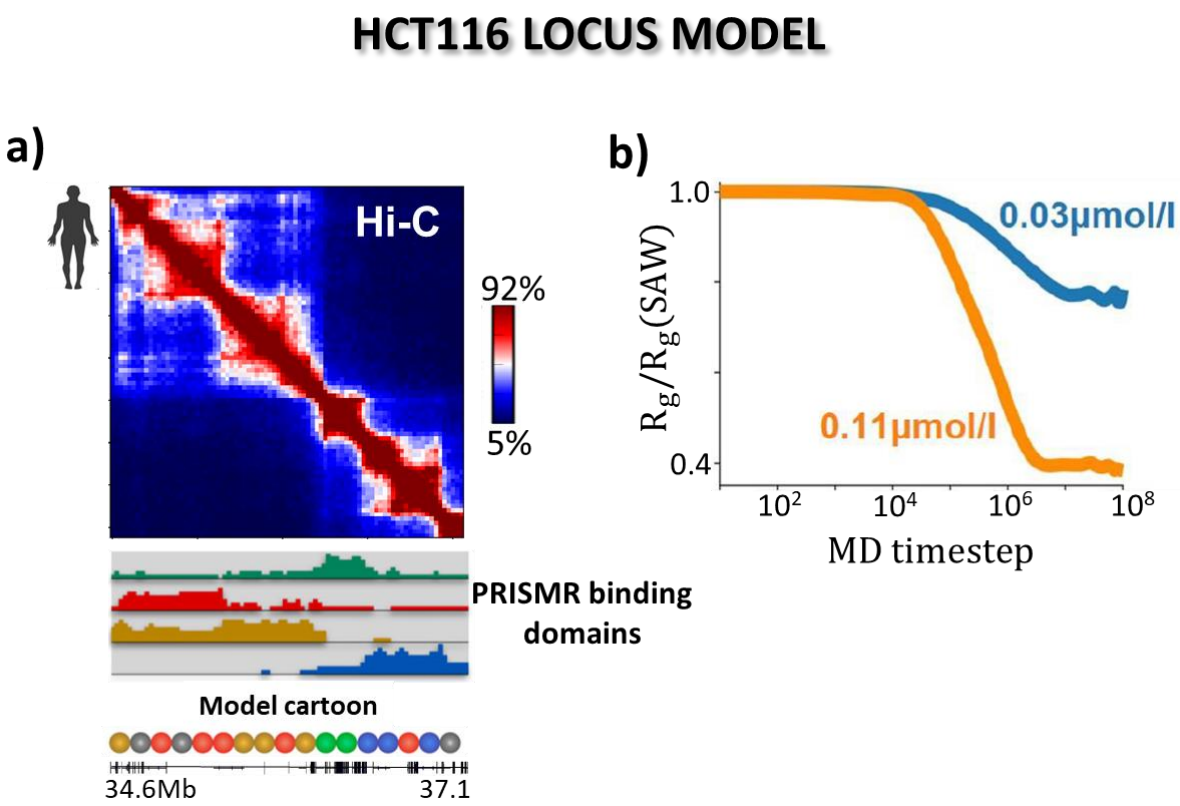


Figure 2.4: a) To derive the best SBS model for the HCT116 locus, PRISMR [96] was applied on the relative Hi-C data at 30 kb [33]. The colorbar indicates the percentiles of the Hi-C heatmap. The best SBS polymer is made of 4 binding domains, visualized in different colors. Their arrangement along the polymer is shown, together with their abundance in each 30kb window. Below, the list of genes of the locus comes from the UCSC Genome Browser.
b) Replicates of the optimal polymer model were placed in random self-avoiding walk configurations, then MD simulations were carried out until stationarity. The average gyration radius of the ensemble of replicates is shown against the MD time for two concentrations of binders. In both cases, the gyration radius undergoes a drop due to the folding action of binders; then a plateau is reached, which marks stationarity and depends on the considered binder concentration. Adapted from [27].

Upon increasing the concentration of binders, the gyration radius, the separation score and the binding energy all undergo a sharp decrease, marking the passage from the coil to the globule phase (**Figure 2.5a**). All three order parameters signal the transition at the concentration of approximately 50nm/l. In the initial coil state, entropic forces are dominant, constraining the polymer in random conformations with high gyration radius and separation score; the binding energy is almost null as attractive interactions are few at low concentration of binders. When the number of binders becomes high enough the attractive forces thermodynamically prevail and the diverse binding domains segregate in separated globules, returning sharply reduced gyration radius and separation

score (the polymer is denser); the binding energy gets deeply negative as binders occupy most of the binding sites. As each binding domain folds overall independently of the others, the transition to the globule state consists of a phase separation process [107]. Importantly, the separation between different globules is not neat as in linear block-copolymers, as the binding domains of the SBS polymer are intertwined along the chain (**Figure 2.4a**), allowing for partial segregations and inter-globules interactions. That hugely increases the degeneracy of states available in the globule phase (**Figure 2.5b**).

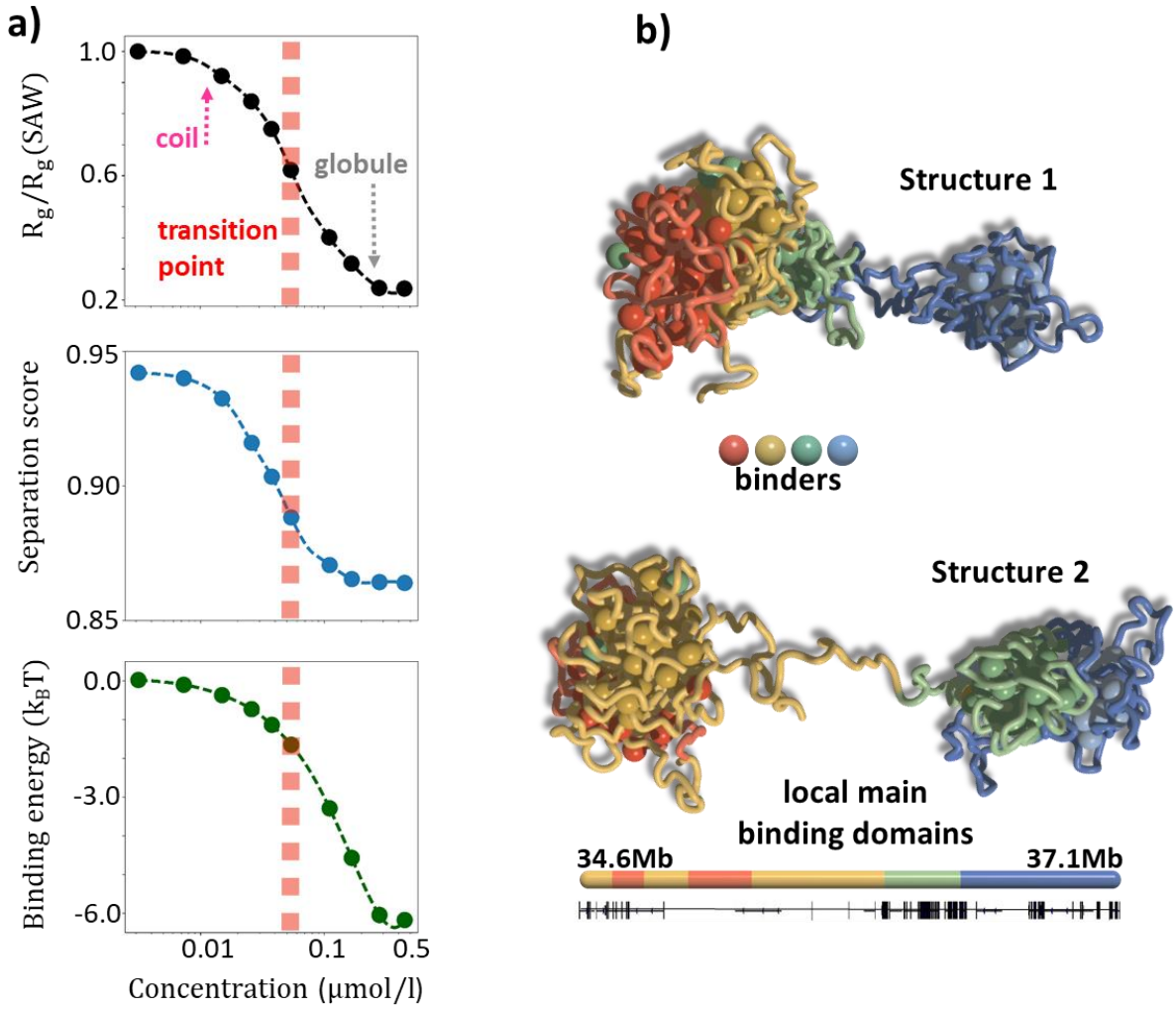


Figure 2.5: a) Upon increasing the concentrations of binders, the SBS polymers undergo a phase transition from the initial coil state to the globule phase-separated state. That is illustrated by the order parameters of the transitions, i.e. the gyration radius (top), the separation score (middle) and the binding energy (bottom), averaged over the ensemble of polymers. All three quantities signal the transition as a sharp drop, at around 50nm/l.
b) Two examples of 3D conformations of our polymers in the globule phase. The color code expresses the more abundant binding domain in each 30kb window (the local main binding domains on the bottom). The phase-separation mechanism of folding can return very different structures, especially because binding domains are overlapped and intertwined along the polymer chain (**Figure 2.4a**). Indeed, the first structure on top presents the green globule interacting with the yellow and red domains, while in the second structure it interacts with the blue domain.
Adapted from [27].

Given such findings, the following analyses will be from the ensemble of 3D configurations corresponding to 0.11µmol/l binder concentration, ensuring the coil-globule transition.

2.4.2 The binding domains found by PRISMR have an epigenetic meaning

We asked whether the binding domains inferred by PRISMR could give biological insights about the folding of the HCT116 locus. To this aim, we investigated their molecular nature, i.e. we compared the genomic location of the 4 binding domains with that of epigenetic factors experimentally observed (see **paragraph 1.4**). Along the locus and in every 30kb window, the abundance of each type of binding site was computed and compared with the enrichment data of various epigenetic features. We used Chip-seq data from [33] and from the ENCODE database [108]. The epigenetic tracks were taken at 30kb resolution, then the Pearson correlation was computed between each of them and the abundance marks of every binding domain. To check the statistical significance of the correlations, we tested them against random controls, obtained from the Pearson correlations between each epigenetic mark and the abundance of binding domains randomly reshuffled along the locus. 100 different reshufflings were realized, so to get a control distribution for every pair of epigenetic track and binding domain. The correlations were considered significant if above the 90th percentile or if lower than the 10th percentile of the corresponding control distributions. That resulted in significant correlation values spanning from -0.6 to 0.6 across all the pairs of binding domains and epigenetic marks considered, as shown in **Figure 2.6**.

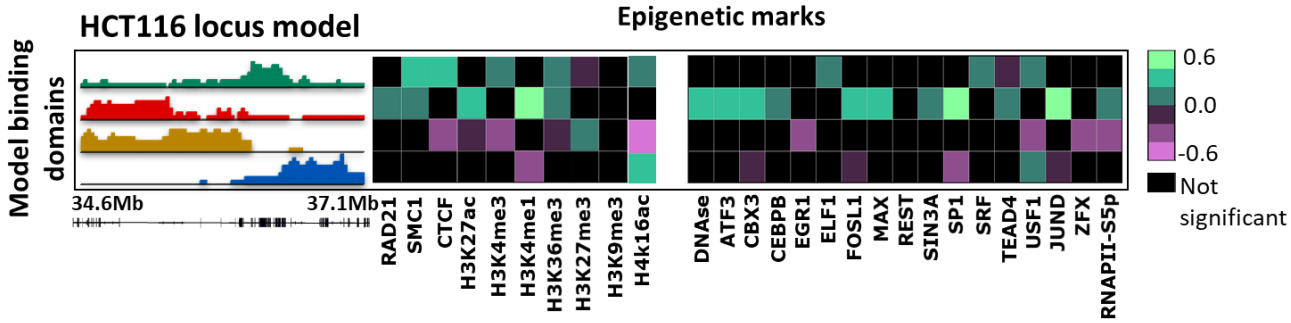


Figure 2.6: Each binding domain inferred by PRISMR was found correlated to combinations of functionally similar epigenetic marks. The first block of epigenetic signatures is from [33], the rightmost group contains additional marks from the ENCODE database [108]. The significance of the correlations was tested against random controls (see **Main Text**). Notably, the green binding domain is strongly associated to SMC1 (a subunit of the cohesin) and CTCF binding sites.

With this procedure, we found that each type of binding domain has statistically significant Pearson correlations with a combination of epigenetic factors known for playing analogous functions (**Figure 2.6**). Interestingly, the first binding domain (colored in green in **Figure 2.6**) is correlated strongly with the CTCF+Smc1 (a component of cohesin) sites; the second domain (red) is associated with epigenetic tracks signaling active chromatin (e.g. H3K27ac, H3K4me3); the third (brown) is more correlated with marks of inactive chromatin (e.g. H3K27me3) and, finally, the fourth domain (blue) is significantly linked with H4K16ac and other specific transcription factors.

Those findings suggest that PRISMR, based only on physics and machine learning, extracts binding domains which are combinations of functionally related epigenetic marks (active, repressive and so

on). In this sense, the binding domains represent effective domains of chromatin states which could help elucidating the *epigenetic code* of DNA.

2.4.3 The model of the HCT116 locus reproduces ensemble and single cell data

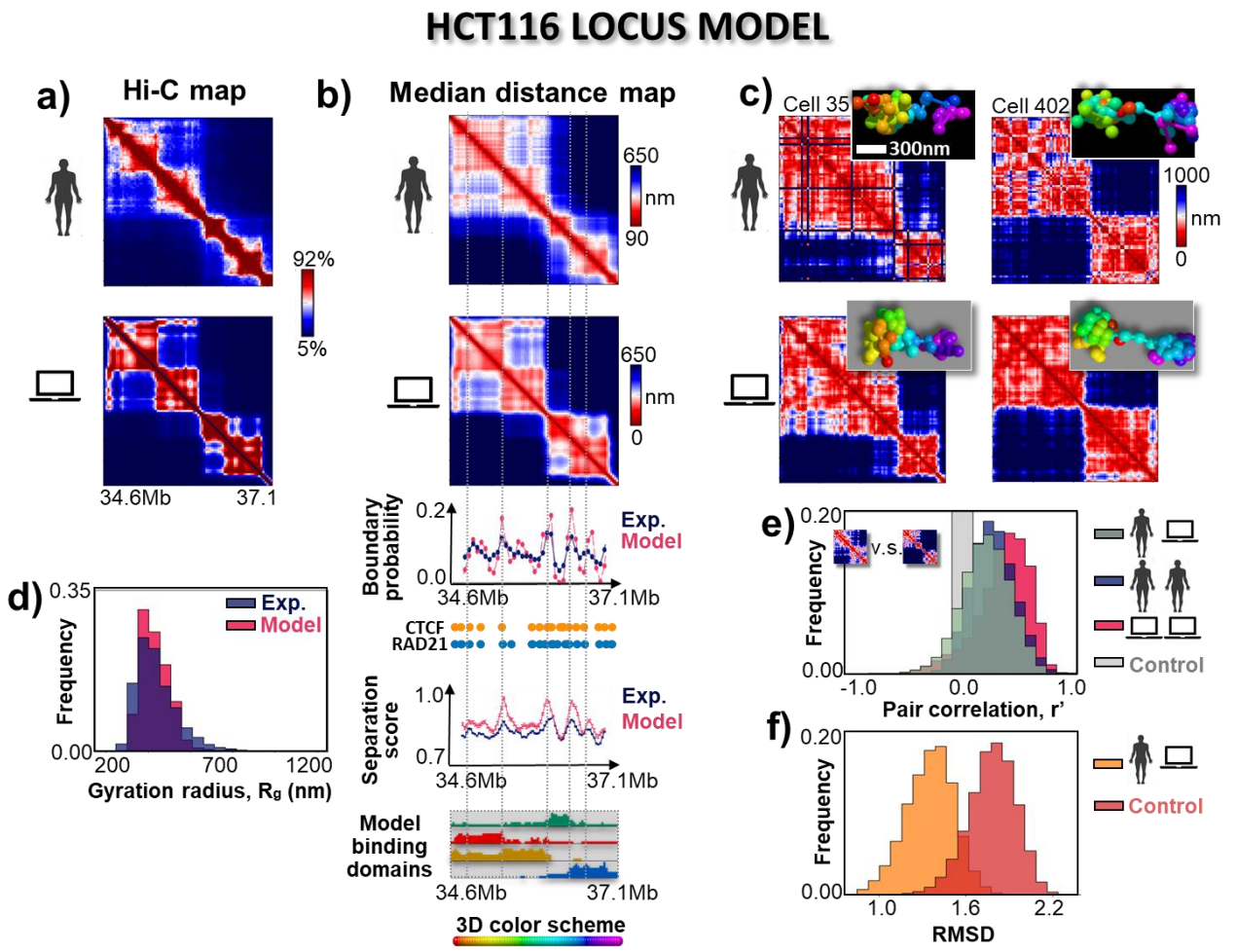


Figure 2.7: a) The ensemble of the model polymer structures returned a contact map (on the bottom) very similar to the experimental Hi-C matrix (top, [33]), with Pearson correlation (r) 0.88 and Pearson distance-corrected correlation (r') 0.68 (see **Main Text**). That is a consistency check, as our polymers were inferred by the same Hi-C data. Colorbars indicate the percentiles of the matrices.
b) The model median distance matrix (bottom) matches the experimental analogue extracted from independent microscopy data (top, [32]), with r=0.95 and r'=0.84. Together with panel a), that indicates our ensemble of polymers is representative of the average structural properties of the HCT116 locus. Below, the boundary probability and separation score derived from the same microscopy data are overall reproduced by those extracted from the model ensemble of structures (r=0.79 and r=0.85). This suggests that the cell-to-cell architectural variability of the locus is well matched by our ensemble of polymers. Under the boundary probability plot, the experimental tracks [32] of CTCF and RAD21 (a subunit of cohesin) binding sites are reported, to show they mostly correspond to the peaks in the boundary probability curves.
c) The structural variability of the HCT116 locus is illustrated by two examples of imaged single-molecule distance maps (top, [32]), exhibiting different TADs. Notably, we report two single polymer distance matrices (bottom) which display the same differences in the TAD pattern. The 3D structures corresponding to each matrix are also shown, colored according to the color code on the bottom of panel b).

d) The gyration radius distribution across the whole collection of 3D structures is plotted both for the imaging [32] and model cases. The distributions are undistinguishable (Mann-Whitney p-value=0.40), with average values of 440nm.
e) The distributions of Pearson distance-corrected correlations (r') are displayed for all possible pairs of single-molecule imaged distance maps (*exp-exp*, blue), of model distance maps (*mod-mod*, pink) and of model and experimental distance maps (*mod-exp*, dark grey). They all overlap with each other more than with a random control distribution, obtained from randomized experimental matrices (light grey). In particular, the *exp-exp* and *mod-exp* distributions are statistically undistinguishable (Mann-Whitney p-value = 0.19, see **Main Text**). The mean values are r'=0.21, 0.27 and 0.36 for, respectively, the *exp-exp, mod-exp* and *mod-mod* distributions, indicating high structural differences in all cases.
f) The imaged 3D configurations of the locus [32] were associated to structures of the polymer model through the RMSD method [27, 75] (**Main Text**). The distribution of the best RMSD for each experimental conformation is not compatible with the distribution from randomly chosen pairs of imaged structures (Mann-Whitney p-value = 0.0), proving the association is significant. Analogous results were found when considering the association of every model structure with the imaged configurations.
The findings illustrated in panels b,c,d,e,f overall prove that our ensemble of phase-separated polymer conformations is a *bona-fide* representation of the real 3D structures of the HCT116 locus in cells. Adapted from [27].

Based on previous studies of the SBS model [96, 97], we hypothesized the possible 3D structures of the HCT116 locus could be represented by our ensemble of polymers in the globule phase. As first check, we verified that the *in-silico* Hi-C matrix (i.e. derived by the polymer population) matches the experimental ensemble Hi-C map [33], with a Pearson correlation (r) as high as 0.88 and a Pearson distance-corrected correlation (r') of 0.68 (**Figure 2.7a**). The r' is a more severe measure of correlation accounting for the average decay of contacts between windows when their genomic separation increases, as, trivially, that decay positively contributes to standard Pearson correlation. Precisely, to compute the r' correlation between two matrices C and D, they are first transformed into the C' and D' matrices according to the following rule: $c'_{ij} = c_{ij} - \langle c_{diagonal} \rangle$, where c'_{ij} and c_{ij} are the entries (i, j) of C' and C, and the bracketed term indicates the average value of all the entries of C belonging to the same matrix diagonal as c_{ij}; the same holds for D'. Then, the Pearson correlation is evaluated between C' and D', returning the r' correlation coefficient. As for the *in-silico* Hi-C matrix, that is computed by counting all the contacts between every pair of beads across the population of polymers, where two beads are considered in contact if their distance is below a given threshold. Then, the contact counts are averaged to have the contact numbers between every pair of 30kb windows (as said, a window is made of 10 beads) and those are arranged in a matrix (see also **Chapter 3**, **paragraph 3.1.1**). The good similarity between the model and experimental Hi-C maps is a consistency check, as the whole ensemble of polymers was inferred by those Hi-C data [33].
Then, we calculated the median distance matrix of the polymers: for each polymer configuration we computed the Euclidean distances between all pairs of windows, then extracted the median values across the polymer ensemble and arranged them in a matrix. The distance between two windows is obtained considering their mass centers. So, to make a more stringent test, we compared our *in-silico* median distance matrix with the median distance map of the locus detected by microscopy [32] in an independent experiment from Hi-C. We found nice similarity, with r=0.95 and r'=0.84 (**Figure 2.7b**).
Those results indicate the SBS model returns the ensemble features of the HCT116 locus, in agreement with previous studies on other loci, cell lines and organisms [38, 95–97, 109].

Importantly, the nice similarity between the model and imaged distance maps also show that the ensemble of polymers in the globule thermodynamic phase naturally yields the emergence of TADs at the population level. That is due to the mechanism of phase separation forming segregated globules in space.
Next, we interrogated the single-molecule properties. The phase separation process of the model binding domains can produce many different polymer conformations, according to the random diffusion of binders. The difference between two independent polymer structures can be as relevant as a diverse arrangement of the phase-separated globules (**Figure 2.5b**). When extracting the distance matrices per each single polymer configuration, the structural degeneracy induced by phase separation is visible as highly variable patterns, including variable TADs (**Figure 2.7c,** pair of matrices on the bottom). That is intriguingly reminiscent of what observed by super-resolution microscopy on the same HCT116 locus [32], where huge TAD variability was revealed across single chromosomes in cells (see **paragraph 1.5.3**). Indeed, in **Figure 2.7c** we show two single-molecule distance maps from microscopy data which are very similar to distance matrices extracted from our polymers. To investigate that quantitatively, we conducted the following analyses.

1) We compared the experimental boundary probability, separation score and gyration radii distribution from [32] with the corresponding model quantities. The experimental and model boundary probabilities (i.e. the frequency whereby a DNA window is a TAD boundary) match with a Pearson correlation of r=0.79 (**Figure 2.7b**). Similarly for the separation score, with a Pearson correlation between model and experiment equal to 0.85 (**Figure 2.7b**). As explained in **Chapter 1**, the experimental boundary probability track reveals the impressive variability of TAD boundaries across cells, albeit preferential locations emerge in correspondence to cohesin and CTCF binding sites (**paragraph 1.5.3**), which inform the TADs seen in the median distance map. Hence, the fact that our model structures can reproduce such a track indicates the phase separation degeneracy is able to recapitulate the observed TAD variability, with the more frequent globules determining the average pattern of TADs. Indeed, in the SBS polymers globules more likely than others are present due to the arrangement of the binding sites. Analogous considerations can be made for the separation score.
Also, the distribution of gyration radii across the polymer ensemble is compatible with that of the imaged structures (Mann-Whitney p-value=0.40, **Figure 2.7d**), after equalizing the average values so to determine the length unit of the model (**paragraph 2.3.2**). The mean gyration radius is thus 440nm for both the model and experimental structures. The gyration radius comparison shows that the imaged conformations exhibit a degree of spatial compaction explainable by the globule phase of SBS polymers.

2) We quantified the variability of the single-molecule distance matrices (and so of the underlying 3D structures) by computing the correlations between all possible pairs of them. To strengthen the value of the analysis, we employed the distance-corrected Pearson correlation coefficient (r'). As already mentioned, that is a measure of correlation accounting for the effect of genomic distance on the contact or distance matrices [96]. Indeed, pairs of DNA windows far apart along the locus tend to be more distant in space than windows genomically close. Such trivial trend may dominate the comparisons based on Pearson correlation, hiding the contributes from more interesting patterns, while it is removed by the r' procedure described above.

We computed the r' correlations between all pairs of experimental single-molecule distance matrices (*exp-exp distribution*) and all pairs of model single-molecule distance matrices (*model-model distribution*); additionally, we extracted the distribution of r' correlations between pairs of model and experimental distance maps (*mod-exp distribution*). Finally, we found the r' correlations for a random control case, i.e., between pairs of bootstrapped experimental distance matrices. The plot of all these distributions is shown in **Figure 2.7e.** The distributions remarkably overlap (more than with the control case) and, specifically, the *exp-exp* and *mod-exp* distributions can be considered undistinguishable (Mann-Whitney p-value: 0.19). That means the model matrices exhibit the same degree of variability of the imaged maps, or, in other words, that the model distance patterns are overall undistinguishable from the experimental analogues. The low average correlation values for all three distributions (r'=0.21, 0.27 and 0.36 for, respectively, the *exp-exp, mod-exp* and *mod-mod* distributions) indicate the differences among the distance maps are relevant in all cases.

