THE UNIVERSITY
of EDINBURGH

Molecular Dynamics Simulations on the Principles Governing Chromosome Organisation

Chi Hang Michael Chiang

Doctor of Philosophy
The University of Edinburgh
August 2021

Abstract

In eukaryotes, chromosomes reside in the crowded environment of the cell nucleus. Understanding the principles or mechanisms governing the three-dimensional (3D) folding of chromosomes and the role that this plays in genome function and disease has been a long-standing challenge in molecular biology and biophysics. Despite recent advances in experimental technologies that can probe chromatin structure at unprecedented resolutions, our knowledge of these principles remains incomplete.

In this thesis, I model chromatin as a coarse-grained polymer and perform molecular dynamics simulations to investigate the mechanisms shaping genome organisation in several contexts. I first study the interplay between chromatin folding and the development of epigenetic patterns using a "recolourable" polymer model. Here, each segment possesses a histone mark or colour that can be updated by "writer" proteins, while "reader" proteins mediate 3D folding by bridging segments with similar marks. Coupling the action of readers and writers leads to the spreading of epigenetic information, facilitated by a change in chromatin conformation. By introducing "genomic bookmarks", factors which associate with chromatin and recruit specific readers and writers, the model produces stable yet plastic epigenetic domains that can be maintained faithfully across replication events, and removal of bookmarks destabilises these domains. Remarkably, with bookmarking the model can reproduce the profile of Polycomb associated modifications along a whole chromosome arm in *Drosophila*.

I then study the mechanisms driving chromatin reorganisation in cellular senescence, a pathological condition in which cells permanently exit from the cell cycle. I consider a minimal model to dissect the role of heterochromatin- and lamina-mediated interactions in oncogene-induced senescence. By varying these two ingredients alone, the model recapitulates typical organisations observed in growing and senescent cells. It demonstrates that the difference in the locality of chromatin interactions in these different conditions can be explained by polymeric phase transitions. It also shows that lamina-associated domains are highly stochastic, as observed in experiments. Crucially, the model offers a biophysical mechanism for the metastability of senescent phenotypes and may explain why it is challenging for senescent cells to return to the growing condition.

Finally, I employ the highly predictive heteromorphic polymer (HiP-HoP) model to examine the elusive link between the spatial interactions of gene regulatory elements and transcription by building a compendium of simulated 3D structures of individual genes within the human genome. The model predicts the frequent associations, interaction topologies, and transcription probability for each regulatory element. It shows that the number of associations correlates significantly with transcription, consistent with the picture of transcription hubs or factories. Interestingly, it indicates that loop extrusion activity is related to the transcriptional variability of a gene across a population of cells. Overall, this pan-genomic analysis offers new insight into the connection between chromatin structure and function.

Lay Summary

DNA is a long molecule that encodes the genetic information governing how a living organism develops and functions. In humans, our genetic instructions are stored as multiple DNA molecules that are located within the nucleus of every single cell of our body. For decades, scientists have been trying to understand how DNA is folded within the nucleus, a remarkable task that requires compacting DNA by as much as 10000 times length-wise. Learning the principles or mechanisms driving this folding is important as it is strongly related to the function of DNA, such as accessing its code to produce proteins, the molecules responsible for carrying out most tasks in life. In addition, the three-dimensional (3D) organisation of DNA is known to play a fundamental role in processes such as development, ageing, and disease.

In this thesis, I use computer simulations to model the 3D structure of the chromatin fibre, the natural form in which DNA exists in the nucleus by packaging with proteins. The objective of this work is to identify some of the principles regulating chromatin structure and function. I will first explore how the spatial organisation of chromatin influences its chemical modifications that are crucial for allowing different tissue cells to perform distinct functions despite sharing the same genetic code. I will then examine the mechanisms driving the extensive structural rearrangement of chromatin in a cellular condition associated with ageing. Lastly, I will investigate the connections between chromatin structure and the regulation of protein production through modelling all chromatin in a human cell.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in:

- D. Michieletto, M. Chiang, D. Colì, A. Papantonis, E. Orlandini, P. R. Cook, D. Marenduzzo. Shaping epigenetic memory via genomic bookmarking. *Nucleic Acids Res.* **46**, 83 (2018)
- M. Chiang, D. Michieletto, C. A. Brackley, N. Rattanavirotkul, H. Mohammed, D. Marenduzzo, T. Chandra. Polymer modeling predicts chromosome reorganization in senescence. *Cell Rep.* **28**, 3212 (2019)
- M. Chiang, G. Forte, N. Gilbert, D. Marenduzzo, C. A. Brackley. Predictive polymer models for 3D chromosome organization. *Methods Mol. Biol.* (in press)
- M. Chiang, C. A. Brackley, D. Marenduzzo, N. Gilbert. Predicting genome organisation and function with mechanistic modelling. *Trends Genet.* (sub judice)

Results presented in Section 4.2 were obtained from my master's thesis and are included in this work for completeness. I gratefully acknowledge C. A. Brackley for assisting the development of the simulation model used in Chapter 6.

(*Chi Hang Michael Chiang, August 2021*)

致謝

Acknowledgements

爸爸、媽媽，感謝您們多年來對我無比的支持，感謝您們給我走遍世界的機會。在海外留學十多年，您們不但沒有怪責我甚少回家，還讓我無拘無束地追尋、鑽研自己的愛好。您們對我的愛如太陽光線，穿越萬里，每天從遠方帶來了溫暖的祝福。

兒子
志行
二零二一年八月

I owe an enormous debt of gratitude to my supervisor, Davide, for his invaluable contribution to my research, for his extensive knowledge and experience, for his unwavering support and encouragement, and for always being so positive and friendly. It has been an absolute privilege to work with and to learn from him over the past six years.

I am extremely grateful to Chris Brackley and Davide Michieletto for their insightful and constructive feedback, for their exceptional guidance and patience, and for always being able to make time and offer help whenever I needed it.

I would like to acknowledge the Carnegie Trust for the Universities of Scotland for funding this PhD studentship.

I want to sincerely thank my flatmates, Michael and Jordan, for their support over the past two years, especially during those difficult days in the pandemic.

Special thanks should go to Emil and Paul – for all the wonderful conversations we had in the office and in pubs. I would also like to thank all my friends from Scotland for their hospitality and for sharing many memorable moments together.

Lastly, I must express my greatest appreciation to my high school teachers, Mr. Inglis and Mr. Butcher, for their inspirational teaching, which has invigorated my interest in understanding the fundamental laws governing the natural world.

Contents

List of Figures

List of Tables

When we really think about the complexity of our genomes, it isn't surprising that we can't understand everything yet. The astonishing triumph is that we understand any of it.

Nessa Carey

1 Introduction

In natural sciences, one of the most intensely scrutinised molecules is deoxyribonucleic acid (DNA). This long and twisted molecule is vital for life, as it encodes the genetic instructions for building and regulating all parts of an organism. The basic molecular structure of DNA, discovered by Crick, Watson, Franklin, and Wilkins in 1953, is that of two strands forming a right-handed double helix [1] (Fig. 1.1A). Each strand is assembled from a sequence of units called nucleotides, and the two strands are coupled by complementary pairing, where each pair of nucleotides is known as a base pair (bp). In eukaryotes, the genetic information is stored in each cell as multiple linear DNA molecules; these molecules are condensed into chromosomes and are located in the specialised compartment known as the cell nucleus [2–4] (Figs. 1.1B–C).

Packaging DNA or chromosomes within the nucleus is an extraordinary feat considering the length scales involved. In its unwrapped, thread-like form, DNA has a diameter of $\sim$2 nm but is extremely long: the DNA in a human cell covers a distance of $\sim$2 m in total [1]. This molecule, however, is confined to the nucleus with a typical size of $\sim$10 μm, and it is compacted by as much as $\sim 10^4$ fold. DNA cannot be packed too tightly though; it has to be folded carefully to facilitate