3) To demonstrate that the SBS phase-separated conformations are a *bona-fide* representation of the imaged configurations of the HCT116 locus, we employed the RMSD method [27, 75], which is a criterion to compare a pair of structures: two spatial configurations are centered and rotated until the root mean square deviation (RMSD) between their corresponding sites is minimized; then, the lower the minimum RMSD is the more similar the structures can be considered (two identical structures would yield zero RMSD).
We compared all the experimental structures of the HCT116 locus [32] with our model configurations, i.e., for each imaged structure, we found the model conformation returning the minimum RMSD. To get adimensional RMSDs, the coordinates of the experimental and model structures were first z-scored. To check that the distribution of minimal RMSD was significant, we tested it against a control case distribution, made of the RMSD between random pairs of imaged structures. The two distributions are clearly separated (Mann-Whitney p-value<10^{-4}, **Figure 2.7f**), so the experimental configurations are significantly associated to the model ones. We did the same on reverse, finding the minimum RMSD across the imaged structures for each model configuration: analogous results were obtained [27]. That proves our polymers in the globule phase are a good proxy of the actual conformations of the HCT116 locus in cells.

Summarizing, the analyses above show that the phase separation mechanism produces polymer 3D conformations which are a *bona-fide* description of those imaged by microscopy for the HCT116 locus [32]. Model architectures return the same ensemble properties (contacts frequencies and pairwise median distances) detected in experiments [32, 33], and, above all, exhibit as much variability as that observed for real structures in single cells [32]. Indeed, the experimental and model boundary probabilities, separation scores and correlations between single-molecule distance matrices are all compatible. In addition, the imaged conformations of the HCT116 locus are significantly associated to our polymer structures according to the RMSD criterion [75]. Ultimately, that indicates the variability of TADs across cells can be a consequence of the thermodynamic degeneracy in phase-separated polymers. In this framework, the average pattern of TADs is simply determined by the more frequent globules, which in turn are established by the arrangement of the binding sites.

Experimentally, it was observed that the most likely TADs are those with CTCF and cohesin binding sites at their boundaries (**Figure 2.7b**). The process whereby CTCF and cohesin designate the preferential boundaries is an open issue and could be related to a diffusive Loop Extrusion mechanism [60], compatible with the equilibrium scenario of phase separation.

2.4.4 A linear block-copolymer model cannot return the complexity of the imaged structures

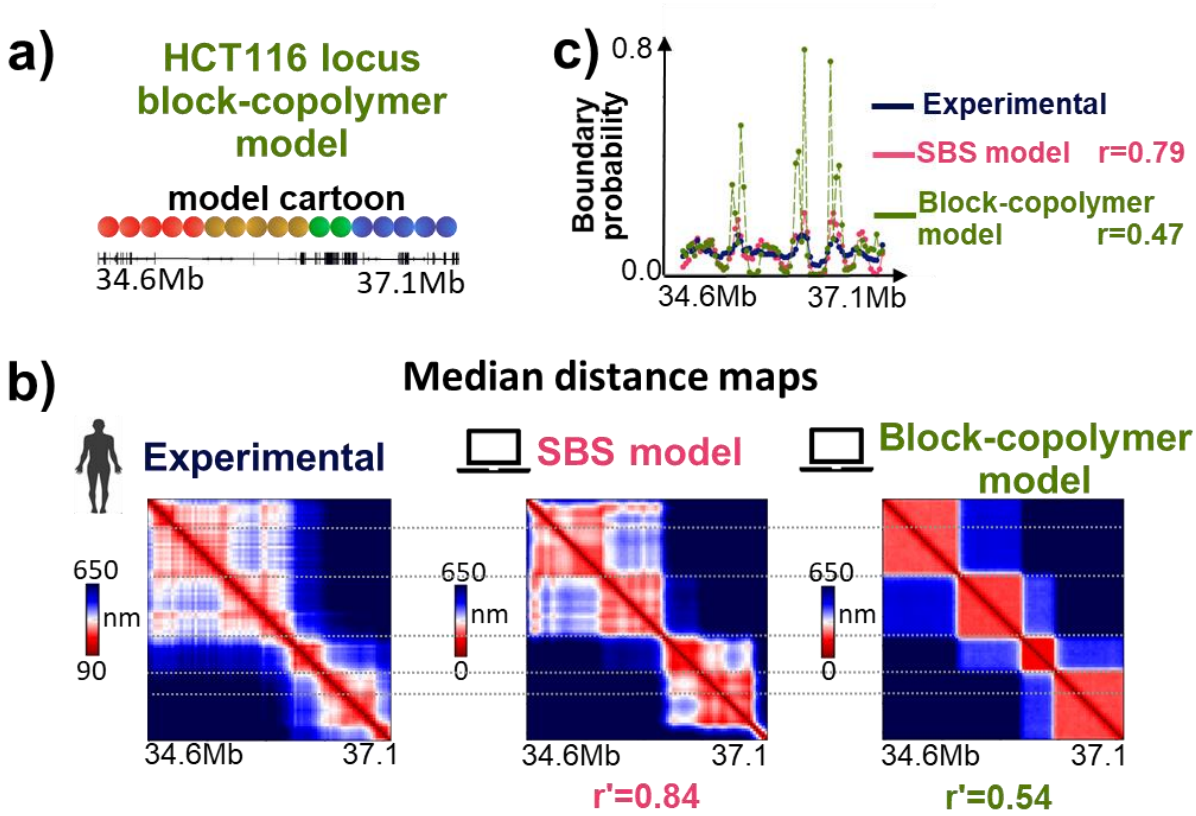


Figure 2.8: a) A block-copolymer model was designed to test whether it could reproduce the architectural features of the HCT116 locus as well as the SBS polymer model. The block-copolymer is made of 4 binding domains which follow each other along the chain of beads, without overlaps.
b) The experimental median distance matrix (left, from [32]) is poorly matched by that extracted from the ensemble of block-copolymers, compared to the SBS analogue. Indeed, the correlation between the experimental and block-copolymer maps is r'=0.54, whereas it is r'=0.84 in the SBS case.
c) The same scenario was observed studying the boundary probability. The SBS-derived boundary probabilities correlate much more with the experimental values than those from the block-copolymers. Indeed, the peaks in the block-copolymer model are about 4 times greater than those experimentally detected.
Adapted from [27].

We asked whether a simpler model than the described SBS system could reproduce equally well the architectural features of the HCT116 locus. To this aim, we prepared a linear block-copolymer, i.e. a polymer whose binding domains are separated along the polymer chain (rather than spread and intertwined according to PRISMR) and placed so to match the TADs and subTADs of the ensemble Hi-C matrix (**Errore. L'origine riferimento non è stata trovata.a**). We ran MD simulations as described above to drive an ensemble of those block-copolymers in the phase-separated state and

extracted the *in-silico* median distance matrix. The r' correlation with the experimental map [32] is 0.54, significantly less than the 0.84 returned by the SBS model matrix (Errore. L'origine riferimento non è stata trovata.**b**). Analogously, the boundary probability of the block-copolymer model poorly fits the experimental counterpart (r=0.47, see **Errore. L'origine riferimento non è stata trovata.c**). Importantly, for the block-copolymer, the peaks of the boundary probability are about 4 times higher than those detected by microscopy and those derived from the SBS picture: the phase-separation mechanism for neatly separated binding domains generates comparatively little degeneracy of conformations and so too stable boundaries across polymers. That accounts for the generally poorer description of the HCT116 locus architecture compared to the SBS framework. This shows that the complexity of the SBS polymer chain as inferred by PRISMR is necessary for an adequate description of the HCT116 locus spatial conformations.

2.5 The effects of cohesin depletion on chromatin conformations are explained by the globule-coil transition of polymers

2.5.1 Binding domains and MD simulations for the HCT116+Auxin locus

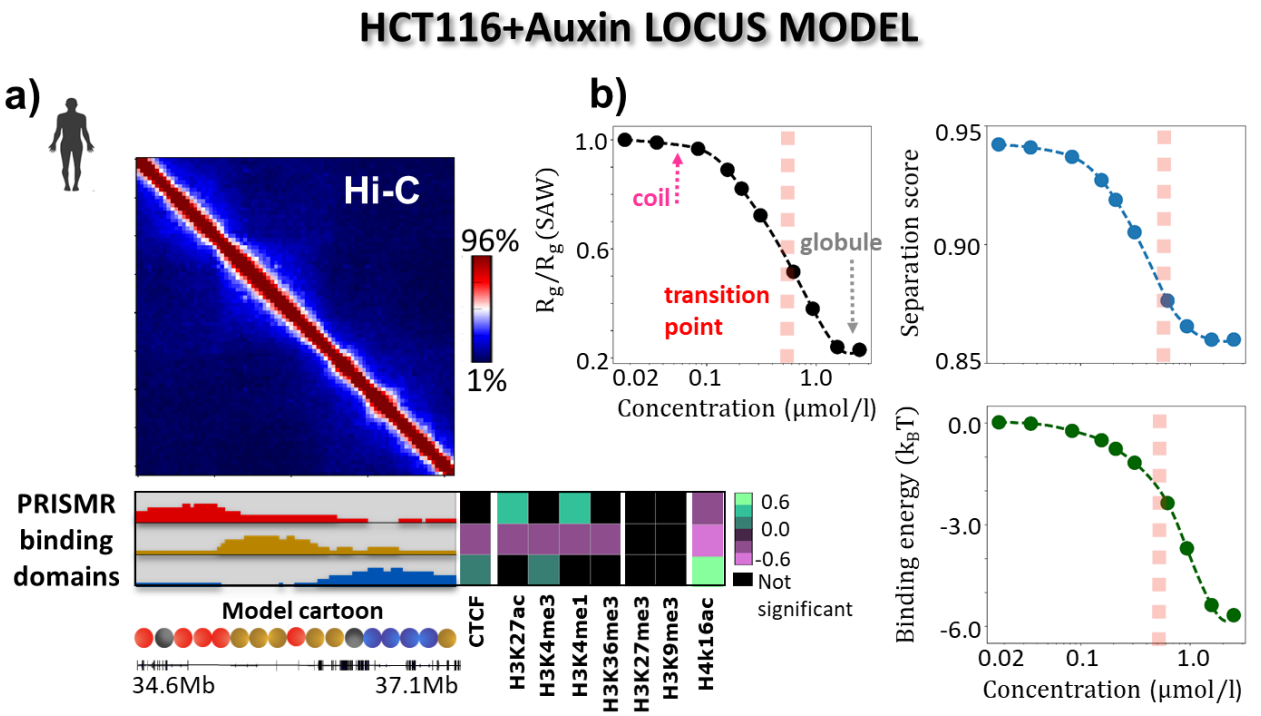


Figure 2.9: a) The Hi-C matrix of the HCT116+Auxin locus [33] is shown, with the colorbar indicating the percentiles. From such data, the optimal SBS polymer was inferred and 3 binding domains were found by PRISMR. Their arrangement along the polymer chain and their abundance in each 30kb window of the locus are reported. Additionally, the correlations with the epigenetic tracks from [33] are illustrated (those from the ENCODE database of **Figure 2.6** are not available for the Auxin treated case). Importantly, the green binding domain of the HCT116 locus model, strongly associated with CTCF and cohesin, is absent here (compare **Figure 2.4a** and **Figure 2.6**). Gene list is from the UCSC Genome Browser.
b) The order parameters of the SBS system (gyration radius, separation score and binding energy) signal the coil-globule transition as a sharp drop at approximately 400nm/l.

We now focus on the same HCT116 locus (chr21:34.6–37.1 Mb) when treated with Auxin and so depleted of cohesin (the *HCT116+Auxin locus*). We derived the best SBS model applying PRISMR on the corresponding Hi-C data at 30kb resolution [33]. The optimal polymer is made again of 830 beads (10 beads per each 30kb window) and, interestingly, presents only three types of binding domains (**Figure 2.9a**), against the four inferred for the wild-type HCT116 locus model. Specifically, we found that the HCT116+Auxin locus model lacks the green binding domain of the HCT116 locus, whereas the other three domains (red, brown and blue) maintain overall the same genomic positions across the two models, albeit in the Auxin case they appear weakened and shrunk (**Figure 2.9a**). Also, proceeding as in **paragraph 2.4.2**, they were found to maintain a similar epigenetic signature to the wild-type case (**Figure 2.9a**)**.** Notably, the green binding domain of the HCT116 locus was associated to CTCF and cohesin locations, so its disappearance in the model of the cohesin-depleted locus proves the consistency of PRISMR and supports the epigenetic interpretation of the binding domains.
The MD simulations were performed as for the wild-type locus, and, again, many concentrations of binders were explored to identify that allowing the coil-globule transition. Analyzing the three order parameters of the transition, we found that a concentration of 400nm/l accomplishes the phase change (**Figure 2.9b**). Hence, the definitive MD simulations were carried out with a 780nm/l concentration of binders, returning an ensemble of 1000 different 3D configurations.

2.5.2 The ensemble and single-cell features of the HCT116+Auxin locus are explained by a mixture of coil and globule polymers

The ensemble Hi-C matrix [33] used to infer the best SBS polymer displays no spatial features, as consequence of cohesin removal (**Figure 2.10a**). That is true also for the median distance matrix observed by microscopy [32] (**Figure 2.10b**). Importantly, we obtained *in-silico* Hi-C and median distance maps highly similar to the experimental matrices using a mixture of polymers in the coil and globule phase (**Figure 2.10a,b**). Specifically, the highest correlations between model and experimental matrices were found for 80% of the polymers taken in the coil phase and the remaining 20% in the phase-separated state (r=0.93, r'=0.33 and r=0.96, r'=0.57 for, respectively, Hi-C and distance data).
Consistently, performing the RMSD comparison (**paragraph 2.4.3**) to match the imaged 3D structures [32] with our polymer conformations, we found that about 80% of the experimental configurations was significantly associated to polymers in the coil phase and 20% to phase-separated polymers (**Figure 2.10f**). That can be interpreted as if the removal of cohesin reverts the thermodynamic phase of chromatin, changing it from globule to coil in most of the cells.
The imaged single-molecule distance matrices show TAD-like features (**Figure 2.10c**) despite the median distance matrix is featureless, as said in **paragraph 1.5.3**. Within our model, that can be explained by the random and rapidly changing domains that form in coil polymers as result of stochastic collisions (**Figure 2.10c**). The casual and unstable nature of such domains causes their disappearance when averaged over thousands of cells.
Next, using the mixed ensemble of polymer structures, we obtained boundary probabilities, separation scores and gyration radius distribution all compatible with the experimental ones. Indeed, the model boundary probabilities and separation scores show an almost flat behavior along the locus, as observed experimentally (**Figure 2.10b**). The flatness of the boundary probability track

also derives from the purely stochastic nature of coil polymer domains, which implies all genomic sites are equally likely to act as domain boundary (and analogously for the separation score).

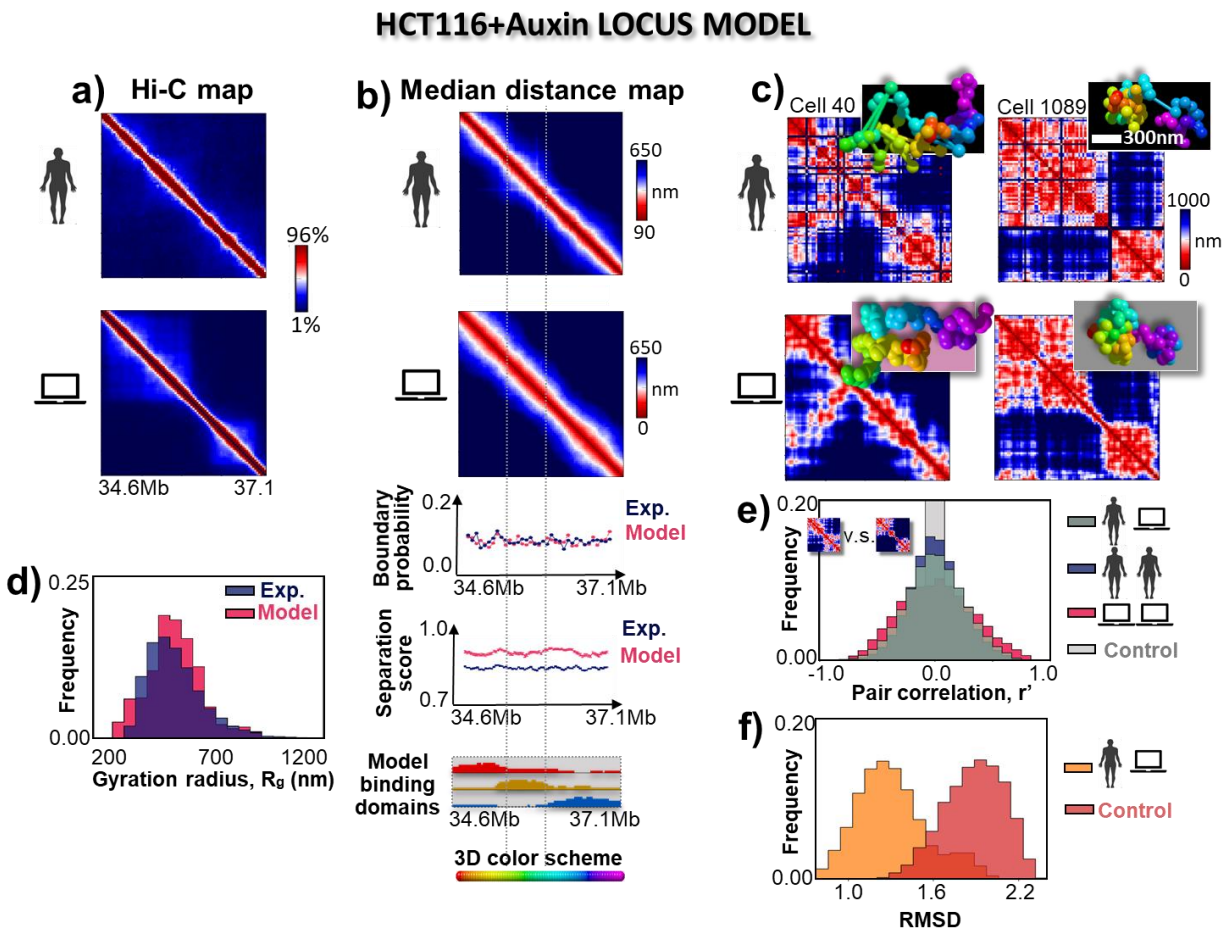


Figure 2.10: a) To model the HCT116+Auxin locus, a mixture of SBS conformations in the coil state and globule state was used (respectively 80% and 20%). As consistency check, the contact map derived from the model ensemble of SBS polymers (bottom) was compared against the experimental Hi-C matrix whereby the SBS model was inferred (top, [33]). The comparison is good (r=0.93 and r'=0.33), although the lack of pattern in the matrices reduces the Pearson distance-corrected correlation value. Colorbars give the percentiles of the matrices.

b) The median distance matrix observed in an independent microscopy experiment [32] is well reproduced by the model one (r=0.96 and r'=0.57). That supports the validity of our polymer structures in returning the average structural features of the HCT116+Auxin locus. On the bottom, the boundary probability and separation score obtained from microscopy data [32] are compared to those returned by the model. In both cases, the model reproduces the almost flat behavior of the experimental tracks (r=0.19 and r=0.41 for the boundary probability and separation score). The random deviations from the flat behavior account for the reduced correlations compared to the wild-type locus (**Figure 2.7b**).

c) Even though the experimental median distance matrix shows no pattern, the imaged single-molecule maps still exhibit variable TAD-like structures, as illustrated by the two examples in the top of the panel [32]. Strikingly, two single polymer distance maps extracted from the coil (left) and globule (right) phases display analogous features. Snapshots of the corresponding 3D structures are shown in all cases, colored according to the scheme on the bottom of panel b.

d) The gyration radius distribution for the imaged conformations is well matched by the model analogue (Mann-Whitney p-value = 0.10). The experimental average value is 540nm, suggesting cohesin depleted structures may be on average more open than in the wild-type condition, where the average gyration radius was 440nm (**Figure 2.7d**).
e) The *exp-exp*, *mod-mod*, *mod-exp* r' distributions (see **Figure 2.7e** and **Main Text**) are shown. They are well overlapped onto each other, more than the control distribution does. In particular, the *exp-exp* and *mod-exp* distributions are undistinguishable (Mann-Whitney p-value = 0.48). The average correlations are zero in all cases, indicating that the structural patterns for this locus are purely random.
f) The best RMSD match was sought for each imaged configuration [32] among the SBS structures in the coil and globule phases. 80% of the best match are in the coil phase, the remaining in the globule state, validating the model mixture. The resulting distribution of optimal RMSD is shown to be significantly different from a random control, made of RMSD between random pairs of experimental structures (Mann-Whitney p-value= 0.0). Analogous results were seen considering the best RMSD for the model conformations.
The results of panels b,c,d,e,f demonstrate that the real conformations of the HCT116+Auxin locus are well represented by an ensemble of almost all coil polymers. Adapted from [32].

The model and experimental gyration radius distributions are significantly overlapped (Mann-Whitney p-value: 0.10, **Figure 2.10d**) with an average value of 540nm in both cases (as above, the model mean gyration radius was equalized to the experimental one to define the length unit of the SBS model). Compared with the average 440nm found for the HCT116 locus, that suggests the cohesin depleted structures may be more open, consistent with the absence of stable globular domains.
We then studied the variability of the model and imaged structures computing the distance-corrected Pearson correlations (r') between all pairs of single-molecule distance maps (see **paragraph 2.4.3**). As done for the HCT116 locus, we computed the *exp-exp*, *mod-mod*, *mod-exp* distributions and verified their overlap is significant compared to a control case from bootstrapped pairs of distance matrices (**Figure 2.10e**). In particular, the *exp-exp* and *mod-exp* distributions are statistically undistinguishable (Mann-Whitney p-value: 0.48), showing that the model structures vary as the experimental configurations of the locus. Notably, the average r' here is zero for all three distributions, against the average correlations found in the HCT116 locus which were around 0.3 (**Figure 2.7e**). Also, the distributions for the cohesin depleted case are broader than those calculated for the wild-type locus. This indicates that the loss of cohesin makes conformational variability across cells higher, to the point that correlations average to zero. Within our model, that happens because of the purely random domains of coil state polymers, as opposed to the preferred globules formed in phase-separated structures which amount to a non-zero average correlation.
Summarizing, the average and single-cell properties of the HCT116+Auxin locus are returned by a mixed ensemble of SBS polymers, with 80% of them in the coil phase and 20% in the phase-separated state. Such a mixture was derived by associating imaged [32] and model configurations of the locus through the RMSD method [27, 75] (**Figure 2.10f**). We obtained *in-silico* Hi-C and median distance matrices well correlated with the corresponding experimental maps [32, 33] (**Figure 2.10a,b**). In addition, we found boundary probabilities, separation scores, gyration radius and correlation distributions matching those extracted from microscopy data [32] (**Figure 2.10b,d,e**). Overall, this shows that the cohesin loss can be interpreted as a drive toward the coil polymeric state: the highly variable TAD-like structures observed in single-molecule distance maps are then generated by the random collisions between segments of coil polymers, as opposed to the specific globules generated by the phase-separation mechanism. The stochasticity of the coil domains accounts for their vanishing when the average conformation of cohesin-depleted loci is detected.

Our analyses elucidate the impressive role that cohesin plays in determining the architecture of chromatin. While models as Loop Extrusion assign cohesin an active, energy-consuming role shaping directly the TADs, in the SBS picture cohesin acts as a thermodynamic switch between the coil and the phase-separated states of polymers. The exact mechanism whereby it may implement the phase switch remains an open question.

2.6 Single-molecule time dynamics

So far, we have employed our models to explain the experimental observations. In this section, we exploit the models to get insights which were not accessible to the considered experiments [32, 33]. For both the HCT116 and the HCT116+Auxin loci, we investigated the time dynamics of the model polymer structures, namely how such structures change with time at stationarity. We computed the r' correlations between single-molecule distance matrices of the same polymer at different timesteps, then averaged over all the available polymers (**Figure 2.11a**) and converted the timesteps in physical temporal units as explained in **paragraph 2.3.2**. For the HCT116+Auxin locus, the r' vs time behavior was computed for polymers in the coil state only and shows a rapid decay to zero.