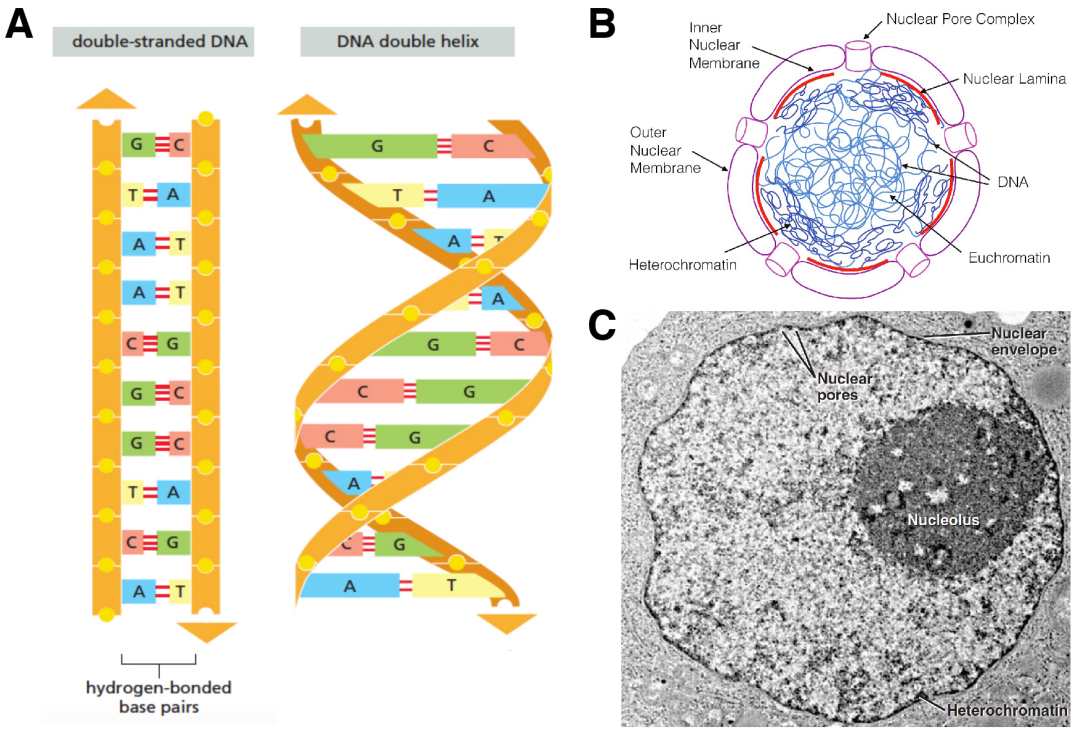

Figure 1.1: DNA and the cell nucleus. (A) Illustrations explaining the structure of double-stranded DNA. *Left*: each strand of this structure is assembled from an array of nucleotides which come in four different types: cytosine (C), guanine (G), adenine (A), and thymine (T). The two strands are complementary to each other, with C pairing with G and A with T, and they are coupled together by hydrogen bonding. Each pair of nucleotides along the double strand is known as a base pair (bp). *Right*: In three dimensions, the two strands coil around each other to form a right-handed double helix due to the hydrophobicity of base pairs. Figure adapted from Ref. [2]. (B) A diagram explaining key components of the cell nucleus [5]. (C) An electron micrograph of a thin cross-section of a cell nucleus, with key features labelled [4].

its function. Cellular machinery needs to access the information along the DNA strands to perform transcription – the copying of a DNA segment (i.e., a gene) into ribonucleic acid (RNA) for producing proteins – and replication – the duplication of this information to be inherited by daughter cells.

A long-standing problem in molecular biology and biophysics is to determine the principles or mechanisms governing the folding of DNA and chromosomes within the nucleus. Addressing this problem is imperative since the three-dimensional (3D) structure of the genome – the complete set of genetic materials of a cell – is closely related to its function [6, 7]. For instance, perturbing the degree of DNA compaction can alter the ability to transcribe or express individual genes, leading to a change in cellular behaviour. More broadly, genome organisation has shown implications in development, cell differentiation, and disease [8]. Basic research

on the structural mechanics of the genome can lead to a better understanding of its function and its role in various cellular processes and can contribute to disease treatments and the development of new technologies such as genome editing. This thesis is, therefore, focussed on deciphering some of the principles regulating genome architecture and its connection to function.

Before taking a deeper look into genome folding and further motivating the work of this thesis, it is worth describing the environment in which the genome resides – the cell nucleus (Figs. 1.1B–C). This organelle is circumscribed by a double-layered membrane, which contains pores to control the transport of materials between the nuclear interior and the rest of the cell. Underneath the membrane is a fibrous layer of proteins called the nuclear lamina, and it provides structural support to the membrane and can interact with the genome. The nuclear interior is a crowded space containing chromosomes, proteins, as well as membraneless foci which perform specific tasks. For instance, a prominent focus is the nucleolus, which is responsible for assembling ribosomes needed for synthesising proteins.

It is generally accepted that DNA is folded within the nucleus in a hierarchical manner (Fig. 1.2A), and its organisation is most understood at the two extreme limits – at the 10-nm and the entire nucleus level. At the 10-nm level, it is known that DNA associates with histone proteins to form nucleosomes, creating a 10-nm thick, "beads-on-a-string" filament called chromatin [2–4] (Fig. 1.2B). Each nucleosome along this fibre contains a histone octamer, and DNA wraps around the octamer in a left-handed manner by $\sim$1.7 turns. Chromatin can be broadly classified into two classes: euchromatin and heterochromatin. The former adopts a more swollen conformation (i.e., less packing into nucleosomes) and is gene-rich, whereas the latter has a more compact structure and is gene-poor. Classically, it is thought that nucleosomes coil together to form a thicker, 30-nm fibre [3]; however, whether this secondary structure exists remains highly debated, as recent *in vivo* electron microscopy experiments have suggested that nucleosomes may be more loosely organised [11].

At the whole nucleus level, it is understood that individual chromosomes occupy their own territories during interphase [12], as demonstrated from "chromosome painting" microscopy experiments using fluorescence *in situ* hybridisation (FISH) [13], a technique to visualise targeted DNA sequences with fluorescence probes (Fig. 1.2C). Studies have shown that these territories are not randomly arranged, with their radial positioning in the nucleus determined by factors such as the gene density [14–16] and size [10, 17] of individual chromosomes.

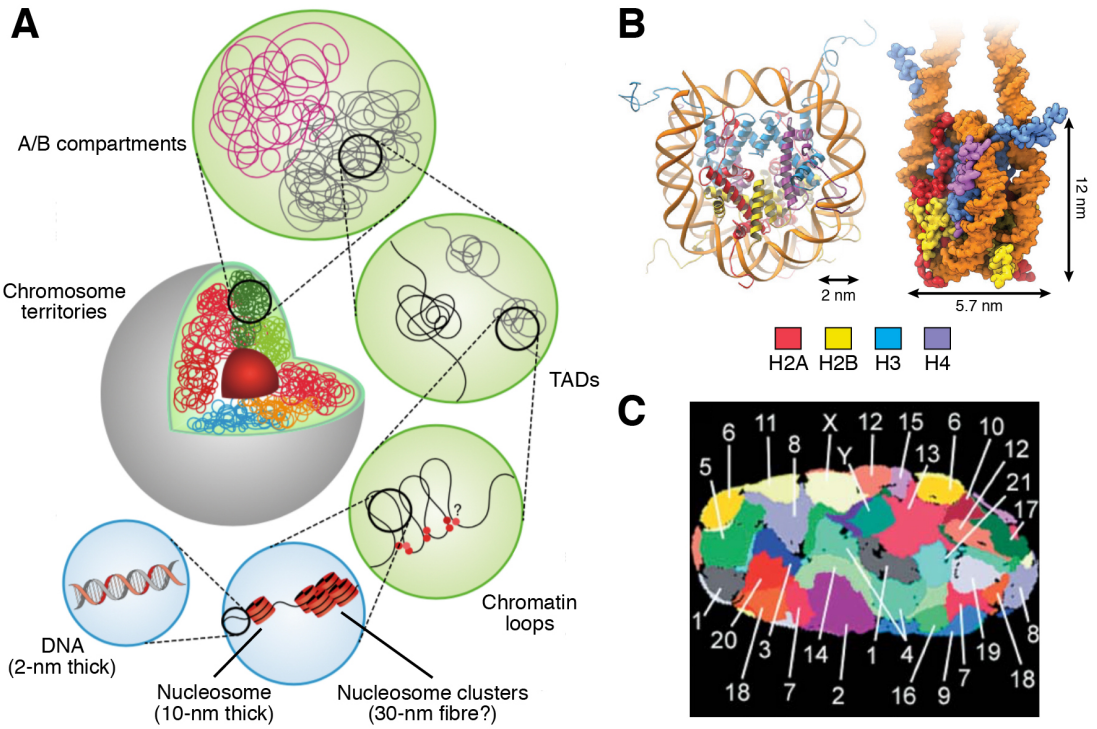

Figure 1.2: The hierarchical folding of a eukaryotic genome. (A) A diagram depicting the architectural features at different levels of compaction of the genome. At the smallest scale, DNA coils around histone proteins to form an array of nucleosomes. Nucleosomes can cluster into a thicker structure, such as a 30-nm fibre, though its existence *in vivo* remains questionable. At a larger scale, the nucleosomal fibre is folded into loops, which provide the structural basis of topologically associating domains (TADs). TADs are then located within a transcriptionally active (A) or inactive (B) compartment. At the scale of the nucleus, individual chromosomes occupy their own territories. Figure adapted from Ref. [9]. (B) Illustrations showing the structure of a nucleosome from a top-down (*left*) and a side view (*right*). A nucleosome contains an octamer which has two copies of four types of histones (H2A, H2B, H3, and H4), and DNA wraps around this octamer in a left-handed manner by ∼1.7 turns (∼146 bp). Figure adapted from Ref. [4]. (C) A false-colour reconstruction of the chromosome territories in a human fibroblast nucleus as identified by FISH microscopy [10].

Recent advances in experimental technologies have helped bridge the gap in the knowledge of genome architecture between the nucleosomal and territorial scales. Notably, methods based on chromosome conformation capture (3C), a molecular technique involving fixation, crosslinking, digestion by restriction enzymes, and ligation, have provided a powerful approach orthogonal to microscopy for probing chromosome organisation [18, 19]. This class of methods quantifies the frequency of spatial interaction, or contact, between pairs of chromatin loci. In particular, Hi-C, a high-throughput variant of 3C with massively parallel sequencing, can provide maps of contact frequency between all loci within the genome at resolutions down to a few kilobase pairs (kbps) [20] (Fig. 1.3).

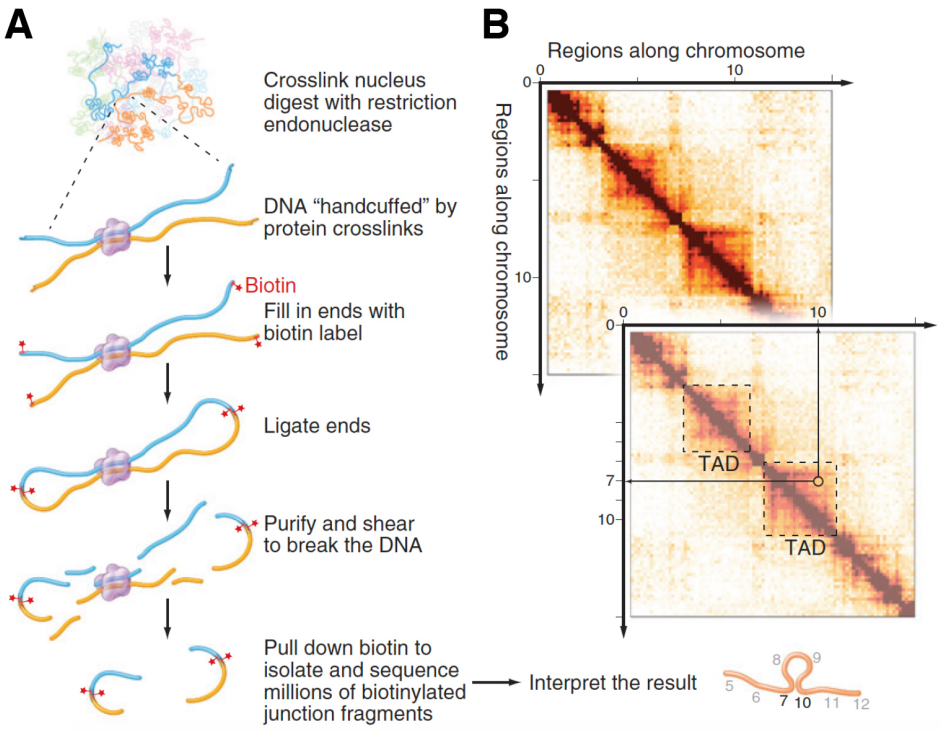

Figure 1.3: Hi-C and resulting contact maps. (A) A flowchart outlining the key procedures in Hi-C. (B) An illustration of contact maps or matrices generated from Hi-C. Each matrix element indicates the frequency of interaction between two genomic loci (a darker colour for a higher contact frequency). TADs can be identified as square domains with enhanced interactions in these maps. Figure adapted from Ref. [4].

3C-based experiments have verified the existence of chromosome territories and have identified other structural features at smaller scales (Fig. 1.2A). Early Hi-C experiments have revealed that within a territory, at the tens of megabase-pair (Mbp) level, the genome is further divided into two major compartments: one associated with transcriptionally active regions (A compartment) and the other with inactive regions (B compartment) [20]. Higher resolution Hi-C data have shown that within compartments there are smaller chromatin regions, spanning hundreds of kbp, called topologically associating domains (TADs), where interactions are more favoured within a domain than between domains [21, 22]. Many TADs are associated with chromatin loops that are mediated by protein complexes binding to specific sequences along DNA [23].

Despite the significant progress made on identifying various architectural features, the principles or mechanisms underlying the formation of these features and their connection to genome function are only starting to become understood. Indeed, it remains challenging for experiments to quantitatively dissect the internal workings of specific genomic components, especially at a single-cell

level. Nevertheless, theoretical modelling has provided valuable contributions to unravelling these structural principles. Modelling enables one to study 3D genome conformation at a spatio-temporal resolution beyond the limits of current experimental techniques and provides a pragmatic approach to exploring relevant biophysical factors systematically. Since chromatin is a long DNA-protein complex, models based on polymer physics have been used extensively. Specifically, coarse-grained polymer models [24] which forgo the details of individual nucleotides and describe chromatin as a chain of beads have proven highly successful in determining the mechanisms shaping large-scale genomic features, such as TADs and compartments. These models have also been useful in reconstructing 3D chromatin structures from 3C data.

Building on the achievements from previous modelling work, this thesis aims to use simulations based on coarse-grained polymer models to shed light on additional principles dictating 3D genome organisation and, more importantly, its relation to function. I will focus on the situation during interphase, the main period of the cell cycle where a cell does not divide, with genes expressed for performing routine tasks; in this way, connections between structure and function are most apparent for investigation.

The work conducted in this thesis is divided into three parts, examining principles related to genome structure and function in different biological contexts. In the first part, I will analyse the connection between chromatin organisation and the epigenetic modifications along this fibre. Epigenetic modifications are heritable modifications of the chromatin fibre which do not alter the underlying genetic code. They are typically associated with the addition of biochemical marks on histones and are crucial for maintaining cell-type-specific expression patterns. The principles of how these marks are regulated, in particular the role played by 3D chromatin folding, are still elusive. The aim here is to use polymer modelling to dissect possible mechanisms at work in this context.

In the second part, I will study the large-scale spatial segregation between euchromatin and heterochromatin within the nucleus. It is known that heterochromatin tends to localise near the periphery, whereas euchromatin resides more towards the interior (Figs. 1.1B–C). However, the biophysical principles orchestrating this organisation are not well understood. To identify these principles, I will examine a pathological condition called cellular senescence where this segregation is severely disrupted. Senescence is a condition in which a cell exits from the cell cycle and remains in a non-proliferative state, and it can be induced by oncogene

activation. In such a scenario, it is observed that heterochromatin typically migrates from the nuclear periphery to the interior and forms large structural foci. The goal is to understand how this reorganisation takes place, which may help uncover the mechanisms governing the conventional positioning of euchromatin and heterochromatin.

In the final part, I will tackle a fundamental problem in genome research – unravelling the mechanistic links between chromatin folding and gene transcription. Classically, it is accepted that gene expression involves a gene interacting with its regulatory DNA sequences, typically by forming 3D loops mediated by DNA-binding proteins. However, it remains unclear how these loops mechanistically facilitate transcription, as well as whether there are other independent principles regulating gene activity. To address these issues, I will conduct a genome-wide analysis by simulating all chromosomes in the human genome to elucidate generic mechanisms linking between structure and transcription.

The remaining chapters of this thesis are organised as follows:

In Chapter 2, I will introduce theoretical principles that are relevant for modelling genome organisation. Specifically, I will discuss some basic concepts in polymer physics and review several coarse-grained polymer models for chromatin folding that are utilised in this thesis.

In Chapter 3, I will explain the molecular dynamics simulation framework employed in this thesis for modelling chromatin. I will describe the potentials and the numerical integration scheme used, as well as the mappings between simulation and physical units.

In Chapter 4, I will discuss the work investigating how chromatin folding contributes to regulating epigenetic modifications. I will dissect the relevant principles using a bead-and-spring polymer model. Beads are "coloured" to represent different epigenetic marks along chromatin, and their colour can change over time to model the action of "writer" proteins modifying the marks on chromatin. This colouring scheme is coupled to 3D chromatin folding by introducing homotypic interactions between beads, modelling bridging mediated by multivalent "reader" proteins. This coupling generates a positive feedback that allows effective spreading of epigenetic information. Crucially, I will show that "genomic bookmarking" – the action of sequence-specific chromatin-binding proteins recruiting particular readers and writers – provides a robust mechanism for establishing and maintaining heterogeneous epigenetic domains, an attribute

important for creating different expression programmes. As a proof of concept, I will employ the model with this bookmarking mechanism to simulate the epigenetic patterns of a whole chromosome arm in *Drosophila*.

In Chapter 5, I will focus on the work examining the large-scale reorganisation of the genome in cellular senescence. In particular, I will study the role played by heterochromatin- and lamina-mediated interactions in this process. To this end, I will consider a model system where a single human chromosome, described as a coarse-grained polymer with heterochromatic and euchromatic regions, interacts with a lamina wall. I will show that varying the two interactions alone drives the system into different polymeric phases, whose conformations resemble those in healthy, growing cells and in senescent cells. I will demonstrate that, by associating the growing and senescent conditions with different phases, the model can explain the change in the locality of chromatin interactions between these conditions. I will also investigate the mechanisms driving the stochasticity of chromatin domains interacting with the lamina. Most importantly, I will study the transition behaviour between different phases, through which I will offer a biophysical explanation for the conformational stability associated with the senescent condition.