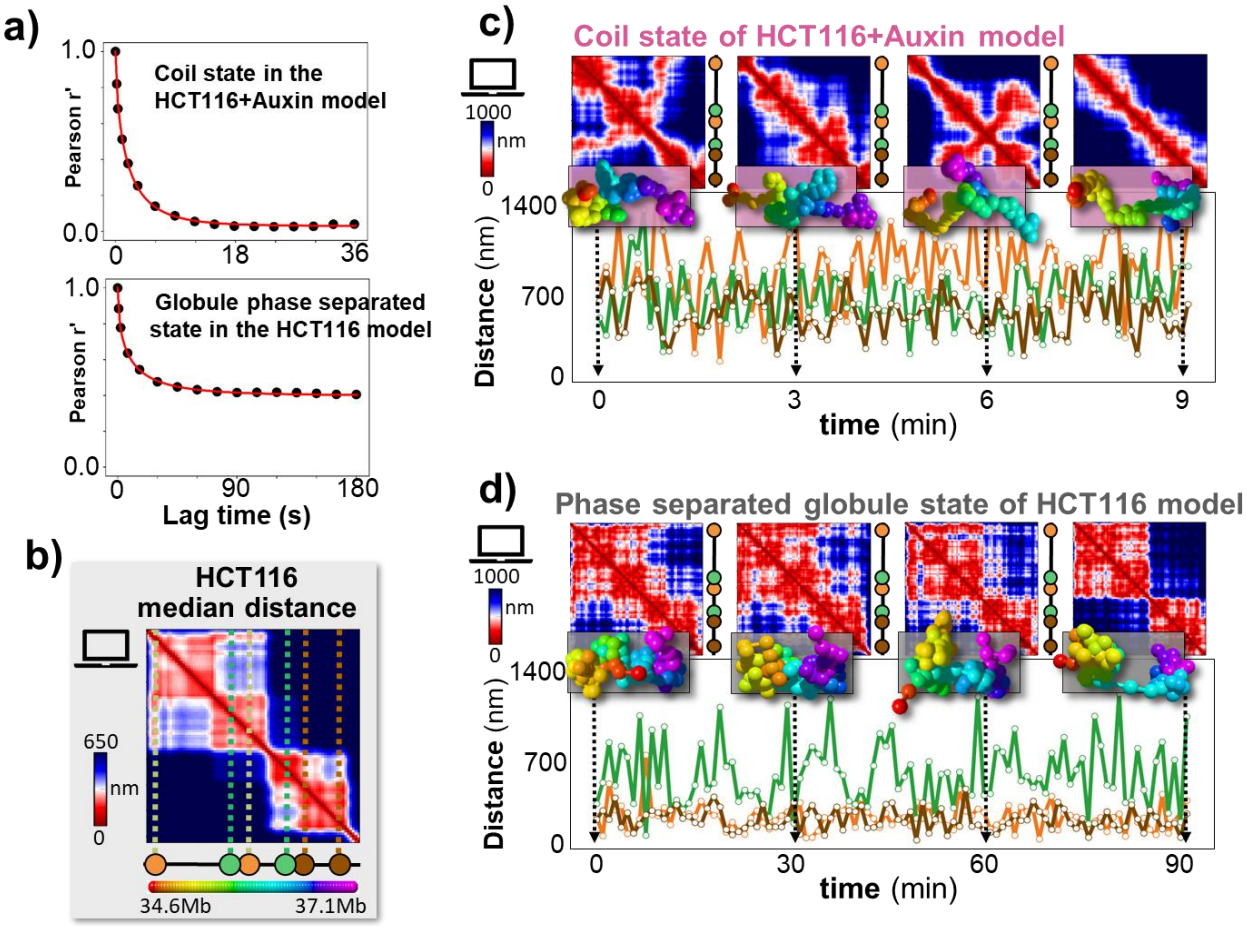


Figure 2.11: a) In the steady state, the distance map of a polymer structure at a given time was correlated to the same map at successive times along Molecular Dynamics. The time dependence was then mediated over several polymers. On top, we show the correlation evolution for the HCT116+Auxin model, using only conformations in the coil state. The r' correlation rapidly decays to zero, indicating that structural patterns randomly form and vanish in coil polymers, due to thermal agitation. The curve is fitted by a stretched

exponential (in red), with characteristic time of 9s. On the bottom, the same plot is shown for the HCT116 locus model, where polymers are all phase-separated. Here, the correlation decays to a not-null plateau (r'=0.39) with a characteristic time of 60s, fitted from a stretched exponential. That is consistent with the fact that in globule polymers self-interacting domains assemble and persist in time due to the abundance of interacting binding sites. The plateau correlation values are consistent with the findings of **Figures 2.7e, 2.10e**.
b) From the model median distance matrix of the HCT116 locus, three pairs of windows were selected: a pair of sites (orange), 1.2 Mb apart, in different subTADs, but same TAD; a pair of 0.6 Mb distant sites (green) with a TAD boundary in between; a pair of sites (brown), almost 0.6 Mb apart and within the same subTAD.
c) The mutual distances for each of those pairs of sites were computed at different times for a coil polymer. For all three pairs a similar temporal behavior was found, with wild oscillations and analogous average values (620nm for the green and brown pairs, 660nm for the orange pair), because none of them is involved in stable interactions and their distances fluctuate due to thermal effects. Snapshots of the polymer distance maps and of the 3D structures are shown at 4 time points. The color code for the 3D structures is described on the bottom of panel b).
d) Same as panel c), but for a phase-separated globule polymer. Here, the brown and orange mutual distances have little fluctuations, with mean values 2.5 times lower than those for the coil polymer (∼280nm). Indeed, in the median distance map, these two pairs of sites are confined in a TAD and a subTAD (panel b). Conversely, the distances of the green pair vary in time as in the coil polymer, because they are on average prevented from interacting by a TAD boundary and their physical proximity is determined only by stochasticity, as in the coil case.
Adapted from [27].

Conversely, in the HCT116 locus case, the r' was calculated for globule phase-separated polymers and the plateau for large lag time is well above zero (r'=0.39). That is in line with the correlations found between pairs of single-molecule distance maps at the same time (*mod-mod* distributions in **Figures 2.7e, 2.10e**). The decay time for the HCT116+Auxin locus is about one order of magnitude smaller than that found for the HCT116 case (9s vs 60s). That roots in the differences between the coil and the globule phases. Indeed, at stationarity, in both cases the 3D conformations fluctuate and breathe according to thermal oscillations, but in the phase-separated state self-interacting globules tend to persist in time, accounting for correlations asymptotically higher than zero and longer decay time. The persistence in time derives from the abundance of cognate binding sites in interaction, compensating for the weak energy affinities which are easily contrasted by thermal agitation. Conversely, in the coil phase contacts and interaction domains are transient, as the small number of binders create few interactions, which are rapidly dismantled by the viscous bath. That determines the average correlation plateauing to zero in shorter time.
Next, we focused on the dynamics of specific pairs of windows. From the model median distance matrix of the HCT116 locus we selected three pairs of sites (**Figure 2.11b**): two sites 1.2Mb apart from each other in different subTADs but same TAD (orange); a pair of 0.6Mb distant sites (green) with a strong median TAD boundary in between; a couple of windows (brown), approximately 0.6Mb apart, segregated in the same subTAD. For all three pairs of windows, we derived the track of their mutual distances across time for a polymer in the coil state and another in the globule state (9min and 90min were explored respectively for the coil and globule case, **Figure 2.11c,d**). In the coil structure, the tracks appear similar for all three pairs (**Figure 2.11c**), with important fluctuations and analogous average values dictated by the linear genomic separation of sites (for instance, the mean distance is 620nm for the green and brown pairs which are both 0.6Mb apart, while it is 660nm for the orange couple of sites, 1.2Mb apart). In the globule case, the tracks become significantly different (**Figure 2.11d**). In particular, the fluctuations of the orange and brown pairs

are much smaller, with average values decreased of a factor 2.5 (∼280nm for both the pairs of sites). That indicates each of the two couples of sites is overall enclosed in the same globule all the time, with thermic oscillations allowing for the moderate fluctuations of the reciprocal distances. Indeed, on the median distance map, both the pairs belong to the same TAD or subTAD. On the other hand, the track of the green pair of sites is similar to that obtained in the coil case, namely the physical distance between the two green sites is still determined by the bare genomic separation. The wild oscillations and the greater average value suggest they rarely come into interaction or that interactions immediately dissolve. Indeed, the green sites are on average separated by a TAD boundary.

In summary, the analysis of the time dynamics of the SBS structures highlighted that globule polymers form persistent domains of interactions, determining greater than zero correlations between the same structures even at very large lag time. The persistence in time of phase-separated globules is explained by the abundance of interacting binding sites, compensating for the weak energy affinities. In the coil state, contacts are fleeting and highly transient in time, as binders are too few and the weak energy of interaction cannot prevail over thermal fluctuations. That results in correlations between structures rapidly decaying to zero with time.

3 COMPARING Hi-C, SPRITE AND GAM TECHNOLOGIES THROUGH POLYMER MODELS OF CHROMATIN

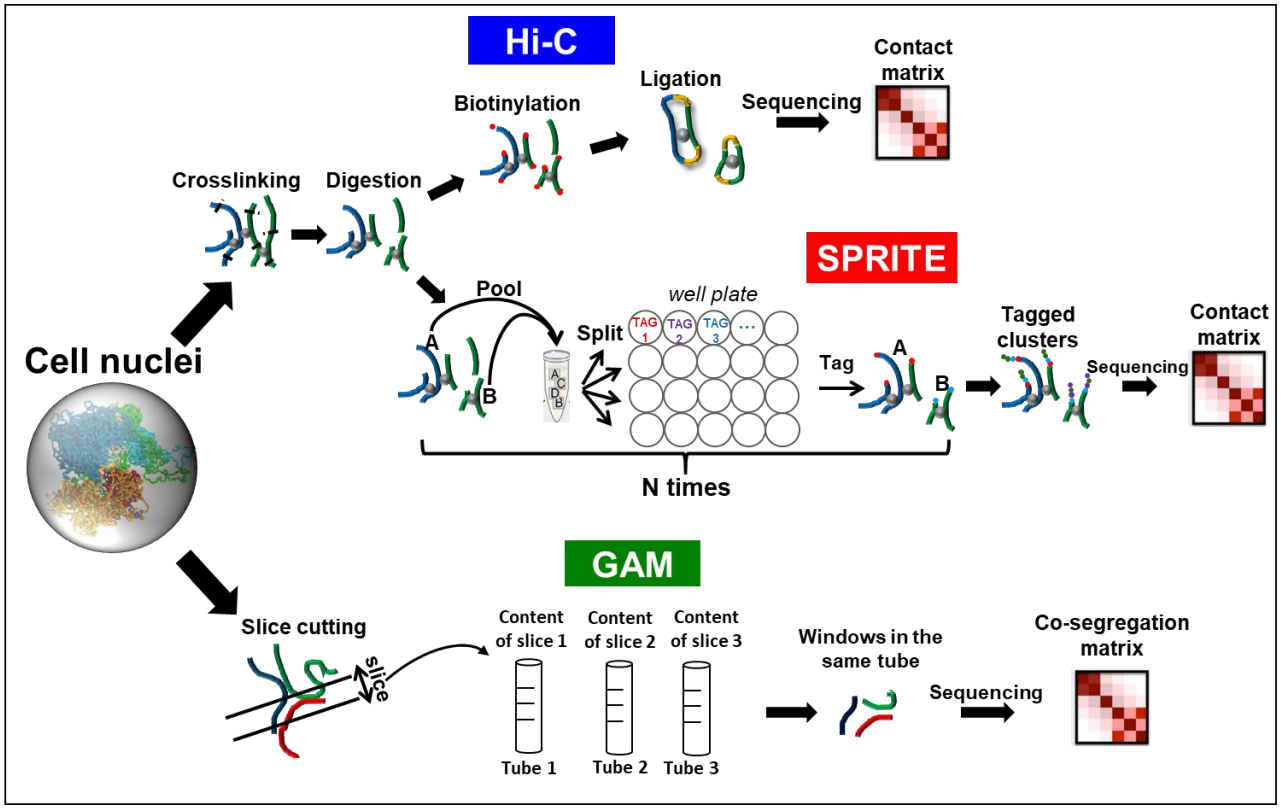


Figure 3.1: Outlook of the main steps of Hi-C [7], SPRITE [9] and GAM [8] experiments.

We have seen how polymer-physics models can make sense of the architectural features of chromatin. As said in **Chapter 1**, many of those striking features have been detected by sequencing technologies (Hi-C, GAM and SPRITE, for example), which up to now represent the most powerful instrument to investigate DNA 3D organization at genome wide level [110]. For this reason, it is crucial that sequencing methods unbiasedly detect the real conformations of chromatin. In principle, that could be argued. Indeed, the protocol of any of those technologies could return a biased visualization of DNA architecture and could reveal only a restricted type of features. Moreover, different sequencing technologies provide diverse measures of chromatin spatial organization and it is not clear how they perform relative to each other. For instance, is Hi-C faithful to the underlying conformation of chromatin as well as GAM? Are Hi-C, SPRITE and GAM equally effective in detecting long and short-range interactions between DNA sites? Rigorous answers to those issues are missing, because, clearly, no benchmark exists.

In this chapter we will illustrate how polymer models can be employed to evaluate the performances of Hi-C [7], SPRITE [9] and GAM [8] sequencing technologies. In a nutshell, Hi-C, SPRITE and GAM experiments are simulated over ensembles of SBS polymer 3D configurations. As the architecture of the model structures is known, they represent a fully controlled reference system to benchmark those technologies, enabling the first quantitative (albeit simplified) assessment of the absolute and relative capacities of Hi-C, SPRITE and GAM. Such investigation can help the design of novel

experiments and, in perspective, shows that polymer-physics may play a role in the setup of experimental studies.
In the first section, we will review the steps of the Hi-C, SPRITE and GAM protocols and describe how they were implemented in simulations over polymer models. Then, in paragraph 3.2, we will discuss whether the simulated technologies are representative of the actual experimental performances and show they are effective in such respect. Hence, in paragraph 3.3 we will assess how faithfully each technology detect the known polymer conformations; how many 3D structures (i.e. cells) are required to have statistically reproducible outputs; how strong the impact of the detection efficiencies is; how the noise-to-signal ratio of the output matrices scales in different conditions. The contents of this chapter are contained in a paper currently under review by *Nature Methods* and most of them are deposited in a BiorXiv preprint [36].

3.1 Review of the Hi-C, SPRITE and GAM technologies

3.1.1 Hi-C

We already described the main steps of a Hi-C experiment in **Chapter 1** (**paragraph 1.3.2**). Here, we recall them adding further details necessary for the comprehension of the next sections.

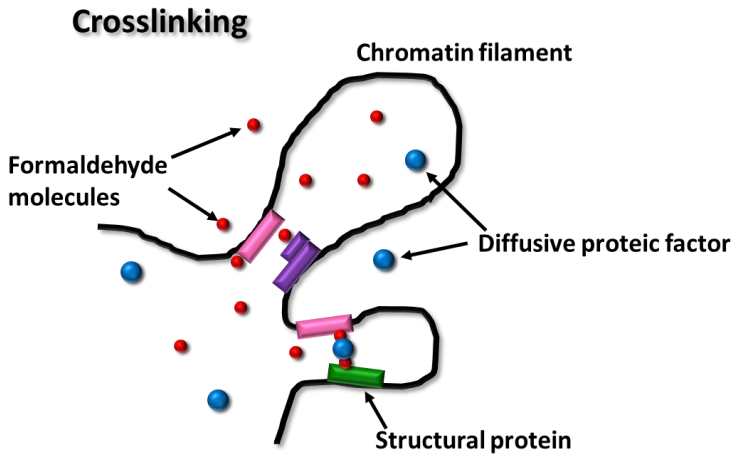


Figure 3.2: Cartoon of the crosslinking process of chromatin. Proteic structural complexes (rectangles) anchored on chromatin filaments (black line) or proteic diffusive factors (blue spheres) offer multiple binding sites for formaldehyde molecules (red dots). Those bind covalently and so indirectly weld together chromatin sites proximal in space.

In the original Hi-C protocol [7] (**Figure 3.1**), formaldehyde is injected in a population of cell nuclei. Formaldehyde molecules can bind to proteic factors or to DNA with a covalent bond [111, 112]. Proteic complexes can have many binding sites for formaldehyde. Since, in chromatin, proteic molecules are anchored all along to DNA filaments or diffuse closely around them (e.g. transcription factors and architectural factors), formaldehyde bonds with proteins result in indirect linkages between DNA sites which are close in space (**Figure 3.2**). Importantly, two DNA segments can come physically close because they directly interact (e.g. enhancers with the target promoters) or because they are segregated inside the same domain or, simply, by chance (e.g. two sites which have short genomic separation are likely to stay close regardless of reciprocal attractions). Hence, a pair of DNA

sites near to each other constitutes, generically, a *contact*, be it random or driven by interaction. Formaldehyde, by linking DNA segments close to each other, defines the contacts and freezes them in a moment in time thanks to the covalent nature of its bondages. Such a process is known as the *crosslinking* of chromatin (**Figure 3.2**).
Next, the nuclei membranes are disrupted and the crosslinked genomes of all the nuclei are digested by restriction enzymes. They are proteic machineries able to cut DNA at specific base sequences, so, eventually, their action reduces chromatin in fragments. According to the type of restriction enzyme used, the median length of a DNA fragment can range from hundreds of bp to units of kb [72]. This is the *digestion* step of the Hi-C experiments.
Chromatin from all the considered nuclei is now composed of crosslinked fragments of DNA. The loose ends of those fragments are subject to *biotinylation*, i.e. they are filled with biotinylated nucleotides. That is done to allow the *ligation* process: DNA-ligase enzymes bind together the biotinylated ends of crosslinked fragments, generating hybrid molecules (the ligation products) made of pairs of DNA pieces which were in contact in their original nucleus (**Figure 3.1**).
Formaldehyde is then removed, all the fragments of DNA are sequenced and those forming a ligation product are counted as in contact. To rationalize the contact information, genome is partitioned in windows of equal size and the number of contacts between every possible pair of windows is derived. The size of the windows defines the resolution of the Hi-C experiment and is chosen based on the number of fragments collected and their median length. Finally, the pairwise contact data are arranged in the format of a matrix, the Hi-C *contact matrix*.
As said, this is the original version of Hi-C [7], also known as *dilution Hi-C.* That is because, after nuclei are disrupted, chromatin is held in a highly diluted solution to avoid that different crosslinked clusters of fragments (possibly from different nuclei) come too close to each other, allowing for spurious ligations. To eradicate such a risk, a new version of Hi-C was implemented, the *in-situ* Hi-C [10]. Here, the chain of steps crosslinking-digestion-biotinylation-ligation is conducted separately in each nucleus and only then the nuclear membranes are lysed and the contacts detected across all the population genomes. Successive variants of the protocol, as the *Micro-C* [73] or the *Low-C* [74], use different types or concentrations of solvent and reagents, or invert the order of specific steps utilized in the *in-situ* Hi-C, yet the core chain of crosslinking-digestion-biotinylation-ligation is overall preserved.
Additionally, as mentioned in **Chapter 1**, *single-cell* Hi-C experiments have been realized, which reveal the contact matrix for a single cellular nucleus. Many variants of single-cell Hi-C presently exists [72, 75–77], however, basically, they all implement the chain of key steps inside a nucleus, isolate the nucleus and only then disrupt the membrane. The capacity to isolate the target nucleus makes the topic difference with the *in-situ* Hi-C.

3.1.2 SPRITE

All the variants of Hi-C rely on the final ligation process. Ligation only permits the identification of contacts between two DNA fragments, while multiple contacts are also expected to exist. That can be considered the crucial methodological limitation of the Hi-C techniques. SPRITE [9] and GAM [8] technologies stemmed mainly from that limitation and were proposed as ligation-free alternatives. Here, we will describe the SPRITE method, while the next paragraph will be focused on GAM.
SPRITE is the acronyms for Split-Pool Recognition of Interactions by Tag Extension. The initial step is crosslinking the chromatin of a population of nuclei, as in Hi-C. In each nucleus, DNA is fragmented

first by sonication and then by digestion through the DNAse restriction enzyme, returning a collection of crosslinked fragments of approximately 150-1000bp [9]. Nuclei membranes are lysed, and chromatin from the entire cell population results in the form of different clusters of crosslinked fragments. That is in line with the digestion process in *in-situ* Hi-C.
Next, the *split-pool tagging* procedure is realized, which is the core of SPRITE (**Figure 3.1**). All the complexes of crosslinked DNA are randomly split among 96 wells of a plate. A sequence of nucleotides is added to every DNA fragment inside a single well (fragments are *tagged*) and, crucially, the sequence is specific for each of the 96 well. The complexes of DNA are re-collected from the plate, pooled together, again randomly shuffled inside the wells and tagged. That is repeated several times, so that in the end all fragments have a barcode made up of the sequence of tags assigned at each split-pool round. Fragments belonging to the same crosslinked cluster are identically barcoded, as they are covalently bound to stay together during the shuffling in the wells. So, when formaldehyde is washed away and DNA fragments are sequenced, the barcodes uniquely identify fragments which were in the same cluster and which, thereby, can be considered in contact with each other. So, pairwise, threewise and, in general, n-wise contacts can be counted for a fixed resolution (defined as in Hi-C). In the pairwise case, a contact is counted for each possible pair of fragments in the same cluster. Then, the size of the clusters is accounted for by dividing each count for $n/2$, where n is the number of fragments composing a complex: this prevents bigger clusters from dominating the contact information. Finally, the normalized counts from all clusters are summed up and arranged in a 2-dimensional matrix.
The number of split-pool tag rounds must be enough to ensure that two different complexes cannot have the same barcode by chance. In the original work [9], 6 rounds were estimated sufficient, as in that way the number of possible barcodes ($\sim 10^{12}$) largely exceeded the number of unique DNA fragments expected from the murine genome ($\sim 10^{9}$).

3.1.3 GAM

The Genome Architecture Mapping (GAM) technology [8] freezes a population of cells embedded in a sucrose solution. Like crosslinking, that stops chromatin architecture in a moment in time and, importantly, fixes the position of nuclei. Then, a thin slice (~220nm) is laser-cut at random orientation from each frozen nucleus; the genomic content from each slice (also called *nuclear profile* or *NP*) is deposited in a tube and sequenced (**Figure 3.1**). That reveals the DNA segments contained in each nuclear profile, which are thus said to have *segregated* in the NP.
As in Hi-C or SPRITE, the genome is visualized as divided into windows of fixed length, which is the resolution of the experiment. Hence, for each tube, all possible DNA windows are assigned 1 if present in that tube, zero otherwise. That constitutes the *segregation table* of a GAM experiment. From the segregation table it is straightforward to extract the segregation frequencies for each window or, notably, the frequencies whereby each possible pair, triplet and, generally, n-plet of windows segregated together in the same slice. These are named *co-segregation frequencies* and, in the pairwise case, are typically arranged in a 2-dimensional co-segregation matrix [8]. Co-segregation frequencies provide a measure of chromatin architecture because windows close in nuclear space are expected to co-segregate much more than far apart windows.
The counting of co-segregating windows is ultimately based on the simple statistics of successful events in N tubes, that allowed the implementation of the SLICE (StatisticaL Inference of Co-sEgregation) algorithm [8]. Based on mathematical modeling and statistics, SLICE infers from GAM

segregation data the probabilities whereby two or more windows interact in a cell. Hence, of all the windows detected close to each other in the nuclear space, only those deriving from real interactions can be identified.

3.1.4 Hi-C, SPRITE and GAM data are difficult to compare

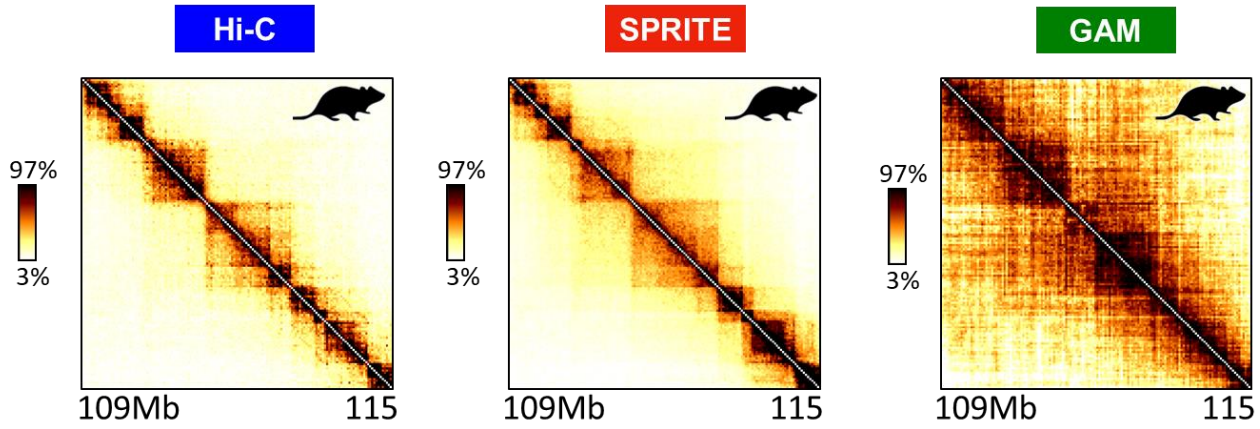


Figure 3.3: The contact matrices obtained from Hi-C, SPRITE and GAM experiments can exhibit differential information on chromatin architecture. Here, maps are shown from the same murine locus (chr11:109-115Mb) at 40kb resolution, from analogous cell lines (staminal mouse cells) [9, 11, 36]. The TAD pattern is overall similar for all three matrices, yet GAM co-segregation frequencies are much more enriched off-diagonal (i.e. for windows far apart along the linear genome) than contacts in Hi-C and SPRITE. Colorbars indicate percentiles of the maps. All three matrices have been normalized to remove technical bias [9, 11, 36].