In Chapter 6, I will present the work on a pan-genomic analysis of the mechanisms linking between chromatin structure and transcriptional activity. I will first discuss the simulations conducted to generate a set of 3D structures of individual chromosomes of the human genome – a key prerequisite for this analysis. These simulations are done using a polymer model which combines multiple well-established mechanisms for chromatin folding. The analysis itself will focus on regulatory elements, or DNA sequences which modulate gene expression, and will be divided into three parts. First, I will study the structural properties of individual elements, examining in detail their spatial interactions with each other. Second, I will quantify the transcriptional output of each element; in particular, I will show that this output can be predicted from simulations, in addition to inferring from transcriptomic experiments. Finally, I will draw connections between the structural and transcriptional properties of these elements by correlating the observables measured for these two aspects.

In Chapter 7, I will draw some conclusions on the different principles learned from the various work conducted in this thesis and discuss how these principles may contribute to future research.

2

Theoretical Principles in Modelling Chromatin

Research on the spatial organisation of chromatin has been a joint venture between experimentalists and theorists. Theoretical work provides elegant models to explain key biophysical principles driving chromatin folding and helps interpret the increasingly complex results from experiments. In this chapter, I introduce theoretical concepts that are essential for modelling the three-dimensional (3D) conformation of chromatin. Motivated by the fact that chromatin is a long molecule, the first part of the chapter is devoted to explaining basic theories in polymer physics relevant to this thesis. The second part then discusses recently developed polymer models for studying chromatin architecture.

2.1 Basic Principles in Polymer Physics

The study of the conformations of DNA and chromatin naturally falls within the realm of polymer physics, which is concerned with understanding the macroscopic

properties of polymers from a statistical mechanics approach [25, 26]. In essence, a polymer is a long molecule consists of repeating elementary units called monomers. These monomers can refer to the actual molecules that are bonded together to create the polymer chain, or, as explained further below, they can represent longer segments of a polymer as a coarser level of description.

A fundamental principle of polymer physics is that, because of the extreme length of a polymer (i.e., the number of monomers $N \gg 1$), the microscopic (chemical) details of individual monomers play little role in governing the large-scale behaviour of the entire polymer. This concept is manifested in the existence of scaling laws [27] that apply to a wide range of polymers, from artificial ones such as polyethylene and polystyrene, which are found in common plastics, to biological ones like DNA and proteins. In the following, I will present a few simple polymer models in order to highlight some scaling relations that are useful for characterising chromatin structure and for understanding the work conducted in this thesis.

2.1.1 Ideal Chains

Similar to the ideal gas which offers a basic model to examine the characteristics of real gases, ideal chains provide a starting point to understand the properties of polymeric systems. This class of chain models assumes that there is no volume interactions between monomers (i.e., monomers can overlap with each other). One of the simplest models in this class is the freely jointed chain (FJC; Fig. 2.1A), where segments connecting two neighbouring monomers have a fixed length σ – also known as the Kuhn length – and the orientation of each segment is random and independent of other segments. By defining a bond vector $\boldsymbol{u}_i = \boldsymbol{r}_{i+1} - \boldsymbol{r}_i$ to represent the segment linking monomers i and $i+1$ at positions $\boldsymbol{r}_i$ and $\boldsymbol{r}_{i+1}$, respectively, the lack of correlation in the bond direction can be expressed as

$$\langle \boldsymbol{u}_i \cdot \boldsymbol{u}_j \rangle = \sigma^2 \delta_{ij}, \tag{2.1}$$

where the brackets $\langle \cdots \rangle$ denote an ensemble average over all possible conformations of the chain.

One approach to describe the chain's conformation is to determine the volume it typically occupies in space, which can be inferred from its mean squared end-to-end distance R^2. This observable is pertinent to examining DNA or chromatin

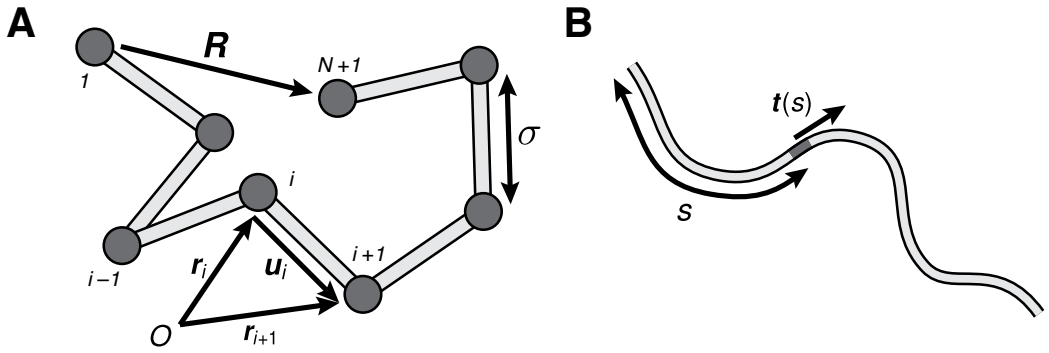

Figure 2.1: Ideal chain models. (A) The freely jointed chain (FJC). This model describes a polymer as a chain of monomers connected by rigid, randomly oriented segments of length σ (Kuhn length). Each monomer, say the ith one, has a position $\boldsymbol{r}_i$, and the segment joining it to the next monomer is given by the bond vector $\boldsymbol{u}_i$. The end-to-end vector of the chain is described by $\boldsymbol{R}$. (B) The worm-like chain (WLC). This model considers a polymer as a smooth, continuous curve whose local degree of bending is small. Each infinitesimal segment is described by its contour distance s from the fibre's initial point, and its direction is denoted by the tangent vector $\boldsymbol{t}(s)$.

structure, as it can be directly measured from experiments, such as microscopy with fluorescence *in situ* hybridisation (FISH; see Chapter 1). To compute R^2 for an FJC of N segments (or $N+1$ monomers), one considers its end-to-end vector

$$\boldsymbol{R} = \boldsymbol{r}_{N+1} - \boldsymbol{r}_1 = \sum_{i=1}^{N} \boldsymbol{u}_i, \tag{2.2}$$

and thus

$$R^2 = \left\langle \boldsymbol{R}^2 \right\rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \boldsymbol{u}_i \cdot \boldsymbol{u}_j \rangle = \sigma^2 N. \tag{2.3}$$

Discarding prefactors, this gives

$$R \sim N^{\nu}, \tag{2.4}$$

with $\nu = 1/2$. This scaling between the end-to-end distance and the number of segments is universal and applies to other ideal chain models as well as real polymers. The exponent ν, known as the Flory exponent, is pivotal to characterising the conformational nature of a polymer. For example, all ideal chains have $\nu = 1/2$, while real polymers can take on other values depending on the interactions between monomers (see Section 2.1.2). Note that the scaling also holds for a subsection of a polymer (due to its fractal nature), as long as that subsection is still sufficiently long (i.e., the number of segments is $\gg 1$).

Another useful way to quantify the 3D size of a polymer is its mean radius of gyration R_g, which, for a chain of N monomers, is defined as

$$R_g^2 = \left\langle \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\boldsymbol{r}_i - \boldsymbol{r}_j)^2 \right\rangle = \left\langle \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{r}_i - \boldsymbol{r}_{\mathrm{cm}})^2 \right\rangle, \tag{2.5}$$

where $\boldsymbol{r}_{\mathrm{cm}}$ is the centre of mass of the chain. It can be shown that for ideal chains $R_g \approx R/\sqrt{6}$ [26], and so R_g exhibits the same scaling with N – i.e., $R_g \sim N^{\nu}$.

An FJC is fully flexible, as its segments can turn freely; yet not all polymers behave in this manner. Many polymers, including DNA and chromatin, are stiff and resist to bending. The worm-like chain (WLC), first proposed by Kratky and Porod, captures this concept of flexibility, and it considers a polymer as a smooth, continuous curve whose local degree of bending is small (Fig. 2.1B). The key property here is the persistence length ℓ_p, which is the characteristic length along the contour over which a polymer loses memory of its orientation. More precisely, letting $\boldsymbol{t}(s)$ be the unit vector tangent to the fibre at contour length s from the fibre's initial point, ℓ_p is defined via the tangent-tangent correlation

$$\langle \boldsymbol{t}(s) \cdot \boldsymbol{t}(s') \rangle = \exp\left(-\frac{|s - s'|}{\ell_p} \right). \tag{2.6}$$

As an example, the persistence length of DNA is around 50 nm (or 150 bp) [28].

One can calculate the typical size of a WLC using the same formalism introduced above for an FJC by taking the continuum limit – i.e., setting $N \to \infty$ and $\sigma \to 0$ while keeping the contour length $L = N\sigma$ finite. In this way, the end-to-end vector becomes

$$\boldsymbol{R} = \int_0^L \mathrm{d}s\, \boldsymbol{t}(s), \tag{2.7}$$

and the mean squared end-to-end distance is given by

$$\begin{aligned} \left\langle \boldsymbol{R}^2 \right\rangle &= \int_0^L \mathrm{d}s \int_0^L \mathrm{d}s'\, \langle \boldsymbol{t}(s) \cdot \boldsymbol{t}(s') \rangle = 2 \int_0^L \mathrm{d}s \int_0^s \mathrm{d}s'\, \exp\left(-\frac{s - s'}{\ell_p} \right) \\ &= 2\ell_p^2 \left[\left(\frac{L}{\ell_p} \right) + \exp\left(-\frac{L}{\ell_p} \right) - 1 \right]. \end{aligned} \tag{2.8}$$

In the limit $L \ll \ell_p$, $R^2 \approx L^2$, and the fibre behaves like a straight rod. When $L \gg \ell_p$, $R^2 \approx 2\ell_p L = 2\ell_p N\sigma$, and one recovers the scaling shown previously [see Eqs. (2.3) and (2.4)]. Importantly, this result indicates that a WLC can be

mapped to an effective FJC, where each rigid segment now has a Kuhn length that is approximately twice the persistence length ($\sigma \approx 2\ell_p$). Indeed, this mapping can be carried out for any ideal chains, with the Kuhn length indicating the scale over which a chain can be viewed as an FJC.

Apart from considering the ensemble average of $\boldsymbol{R}^2$, one can also examine the underlying distribution $\mathcal{P}(\boldsymbol{R}; N)$, which gives the probability density of observing an end-to-end vector $\boldsymbol{R}$ for a chain of N segments. For an FJC (and for any ideal chain mapped to an effective one), its contour is equivalent to the path of a random walk, and $\boldsymbol{R}$ is a sum of random variables [Eq. (2.2)]. Hence, when the number of segments is large ($N \gg 1$), $\mathcal{P}(\boldsymbol{R}; N)$ approaches a Gaussian by the central limit theorem [25], i.e.,

$$\mathcal{P}(\boldsymbol{R}; N) = \left(\frac{2\pi N\sigma^2}{d}\right)^{-d/2} \exp\left(-\frac{d\boldsymbol{R}^2}{2N\sigma^2}\right), \tag{2.9}$$

where d is the dimensionality of space. This equation also holds for a subsection of an ideal chain, since the subsection also exhibits ideal behaviours (provided that it is long enough). A useful quantity that can be inferred from this distribution is the contact or looping probability $P_c(s)$, which is the likelihood of finding two segments separated by contour length s to be within a small spatial distance of r_c; in other words,

$$P_c(s) = \int \mathrm{d}\boldsymbol{R}\, \Theta(r_c - |\boldsymbol{R}|)\mathcal{P}(\boldsymbol{R}; s/\sigma), \tag{2.10}$$

where $\Theta(x)$ is the Heaviside step function (i.e., $\Theta(x) = 1$ if $x > 0$ and zero otherwise). This probability is relevant to chromosome conformation capture (3C) experiments (see Chapter 1), whose results provide information on the frequency of contact between pairs of genomic loci. Substituting Eq. (2.9) into Eq. (2.10) and assuming that $r_c \ll (s\sigma)^{1/2}$, one arrives at

$$P_c(s) \approx \left(\frac{s\sigma}{d}\right)^{-d/2} \int_0^{r_c} \mathrm{d}R\, R^{d-1} \exp\left(-\frac{dR^2}{2s\sigma}\right) \approx \frac{r_c^d}{d}\left(\frac{s\sigma}{d}\right)^{-d/2}, \tag{2.11}$$

where the exponential has been approximated as unity when evaluating the integral. Hence, this contact probability exhibits the scaling

$$P_c(s) \sim s^{-\alpha}, \tag{2.12}$$

where α is known as the contact exponent and $\alpha = d/2$ for ideal chains. Similar to ν, this exponent also encapsulates the conformational behaviour of a polymer, and it has been measured from Hi-C data to understand the folding nature of chromatin [20, 29].

2.1.2 Real Chains

Different from ideal chains, real polymers have a finite size, and their monomers cannot overlap. A polymer whose monomers interact by steric repulsion is known as a self-avoiding walk (SAW). The space occupied by a SAW is generally greater than that by an ideal chain due to the excluded volume of individual monomers, and this is manifested in a change in the exponent ν. Developed by Flory, one method to approximate ν in this case involves minimising the Helmholtz free energy $F(R)$ of a chain (of N monomers) whose two ends are separated by distance R. Two contributions to F here are an elastic contribution $F_{\text{el}}(R)$ and another due to volume interactions between monomers $F_{\text{int}}(R)$.

The elasticity of a chain (in an athermal condition) originates from its conformational entropy $S(R) = k_B \ln \Omega(R)$, where k_B is the Boltzmann constant, and $\Omega(R)$ is the total number of configurations of the chain with an end-to-end distance R and is proportional to the probability $\mathcal{P}(\boldsymbol{R}; N)$. Using Eq. (2.9), one finds

$$\frac{F_{\text{el}}(R)}{k_B T} = -\ln \Omega(R) \approx \frac{dR^2}{N\sigma^2}, \tag{2.13}$$

where T is the temperature of the system, and constant terms which do not involve R are ignored for simplicity. This expression is analogous to that for a Hookean spring, with a spring constant proportional to $dk_BT/(N\sigma^2)$. Qualitatively, stretching a chain (increasing R) reduces its conformational entropy, as there are fewer possible configurations. This gives rise to an elastic force which restores the chain to a conformational state with a higher entropy by reducing R.

The free energy due to volume interactions can be estimated as follows [26]. Assuming that the monomers are evenly distributed within the volume R^d occupied by the chain, the number density of monomers is N/R^d. Then, the chance of finding another monomer within the excluded volume v of a monomer is around vN/R^d. The energetic cost is roughly k_BT per an actual exclusion event, and so the average cost per exclusion is k_BTvN/R^d. Since there are N

monomers in total, the overall cost is N times higher, giving

$$\frac{F_{\text{int}}}{k_B T} \approx \frac{vN^2}{R^d}. \tag{2.14}$$

Adding these two contributions together and minimising the result with respect to R leads to the scaling

$$R \sim N^{\nu} \quad \text{with} \quad \nu = \frac{3}{d+2}. \tag{2.15}$$

This expression for ν is surprisingly accurate for a SAW despite the crude approximations made above. It agrees with actual results for $d = 1$ ($\nu = 1$) and $d = 2$ ($\nu = 3/4$), and is close to the current best estimate for $d = 3$ (i.e., $\nu = 3/5$ compared to 0.588 from more advanced theories [30]). When $d = 4$, it indicates that a SAW becomes an ideal chain with $\nu = 1/2$. This is because the increase in dimensions reduces the chance of chain segments "seeing" each other, and correlations between segments become less important (which is what is assumed in the ideal limit). When $d > 4$, the formula gives $\nu < 1/2$, but this is unrealistic as the excluded volume effect cannot make a polymer smaller than that of an ideal chain (whose monomers are already point-like). Therefore, $\nu = 1/2$ for $d > 4$.

Apart from the short-range repulsion due to excluded volume, monomers can also experience long-range attraction with each other, for example due to intrinsic chemical affinities between monomers or to solvent conditions. When this attraction is strong, a polymer can collapse into a globular structure. Performing an analysis similar to the one above shows that $\nu = 1/d$ in this scenario [25, 26]. An intuitive way to obtain this exponent is to observe that when a polymer is compact, its density of monomers $N/V \approx N/R^d$ is roughly constant, and thus $R \sim N^{1/d}$. Interestingly, when the strengths of the repulsive and attractive interactions are similar (also known as the Θ condition), effects from both kinds of interactions can cancel out, and the polymer behaves like an ideal chain with $\nu = 1/2$. The transition from a swollen, self-avoiding conformation to a compact, globular one is known as the coil-to-globule transition, and comprehensive theories have been developed for characterising this transition [25, 26].

2.2 Polymer Models of Chromatin

Given that chromatin is a long fibre, a standard theoretical approach to study its architecture is to construct a suitable polymer model for this system. As demonstrated from the previous section, it is not necessary for a model to incorporate all microscopic details of a polymer in order to capture its overall behaviour (recall, e.g., the mapping of different ideal chain models to the FJC), and this concept also applies to modelling chromatin. More generally, the procedure of removing details that are irrelevant to the length scale of interest is known as coarse-graining (CG). A common CG scheme in modelling chromatin is to describe it as a bead-and-spring polymer, where each bead represents a number of nucleosomes (Fig. 2.2A). As a result, the exact structure of chromatin within a bead is not resolved, and the amount of chromatin covered by a single bead can depend on, for example, the size of the system to be modelled. This scheme is particularly successful at capturing the large-scale folding of chromatin (i.e., at resolutions above hundreds of bp), and two modelling approaches which make use of this scheme are inverse (or data-driven or "top-down") and mechanistic (or first-principles or "bottom-up") modelling [24]. Below, I will briefly summarise the inverse approach before discussing more in detail the mechanistic approach, which is employed in this thesis.