Hi-C, SPRITE and GAM provide different measures of chromatin architecture. Focusing on the pairwise level, Hi-C and SPRITE produce both a contact matrix, but the contacts are detected differently, i.e. as number of ligation products (the former) and as number of pairs in crosslinked clusters (the latter). On the other hand, GAM generates a co-segregation matrix.
SPRITE contact matrices from mouse cells showed common architectural features to those from Hi-C maps, e.g. compartments, TADs, loops and so on. Also, the location of TAD boundaries appears conserved between the two methods [9] (**Figure 3.3).** GAM co-segregation matrices in mouse returned patterns which differ from those of Hi-C especially when long-range contacts are observed (**Figure 3.3**). Specifically, GAM shows quite relevant co-segregation frequencies between windows far apart along the genome, while Hi-C contacts between such pairs are usually negligible. Are GAM long-range co-segregations originated from noise? Or are they spatial features of chromatin that Hi-C fails to detect? On the contrary, TADs are overall conserved between GAM and Hi-C [8], albeit more differences exist than those between SPRITE and Hi-C.
To investigate further the differences or similarities between Hi-C, SPRITE and GAM outcomes, huge resources would be required, as each kind of experiment should be conducted many times on the same sample of cells to acquire statistically relevant results. Moreover, as mentioned, a rigorous assessment is difficult without an independent benchmark, as there is no first principle motivation to use, e.g., Hi-C to benchmark SPRITE rather than vice versa (microscopy methods are still unable to detect 3D information genome wide). In the following, we will thus employ a computational approach.

3.2 Modeling Hi-C, SPRITE and GAM technologies

We will describe "computational experiments" where Hi-C [7], SPRITE [9] and GAM [8] methods are simulated over 3D configurations of polymer models. We will be concerned only about pairwise architectural information, as Hi-C is limited to that and because it is still the most explored level of spatial organization. Additionally, we will focus on the applications of the technologies over mammal cells. Here, the genome is diploid, i.e. two copies of each chromosome are present (the *alleles*). So, in general, the architectural information about a specific DNA locus in a cell derives from two allelic conformations. This means, for instance, that *single-cell* Hi-C experiments reveals the average contact pattern of two alleles. Variants of the experimental protocols able to distinguish the alleles in cells (*phased experiments*) will not be considered in the present work.

We prepared algorithms that take the spatial coordinates of a polymer structure in input, simulate the main steps of the Hi-C, SPRITE and GAM technologies and return their outputs. The polymer models employed for the study are the SBS models of specific DNA regions (**paragraph 2.1**), but any kind of polymer 3D structures could have been used as well. To account for the mammal diploidy, in our computational experiments a cell is represented by a pair of different polymer structures, hereby called an *in-silico* cell. The Hi-C, SPRITE and GAM algorithms are devised to work on a single *in-silico* cell at a time, so, to simulate experiments carried over a population of cells, iterations are conducted.

We dedicate this section to illustrating how we modelled each experimental technique. The following descriptions are taken from [36], with some adaptations where necessary.

3.2.1 Simulating Hi-C

For *in-silico* Hi-C experiments, we implemented a proxy of the key steps of a Hi-C protocol, i.e. crosslinking, digestion, biotinylation, ligation and contact matrix generation. As said, those steps are conducted in every *in-silico* cell, as in real *in-situ* Hi-C [10]. Since in computer simulations we always have control of the *in-silico* cell studied, we can also simulate single-cell Hi-C experiments [72, 75–77, 113].

Crosslinking - During real Hi-C crosslinking, DNA contacting sites are bound together with formaldehyde to fix the overall 3D structure. Formaldehyde binds to DNA-protein complexes, and consequently fixes DNA sites through protein bridges (**paragraph 3.1.1**).

In our SBS polymers only cognate binding sites can interact with each other through a binder and only if they are closer than a threshold distance, d, fixed by the interaction energy cutoff (**paragraph 2.1** and in **paragraph 2.3.1, equation 4**). So, in *in-silico* Hi-C, we identify as "crosslinked" those beads which are of the same type and are closer than d. Note that beads realizing such condition are not necessarily interacting, as they could be near to each other without a cognate binder bridging the interaction: as in Hi-C, we generically derive contacts between polymer beads. This is done with an efficiency p_c, simulating the efficiency whereby formaldehyde creates covalent bonds. To identify the sets of crosslinked beads, a customized version of the DBSCAN clustering algorithm [114] is employed.

Digestion - After crosslinking, DNA is digested (**paragraph 3.1.1**). In standard Hi-C experiments, digestion fragments have a median length in the range from few hundreds of bp to some kb, depending on the restriction enzyme used [72].

Given the SBS polymer model of a genomic locus, the genomic length of each bead is estimated dividing the entire length of the locus by the number of beads. Importantly, in all the models considered for the present study the genomic lengths of beads were found comparable to the digestion fragment typical sizes. Therefore, single beads well represent the digestion fragments and digestion is implemented by splitting the polymer chain into its own single beads. That results in a set of different clusters consisting of crosslinked beads.

Biotinylation - The next step in Hi-C is biotinylation, where DNA fragments in each crosslinked complex are marked with biotin. Unmarked fragments cannot be ligated and, so, detected.
In our algorithm that is implemented by removing beads from their clusters with probability 1-p_b, modeling the efficiency of the biotinylation process.

Ligation - In Hi-C, crosslinked and biotinylated pairs of DNA fragments are randomly linked together. In our algorithm, we randomly select pairs of beads from the same crosslinked cluster within the above threshold distance d and call them ligated. To account for the experimental ligation efficiency, each selected bead is ligated only with a probability p_l, otherwise is discarded.

Contact matrix generation - Next in Hi-C, ligated fragments are sequenced and a contact is counted between their corresponding windows, defined by the resolution of the experiment. Eventually, a Nwindow x Nwindow contact matrix is produced.
Similarly, in our algorithm, we produce a 2-dimensional Nwindow x Nwindow matrix. For each polymer structure in input, "ligated" beads are counted as a contact with a given detection probability p_d - modeling the sequencing efficiency of real experiments - and their corresponding matrix entry is incremented by 1. Note that in general several beads compose a window (see **paragraph 2.2**, the r parameter of PRISMR), so each of them can contribute incrementing the same matrix entry. The procedure is iterated over the N simulated cells, and the final *in-silico* matrix yields the total count of contacts between each possible pair of windows.

3.2.2 Simulating SPRITE

For SPRITE, the main steps of its protocol were considered, i.e. crosslinking, digestion, split-pool tagging and contact matrix generation.

Crosslinking - In SPRITE experiments, crosslinking is carried out as in Hi-C, so the same procedure described above for our *in-silico* Hi-C is employed.

Digestion - After crosslinking, in SPRITE experiments DNA is fragmented first by sonication and then by DNAse digestion, resulting in a collection of crosslinked fragments of approximately 150-1000bp, similarly to the restriction fragments produced by digestion in Hi-C (**paragraph 3.2.1**). Hence, we implement SPRITE chromatin digestion as in *in-silico* Hi-C.

Split-pool tagging - The split-pool tagging procedure allows to identify DNA fragments belonging to the same crosslinked cluster. In our *in-silico* procedure, the beads composing a given cluster are known, so an explicit split-pool tagging implementation is not required. However, since in real

experiments some fragments may not be tagged successfully, we remove beads from their clusters with a probability 1-p_s.

Contact matrix production - Experimentally, fragments with the same barcode are sequenced and assigned to their corresponding genomic windows. Then, a contact is counted for every possible pair of windows associated to the same cluster. The count is divided by a corrective factor $n/2$ accounting for the cluster size.
In-silico, each fragment is represented by a polymer bead. All beads of a cluster are assigned to their corresponding windows and a contact is counted for every window pair. Each bead is detected only with a given probability, p_d, modeling the sequencing efficiency. Contact counts from every cluster are divided by the $n/2$ factor, then summed across all the *in-silico* cells and finally collected in a Nwindow x Nwindow matrix.

3.2.3 Simulating GAM

In GAM experiments, a nuclear slice is extracted at random orientation from each nucleus of a cellular population, the DNA windows from each slice are sequenced and their co-occurrence across all slices measured to construct a GAM co-segregation matrix.

Slice cutting - We model a cell nucleus as a sphere containing two different, randomly placed polymer structures of the locus of interest, accounting for diploidy. For each *in-silico* cell, we generate a randomly oriented slice passing through the sphere and all the polymer beads inside it are counted as co-segregating. The simulated slices can happen to be empty of polymer beads, as in a real GAM experiment cellular slices only contain a fraction of the nuclear volume and could miss the locus of interest. To account for the experimental sequencing efficiency, beads inside a simulated slice are counted only with a certain probability p_d.

Co-segregation matrix production - In GAM, windows found in the same slice are counted as co-segregating. Co-segregation frequencies are then arranged in a 2-dimensional Nwindow x Nwindow matrix.
Similarly, in our algorithm we build a 2-dimensional Nwindow x Nwindow sized matrix. Each bead segregated in a slice is assigned to its corresponding window, all the possible pairs of windows found in the same slice are counted and the counts are added to the corresponding entries in the co-segregation matrix. We finally normalize the matrix by the number of slices employed to obtain co-segregation frequencies. For simplicity, later in the text we will use the expression *contact matrix* also to indicate the GAM co-segregation matrix, unless otherwise required.

3.2.4 Setting parameters in the algorithms

For the algorithms to run, some parameters need to be set. In *in-silico* Hi-C and SPRITE, the distance within which crosslinking is allowed is put equal to 2 times the cutoff distance of the Lennard-Jones potential (**equation 4**). That is used also as threshold distance for the simulated ligation. In *in-silico* GAM, the cellular nuclei are represented by identical spheres, so their radius must be fixed. In SBS polymers the length unit is the diameter of a bead, σ, and, as explained in **paragraph 2.3.2**, its physical value can be deduced according to the modelled DNA locus. For all the models of loci considered in the following, the value of σ was computed using 70bp/nm as estimate for the

average chromatin compaction (**paragraph 2.3.2**), which is an average value between the 30nm fiber and the naked DNA [28, 115]. Hence, in units of σ, the sphere radius was set to match the experimental estimates of nuclear size in the cell line of each locus [8, 116, 117]. We set analogously the thickness of simulated slices: as in real GAM experiments slices are $\sim$220nm thick, we used the corresponding value in units of σ.

In Hi-C, SPRITE and GAM algorithms every process presents a finite efficiency, i.e. a probability of success. By construction, each step is independent from the other, so the total efficiency ε is given by the product of all single-step efficiencies and the average output of each algorithm is expected to depend only on ε, irrespective of the values for each single step. So, in the following sections, we will discuss only the total efficiency of the algorithms, omitting the adjective "total" unless necessary. Importantly, that is enough to compare with real experimental conditions, because, in experiments, the efficiency of the whole procedure is typically estimated. For instance, in Hi-C, the number of ligation products detected is compared to those expected from an entire genome [72], returning an estimation of global efficiency of the protocol.

We will illustrate the performances of the *in-silico* technologies for several efficiency values, with a specific focus for those typically used in real experiments.

3.3 Comparing Hi-C, SPRITE and GAM technologies in computational experiments

To compare Hi-C, SPRITE and GAM, we performed computational experiments on the SBS polymer models of 4 different loci: the *Sox9* locus (chr11:109-115Mb) and *HoxD* locus (chr2:71-78Mb) at 40kb resolution from murine staminal cells (mESC); the *Epha4* locus (chr1:73-79Mb) at 10kb resolution from the CHLX-12 murine cell line; the human HCT116 locus (chr21:34.6–37.1 Mb) we already discussed in **Chapter 2**. All such models were obtained from PRISMR applied on Hi-C data[27, 95–97] and their ensembles of 3D structures were produced by Molecular Dynamics simulations [27, 95–97], as described previously in this work . In principle, the connection of those conformations with specific DNA regions is not necessary for our scope, as we only require tridimensional structures which are fully known, so to benchmark the effectiveness of the simulated technologies. In this sense, whatever geometry could be used, unrelated to real-life cases. However, for our comparison to be meaningful, the used polymer structures should be as complex as the real conformations of chromatin and, preferably, share some of their main architectural features. Otherwise, we could draw conclusions on Hi-C, SPRITE and GAM arguably reproducible in experiments. That is the reason we recurred to SBS models of real DNA loci, because they have been proven to reproduce well enough average and single-cell organizational features of chromatin [27, 95–97] and, in the case of the HCT116 locus, they were even validated against imaging data, as seen before. That ensures the *in-silico* technologies are tested on architectures comparable to those encountered in cellular nuclei.

For definiteness, we will consider the *Sox9* locus as case study and will report in full details only the analyses conducted on the relative polymers, while we will simply summarize those for the other three loci which, importantly, returned analogous results. In the next paragraphs, we will first prove that our algorithms simulate effectively the real technologies; then we will investigate whether all three technologies detect faithfully the average and single-cell structures of polymers; we will compute how many *in-silico* cells are required for an experiment to yield statistically reproducible outputs and, finally, we will extract the noise-to-signal behavior for varying efficiencies, number of cells and genomic separation.

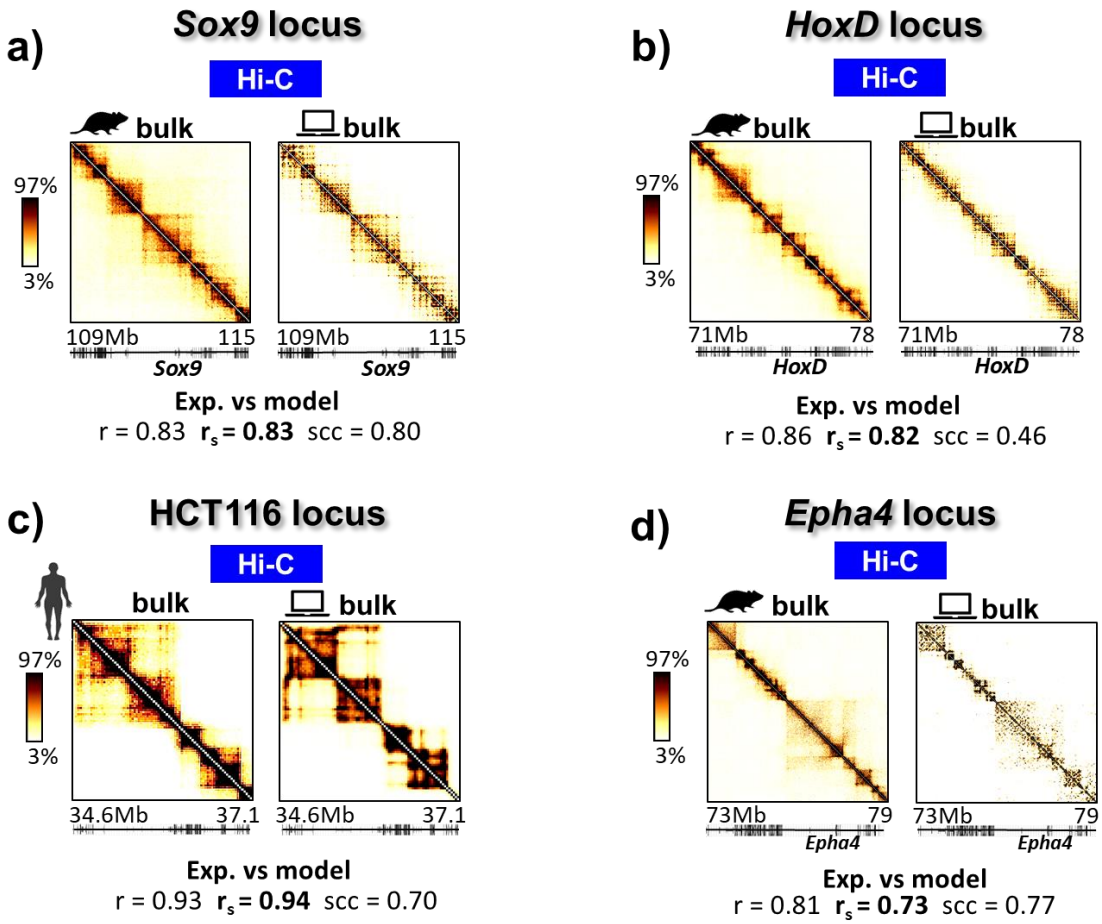


Figure 3.4: **a)** The experimental Hi-C map [11] of the *Sox9* locus (mESC, chr11:109-115Mb, 40kb, on the left) is well matched by the *in-silico* Hi-C matrix derived from the ensemble of model polymer conformations from [95] (on the right). To compare with the experimental contact map, the *in-silico* Hi-C matrix was obtained from a large sample of simulated cells (bulk condition, see **Main Text**). The colorbar reports the percentiles of the maps. Below, the Pearson (r), Spearman (r_S) and HiCRep (scc) correlations between the two Hi-C matrices quantitatively confirm the visual similarity. The list of genes is from the UCSC Genome Browser.
b) Same as panel a) but for the *HoxD* locus (mESC, chr2:71-78Mb, 40kb), with data from [11] and polymer structures from [97].
c) Same as previous panels but for the HCT116 locus (HCT116 human cell line, chr21:34.6-37.1Mb, 30kb). The experimental map is from [33] and the polymer model from [27].
d) As in previous panels, for the *Epha4* chromatin region (mouse CHLX-12 cells, chr1:73-79Mb, 10kb). Experimental data from [10] and ensemble of polymers from [96].
The good similarity between model and experiment in all cases validates our simulation of the Hi-C protocol and represents a consistency check, as each ensemble of polymer structures was inferred by the considered Hi-C data.

3.3.1 The simulations of Hi-C, SPRITE and GAM provide a good proxy of real experiments

In the *Sox9* locus case, the SBS polymer model was inferred by PRISMR from Hi-C data at 40kb in mESC [11] and an ensemble of 500 independent 3D configurations was derived [95]. We performed the Hi-C algorithm over the ensemble of polymer structures and compared the output contact matrix with the experimental Hi-C map (**Figure 3.4a**). The experimental matrix was extracted from a population of cells (bulk Hi-C), i.e. a statistically wealthy sample of DNA. To match closely the experimental condition, we iterated *in-silico* Hi-C over 10000 different *in-silico* cells, made of

different pairs of polymer structures. Indeed, we checked that the outputs obtained from 10000, 50000 or 100000 simulated cells are nearly identical, signaling that 10000 *in-silico* cells already achieve the bulk condition, when statistics is saturated (see below).

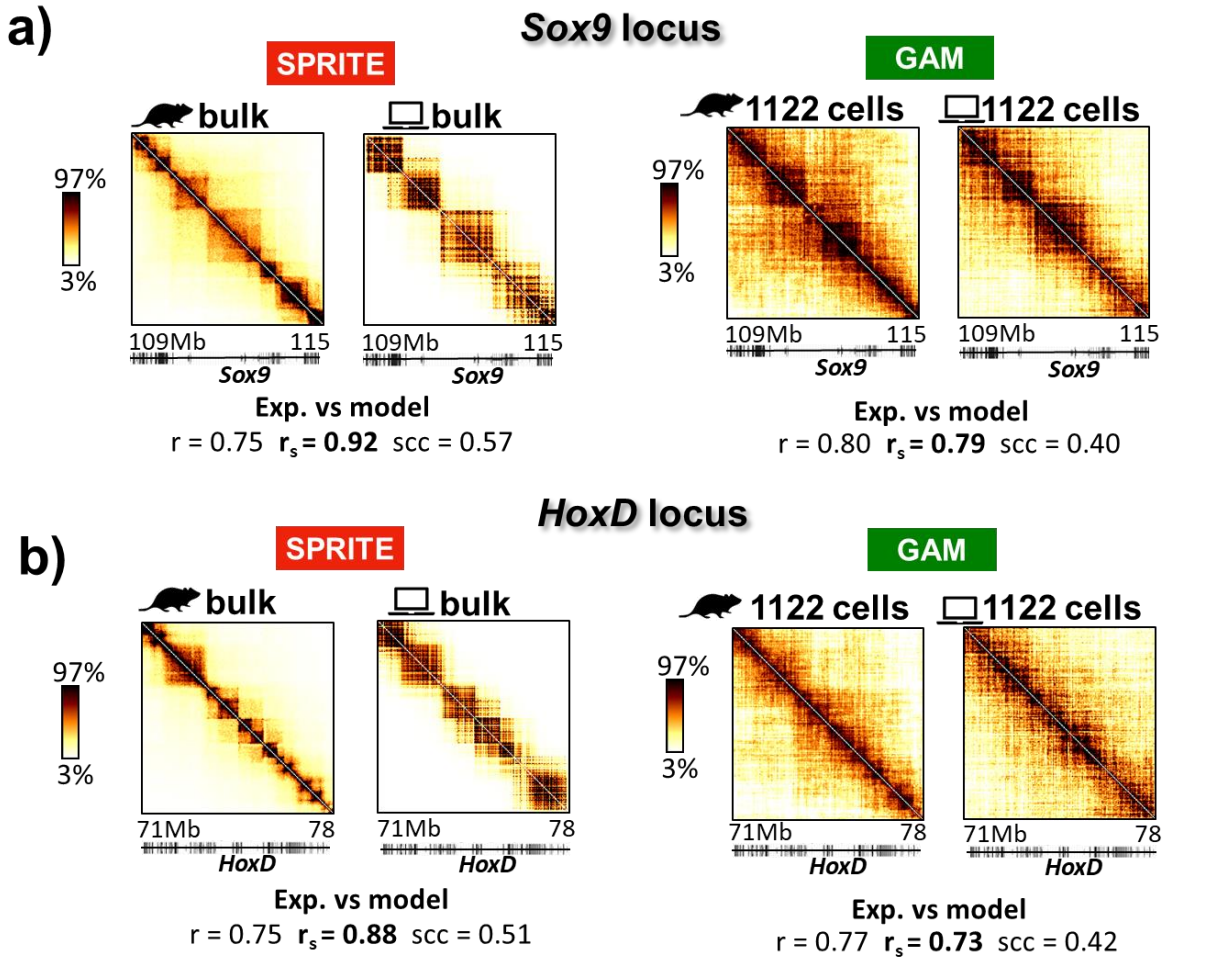


Figure 3.5: a) (left) For the *Sox9* locus, the *in-silico* SPRITE matrix obtained in bulk condition (large number of simulated cells) is very similar to the experimental SPRITE map measured on the same locus [9]. Below, the Pearson (r), Spearman (r_S) and HiCRep (scc) correlations between the maps are reported. All in all, the coefficients express good similarity. (right) The simulated GAM matrix is also similar to the experimental GAM map extracted from 1122 murine staminal cells [36]. The *in-silico* matrix was obtained from 1122 *in-silico* cells as well. Below, the three measures of correlations are shown, with overall nice values of similarity.
b) Same as panel a), but for the *HoxD* locus. The experimental data are the same as those used for the *Sox9* locus. The comparison between model and experimental matrices is nice also in this case.
The good match between model and experimental maps for SPRITE and GAM shows the effectiveness of our simulations to reproduce the two technologies.

From **Figure 3.4a** we see that the model and experimental Hi-C maps are very similar and their Spearman correlation (r_S) is high (r_S=0.83). To get a more robust quantification of the resemblance, we also computed the Pearson (r) correlation and the HiCRep Stratum-Adjusted Correlation Coefficient (scc), obtaining similarly high scores, i.e. r=0.83 and scc=0.80. The HiCRep correlation [118] is a sophisticated variant of the Pearson measure and was designed specifically to compare two Hi-C matrices. It accounts for effects which could artificially bias the Pearson or Spearman

correlations, e.g. the decay of contact frequencies with increasing genomic separation between DNA sites (it is conceptually similar to the r' correlation introduced in the previous chapter). We will use the HiCRep assessment all over the current work to cross-validate the other correlation measures.

Analogously nice comparisons between model and experimental Hi-C maps were found for the *HoxD, Epha4* and HCT116 loci (**Figure 3.4b,c,d**). Overall, this indicates that the Hi-C algorithm correctly simulates the key steps of a Hi-C experiment, as it returned contact matrices similar to those of real loci. Collaterally, that is also a consistency check of the polymer 3D structures, as they were inferred from Hi-C data.

Next, we realized *in-silico* SPRITE and GAM experiments over the *Sox9* polymer configurations. We tested their output matrices against mESC SPRITE bulk data [9] and F123 mESC GAM data from 1122 slices [36], both at 40kb resolution. To match the experimental matrices, the *in-silico* SPRITE map was obtained in bulk condition, as done for Hi-C, while GAM was simulated exactly from 1122 *in-silico* cells. Strikingly, the simulated and experimental matrices are all similar to each other, with r_s=0.92 and r=0.75 for SPRITE and r_s=0.79 and r=0.80 for GAM (**Figure 3.5a**). Albeit HiCRep was conceived to compare Hi-C maps, we computed its scores also in the SPRITE and GAM cases, with scc=0.57 and scc=0.40 respectively. As no established benchmarks are available for HiCRep on SPRITE and GAM matrices, we checked that those correlations are significant. They were tested against random control distributions of scc correlations extracted from randomized model and experimental matrices and were both found above the 90th percentiles of the controls, so significantly high (**Figure 3.6**). Analogous findings hold for the *HoxD* locus (**Figure 3.5b**; SPRITE and GAM data are not available for the cell lines of the *Epha4* and HCT116 loci). That supports the effectiveness of our simulated SPRITE and GAM, in that they reproduce the corresponding experimental outcomes. Additionally, that is a strong endorsement toward the validity of our polymer conformations, because, albeit inferred from Hi-C, they can return completely independent experimental data, like SPRITE and GAM.