2.2.1 Inverse Modelling

The inverse method utilises 3C-based data (e.g., Hi-C) as an input to reconstruct the spatial structure of chromatin (Fig. 2.2B). In some early models, constraints for the separation between chromatin loci were inferred from a 3C interaction matrix (or contact map) and a single "averaged" conformation which best satisfies the constraints was computed [32]. Recognising the ensemble nature of conventional Hi-C data, other work considered sampling multiple conformations based on the same set of constraints but starting from different initial conditions [33, 34]. There were also studies that attempted to generate and iteratively improve a family of structures that, when measured collectively, reproduce the experimental data [35, 36]. More recently, some inverse models have been first trained against some existing data in a well-studied condition, and then employed to estimate the 3D conformations in an alternative condition where there may be less information available, such as when the genome undergoes mutations or rearrangements [37].

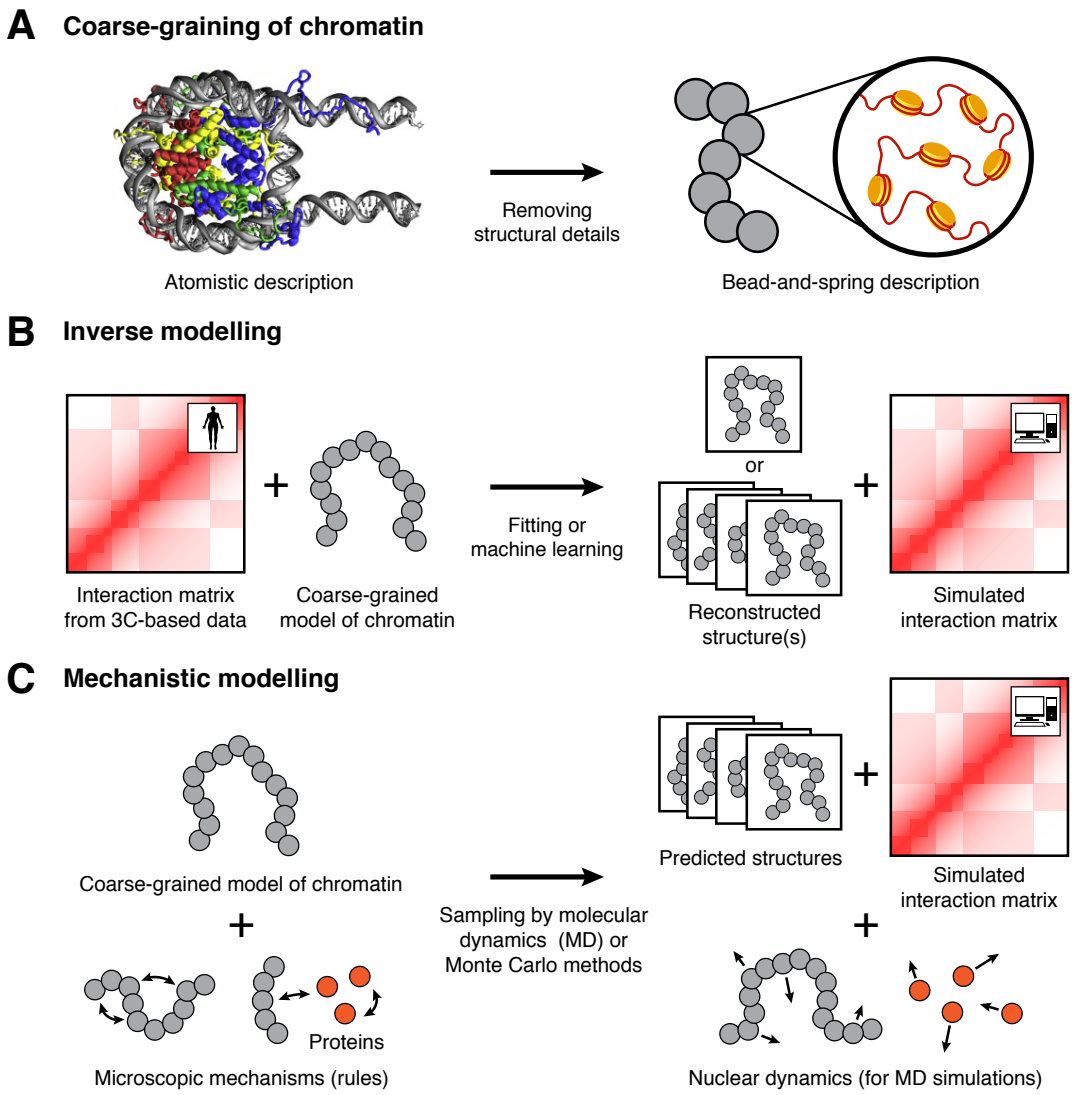

Figure 2.2: Coarse-grained modelling of chromatin. (A) The process of constructing a coarse-grained model for chromatin involves removing structural details, such as the positions of individual atoms in a nucleosome (*left*; figure adapted from Ref. [31]), that are irrelevant to the length scale of interest. A common scheme is to represent the chromatin fibre as a bead-and-spring polymer (*right*), where each bead contains a number of nucleosomes. (B and C) Two approaches which use coarse-grained models to understand 3D genome folding are inverse and mechanistic modelling. (B) In the inverse approach, 3C-based data are used as input to fit or train model parameters (e.g., interaction strength between chromatin loci). The output of the model is a population-averaged structure or an ensemble of structures, alongside a simulated interaction matrix. (C) In the mechanistic approach, the input is a set of microscopic mechanisms or "rules" postulated to be important in regulating genome organisation. Molecular dynamics (MD) simulations or Monte Carlo algorithms implementing these rules are then used to sample possible conformations. As before, the output is a set of 3D structures and a simulated interaction matrix. Additionally, models implemented using MD simulations can also give information on nuclear dynamics.

2.2.2 Mechanistic Modelling

The mechanistic method focusses on understanding the physical and molecular principles that give rise to prominent features of genome architecture [38], an ethos which aligns with the overarching objective of this thesis. Fundamentally, this approach starts from a minimal representation of chromatin – i.e., a bead-and-spring polymer – and explores possible microscopic mechanisms that can explain the folding features. In practice, the model input is a set of "rules" based on known biophysical processes or hypotheses inspired from empirical observations (Fig. 2.2C). For example, these could describe how chromatin loci interact with one another and with proteins and other nuclear landmarks. Experimental data such as those on histone modifications (i.e., biochemical marks on histone tails) may be used, but unlike the inverse approach, no fitting or training (e.g., to 3C-based data) is required. The result is then a set of predicted 3D structures and a simulated interaction matrix, and modelling performed using molecular dynamics (see Chapter 3) also gives information on chromatin and protein dynamics.

2.3 Examples of Mechanistic Models

In the mechanistic approach, the simple bead-and-spring model, without additional ingredients, can already provide fruitful insight into genome organisation. For instance, simulations of chromosome decompaction after mitosis reveal that the formation of territories may be attributed to the slow relaxation dynamics arising from polymeric topological constraints [39]. Entropic effects due to the flexibility of the chromatin fibre can partly account for the difference in radial positioning between chromosomes [40]. In addition, simulations have suggested that the 3D conformation of chromatin is to some extent consistent with a fractal globule structure, which is self-similar and knot-free [29]. Nevertheless, this generic polymer model can only capture limited aspects of nuclear arrangement. In the following, I discuss several mechanistic models which extend beyond the basic framework and have successfully reproduced other phenomena discovered in recent Hi-C and microscopy experiments.