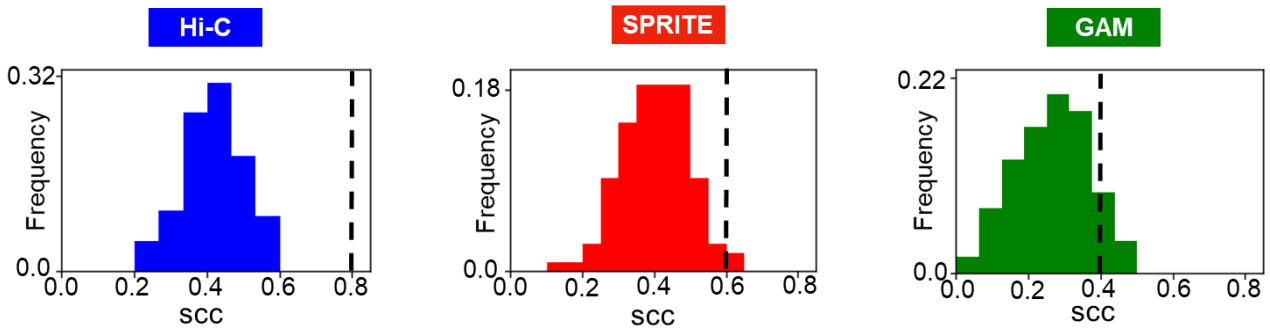
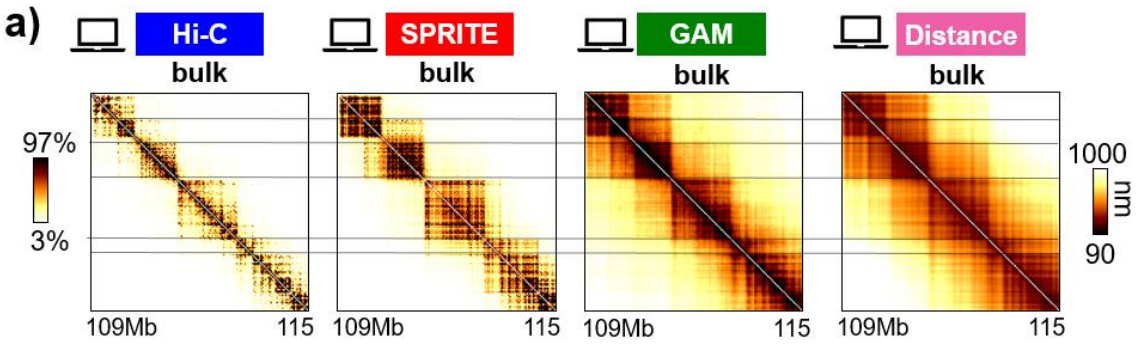


Figure 3.6: In the *Sox9* locus case, the significance of the HiCRep correlations between model and experimental maps (**Figures 3.4a, 3.5a**) is tested against a control distribution, for each of the three technologies. In the Hi-C case on the left, the model-experiment correlation (dashed line) is above the 90th percentile of the control, made of the scc between 100 pairs of randomized model and experimental Hi-C maps: that proves the significance of the similarity. Analogous results were found for SPRITE (middle) and GAM (right).

3.3.2 Bulk Hi-C, SPRITE and GAM return faithfully the underlying architecture



b)

Correlation table

bulk	Distance		
Hi-C	-0.54	-0.89	-0.73
SPRITE	-0.72	-0.94	-0.52
GAM	-0.94	-0.99	-0.93
	r	r_S	scc

Figure 3.7: a) In the *Sox9* locus case, the *in-silico* Hi-C, SPRITE and GAM contact maps computed in ideal conditions (bulk limit and 100% efficiency) are compared against the average distance matrix derived from the ensemble of polymer structures (rightmost matrix). The TAD pattern detected by all three methods is overall similar, consistent with experimental findings [8, 9]. Strikingly, those patterns are also compatible with the distance map, showing that Hi-C, SPRITE and GAM, in ideal conditions, faithfully detect the average architecture of the considered conformations. However, the GAM matrix reproduces better the long-range and interTAD features of the distance map. The horizontal lines highlight the similarities of the TADs across all matrices. The colorbar for the contact maps indicates the percentiles.
b) The Pearson, Spearman and HiCRep correlations are reported between each *in-silico* contact map and the average distance matrix. Coefficients are overall very high in absolute value. Notably, GAM exhibits the highest correlations, as it captures more effectively the long-range features of the distance map.
Adapted from [36].

We asked whether Hi-C, SPRITE and GAM detect faithfully the spatial organization of chromatin or return only partial and different aspects of it. At the pairwise level, the organization of DNA in space is conveyed by the distances between all pairs of sites. Indeed, all three technologies ultimately aim to reveal the pattern of physical distances between chromatin sites, by providing data differently related to spatial proximity.
So, in the *Sox9* locus model case, we computed the average distance matrix, reporting the physical distances between each possible pair of windows and averaged across all the available polymer structures. Then, we tested how *in-silico* Hi-C, SPRITE and GAM matrices compare against it (**Figure 3.7**). As the goal is to unveil possible bias intrinsic to the protocols, all three maps were computed in ideal conditions, i.e. with efficiency 1 and from a large number of *in-silico* cells (bulk condition, see previous paragraph).
First, we note that the patterns emerging from the three *in-silico* matrices are overall similar (**Figure 3.7a**). In particular, they identify the same TAD structure, in agreement with experimental investigations [8, 9]. Nonetheless, the GAM matrix exhibits more evident inter-TADs and, generally,

long-range co-segregation frequencies compared to the contact frequencies visible in SPRITE and Hi-C. This is again consistent with experimental observations on long-range interactions in DNA loci [8].

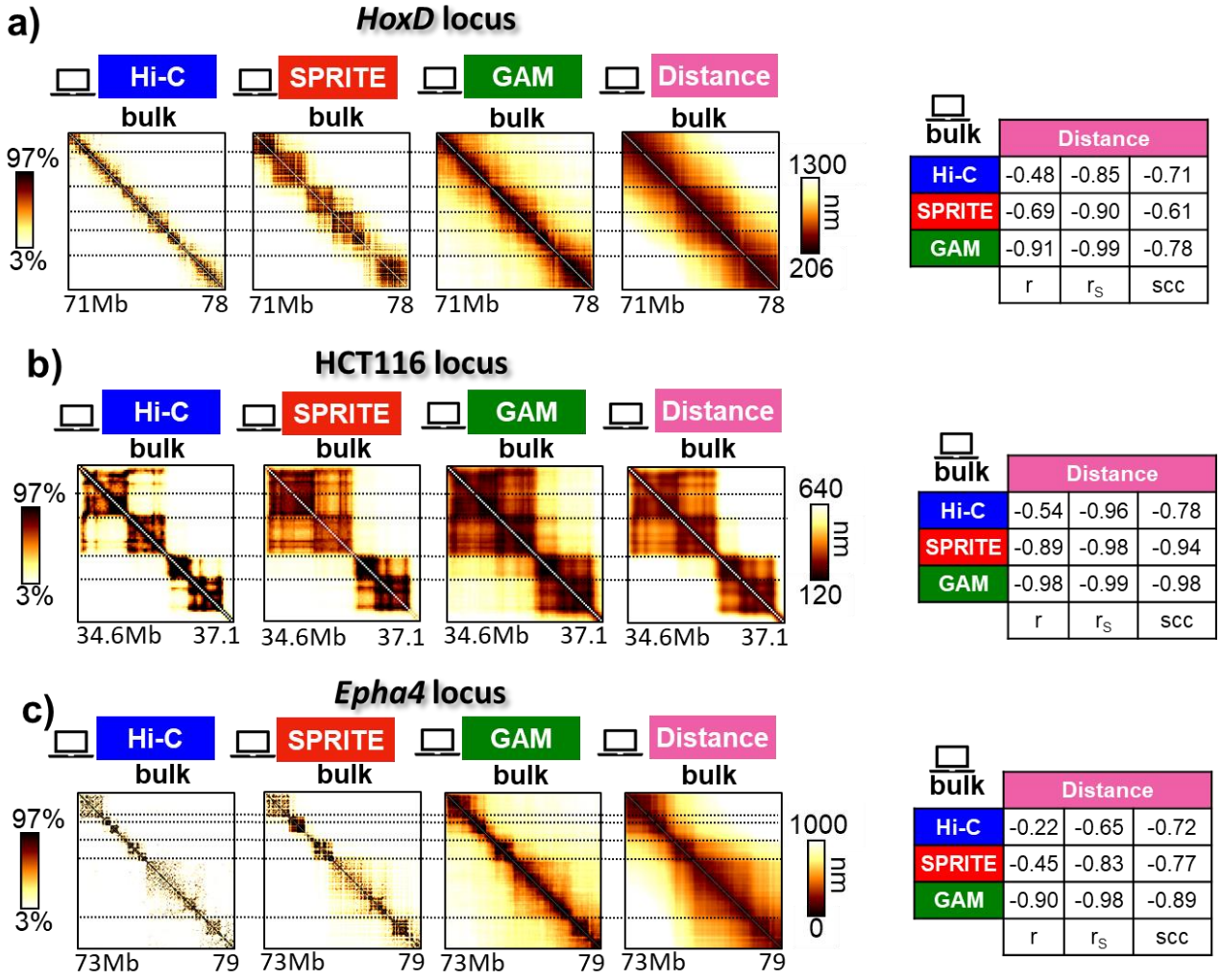


Figure 3.8: Analogous results to those discussed in **Figure 3.7** for the *Sox9* locus were found for the *HoxD* locus (panel a), the HCT116 locus (panel b) and the *Epha4* locus (panel c). Indeed, in all cases the *in-silico* contact maps (in the bulk limit and at 100% efficiency) display overall analogous features and reproduce faithfully the average pattern of distances. The tables on the right show very high (in absolute value) Pearson (r), Spearman (r_S) and HiCRep (scc) correlations between each *in-silico* contact map and the distance matrix, confirming the visual inspection. In all three loci, the GAM map results the most effective in detecting the long-range and interTADs architectures revealed by the distance matrix, as also shown by the correlations overall higher than those for Hi-C and SPRITE.

Second, the patterns detected by Hi-C, SPRITE and GAM are all consistent with the average distance matrix, as quantified by the high absolute Spearman correlation values (r_s <-0.89 for all three methods, the correlations are negative because of the inverse relation between contacts and distances). Pearson and HiCRep correlations yield a similar scenario (**Figure 3.7b**). Hence, all three technologies, in ideal condition, catch unbiasedly the average architecture of the underlying structures. Notably, the GAM matrix has the highest correlations with the average distance map. Indeed, the enriched long-range co-segregation frequencies across TADs and between windows

more than 1Mb apart closely reproduce the pattern of distances, better than the contacts of Hi-C and SPRITE. That is most likely because contacts are detected within a strict threshold distance (fixed by crosslinking) while co-segregations are found in slices spanning the whole nuclear length scale. From this perspective, our analysis supports the meaningfulness of the long-range patterns identified by GAM experiments and indicates that the co-segregation pattern is slightly more faithful to the distances than the contacts in Hi-C and SPRITE.
In summary, we showed that all three technologies, in ideal conditions, return overall the same description of the average spatial conformation investigated and that such description is faithful to the average pattern of distances. Additionally, we found that GAM, based on nuclear slicing, is slightly more accurate in detecting the long-range, interTADs structural features.
Analogous results were obtained for the *HoxD, Epha4* and HCT116 loci (**Figure 3.8**).

3.3.3 Single-cell Hi-C, SPRITE and GAM poorly return the underlying architecture

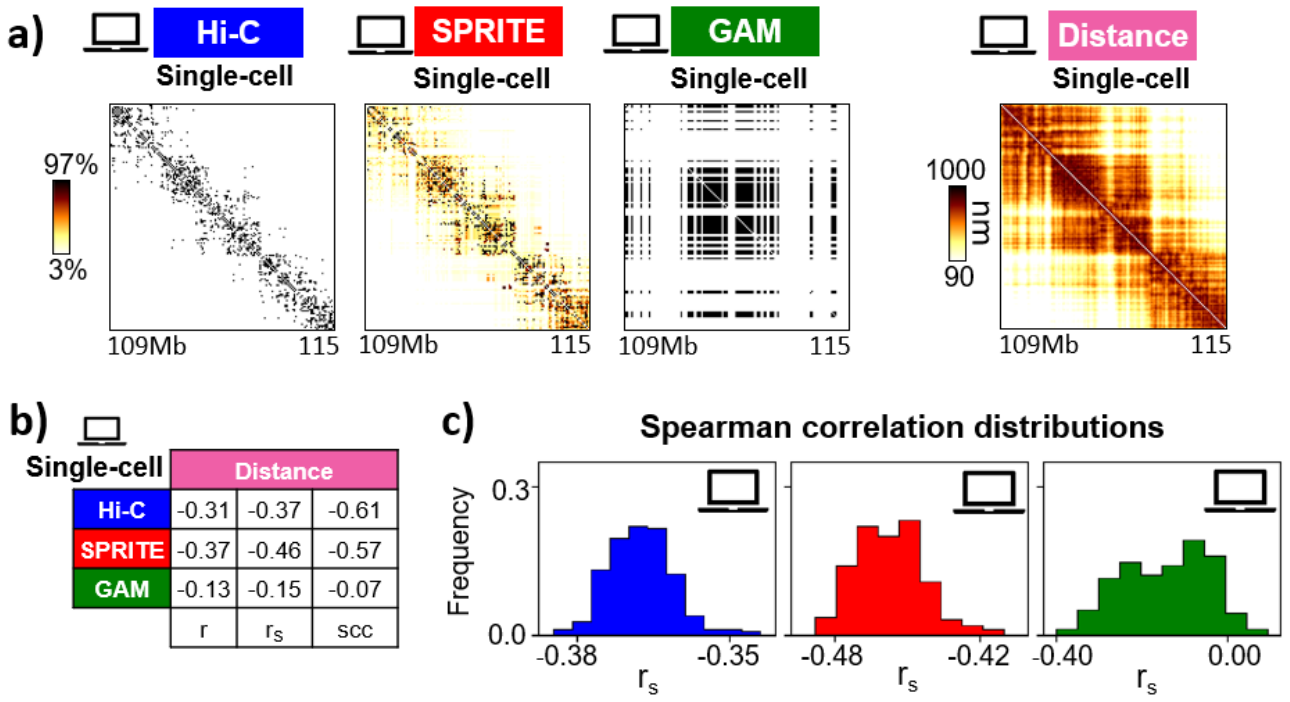


Figure 3.9: **a)** In the *Sox9* locus case, the *in-silico* Hi-C, SPRITE and GAM contact maps from a single simulated cell poorly reproduce the distance matrix for that same *in-silico* cell (rightmost matrix), albeit they were computed at 100% efficiency.
b) Indeed, calculating the correlations between the single-cell contact and distance maps from 250 *in-silico* cells and then averaging, we obtained mean Pearson (r), Spearman (r_S) and HiCRep (scc) correlations lower than those seen in the bulk case, for each technology (**Figure 3.7b**).
c) The distributions of Spearman correlations between single-cell contact and distance maps are shown for all three methods. GAM has the broadest distribution.
As the contact maps were obtained for efficiency 1, the results shown here illustrate that at single-cell level Hi-C, SPRITE and GAM protocols have intrinsic limitations in detecting the underlying architecture. In particular, GAM exhibits the worst performance, as its protocol was designed for populations of cells.
Adapted from [36].

In the previous paragraph we interrogated the ability of Hi-C, SPRITE and GAM to detect the average chromatin architecture from an ensemble of cells. Here, we explore their effectiveness in reproducing the spatial organization of chromatin in a single cell. As already mentioned, in mammals, that amounts at detecting the average conformation of DNA across two allelic copies.
For the *Sox9* locus model, we selected 250 *in-silico* cells made of randomly chosen pairs of polymer structures. For each simulated cell we extracted the single-cell distance map and computed the *in-silico* Hi-C, SPRITE and GAM single-cell matrices. We used a 100% efficiency in all three cases to simulate perfect experiments. The *in-silico* maps of all three methods appear much less similar to their corresponding distance matrices than observed in the bulk case (**Figure 3.9a**). Indeed, the average Spearman correlations between each single-cell contact map and the relative distance matrix are r_s=-0.37, r_s= -0.46 and r_s=-0.15 for respectively Hi-C, SPRITE and GAM (similar values were observed with Pearson and HiCRep correlations, **Figure 3.9b**). Such correlation coefficients are much lower than those found for the bulk matrices (**Figure 3.7b**). Since a 100% efficiency was considered, the worsened performance of all three technologies at single-cell level highlights the intrinsic limitations of the protocols. Interestingly, while GAM was found the best performing technology in reproducing the average distance pattern in bulk conditions, at single-cell level it becomes the worst. That is because single-cell Hi-C and SPRITE collect the contacts from the entire cellular nucleus, whereas single-cell GAM reveals the co-segregation events from a single slice only, i.e. a tiny portion of the nuclear volume. In this sense, the GAM protocol per se was not conceived for single-cell exploration. Indeed, examining the distributions of Spearman correlations between single-cell distance and contact matrices (**Figure 3.9c**), the GAM case presents the greatest broadness, as its output is hugely dependent on the random orientation of the slice used. Hi-C and SPRITE performances are overall comparable, with SPRITE having slightly better correlations.
Although we will systematically study the effect of the efficiency on contact matrices in the following sections, here we anticipate that for efficiency values typically encountered in real experiments (0.05 for Hi-C and SPRITE, 0.5 for GAM, see **paragraph 3.3.4**), the mean Spearman correlations between single-cell contact and distance maps all drop, in absolute value, below 0.2. This suggests that the architectural features revealed by single-cell experiments must be taken cautiously.
Summarizing, the present analyses illustrated that single-cell Hi-C, SPRITE and GAM maps return a poorer description of chromatin architecture than in the bulk case, even for 100% efficiency. That roots in the intrinsic limitations of the respective protocols which, are, instead, overcome in the bulk limit, as seen in **paragraph 3.3.2**. Additionally, since only single-cell Hi-C experiments have been realized so far [72, 75–77, 113], our investigations of the performances for single-cell SPRITE and GAM may be used in a predictive perspective. They showed, in particular, that SPRITE may capture single-cell chromatin conformations comparably to Hi-C and that GAM, as it is, is not designed for single-cell experiments [8].

3.3.4 Reproducibility of the Hi-C, SPRITE and GAM contact maps

Contact matrices (Hi-C, SPRITE or GAM) extracted from cells of the same type, organism and at the same experimental conditions are nonetheless expected to be different from each other. That is because of two main sources of noise: the experimental efficiency and the variability of chromatin conformation across cells. Indeed, as we have seen in **Chapter 2,** the spatial configurations of a given DNA locus can significantly vary across single cells. Hence, even for 100% experimental efficiencies, contact matrices from homologous populations of cells will, in general, differ because different

architectures possibly compose the respective populations. From now on, we will use the expression *replicate* contact matrices for the outcomes of experiments conducted in the same conditions, i.e. with equal efficiency and same type, organism and number of cells considered.
To make replicate contact matrices similar, large samples of cells must be used, as in this case cell-to-cell structural variations and efficiency effects are expected to average out. When the sample of cells is big enough that the noise level is negligible and, thus, replicate contact matrices are highly similar to each other, the experiment is said *reproducible*, i.e. the bulk limit is approached. Notably, if a batch of cells makes an experiment reproducible, a greater sample will negligibly improve the quality of the output contact matrix, as noise effects are already nearly suppressed: this is the definition of bulk limit we used in **paragraph 3.3.1**. In summary, the reproducibility condition ensures that an experiment returns statistically robust output data.
In standard Hi-C and SPRITE experiments, cells are cultivated *in-vitro* and populations of millions of cells are typically used, which ensure the bulk limit. However, when real tissues are studied to get a picture of chromatin in its native environment (e.g. in biopsies), the availability of cells is limited and can be far less than millions. In such relevant cases, the knowledge of the *minimal number of cells for reproducibility* is crucial to assess if Hi-C or SPRITE experiments can be safely conducted. That is even more important for GAM, which is normally applied on samples of hundreds or few thousands of cells [8]. In this paragraph we will present an approach to estimate the minimal number of cells to reproducibility for Hi-C, SPRITE and GAM, based on our computational experiments.
From the ensemble of 3D configurations of the *Sox9* locus model, we generated Hi-C, SPRITE and GAM contact maps at different number of *in-silico* cells, at efficiency 1 (**Figure 3.10a**). The quality of the contact maps clearly improves as the input number of simulated cells is increased, signaling that the variability of our polymer structures produces significative noise, consistently with the findings of **Chapter 2**. Additionally, it appears that the velocity whereby the contact maps stabilize and get to the bulk limit is different for the three technologies. Then, contact maps at different number of *in-silico* cells were produced at realistic efficiencies (**Figure 3.10b**), i.e. efficiencies similar to those employed in real experiments. Specifically, we posed for Hi-C an efficiency of 0.05, which is in the range of those measured in real applications [119]; for SPRITE the same 0.05 efficiency was assumed; for GAM we selected a 0.5 efficiency, because GAM experiments typically achieve efficiencies an order of magnitude higher than those of Hi-C and SPRITE [8]. Clearly, the limited efficiency worsens the quality of the matrices, slowing the approach to reproducibility with the increasing number of cells. Overall, these observations suggest our computational environment is a good proxy of real experimental conditions, where the same kind of noise effects are expected.
To quantify all those impressions, we elaborated the following procedure. For each technology, computational experiments at the same number of *in-silico* cells and efficiency are repeated many times, to produce replicate contact maps. The similarity of these replicate matrices is assessed computing the average Pearson correlation over all the possible pairs. Since reproducibility is achieved when replicate contact maps can be considered highly similar, we assumed that corresponds to an average Pearson correlation of $r_t = 0.90$. So, within such approach, the minimal number of cells for reproducibility is identified as that number returning an average Pearson correlation of 0.90 among replicate contact maps. For brevity, in what follows we indicate the number of *in-silico* cells used as N and the minimal number for reproducibility as M.
We computed the average Pearson correlation among replicate contact maps for several values of N at the fixed efficiency of 0.1 (**Figure 3.10c**). As expected, for all three technologies, the replicate

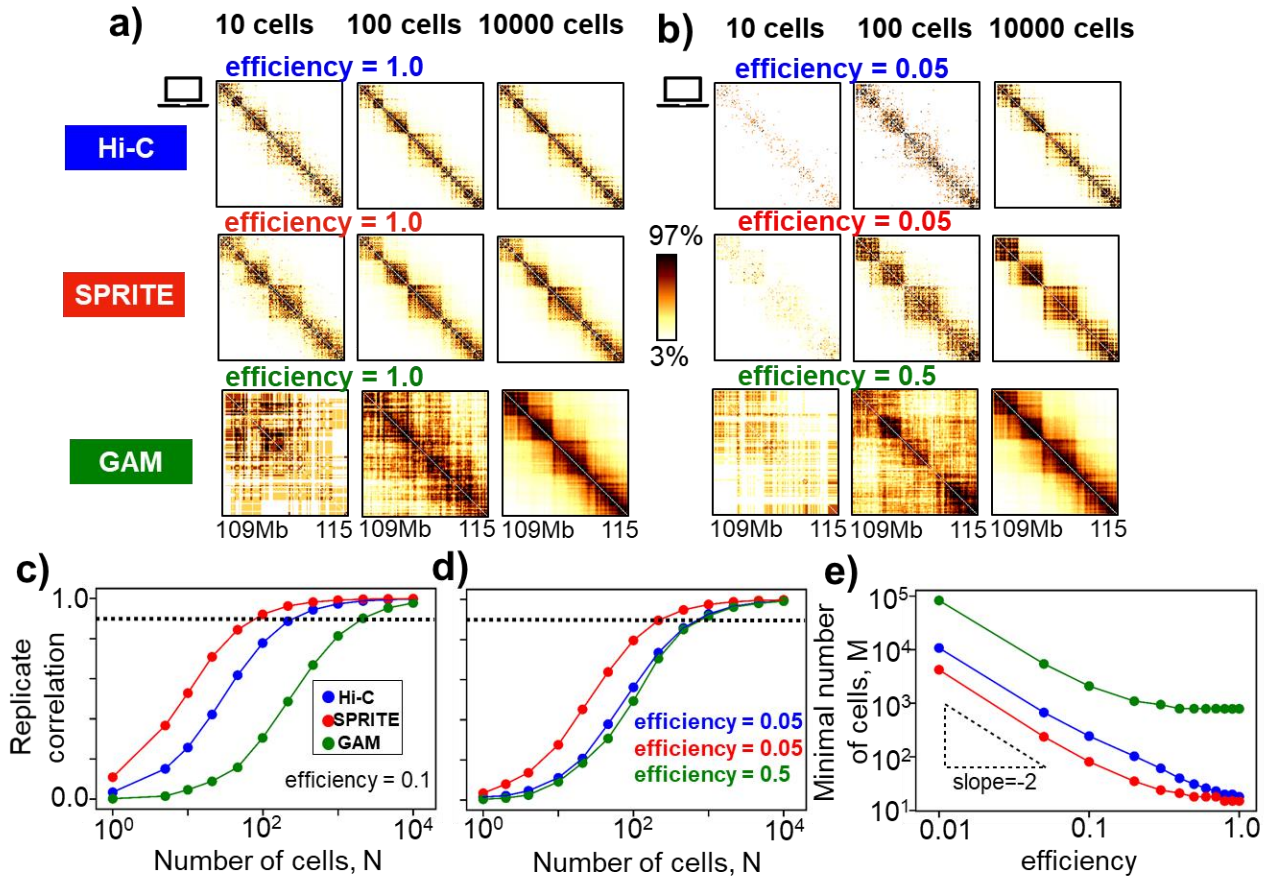


Figure 3.10: a) The *in-silico* Hi-C, SPRITE and GAM contact maps of the *Sox9* locus depend on the number of *in-silico* cells, N, considered in the experiment (here is shown the case with efficiency equal to 1). Color scale indicates the percentiles of each map.
b) Results analogous to those in panel a) are shown in the case where efficiencies similar to those found in real experiments are considered: here, for Hi-C and SPRITE the *in-silico* efficiency is set equal to 0.05, and for GAM equal to 0.5 (see **Main Text**). As expected, the limited efficiency impacts on the quality of the contact maps.
c) The Pearson correlation is shown between replicate contact maps as a function of N for Hi-C, SPRITE and GAM at a given efficiency (0.1). The dashed line is the threshold correlation value r_t=0.90, marking the minimal number of cells to reproducibility (M). Importantly, the curves of the three technologies achieve reproducibility for different values of N (M is about 200, 100, 2000 for respectively Hi-C, SPRITE and GAM).
d) Results analogous to those in panel d) are shown in the case of the realistic efficiencies reported in panel b). As GAM has the highest efficiency, its corresponding curve becomes closer to that of Hi-C and SPRITE. Here, M is approximately 650, 250 and 800 for respectively Hi-C, SPRITE and GAM.
e) The value of M is shown for Hi-C, SPRITE and GAM as a function of the efficiency. M increases as the efficiency is reduced and grows approximately as an inverse squared power law at small efficiencies. For all the efficiencies, M is the smallest in SPRITE, a factor of two higher in Hi-C and an order of magnitude higher in GAM. Given also the findings of panels c) and d), that supports SPRITE as the best method to employ when the available sample of cells is limited, e.g. for *in-vivo* experiments.
Adapted from [36].

correlation grows with N, until plateauing to the maximum correlation r=1. The plateau quantitatively defines the bulk condition and, in particular, at 10000 *in-silico* cells all three technologies are found in the bulk limit. Importantly, Hi-C, SPRITE and GAM correlation curves get to the 0.90 threshold for different values of N. Indeed, at the selected 0.1 efficiency, M is 200, 100

and 2000 for, respectively, Hi-C, SPRITE and GAM. However, in real applications the efficiencies can be very different across the technologies. We then studied Hi-C, SPRITE and GAM at the realistic efficiencies introduced before (0.05 for Hi-C and SPRITE, 0.5 for GAM). We found an analogous scenario of replicate correlations (**Figure 3.10d**), albeit the estimated M values become more similar to each other: M is 650, 250 and 800 respectively for Hi-C, SPRITE and GAM.
Next, as other efficiencies may be used in real applications, we computed M for various efficiency values ranging from 0.01 to 1. M increases when the efficiency gets smaller (**Figure 3.10e**) and, interestingly, for small efficiencies the increase of M follows approximately an inverse squared relation for all the technologies: halving the efficiency requires a quadruple number of cells for an experiment to maintain reproducibility. Notably, at all the efficiencies explored, SPRITE was found the technology with the least number of cells needed to reproducibility, GAM requires the highest and Hi-C is in the middle, although its M values get very similar to those of SPRITE for nearly 1 efficiencies.

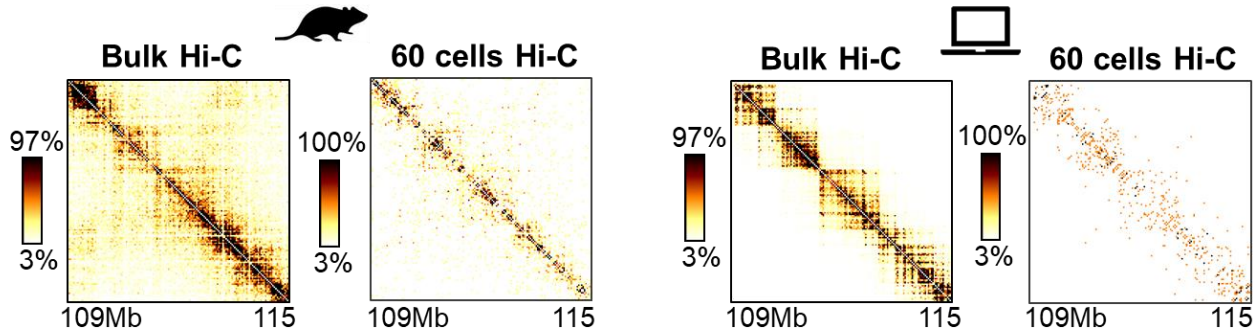


Figure 3.11: On the left, Hi-C data of the *Sox9* locus from murine CD4-T_H1 cells at 40kb resolution are shown [72]. Specifically, the bulk Hi-C map is compared to the Hi-C matrix measured from 60 cells, with a Spearman correlation of r_S=0.33. On the right, we produced the *in-silico* analogues of those two matrices from our ensemble of polymer structures, relative to the *Sox9* locus in mESC cells at 40kb. We used as efficiency value the estimate reported in [72], i.e. 0.025. The Spearman correlation between the simulated Hi-C matrices is r_S=0.27, which is similar to that found in the experimental case, for a different cell line. That suggests our simulated matrices depend on the number of cells and efficiency analogously to real experimental data. Colorbars indicate the percentiles of the maps. Adapted from [36].

Overall, our approach indicates that SPRITE could be the most effective method to probe chromatin architecture in small samples of cells, as in case of biopsies. Indeed, for equal conditions, SPRITE has the lowest M at all the efficiencies explored, while GAM the highest, with M values almost 10 times bigger. However, in real applications the experimental efficiencies are typically very different across the methods, with GAM efficiencies about an order of magnitude greater than those used in Hi-C and SPRITE experiments [8, 119]. In such realistic scenario, the minimal numbers of cells to reproducibility get closer, especially between Hi-C and GAM (650 and 800 at 0.05 and 0.5 efficiencies, respectively), with SPRITE still having the lowest M (250).
We checked the validity of our results by comparing against available experimental findings. First, we found that the correlation between bulk Hi-C and Hi-C data extracted from 60 cells in the CD4-T_H1 cell line at 0.025 efficiency [72] is similar to that between our corresponding *in-silico* Hi-C maps in mESC at the same efficiency (r_S=0.33 against r_S=0.27, **Figure 3.11**).

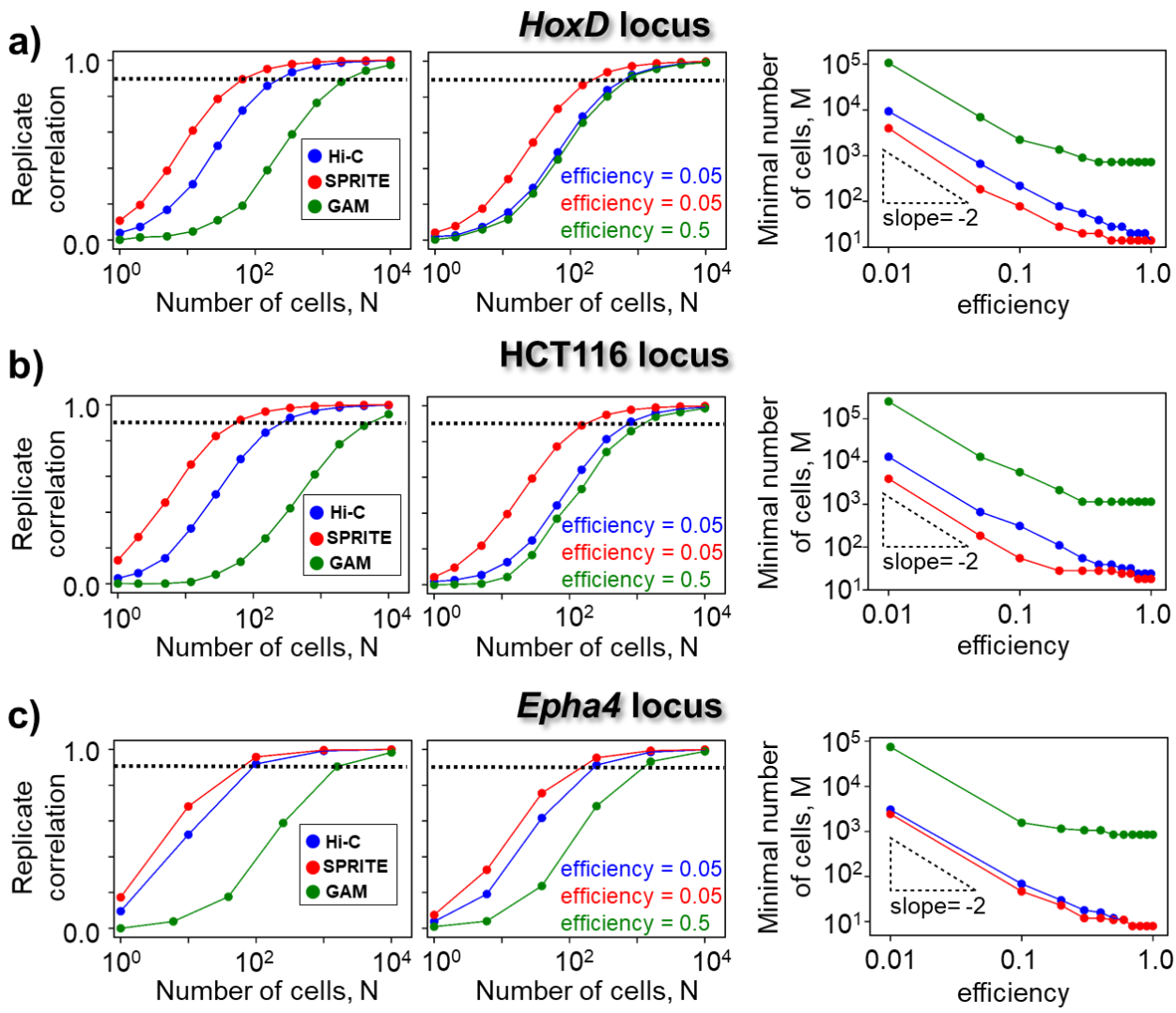


Figure 3.12: a) (left) For the *HoxD* locus, the Pearson correlation between replicate contact matrices is shown against the number of *in-silico* cells (N) for all three technologies and at given efficiency (0.1). The dashed line marks the threshold correlation r_t=0.90, signaling the reproducibility condition. At the used efficiency, the three curves approach very differently to the threshold and so the minimal cell numbers to reproducibility, M, are diverse for each technology. (middle) Same as in the plot on the left, but for realistic efficiencies (0.05 for Hi-C and SPRITE, 0.5 for GAM, see **Main Text**). In this case, the correlation curve of GAM gets closer to that of Hi-C thanks to the higher efficiency. (right) The M values are shown for several efficiency values. In the small efficiency range, M grows with decreasing efficiencies approximately as an inverse square power law, for all three technologies. At each given efficiency, SPRITE has the lowest value of M, GAM the highest (almost 10 times greater) and Hi-C exhibits M values approximately 2 times bigger than those of SPRITE. Importantly, all these results are analogous to those found for the *Sox9* locus (**Figure 3.10c,d,e**). Similar findings were obtained for the HCT116 locus and the *Epha4* locus, as illustrated in panels b) and c). The consistency of the results across all the investigated loci supports their general validity and robustness.

That supports the correlation numbers we get between our *in-silico* maps are meaningful and, also, that our definition of the bulk limit is effective. Next, to validate the estimates of M, we considered the data available from a Low-C experiment (a variant of Hi-C) on mESC [74]. Here, for a 10Mb long locus, a sample of 1000 cells was found enough to return a contact matrix compatible with the bulk

one (Pearson correlation r=0.95). Such experimental estimate of the minimal number of cells to reproducibility is overall consistent with that we obtained for Hi-C at realistic efficiency (M=650). Importantly, all the findings we described for the *Sox9* locus were generally confirmed in the *HoxD, Epha4* and HCT116 loci (**Figure 3.12**), supporting their robustness and generality.

3.3.5 Reproducibility of the SLICE interaction maps

As said, GAM experiments return the co-segregation frequencies for all the window pairs of a locus. Based only on statistics, the SLICE (StatisticaL Inference of Co-sEgregation) computational tool derives from co-segregation data the probabilities of interaction between DNA windows in single cells [8]. So, the output of SLICE is a probability of interaction (PI) matrix. We extended the analysis of reproducibility described above to investigate the performance of SLICE. Specifically, the simulated PI matrices were obtained applying SLICE to the *in-silico* co-segregation data extracted from the *Sox9* model polymers.

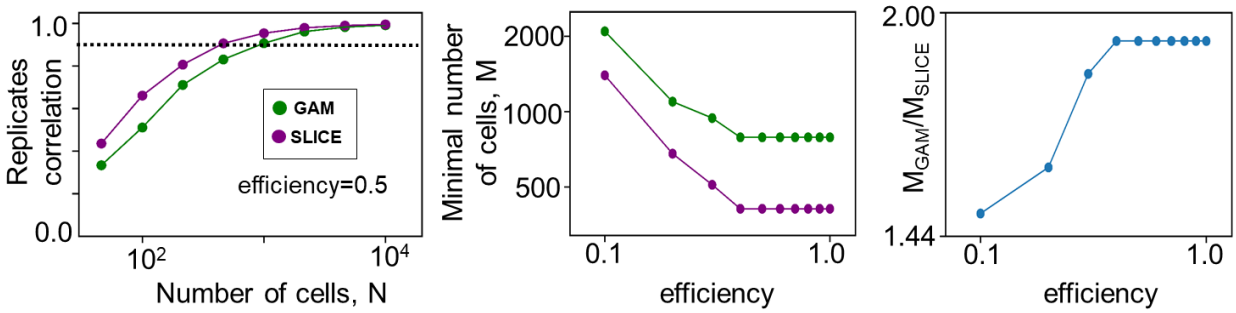


Figure 3.13: (left) For the *Sox9* locus case, the Pearson correlations between replicate *in-silico* SLICE matrices and between replicate *in-silico* GAM matrices are shown against the considered number of simulated cells (compare **Figure 3.10c**). The realistic GAM efficiency (0.5) was employed (**Main Text**) and the dashed line indicates the threshold correlation r_t=0.90. The SLICE curve is always above the GAM curve, namely the SLICE maps achieve faster the reproducibility condition. (middle) Consistently, the values of M for SLICE are smaller than those found for GAM at all the efficiencies explored. (left) Specifically, the minimal number of cells for reproducibility of SLICE is almost half than that of GAM, for a given efficiency. The reduction of M indicates that the SLICE computational tool can significantly clean the noise level present in GAM matrices. Adapted from [36].

The average Pearson correlation among replicate PI maps increases with N as observed for Hi-C, SPRITE and GAM, with a plateau to 1 signaling the bulk limit (**Figure 3.13**). Interestingly, for the realistic 0.5 efficiency introduced before, the curve of correlations of SLICE plateaus faster than that of GAM. That is observed also for other efficiencies: the minimal number of cells to reproducibility results lower for SLICE than for GAM at all the efficiencies explored (**Figure 3.13**). Specifically, the value of M for the PI maps is approximately 2 times smaller than that for GAM co-segregation maps, at the same efficiency (**Figure 3.13**). That indicates the application of SLICE on GAM co-segregation data drastically reduce the noise effects, allowing for halved values of M. This is consistent with the scope of SLICE, as it identifies only window pairs truly interacting. For an efficiency close to that used in real applications (0.5), while the minimal cell number for GAM data is approximately 800 (**Figure 3.10d**), for SLICE it drops to about 400. Such result suggests that GAM in combination with SLICE may be an adequate instrument to dissect chromatin architecture on *in-vivo* samples of cells.

3.3.6 Noise-to-signal analysis in Hi-C, SPRITE and GAM contact maps

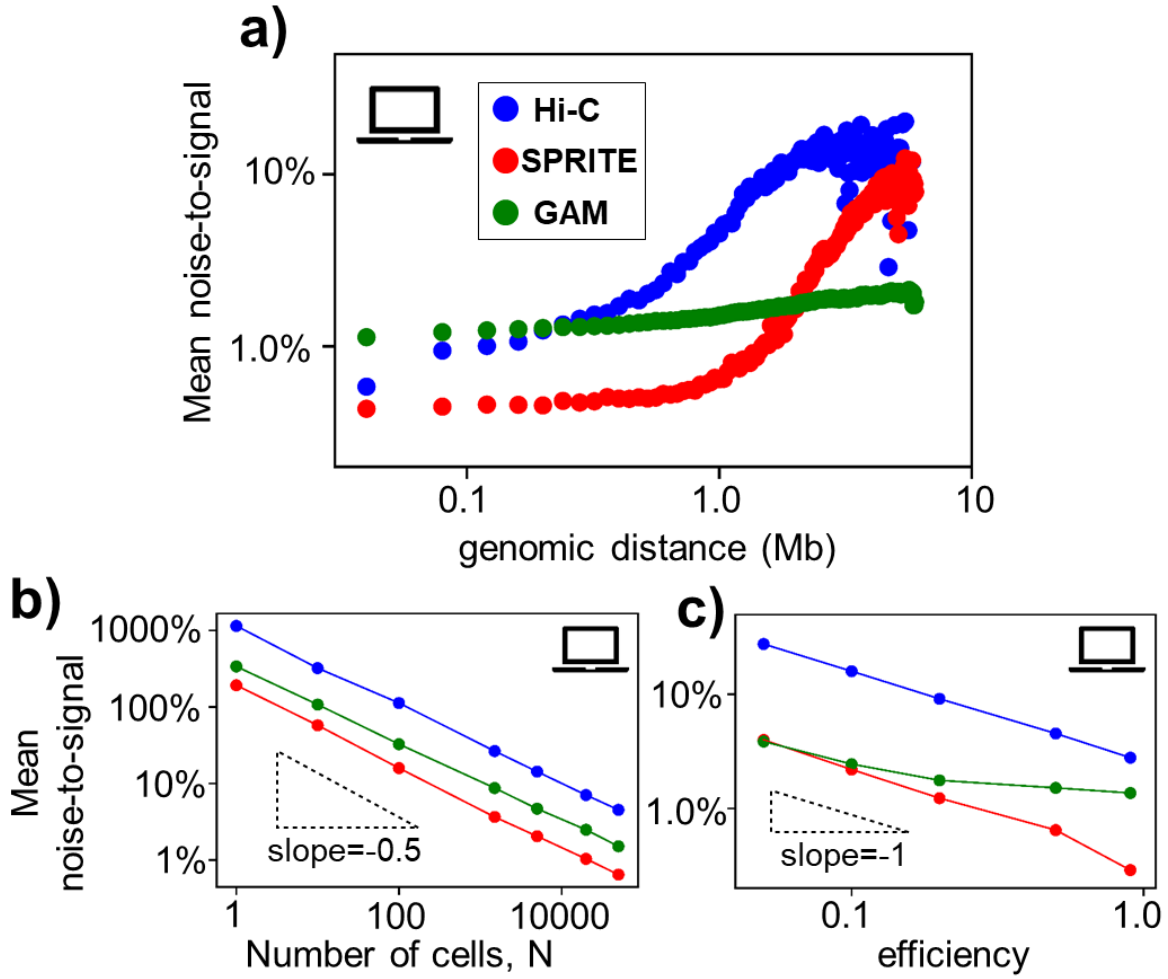


Figure 3.14: **a)** The average noise-to-signal ratio (**Main Text**) is shown against the genomic distance, in the case of the *Sox9* locus model and for fixed number of *in-silico* cells (50000, bulk limit) and efficiency (0.5). The Hi-C and SPRITE ratios are the lowest for very short genomic distances (<100kb), indicating they are the most accurate methods to detect intraTADs contacts. Then the Hi-C curve rises and becomes comparable to that of GAM, while SPRITE maintains the smallest noise-to-signal ratio until 1Mb. Here, both Hi-C and SPRITE curves have a steep increase, signaling that their detection of contacts gets noisier for supermegabase genomic distances. Conversely, the noise-to-signal ratio for GAM is overall constant across all the genomic separations and is the lowest above 1Mb, showing that GAM is the most effective technology to reveal long-range contacts, in agreement with the findings of **Figure 3.7.**
b) The average noise-to-signal ratios are plotted for Hi-C, SPRITE and GAM against the number of *in-silico* cells, N, for fixed efficiency (0.5) and at the genomic distance of 1Mb. The ratios decrease with N as an inverse square root relation, accordingly to statistical expectations based on the Central Limit Theorem.
c) The mean noise-to-signal ratio is shown against the efficiency, for N=50000 and genomic separation 1Mb. For small efficiencies and for all three technologies, the average noise-to-signal decreases approximately as an inverse relation.
Adapted from [36].

The study of reproducibility indirectly elucidated the noise level of the Hi-C, SPRITE and GAM contact maps. Here, we aim at quantifying the amount of noise affecting the contact matrices in different conditions.

That was done in the following way. For a given number of *in-silico* cells and efficiency value, several replicate experiments are performed, returning corresponding replicate contact matrices. Then, for each matrix entry (i,j) the standard deviation σ_{ij} and average μ_{ij} are computed across the ensemble of replicate matrices. The noise-to-signal ratio for the entry (i,j) is defined as σ_{ij}/μ_{ij} and evaluates the amount of fluctuations affecting the entry.

First, we investigated how the noise-to-signal ratios vary with the genomic separation between i and j, where, for a locus model at resolution *res*, the genomic distance between the windows i and j is given by $res * |i-j|$ (**Figure 3.14a**). To that aim, for the *Sox9* locus model and for each technology, we generated the replicate contact maps from 50000 *in-silico* cells and at efficiency 0.5: that ensures all the technologies are evaluated in the bulk limit (see **Figure 3.10**). Then, we studied the average noise-to-signal ratio over all window pairs with fixed genomic separation. We observed that the mean noise-to-signal behave very differently across Hi-C, SPRITE and GAM, consistently with the findings about M (**paragraph 3.3.4**). Specifically, Hi-C and SPRITE present the lowest noise level at very short genomic distances (less than 1% for distances < 100Kb). Then, Hi-C noise-to-signal ratio starts rising and gets comparable to that observed for GAM, while the noise level of SPRITE keeps being the lowest within 1Mb. At 1Mb both Hi-C and SPRITE noise-to-signal ratios have a steep increase, plateauing at approximately 10% when the total length of the locus is reached (6Mb). That is interesting in that 1Mb is the scale of TADs: consistent with the findings of **paragraph 3.3.2** (where contact maps were compared with the average distance map), such behavior of the noise-to-signal indicates Hi-C and SPRITE observations become noisier for long-range contacts, e.g. for interTADs contacts. On the other hand, Hi-C and SPRITE result the most effective in revealing contacts at short genomic separation, e.g. inside TADs. As for GAM, the noise-to-signal mean ratio is almost constant along all the genomic distances and is the lowest after 1Mb. Hence, in agreement with **paragraph 3.3.2**, the GAM method appears the most effective to capture organizational features of chromatin involving windows far apart along the genome.

Next, we explored the dependence of the noise-to-signal ratio with the number of simulated cells, N. We produced replicate *in-silico* experiments for several cell numbers, at efficiency 0.5 and focused on the noise-to-signal mean value at the genomic distance of 1Mb. For all three technologies, we obtained a power law behavior (**Figure 3.14b**), that is the average noise-to-signal ratio scales with N^(-0.5). That is expected from the Central Limit Theorem. Consistent with the observations and findings described in **paragraphs 3.3.3** and **3.3.4**, for 1 *in-silico* cell (single-cell experiment) the noise level can get as high as 1000% in Hi-C, due to the variability of 3D conformations across cells. For N above 10000, the noise-to-signal ratios are all under 10%, as the bulk limit is reached: the reproducibility is clearly associated to the reduction of the contact maps noise level.

Finally, the noise-to-signal ratio was analyzed for various values of efficiency, fixing the number of cells at 50000 and the genomic separation at 1Mb. For all three methods, an inverse power law is observed for small efficiencies (**Figure 3.14c**).

All those investigations were conducted also for the *HoxD, Epha4* and HCT116 loci, obtaining analogous results (**Figure 3.15**). That supports our findings have general validity and are not influenced by the cell line or the resolution or the locus size considered.