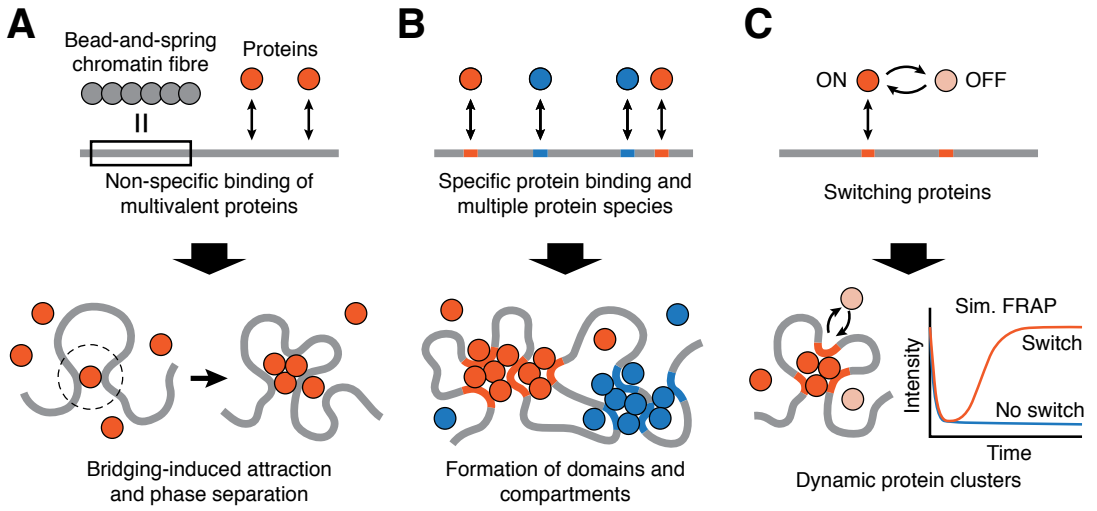

Figure 2.3: The transcription factor model. (A) The simplest form of this model considers a single species of multivalent proteins (or binders or transcription factors) which bind non-specifically to anywhere along the chromatin fibre (modelled as a bead-and-spring chain). An emergent phenomenon is the "bridging-induced attraction". Here, chromatin bridging mediated by binders increases the local chromatin density, leading to more binders being attracted to the region (dashed circle) to facilitate more bridging, which drives further increase in density and so on, creating a positive feedback loop which results in phase separation. (B) By introducing specific binding sites along the fibre and multiple species of binders, the model yields chromatin domains and compartments. (C) With binders switching stochastically between a binding and a non-binding state, their clusters become dynamic, and the model reproduces trends in FRAP experiments for nuclear bodies.

2.3.1 Transcription Factor Model

One simple extension to the bead-and-spring model is the inclusion of diffusive chromatin-binding molecules, representing protein complexes which associate with chromatin – i.e., transcription factors (TFs). This class of models has been known as the strings and binders switch (SBS) model [41] or the TF model [42] (Fig. 2.3). The binding proteins (also referred to as binders or TFs) are typically described as spherical beads for simplicity, and a rudimentary form of the model allows these binders to bind non-specifically to any sites along the chromatin fibre. Crucially, the binders are multivalent[1], so they can form molecular bridges between two or more sites. The model successfully recapitulates Hi-C results on the decay in the contact probability between two chromatin segments as a function of their genomic separation [41] [see also Eq. (2.12)].

[1]Multivalency is a natural assumption, as each binder typically represents a complex of DNA-binding proteins. Hence, even if each of the single proteins is monovalent, the complex is still multivalent.

An interesting phenomenon which emerges from this model is that when the interaction with chromatin is strong enough, the binders themselves, which are not directly attracted to one another, tend to cluster and form large aggregates. This effect is known as "bridging-induced attraction" (BIA) [42], and it briefly works as follows (Fig. 2.3A). First, diffusive multivalent binders associate with several chromatin loci, leading to a local increase in chromatin density. This effect, in turn, recruits more binders to that region, and they facilitate more bridging between chromatin segments, resulting in a further increase in chromatin density. This chain of events initiates a positive feedback loop which ultimately leads to phase separation [43], whereby a local high concentration of proteins is developed. Recently, microscopy experiments examining DNA-protein interactions have provided evidence of BIA in action, both *in vitro* and *in vivo* [44]. BIA also offers an appealing mechanism to explain the biogenesis of membraneless organelles seen within the nucleus, or nuclear bodies [45].

When there is only a single kind of non-specific binders, their clusters arising through BIA coarsen indefinitely, and eventually only one remains in steady state. This macroscopic phase separation behaviour is inconsistent with experimental data, which show that nuclear bodies have a well-defined size and generally do not grow beyond this limit [45]. A refinement to the model which reconciles this discrepancy is to incorporate specific (e.g., sequence-dependent) protein binding sites along the chromatin fibre [46] (Fig. 2.3B). In this way, binders can only bridge between these specific sites, thereby generating chromatin loops. The thermodynamic (entropic) cost associated with looping arrests coarsening and results in the formation of smaller, size-limiting clusters, qualitatively similar to nuclear bodies [47].

Protein clustering is accompanied by the formation of chromatin domains, or regions of chromatin enriched in intra-chromosomal (or *cis*-) interactions. As with clusters, typically there are multiple chromatin domains, and these are enriched in different histone modifications. This motivates the introduction of multiple species of binders to the TF model [46, 48, 49], each binding to different types of chromatin (i.e., chromatin with different histone modifications; Fig. 2.3B). With this additional feature, multiple chromatin domains form spontaneously; in particular, domains mediated by different types of binders strongly segregate in space, a phenomenon akin to compartmentalisation [43]. Remarkably, by including only two kinds of binders – one binding to transcriptionally active sites and the other to inactive sites – and by using experimental data (e.g., chromatin

immunoprecipitation with sequencing [ChIP-seq][2] for histone modifications) to determine precise locations of the binding sites, this model can accurately predict up to 85% of the domain boundaries found in Hi-C data [46, 48].

It is also possible to use the TF model to study the dynamical properties of protein clusters or nuclear bodies. Experiments such as fluorescence recovery after photobleaching (FRAP)[3] have shown that these protein foci are macroscopically stable but highly dynamic – there is a constant turnover of constituents within the foci with those from the soluble pool [47]. This feature is not captured by the basic TF model, where clusters are relatively static once established. An improvement made to the model is to allow binders to "switch" stochastically between a binding and a non-binding state [47] (Fig. 2.3C). Switching mimics active (i.e., ATP-dependent) processes such as post-translational modifications or protein degradation, driving the model out-of-equilibrium – a ubiquitous characteristic of biological systems. With switching, protein clusters can still nucleate via BIA, while their compositions become more dynamical.

2.3.2 Loop Extrusion Model

Another cutting-edge development in mechanistic modelling is the loop extrusion (LE) model [50, 51]. Extending the basic polymer framework, the model posits that there are protein complexes which can attach to and translocate along chromatin or DNA to create loops. Loop extrusion was proposed long ago as a mechanism for chromosome compaction during mitosis [52, 53], but recently this concept has been revived thanks to higher resolution data on chromatin architecture available from experiments. In the context of genome organisation during interphase, Hi-C results [23] revealed that many topologically associating domains (TADs) are supported by chromatin loops, and domain boundaries are often colocalised with subunits of structural maintenance of chromosome (SMC) complexes[4] as well as the CCCTC-binding factor (CTCF)[5], whose binding motif

[2]ChIP-seq is a technique which uses antibody binding and next generation sequencing to measure genome-wide the enrichment of a specific protein or histone modification along the chromatin fibre.

[3]FRAP is a method for interrogating the dynamics of the internal constituents of a cellular or nuclear focus by inactivating fluorophores within the focus (i.e., bleaching) and analysing the rate of recovery of fluorescence signal afterwards.

[4]SMC complexes are a class of architectural proteins which mediate the 3D compaction of chromosomes. Two well-known SMC complexes are cohesin and condensin.

[5]CTCF is a transcription factor with zinc finger domains which is important for chromatin insulation and for regulating transcription and genome structure.

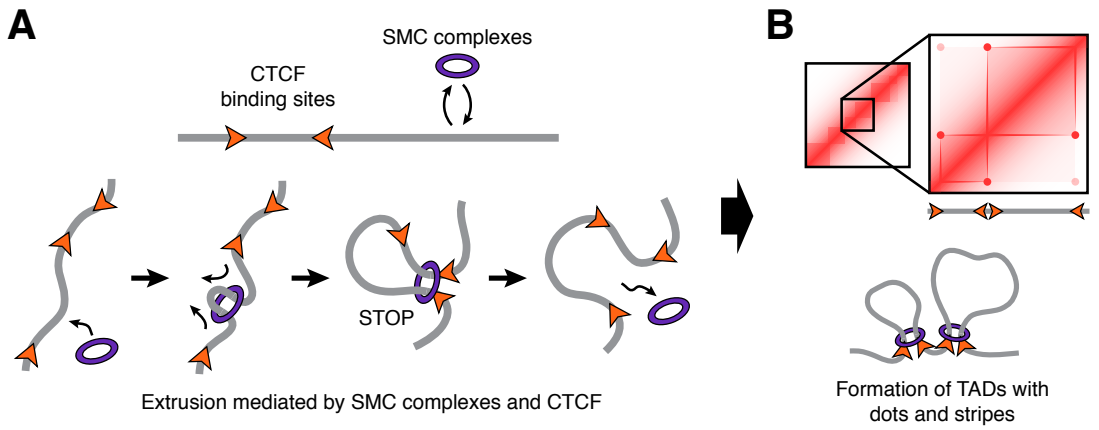

Figure 2.4: The loop extrusion model. (A) This model postulates that there are active factors (or extruders), such as structural maintenance of chromosome (SMC) complexes, which bind to the DNA or chromatin fibre and translocate outwards to form loops. These extruders are thought to move along the fibre until reaching a CTCF site that is oriented opposite to its direction of travel or colliding with another extruder (not depicted here). (B) The model produces chromatin domains similar to TADs seen in contact maps from Hi-C, with dots and stripes near domain boundaries as a result of extruded loops bringing chromatin loci together.

on DNA is orientation-specific. Surprisingly, most loop anchors are associated with convergent CTCF binding sites, and this bias is incompatible with a model where simple diffusive binders mediate the loops (i.e., the TF model).