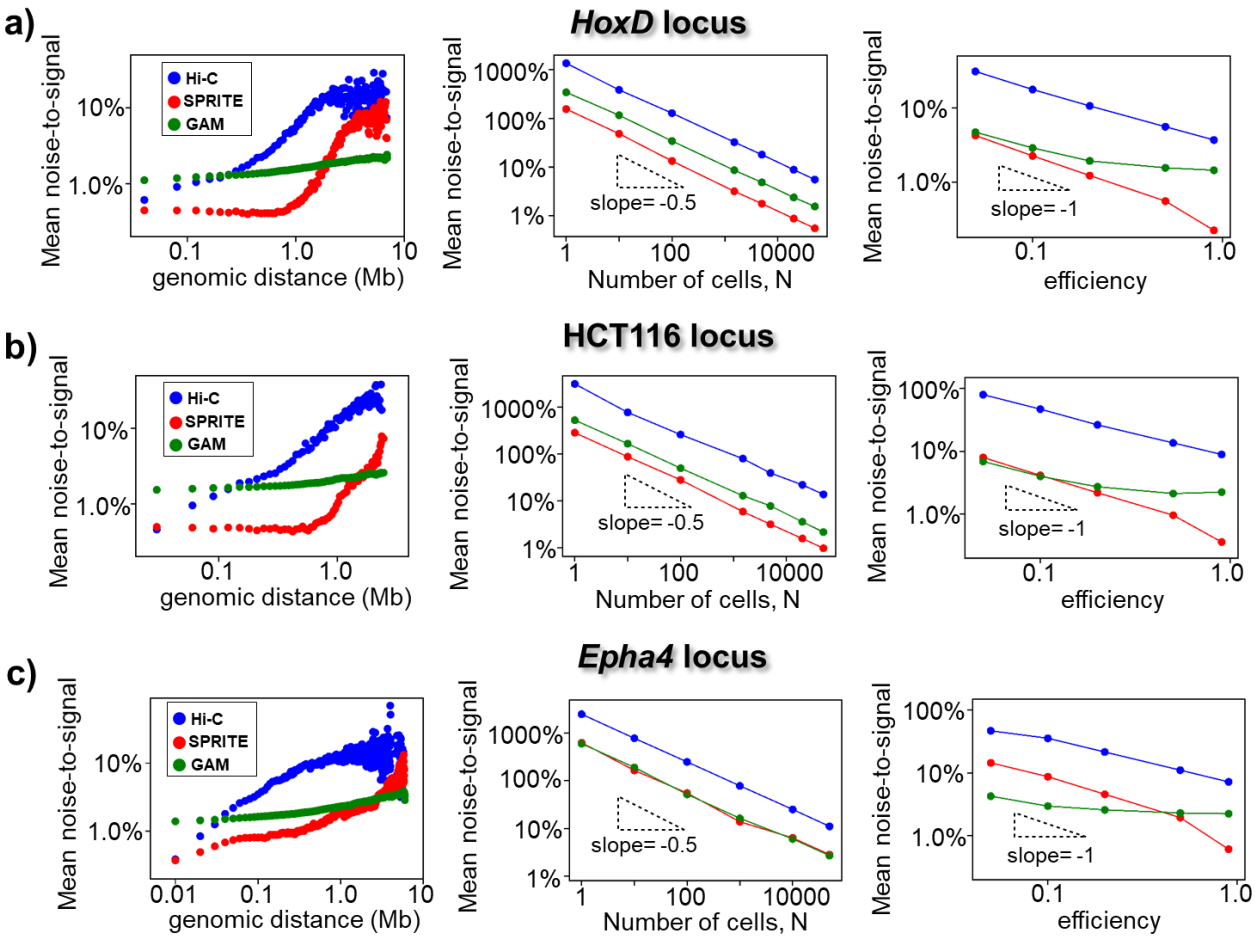


Figure 3.15: Analogous results to those shown in **Figure 3.14** for the *Sox9* locus model were found for the *HoxD* locus (panel a), the HCT116 locus (panel b) and the *Epha4* locus (panel c) models. Indeed, in all cases Hi-C and SPRITE present the lowest noise-to-signal ratio at short genomic distances, while for genomic separations above 1Mb GAM has the smallest noise level (plots on the left). The average noise-to-signal ratios scale with the number of *in-silico* cells, N, as N^(-0.5) (middle plots) and decrease approximately with an inverse relation with the efficiency (plots on the right). In all cases the same conditions described in **Figure 3.14** were used. Overall, that strongly supports the robustness of our analyses.

3.3.7 Investigations on a GAM-inferred polymer model and on a toy model confirm the robustness of the approach

All the considered polymer models were derived by PRISMR from Hi-C data (**paragraph 2.2**). However, we have shown that the polymer models can reproduce faithfully SPRITE and GAM data (**paragraph 3.3.1**) and, in the case of the HCT116 locus, also imaging data. That strongly supports the analyses on the performance of Hi-C, SPRITE and GAM are not biased by the procedure used to generate our polymers. Additionally, as specified above, for our study to be meaningful the only requirement is that the polymer conformations must be comparable in complexity and variability to those observed for DNA filaments. Nonetheless, to prove further the robustness of our conclusions, we repeated all the analyses exposed before on a polymer model [100] inferred from GAM data [8], namely from the GAM data of the same *Sox9* locus we used as case study (chr11:109-115Mb, 40kb, mouse staminal cells). To do so, PRISMR was modified so to run on GAM matrices [100].

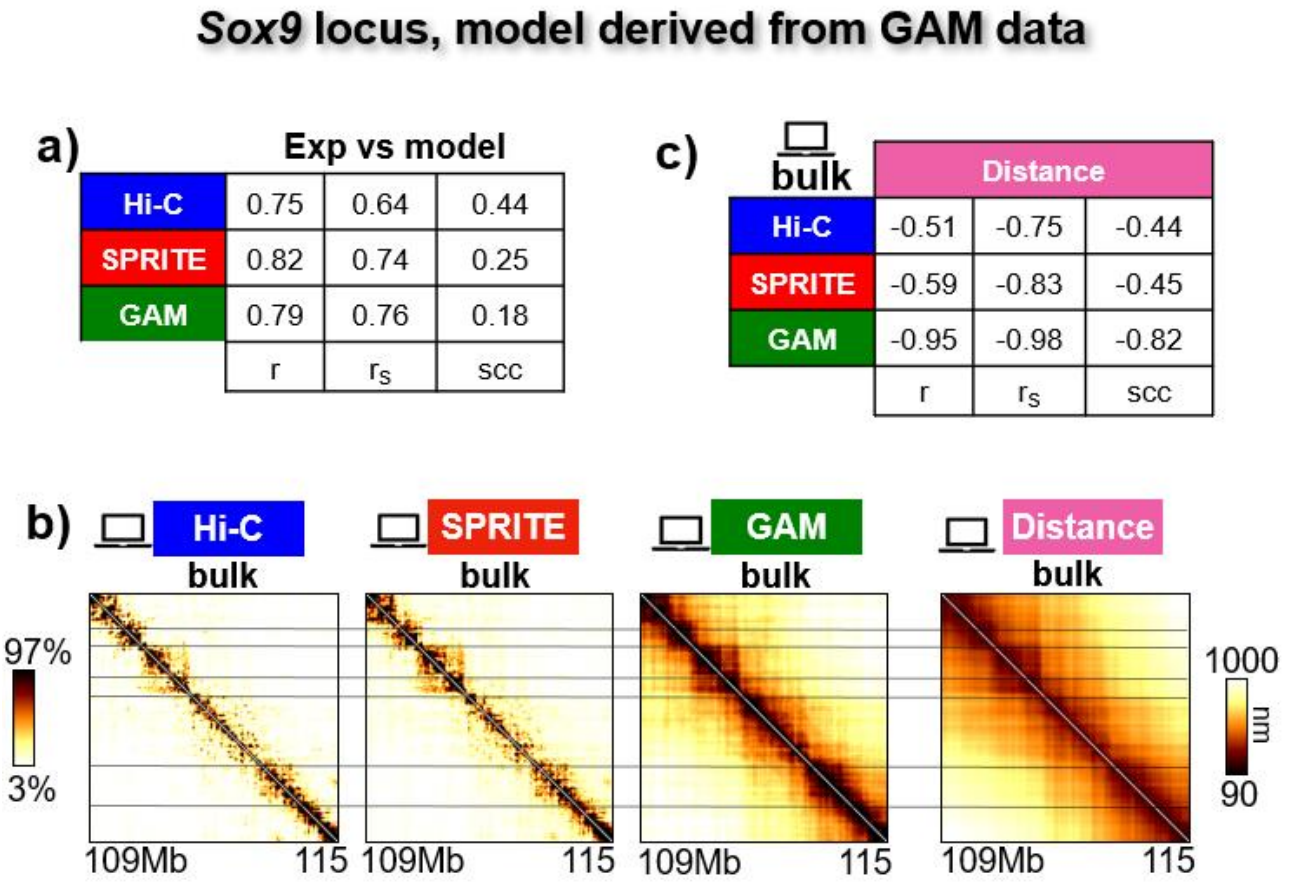


Figure 3.16: a) We considered 3D structures for the mESC *Sox9* locus derived from GAM data [8, 100]. The corresponding Hi-C, SPRITE and GAM *in-silico* contact maps were compared to the experimental data (the same used in **Figures 3.4, 3.5**) and their Pearson, Spearman and HiCRep correlations are reported.
b) The *in-silico* bulk contact maps, at efficiency 1, are compatible with the average distance pattern obtained from the ensemble of GAM-derived 3D conformations. Horizontal lines are drawn to highlight the patterns detected across the contact and the distance maps. For the contact maps, color scale indicates the percentiles. The GAM matrix returns better the long-range structural patterns than Hi-C and SPRITE. Such findings are in line with those seen for all the other loci (**Figures 3.7, 3.8**).
c) Pearson, Spearman and HiCRep correlations are reported between each bulk contact map and the average distance map. The correlations are generally good and highest for GAM, confirming the visual inspection.
Adapted from [36].

First, we verified that the *in-silico* Hi-C, SPRITE and GAM contact maps reproduce the corresponding experimental matrices (the same employed for the Hi-C derived case). Correlations values are overall good (**Figure 3.16a**), yet on average lower than those found for the Hi-C inferred model. That can be explained through our own analysis about reproducibility. Indeed, the GAM dataset employed to derive the polymer ensemble was extracted from 408 cells [8]. As said in **paragraph 3.3.4**, at the typical experimental efficiencies 408 is almost half of the number of cells needed for reproducibility, so we expect the polymer model to be affected by the noise level still present in the data. That explains the comparatively reduced correlations of the *in-silico* maps with Hi-C or SPRITE experimental matrices [9, 11] and also with the GAM experimental map, derived from 1122 cells in a more recent experiment [36]. So, the ensemble of polymer 3D configurations inferred from GAM can be considered a proxy of possible chromatin architectures and thus can be used to benchmark Hi-C, SPRITE and GAM performances.

We compared the *in-silico* Hi-C, SPRITE and GAM maps in bulk condition with the average distance matrix (**Figure 3.16b,c**), studied the reproducibility and computed the noise-to-signal ratio (**Figure 3.17a,b**) as described before. Strikingly, for all the analyses analogous results were found to those obtained from the Hi-C derived loci. That strengthens the generality of our conclusions, showing they are not affected by the input of PRISMR.

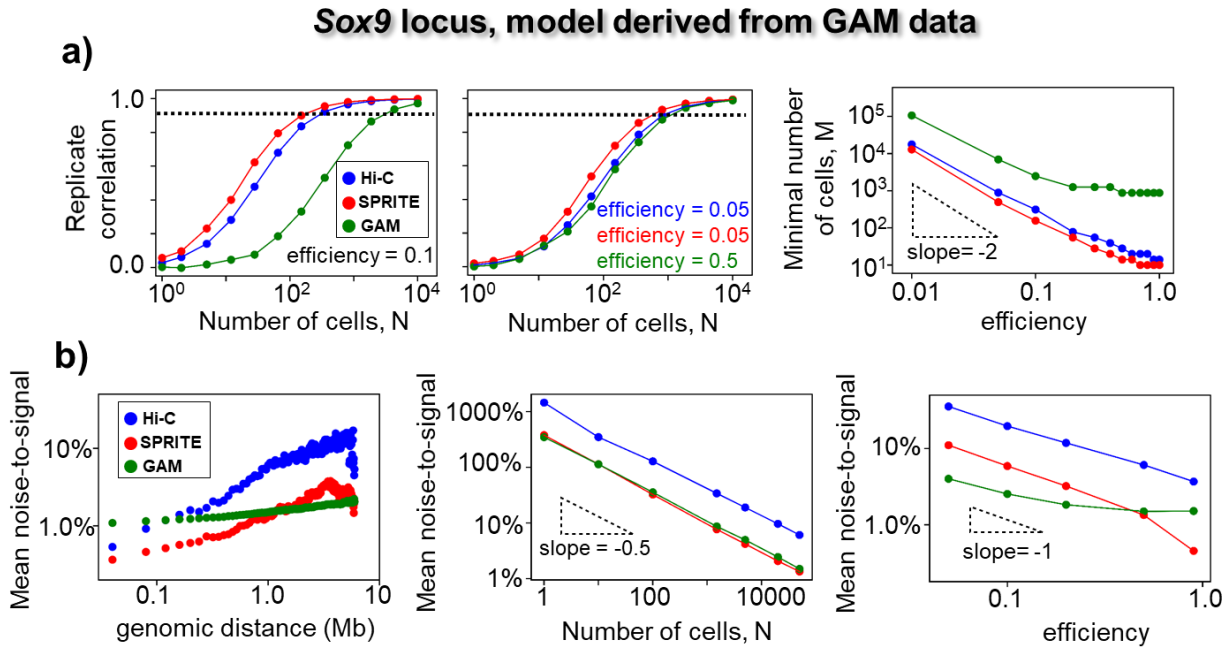


Figure 3.17: a) (left) For the *Sox9* locus model inferred by GAM data, the Pearson correlation between replicate *in-silico* contact maps is plotted against the considered number of simulated cells, N, at fixed 0.1 efficiency. The horizontal dashed line marks the threshold for reproducibility, r_t=0.90. The curves of correlations grow at different speeds, eventually reaching a plateau. That results in different M for each technology, where M is the minimal number of cells corresponding to reproducibility. (middle) The same plot is proposed using efficiencies typical of real experiments (0.05 for Hi-C and SPRITE, 0.5 for GAM). Here, the values of M get closer across the three methods. (right) The M values are shown for different efficiency values: we found that, at small efficiencies, M increases approximately as an inverse square law for all the technologies. Notably, at all efficiencies, SPRITE has the lowest value of M.
b) (left) The average noise-to-signal ratio is studied against the genomic separation, for fixed efficiency and *in-silico* cell number (0.5 and 50000, respectively). The ratio of SPRITE is the lowest within 1Mb, then sharply increase. Similarly for Hi-C, with the increase starting before 1Mb. GAM has an overall constant noise-to-signal ratio, which is the lowest above 1Mb. (middle) For each technology, the mean noise-to-signal ratio against the number of *in-silico* cells, for given efficiency (0.5) and genomic distance (1Mb), decreases as an inverse root relation, in agreement with expectations based on the Central Limit Theorem. (right) For fixed genomic distance (1Mb) and 50000 *in-silico* cells and for all three methods, the mean noise-to-signal ratio is approximately inversely proportional to the efficiency, when that is small enough. All such findings are fully compatible with those extracted from the Hi-C inferred polymer models (**Figures 3.10,3.12,3.14,3.15**).

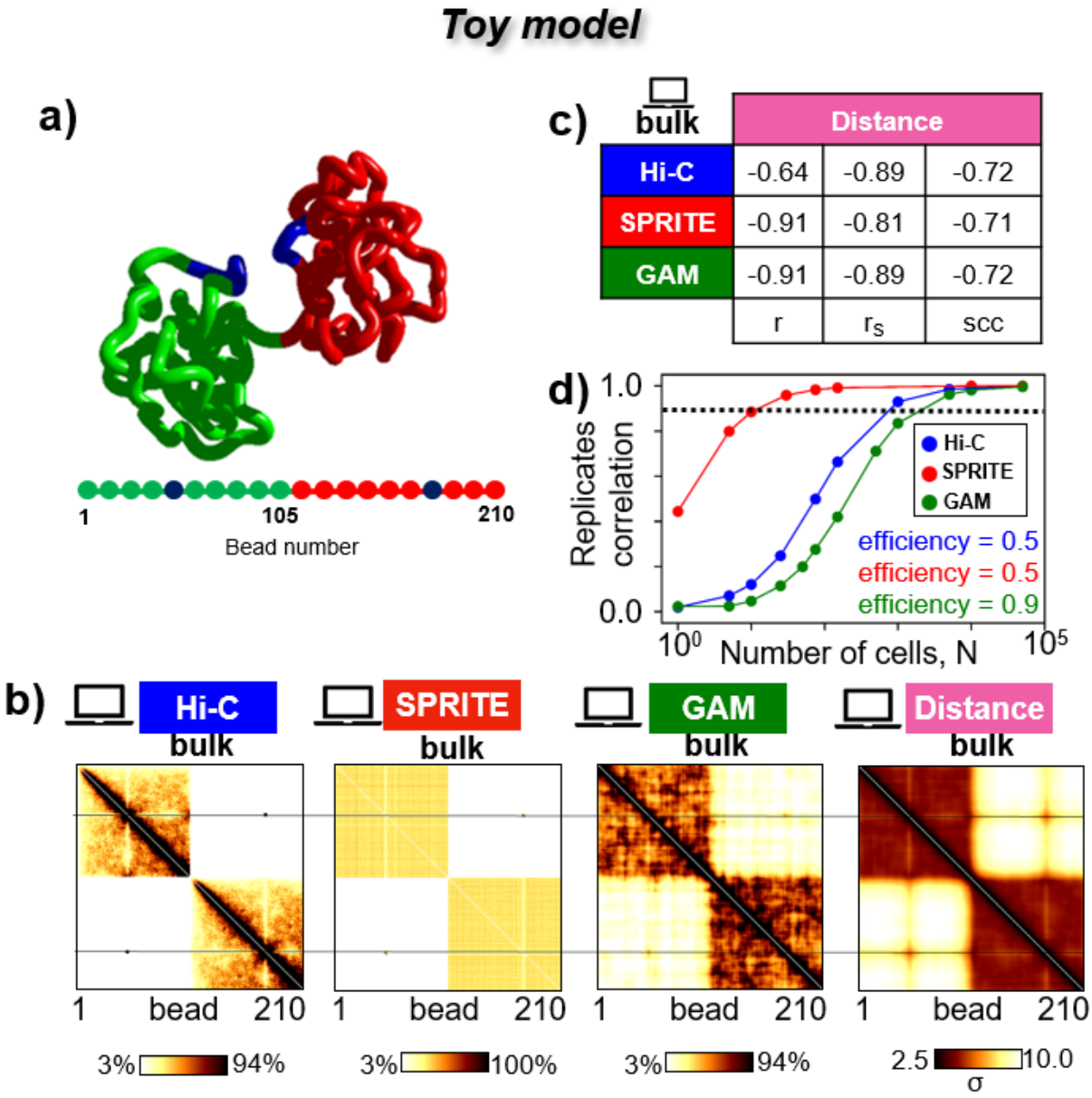


Figure 3.18: **a)** We considered a simple block copolymer model made of 210 beads and with three types of binding sites, unrelated to any real genomic locus. Two types of binding sites occupy each the first and the second half of the polymer chain (visualized in green and red). A single binding site of the third type (visualized in blue) is then placed inside each of the two halves of the chain. The example of a phase-separated 3D structure is shown.

b) *In-silico* Hi-C, SPRITE and GAM bulk contact maps at 100% efficiency all yield contact patterns compatible with the average distance matrix derived from our ensemble of conformations. The horizontal lines are a guide to the eye. Color scale for the contact maps indicates the percentiles, while, for the distance map, color scale is given in units of σ, the diameter of a polymer bead.

c) The Pearson, Spearman and HiCRep correlations between each bulk contact matrix and the average distance map are reported.

d) Replicate Pearson correlations are plotted v.s. the number of cells N, for efficiencies equal to 0.5 for Hi-C and SPRITE, 0.9 for GAM. The curves approach the reproducibility threshold (dashed line at r_t=0.9) at very different paces, with consequent different values of M, the minimal cell number to reproducibility.

All the above results are overall comparable to those obtained from the models of all the other loci analyzed (**Figures 3.7,3.8,3.10,3.12**).

Taken from [36].

Finally, we considered a simple block-copolymer model unrelated to any real genomic region. The polymer was created with two separated blocks of different binding sites (visualized as green and red in **Figure 3.18a**). Inside each block, a single binding site of another kind was inserted (colored in blue in **Figure 3.18a**). At stationarity, the average folded conformation of such a polymer displays two separated self-interacting globules; the blue binding sites stretches out from the globular domains to interact with each other, forming a long-range contact (**Figure 3.18a**). The ensemble of polymer 3D structures was derived by Molecular Dynamics (**paragraph 2.3.3**) and the average distance matrix computed, featuring two TAD-like domains and a pointwise long-range contact, as expected. Since this is an architectural pattern resembling that of real chromatin loci, the ensemble of block-copolymers can be used to conduct meaningful computational experiments.
We generated *in-silico* bulk Hi-C, SPRITE and GAM contact maps to compare against the average distance map, getting overall similar results to those found for the models of real loci (**Figure 3.18b,c**). Similarly, we calculated the average Pearson correlation between replicates for different number of *in-silico* cells, at efficiencies 0.5, 0.5 and 0.9 for respectively Hi-C, SPRITE and GAM. The scenario is analogous to those seen for the other polymer models, with SPRITE the fastest to gain reproducibility and GAM the slowest (**Figure 3.18d**). That highlights the robustness of the approach presented here and strongly suggests our results are not dependent on the usage of the SBS polymer models of chromatin.

CONCLUSIONS

The comprehension of chromatin spatial organization in cell nuclei is a major challenge of contemporary biology. Far from being random, the architecture of chromosomes is associated to the functionality of the genome code [16–19] and its misfolding can cause serious diseases [23–25]. Hence, fully understanding the molecular mechanisms controlling DNA organization in space can be crucial for biomedicine. To tackle such a conundrum, powerful technologies have been designed, from super-resolution microscopy [1–3] to sequencing tools [4–9], which unveiled key architectural features. From the theoretical viewpoint, polymer physics models and concepts have been enrolled [26–31]. Indeed, in the present work, we illustrated two remarkable approaches whereby polymer-physics can contribute to understand the architecture of chromatin.

(i) First, we showed that polymer-physics models can elucidate the molecular processes driving the organizational features detected by experiments [27]. We considered a textbook polymer model of chromatin, the *Strings&Binders Switch* (SBS) model [37, 38], where contacts between distal binding sites of the polymer are mediated by diffusing cognate molecules. Binding sites are located along the polymer chain based on Hi-C ensemble data [33] and are arranged so that different types of binding domains are significantly intertwined, as opposed to simple linear block-copolymers. We showed that the SBS polymers undergo a coil-globule phase transition and proved that their 3D conformations in the globule steady state recapitulate the imaged structures of a chromatin locus in human cells [32]. Specifically, the cell-to-cell variability of TADs and subTADs observed experimentally [32] was explained as byproduct of the phase separation mechanism driving the coil-globule transition. Indeed, the diverse types of binding sites and, above all, their overlap along the polymer chain imply that numerous arrangements of globules can assemble during phase-separation, resulting in high degeneracy and, so, structural variability. The abundancy of specific binding sites in specific locations of the polymer determines the more frequently assembled globules and accounts for the average TAD pattern emerging in experiments mediated over population of cells [32, 33]. Additionally, we found that the effects of cohesin depletion revealed by experimental investigations [32, 33] are also explainable in the framework of the coil-globule phase transition, as cohesin can be interpreted as a thermodynamic switch. Upon its removal, most of the SBS polymers are reverted to the coil phase, where interactions between binding sites are few and overwhelmed by thermal fluctuations, so that domains of contacts form out of random collisions and rapidly dissolve. We showed that such mechanism describes well the 3D conformations imaged in cohesin-depleted human cells [32] and also explains the featureless architecture detected averaging over many cells.

Overall, our study reveals that the Topologically Associated Domains [11, 12] could arise spontaneously from the phase separation of classical polymers in the globule state. The role of cohesin may be that to enforce the globule thermodynamic phase and select the preferred domains, albeit the exact way it manages in doing so should be further explored. This picture ultimately roots chromatin 3D organization in the diffusion of binder molecules leading to stationary thermodynamic configurations. Thus, it offers a different perspective than the off-equilibrium scenario of the Loop Extrusion model of chromatin [29, 34, 35], where ATP-pumped processes drive the folding. However, interestingly, a diffusive variant of the extrusion could co-exist with diffusion-based, steady-state polymers [60]. So, one could explore combined sophisticated models where both phase separation and extrusion mechanisms are at play or where variants of them interplay to return chromatin organization at multiple scales [86].

(ii) Second, we used polymer-physics models of chromatin to benchmark the performances of technologies designed to detect DNA 3D organization. In this approach, model polymeric structures are employed to test and evaluate the quality of the data generated by experiments, in a simplified yet rigorous way [36]. We implemented algorithms to simulate Hi-C [7], SPRITE [9] and GAM [8] experiments on 3D conformations of polymer models and, first, verified their effectiveness in reproducing real experimental data. Then we realized computational experiments on models of many different chromatin regions, at different resolutions and from diverse cell lines and organisms, and also on a simple block-copolymer unrelated to any real DNA locus. In all cases, we found analogous results, proving their generality. Specifically, we found that all three technologies, in bulk condition, are overall effective in detecting the average architecture of chromatin from a population of cells, but, importantly, GAM was showed to be the most suited method to capture long-range spatial organizations. We illustrated that single-cell contact maps, even for ideal 100% efficiencies, poorly reproduce the pattern of distances between DNA sites in a single nucleus, which necessarily roots in the limitations intrinsic in Hi-C, SPRITE and GAM protocols. As for GAM, its single cell maps are the least reliable, because the GAM protocol was designed only for populations of cells. Then, we studied the reproducibility of Hi-C, SPRITE and GAM experiments, i.e. the minimal number of cells required to return contact matrices negligibly affected by noise. We found the three technologies approach reproducibility with significantly different speeds and, specifically, we found that for equal conditions SPRITE is the method requiring the smallest sample of cells, GAM the highest. For efficiencies close to those typically used in real experiments [8, 119], the minimal number of cells to reproducibility is 650, 250, 800 for, respectively, Hi-C, SPRITE and GAM. Next, we quantified the noise level which affects contact matrices studying the noise-to-signal ratio across replicate computational experiments. We showed that the average noise-to-signal ratio strongly varies with the genomic distance: Hi-C and SPRITE have the lowest noise-to-signal below 1Mb, then it sharply increases; the noise-to-signal is instead almost constant for GAM and is the lowest above the 1Mb scale, confirming its effectiveness in revealing long-range contacts.

Such usage of the polymer models of chromatin can guide the design of novel experiments, by elucidating the contexts where each technology is most effective. Importantly, the identification of the optimal experimental setup on computational basis can save remarkable amounts of money and materials, as the costs of biological experiments as GAM or Hi-C are still challenging.

In conclusion, we presented two approaches whereby polymer physics can play an important role in the dissection of chromatin architecture. Notably, such approaches are complementary. In the case of (i), we showed how polymer-physics models can be effectively used to make sense of the data collected in experimental investigations and to return a coherent a picture of all the architectures observed [27]. In the case (ii), we employed polymer models to test the quality of the data measured by experimental technologies and to benchmark their characteristics [36]. The intriguing perspective could be a future where polymer models of chromatin may be used to guide the preparation of novel experiments and then to explain the data produced by them, unveiling the underlying physical mechanisms.

REFERENCES

1. Cattoni, D.I., Gizzi, A.M.C., Georgieva, M., Di Stefano, M., Valeri, A., Chamousset, D., Houbron, C., Déjardin, S., Fiche, J.B., González, I., Chang, J.M., Sexton, T., Marti-Renom, M.A., Bantignies, F., Cavalli, G., Nollmann, M.: Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. Nat. Commun. 8, 1–10 (2017). https://doi.org/10.1038/s41467-017-01962-x.
2. Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.T., Zhuang, X.: Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature. (2016). https://doi.org/10.1038/nature16496.
3. Cardozo Gizzi, A.M., Cattoni, D.I., Fiche, J.B., Espinola, S.M., Gurgo, J., Messina, O., Houbron, C., Ogiyama, Y., Papadopoulos, G.L., Cavalli, G., Lagha, M., Nollmann, M.: Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. Mol. Cell. 74, 212-222.e5 (2019).
4. Dekker, J., Rippe, K., Dekker, M., Kleckner, N.: Capturing chromosome conformation. Science (80-.). 295, 1306–1311 (2002). https://doi.org/10.1126/science.1067799.
5. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., De Laat, W.: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. 38, 1348–1354 (2006). https://doi.org/10.1038/ng1896.
6. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D., Dekker, J.: Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. 16, 1299–1309 (2006). https://doi.org/10.1101/gr.5571506.
7. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). 326, 289–293 (2009).
8. Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., Fraser, J., Dostie, J., Game, L., Dillon, N., Edwards, P.A.W., Nicodemi, M., Pombo, A.: Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 543, 519–524 (2017).
9. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M.: Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell. 174, 744-757.e24 (2018).
10. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 159, 1665–1680 (2014).
11. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 485, 376–380 (2012).

12. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., Heard, E.: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 485, 381–385 (2012).
13. Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., Xie, S.Q., Morris, K.J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A.R., Semple, C.A., Dostie, J., Pombo, A., Nicodemi, M.: Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. 11, 852 (2015).
14. Cremer, T., Cremer, M.: Chromosome territories. Cold Spring Harb. Perspect. Biol. 2, (2010). https://doi.org/10.1101/cshperspect.a003889.
15. van Steensel, B., Belmont, A.S.: Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell. 169, 780–791 (2017). https://doi.org/10.1016/j.cell.2017.04.022.
16. Bickmore, W.A.: The Spatial Organization of the Human Genome. Annu. Rev. Genomics Hum. Genet. 14, 67–84 (2013).
17. Dekker, J., Misteli, T.: Long-range chromatin interactions. Cold Spring Harb. Perspect. Biol. 7, a019356 (2015).
18. Pombo, A., Dillon, N.: Three-dimensional genome architecture: Players and mechanisms. Nat. Rev. Mol. Cell Biol. 16, 245–257 (2015). https://doi.org/10.1038/nrm3965.
19. Dekker, J., Mirny, L.: The 3D Genome as Moderator of Chromosomal Communication. Cell. 164, 1110–1121 (2016).
20. Dixon, J.R., Gorkin, D.U., Ren, B.: Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell. 62, 668–680 (2016). https://doi.org/10.1016/j.molcel.2016.05.018.
21. Spielmann, M., Lupiáñez, D.G., Mundlos, S.: Structural variation in the 3D genome. Nat. Rev. Genet. 19, 453–467 (2018).
22. Finn, E.H., Misteli, T.: Molecular basis and biological function of variability in spatial genome organization. Science (80-.). 365, eaaw9498 (2019).
23. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S.: Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 161, 1012–1025 (2015). https://doi.org/10.1016/j.cell.2015.04.004.
24. Valton, A.L., Dekker, J.: TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. 36, 34–40 (2016). https://doi.org/10.1016/j.gde.2016.03.008.
25. Weischenfeldt, J., Dubash, T., Drainas, A.P., Mardin, B.R., Chen, Y., Stütz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B., Efthymiopoulos, T., Erkek, S., Siegl, C., Brenner, H., Brustugun, O.T., Dieter, S.M., Northcott, P.A., Petersen, I., Pfister, S.M., Schneider, M., Solberg, S.K., Thunissen, E., Weichert, W., Zichner, T., Thomas, R., Peifer, M., Helland, A., Ball, C.R., Jechlinger, M., Sotillo, R., Glimm, H., Korbel, J.O.: Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. (2017). https://doi.org/10.1038/ng.3722.
26. Jost, D., Carrivain, P., Cavalli, G., Vaillant, C.: Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. 42, 9553–61 (2014).
27. Conte, M., Fiorillo, L., Bianco, S., Chiariello, A.M., Esposito, A., Nicodemi, M.: Polymer physics indicates chromatin folding variability across single-cells results from state

degeneracy in phase separation. Nat. Commun. 11, 3289 (2020).
28. Brackley, C.A., Taylor, S., Papantonis, A., Cook, P.R., Marenduzzo, D.: Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc. Natl. Acad. Sci. U. S. A. 110, E3605–E3611 (2013).
29. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., Mirny, L.A.: Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 15, 2038–2049 (2016). https://doi.org/10.1016/j.celrep.2016.04.085.
30. Rosa, A., Zimmer, C.: Computational models of large-scale genome architecture. In: International Review of Cell and Molecular Biology. pp. 275–349. Elsevier Inc. (2014). https://doi.org/10.1016/B978-0-12-800046-5.00009-6.
31. Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G., Marti-Renom, M.A.: Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol. 13, e1005665 (2017).
32. Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., Zhuang, X.: Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science (80-.). 362, eaau1783 (2018).
33. Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S., Aiden, E.L.: Cohesin Loss Eliminates All Loop Domains. Cell. 171, 305-320.e24 (2017). https://doi.org/10.1016/j.cell.2017.09.026.
34. Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., Aiden, E.L.: Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A. 112, E6456–E6465 (2015).
35. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., Mirny, L.A.: Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harb. Symp. Quant. Biol. (2017). https://doi.org/10.1101/sqb.2017.82.034710.
36. Fiorillo, L., Musella, F., Kempfer, R., Chiariello, A.M., Bianco, S., Kukalev, A., Irastorza-Azcarate, I., Esposito, A., Conte, M., Prisco, A., Pombo, A., Nicodemi, M.: Comparison of the Hi-C, GAM and SPRITE methods by use of polymer models of chromatin. bioRxiv. 2020.04.24.059915 (2020).
37. Nicodemi, M., Prisco, A.: Thermodynamic pathways to genome spatial organization in the cell nucleus. Biophys. J. 96, 2168–2177 (2009).
38. Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.M., Dostie, J., Pombo, A., Nicodemi, M.: Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl. Acad. Sci. U. S. A. 109, 16173–16178 (2012).
39. Annunziatella, C., Chiariello, A.M., Esposito, A., Bianco, S., Fiorillo, L., Nicodemi, M.: Molecular Dynamics simulations of the Strings and Binders Switch model of chromatin. Methods. 142, 81–88 (2018). https://doi.org/10.1016/j.ymeth.2018.02.024.
40. Bulger, M., Groudine, M.: Functional and mechanistic diversity of distal transcription enhancers. Cell. 144, 327–339 (2011). https://doi.org/10.1016/j.cell.2011.01.024.
41. Calo, E., Wysocka, J.: Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell. 49, 825–837 (2013). https://doi.org/10.1016/j.molcel.2013.01.038.
42. Ong, C.T., Corces, V.G.: Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. 12, 283–293 (2011). https://doi.org/10.1038/nrg2957.
43. Schalch, T., Duda, S., Sargent, D.F., Richmond, T.J.: X-ray structure of a tetranucleosome and

its implications for the chromatin fibre. Nature. 436, 138–141 (2005). https://doi.org/10.1038/nature03686.
44. Fraser, J., Williamson, I., Bickmore, W.A., Dostie, J.: An Overview of Genome Organization and How We Got There: from FISH to Hi-C. Microbiol. Mol. Biol. Rev. 79, 347–372 (2015). https://doi.org/10.1128/mmbr.00006-15.
45. Grigoryev, S.A., Woodcock, C.L.: Chromatin organization - The 30nm fiber. Exp. Cell Res. 318, 1448–1455 (2012). https://doi.org/10.1016/j.yexcr.2012.02.014.
46. Razin, S. V, Gavrilov, A.A.: Chromatin without the 30-nm fiber: constrained disorder instead of hierarchical folding. Epigenetics. 9, 653–7 (2014). https://doi.org/10.4161/epi.28297.
47. Tremethick, D.J.: Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. Cell. 128, 651–654 (2007). https://doi.org/10.1016/j.cell.2007.02.008.
48. Hsieh, T.H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Rando, O.J., Tjian, R., Darzacq, X.: Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol. Cell. 78, 539-553.e8 (2020). https://doi.org/10.1016/j.molcel.2020.03.002.
49. Zhang, H., Emerson, D.J., Gilgenast, T.G., Titus, K.R., Lan, Y., Huang, P., Zhang, D., Wang, H., Keller, C.A., Giardine, B., Hardison, R.C., Phillips-Cremins, J.E., Blobel, G.A.: Chromatin structure dynamics during the mitosis-to-G1 phase transition. Nature. 576, 158–162 (2019). https://doi.org/10.1038/s41586-019-1778-y.
50. Chiariello, A.M., Bianco, S., Marieke, A., Higgs, D.R., Hughes, J.R., Nicodemi, M., Marieke Oudelaar, A., Esposito, A., Annunziatella, C., Fiorillo, L., Conte, M., Corrado, A., Prisco, A., Larke, M.S.C., Telenius, J.M., Sciarretta, R., Musella, F., Buckle, V.J.: A Dynamic Folded Hairpin Conformation Is Associated with α-Globin Activation in Erythroid Cells. Cell Rep. 30, (2020). https://doi.org/10.1016/j.celrep.2020.01.044.
51. Maass, P.G., Barutcu, A.R., Weiner, C.L., Rinn, J.L.: Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. Mol. Cell. 69, 1039-1045.e3 (2018). https://doi.org/10.1016/j.molcel.2018.02.007.
52. Fiorillo, L., Bianco, S., Esposito, A., Conte, M., Sciarretta, R., Musella, F., Chiariello, A.M.: A modern challenge of polymer physics: Novel ways to study, interpret, and reconstruct chromatin structure. Wiley Interdiscip. Rev. Comput. Mol. Sci. e1454 (2019).
53. Tiana, G., Giorgetti, L.: Modeling the 3D Conformation of Genomes. CRC Press, Taylor & Francis Group (2019). https://doi.org/10.1201/9781315144009.
54. Nicodemi, M., Pombo, A.: Models of chromosome structure. Curr. Opin. Cell Biol. 28, 90–95 (2014).
55. Dans, P.D., Walther, J., Gómez, H., Orozco, M.: Multiscale simulation of DNA. Curr. Opin. Struct. Biol. 37, 29–45 (2016). https://doi.org/10.1016/j.sbi.2015.11.011.
56. Portillo-Ledesma, S., Schlick, T.: Bridging chromatin structure and function over a range of experimental spatial and temporal scales by molecular modeling. Wiley Interdiscip. Rev. Comput. Mol. Sci. 10, (2020). https://doi.org/10.1002/wcms.1434.
57. Collepardo-Guevara, R., Schlick, T.: Chromatin fiber polymorphism triggered by variations of DNA linker lengths. Proc. Natl. Acad. Sci. U. S. A. 111, 8061–8066 (2014). https://doi.org/10.1073/pnas.1315872111.
58. Fan, Y., Korolev, N., Lyubartsev, A.P., Nordenskiöld, L.: An Advanced Coarse-Grained Nucleosome Core Particle Model for Computer Simulations of Nucleosome-Nucleosome Interactions under Varying Ionic Conditions. PLoS One. 8, (2013). https://doi.org/10.1371/journal.pone.0054228.
59. Sun, T., Mirzoev, A., Minhas, V., Korolev, N., Lyubartsev, A.P., Nordenskiöld, L.: A multiscale analysis of DNA phase separation: from atomistic to mesoscale level. Nucleic Acids Res. 47, 5550–5562 (2019). https://doi.org/10.1093/nar/gkz377.

60. Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R., Marenduzzo, D.: Nonequilibrium Chromosome Looping via Molecular Slip Links. Phys. Rev. Lett. (2017). https://doi.org/10.1103/PhysRevLett.119.138101.
61. Rosa, A., Everaers, R.: Structure and dynamics of interphase chromosomes. PLoS Comput. Biol. 4, e1000153 (2008). https://doi.org/10.1371/journal.pcbi.1000153.
62. Nelson, P.: Transport of torsional stress in DNA. Proc. Natl. Acad. Sci. U. S. A. 96, 14342–14347 (1999). https://doi.org/10.1073/pnas.96.25.14342.
63. Trask, B.J.: Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. Trends Genet. 7, 149–154 (1991). https://doi.org/10.1016/0168-9525(91)90378-4.
64. Speicher, M.R., Carter, N.P.: The new cytogenetics: Blurring the boundaries with molecular biology. Nat. Rev. Genet. 6, 782–792 (2005). https://doi.org/10.1038/nrg1692.
65. Volpi, E. V., Bridger, J.M.: FISH glossary: An overview of the fluorescence in situ hybridization technique. Biotechniques. 45, 385–409 (2008). https://doi.org/10.2144/000112811.
66. Williamson, I., Eskeland, R., Lettice, L.A., Hill, A.E., Boyle, S., Grimes, G.R., Hill, R.E., Bickmore, W.A.: Anterior-posterior differences in HoxD chromatin topology in limb development. Dev. 139, 3157–3167 (2012). https://doi.org/10.1242/dev.081174.
67. Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R.S., Paquette, D., Dostie, J., Bickmore, W.A.: Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev. 28, 2778–2791 (2014). https://doi.org/10.1101/gad.251694.114.
68. Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., Shiroishi, T.: Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. Dev. Cell. 16, 47–57 (2009). https://doi.org/10.1016/j.devcel.2008.11.011.
69. Chambeyron, S., Bickmore, W.A.: Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes Dev. 18, 1119–1130 (2004). https://doi.org/10.1101/gad.292104.
70. Branco, M.R., Pombo, A.: Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. 4, 780–788 (2006). https://doi.org/10.1371/journal.pbio.0040138.
71. Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., Cavalli, G.: Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 171, 557-572.e24 (2017). https://doi.org/10.1016/j.cell.2017.09.043.
72. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P.: Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 502, 59–64 (2013).
73. Hsieh, T.H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., Rando, O.J.: Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 162, 108–119 (2015). https://doi.org/10.1016/j.cell.2015.05.048.
74. Díaz, N., Kruse, K., Erdmann, T., Staiger, A.M., Ott, G., Lenz, G., Vaquerizas, J.M.: Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nat. Commun. 9, 1–13 (2018). https://doi.org/10.1038/s41467-018-06961-0.
75. Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A.J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M.G.S., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B.,

Klenerman, D., Laue, E.D.: 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature. 544, 59–64 (2017). https://doi.org/10.1038/nature21429.
76. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., Shendure, J.: Massively multiplex single-cell Hi-C. Nat. Methods. 14, 263–266 (2017). https://doi.org/10.1038/nmeth.4155.
77. Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L.A., Tachibana-Konwalski, K.: Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature. 544, 110–114 (2017). https://doi.org/10.1038/nature21711.
78. Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., Tanay, A.: Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature. 540, 296–300 (2016). https://doi.org/10.1038/nature20158.
79. Berger, S.L.: The complex language of chromatin regulation during transcription. Nature. 447, 407–412 (2007). https://doi.org/10.1038/nature05915.
80. Kouzarides, T.: Chromatin Modifications and Their Function. Cell. 128, 693–705 (2007). https://doi.org/10.1016/j.cell.2007.02.005.
81. Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., Patel, D.J.: How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. Nat. Struct. Mol. Biol. 14, 1025–1040 (2007). https://doi.org/10.1038/nsmb1338.
82. Hsieh, T.H.S., Fudenberg, G., Goloborodko, A., Rando, O.J.: Micro-C XL: Assaying chromosome conformation from the nucleosome to the entire genome. Nat. Methods. 13, 1009–1011 (2016). https://doi.org/10.1038/nmeth.4025.
83. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., Bruneau, B.G.: Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell. (2017). https://doi.org/10.1016/j.cell.2017.05.004.
84. Nasmyth, K.: Disseminating the genome: Joining, resolving, and separating sister chromatids during mitosis and meiosis. Annu. Rev. Genet. 35, 673–745 (2001). https://doi.org/10.1146/annurev.genet.35.102401.091334.
85. Alipour, E., Marko, J.F.: Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Res. 40, 11202–11212 (2012). https://doi.org/10.1093/nar/gks925.
86. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., Mirny, L.A.: Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. (2018). https://doi.org/10.1073/pnas.1717730115.
87. Goloborodko, A., Imakaev, M. V., Marko, J.F., Mirny, L.: Compaction and segregation of sister chromatids via active loop extrusion. Elife. 5, (2016). https://doi.org/10.7554/eLife.14864.
88. Tedeschi, A., Wutz, G., Huet, S., Jaritz, M., Wuensche, A., Schirghuber, E., Davidson, I.F., Tang, W., Cisneros, D.A., Bhaskara, V., Nishiyama, T., Vaziri, A., Wutz, A., Ellenberg, J., Peters, J.M.: Wapl is an essential regulator of chromatin structure and chromosome segregation. Nature. 501, 564–568 (2013). https://doi.org/10.1038/nature12471.
89. Gassler, J., Brandão, H.B., Imakaev, M., Flyamer, I.M., Ladstätter, S., Bickmore, W.A., Peters, J., Mirny, L.A., Tachibana, K.: A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. EMBO J. 36, 3600–3618 (2017). https://doi.org/10.15252/embj.201798083.
90. Gerlich, D., Koch, B., Dupeux, F., Peters, J.M., Ellenberg, J.: Live-Cell Imaging Reveals a Stable Cohesin-Chromatin Interaction after but Not before DNA Replication. Curr. Biol. 16, 1571–

1578 (2006). https://doi.org/10.1016/j.cub.2006.06.068.
91. Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., Darzacq, X.: CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Elife. 6, (2017). https://doi.org/10.7554/eLife.25776.001.
92. Terakawa, T., Bisht, S., Eeftens, J.M., Dekker, C., Haering, C.H., Greene, E.C.: The condensin complex is a mechanochemical motor that translocates along DNA. Science (80-.). (2017). https://doi.org/10.1126/science.aan6516.
93. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., Spitz, F.: Two independent modes of chromatin organization revealed by cohesin removal. Nature. (2017). https://doi.org/10.1038/nature24281.
94. Negri, M., Gherardi, M., Tiana, G., Cosentino Lagomarsino, M.: Spontaneous domain formation in disordered copolymers as a mechanism for chromosome structuring. Soft Matter. 14, 6128–6136 (2018). https://doi.org/10.1039/c8sm00468d.
95. Chiariello, A.M., Annunziatella, C., Bianco, S., Esposito, A., Nicodemi, M.: Polymer physics of chromosome large-scale 3D organisation. Sci. Rep. 6, 29775 (2016).
96. Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., Mundlos, S., Nicodemi, M.: Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. 50, 662–667 (2018).
97. Bianco, S., Annunziatella, C., Andrey, G., Chiariello, A.M., Esposito, A., Fiorillo, L., Prisco, A., Conte, M., Campanile, R., Nicodemi, M.: Modeling Single-Molecule Conformations of the HoxD Region in Mouse Embryonic Stem and Cortical Neuronal Cells. Cell Rep. 28, 1574-1583.e4 (2019).
98. Lyu, H., Liu, E., Wu, Z.: Comparison of normalization methods for Hi-C data. Biotechniques. 68, 56–64 (2020). https://doi.org/10.2144/btn-2019-0105.
99. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092 (1953). https://doi.org/10.1063/1.1699114.
100. Fiorillo, L., Bianco, S., Chiariello, A.M., Barbieri, M., Esposito, A., Annunziatella, C., Conte, M., Corrado, A., Prisco, A., Pombo, A., Nicodemi, M.: Inference of chromosome 3D structures from GAM data by a physics computational approach. Methods. 181–182, 70–79 (2020). https://doi.org/10.1016/j.ymeth.2019.09.018.
101. Plimpton, S.: Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. 117, 1–19 (1995).
102. Weeks, J.D., Chandler, D., Andersen, H.C.: Role of repulsive forces in determining the equilibrium structure of simple liquids. J. Chem. Phys. 54, 5237–5247 (1971). https://doi.org/10.1063/1.1674820.
103. Kremer, K., Grest, G.S.: Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. J. Chem. Phys. 92, 5057–5086 (1990).
104. De Gennes, P.G.: Scaling concepts in polymer physics. Cornell university press. Ithaca N.Y.,. 324 (1979).
105. Bates, F.S., Fredrickson, G.H.: Block copolymer thermodynamics: Theory and experiment. Annu. Rev. Phys. Chem. 41, 525–557 (1990). https://doi.org/10.1146/annurev.pc.41.100190.002521.
106. Hamley, I.W.: The Physics of Block Copolymers (Oxford Science Publications). (1999).
107. Matsushita, Y.: Microphase Separation (of Block Copolymers) BT - Encyclopedia of Polymeric Nanomaterials. Presented at the (2014). https://doi.org/10.1007/978-3-642-36199-9_149-1.

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

Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D.L., Byron, R., Canfield, T.K., Diegel, M.J., Dunn, D., Ebersol, A.K., Frum, T., Garg, K., Gist, E., Hansen, R.S., Boatman, L., Haugen, E., Humbert, R., Johnson, A.K., Johnson, E.M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M.T., Neph, S.J., Neri, F. V., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Rynes, E., Sanchez, M.E., Sandstrom, R.S., Shafer, A.O., Stergachis, A.B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M.A., Yan, Y., Zhang, M., Akey, J.M., Bender, M., Dorschner, M.O., Groudine, M., MacCoss, M.J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N.M., Sobral, D., Vaquerizas, J.M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M.W., Schaub, M.A., Miller, W., Bickel, P.J., Banfai, B., Boley, N.P., Huang, H., Li, J.J., Noble, W.S., Bilmes, J.A., Buske, O.J., Sahu, A.D., Kharchenko, P. V., Park, P.J., Baker, D., Taylor, J., Lochovsky, L.: An integrated encyclopedia of DNA elements in the human genome. Nature. 489, 57–74 (2012). https://doi.org/10.1038/nature11247.

109. Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A.M., Jerković, I., Harabula, I., Guckelberger, P., Pechstein, M., Wittler, L., Chan, W.L., Franke, M., Lupiáñez, D.G., Kraft, K., Timmermann, B., Vingron, M., Visel, A., Nicodemi, M., Mundlos, S., Andrey, G.: Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. 50, 1463–1473 (2018).
110. Kempfer, R., Pombo, A.: Methods for mapping 3D chromosome architecture. Nat. Rev. Genet. 21, 207–226 (2019). https://doi.org/10.1038/s41576-019-0195-2.
111. Hoffman, E.A., Frey, B.L., Smith, L.M., Auble, D.T.: Formaldehyde crosslinking: A tool for the study of chromatin complexes. J. Biol. Chem. 290, 26404–26411 (2015). https://doi.org/10.1074/jbc.R115.651679.
112. Gavrilov, A., Razin, S. V, Cavalli, G.: In vivo formaldehyde cross-linking: It is time for black box analysis. Brief. Funct. Genomics. 14, 163–165 (2015). https://doi.org/10.1093/bfgp/elu037.
113. Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., Tanay, A.: Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature. 547, 61–67 (2017). https://doi.org/10.1038/nature23001.
114. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996).
115. Bystricky, K., Heun, P., Gehlen, L., Langowski, J., Gasser, S.M.: Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proc. Natl. Acad. Sci. U. S. A. 101, 16495–16500 (2004). https://doi.org/10.1073/pnas.0402766101.
116. Tahara, M., Inoue, T., Miyakura, Y., Horie, H., Yasuda, Y., Fujii, H., Kotake, K., Sugano, K.: Cell diameter measurements obtained with a handheld cell counter could be used as a surrogate marker of G2/M arrest and apoptosis in colon cancer cell lines exposed to SN-38. Biochem. Biophys. Res. Commun. 434, 753–759 (2013). https://doi.org/10.1016/j.bbrc.2013.03.128.
117. Yang, F., Yang, X., Jiang, H., Bulkhaults, P., Wood, P., Hrushesky, W., Wang, G.: Dielectrophoretic separation of colorectal cancer cells. Biomicrofluidics. 4, (2010). https://doi.org/10.1063/1.3279786.
118. Yang, T., Zhang, F., Yardımci, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., Li, Q.:

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 27, 1939–1949 (2017).
119. Lando, D., Stevens, T.J., Basu, S., Laue, E.D.: Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. Nucleus. 9, 190–201 (2018). https://doi.org/10.1080/19491034.2018.1438799.