This finding has prompted the development of a specific LE model to explain chromatin looping in interphase [54, 55] (Fig. 2.4A). In this model, SMC complexes, such as cohesin and condensin, are postulated to act as extruding factors with motor activity. Once bound to the DNA or chromatin fibre at a single locus (i.e., two adjacent beads), subunits of these complexes move independently in opposite directions along the fibre, thereby generating loops. Extrusion activity is halted when the complex collides with another one, or when it reaches a CTCF protein whose binding site is oriented against its direction of travel (but it can move past a CTCF site that is parallel to its movement). These rules naturally give rise to the preference for establishing convergent CTCF loops. Additionally, they allow the model to recapitulate architectural features, such as “dots” and “stripes”, that coincide with TAD boundaries as seen in Hi-C contact maps (Fig. 2.4B).

Since the proposal of this model, there has been growing evidence for extrusion mediated by SMC complexes and for their role in establishing topological domains with CTCF. For example, genome editing experiments have shown that

manipulating CTCF binding sites can alter chromatin domain patterns [54, 56]. Domains are also affected by degradation of CTCF [57], cohesin [58, 59], and complexes involved in cohesin loading and unloading [60, 61]. More recently, a plethora of single-molecule experiments have provided direct evidence of the extrusion activity of cohesin and condensin on DNA *in vitro* [62–66]. Nevertheless, the exact mechanism of how these factors perform extrusion is still much debated [50, 51].

2.3.3 Block Copolymer Model

Another class of mechanistic model is the copolymer or block copolymer model. Here, beads along the polymer are assigned to one of several different species according to, for example, the local histone modifications of the chromatin fibre. Attractive interactions are added between beads of the same species, and they either represent direct chromatin-chromatin interactions (mediated by histone tails in different nucleosomes or by the linker histone H1) or are an approximation to the TF model (where the number of binders present in the system is saturating, and individual binders are readily available to mediate chromatin bridging). This model naturally gives rise to phase separation between different species of chromatin beads when the strength of the homotypic interactions is sufficiently strong, and it has been able to capture the Hi-C patterns in *Drosophila* [67]. This type of modelling has also been successful in explaining epigenetic memory [68–70] (see Chapter 4) and, with the incorporation of the nuclear lamina, the large-scale reorganisation of the genome in rod cells of nocturnal mammals [71] and in cellular senescence [72, 73] (see Chapter 5).

The mechanistic models presented above have each been successful in capturing various aspects of chromatin folding. In recent years, studies have also worked on combining these different models. For example, the block copolymer and the TF models have each been coupled with the LE model to understand the combined effect of LE activity and phase separation, as arising from (direct or TF-mediated) homotypic chromatin interactions, on genome architecture [74, 75]. These studies have shown that both mechanisms complement each other, with LE driving the formation of smaller-scale features such as TADs and phase separation leading to larger-scale structures such as compartments.

3

Modelling Chromatin Using Molecular Dynamics Simulations

Computer simulations have been an important methodology in scientific research since the dawn of electronic computing. Simulations provide a practical means to test theories against experiments, and they are particularly useful for modelling complex systems, such as chromatin and other biomolecules. A popular and powerful simulation technique is molecular dynamics (MD) [76], which is utilised extensively in this thesis for understanding genome organisation. In this chapter, I provide an overview of MD simulations and outline the general framework of modelling chromatin using this technology. I will discuss the specific details of individual simulation models in subsequent chapters.

3.1 Molecular Dynamics Simulations

Molecular dynamics (MD) is a simulation method designed for studying the physical properties of a many-body system. It computes the time evolution

of the positions and momenta of a set of interacting particles which obey the classical equations of motion. The method has found applications in various disciplines including material science, chemistry, biophysics, and astrophysics. MD simulations were developed in the 1950s when electronic computers started to become available. They were first used in studying the properties of hard-sphere systems [77] and subsequently of simple liquids [78]. In the 1970s, the method was applied to investigating the structure and dynamics of biomolecules [79] and has since remained a vital tool in this research area [80].

There are two main purposes for conducting MD simulations. First, they allow one to probe the *dynamical* properties of a system, such as transport coefficients and time-dependent responses. Second, similar to Monte Carlo (MC) simulations, they can be applied to determining the thermodynamic, *equilibrium* properties of the system, provided that the underlying dynamics obey the ergodic hypothesis (that every microstate of the system is accessible from all other microstates). In other words, one can estimate the statistical or ensemble average of an observable of the system (e.g., the internal energy or radius of gyration) by calculating the time average of that quantity in the simulation.

A key consideration when setting up an MD simulation is to decide the appropriate level of coarse-graining (CG; see Section 2.2), and this depends on the main objective of the research. On the one hand, high-resolution simulations at the atomistic level tend to be very accurate, but they are computationally expensive and are only feasible for short time scales. On the other hand, more coarse-grained simulations, despite the reduced level of details, allow one to examine the long-time behaviour of the system, which may be advantageous for exploring the full configuration space of the system and calculating thermodynamic quantities.

In this thesis, I employ a common CG approach which models chromatin as a bead-and-spring polymer, as discussed in Section 2.2 (Fig. 2.2A). Specifically, each bead has a diameter σ and represents a segment of chromatin (i.e., a number of nucleosomes) whose internal structure is not resolved within the model. The resolution of each bead depends on the specific problem under consideration. For instance, this could be 1 kbp of chromatin when modelling a particular gene locus, or 10 kbp when looking at the organisation of a whole chromosome. The level of description here is sufficient to capture the polymeric nature of chromatin while ignoring the chemistry of individual base pairs, which is less important in shaping the large-scale structure of chromatin. Furthermore, beads can be divided into

different types, or "colours", with different interaction properties based on the underlying one-dimensional (1D) genomic data (see, e.g., Section 2.3.3). This framework enables coupling between the three-dimensional (3D) organisation of chromatin and its 1D information.

3.1.1 A Basic Framework of Molecular Dynamics

In their elementary form, MD simulations attempt to solve numerically the classical (Newtonian) equations of motion for a system of N particles:

$$\frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t} = \boldsymbol{v}_i \tag{3.1}$$

$$m_i \frac{\mathrm{d}\boldsymbol{v}_i}{\mathrm{d}t} = \boldsymbol{F}_i = -\boldsymbol{\nabla}_i \mathcal{U}(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N), \tag{3.2}$$

where m_i, $\boldsymbol{r}_i$, and $\boldsymbol{v}_i$ are the mass, position, and velocity, respectively, of particle i (with $i = 1, ..., N$). $\boldsymbol{F}_i$ is the net force acting on the particle due to the interactions with all other particles, and this is derived from the potential energy $\mathcal{U}$ of the system. The basic framework of running an MD simulation can be summarised as follows:

1. Specify the simulation box details (e.g., box size and boundary conditions).
2. Define the initial position and velocity of each particle.
3. Establish the potentials governing the interactions between particles, and those between particles and the boundary walls (if they are present).
4. Calculate the net forces acting on the particles and thus their accelerations at the current time step.
5. Compute the velocities and positions of the particles at the next time step by numerically integrating Eqs. (3.1) and (3.2).
6. Measure physical quantities of interest (at sampling time steps).
7. Repeat Steps 4–6 using the updated velocities and positions for as many times as required.

A critical part of an MD simulation scheme is the set of potentials that describe the interactions between particles (and those between particles and the boundary walls) in the system. These can be directly based on the physical interactions governing the particles (e.g., using the Coulomb potential for charged particles),

or they can be purely phenomenological. In this work, the potentials used are divided into two classes: bonded and non-bonded potentials. The former models particles that are connected together at a molecular level, such as nucleotides along a DNA polymer. The latter deals with attractive and repulsive interactions between non-connected particles, such as excluded volume effects and chromatin-protein binding. The following discusses standard potentials for polymer simulations that are considered in this thesis.

3.1.2 Bonded Potentials

Modelling a polymer requires linking monomers (or beads) together. A common potential to enforce the connectivity between consecutive beads is the finite extensible non-linear elastic (FENE) potential [81]

$$\mathcal{U}_{\mathrm{FENE}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = -\frac{K_F R_0^2}{2} \ln\left[1 - \left(\frac{r_{ij}}{R_0}\right)^2\right] \tag{3.3}$$

for $r_{ij} < R_0$ and $\mathcal{U}_{\mathrm{FENE}} = \infty$ otherwise. Here, $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is the distance between the beads, R_0 is the maximum length of the bond, and K_F is its stiffness. The advantage of this potential is that it limits how far the bond can be extended, which can help reduce chain crossing and entanglements. In all simulations, $R_0 = 1.6\sigma$ and $K_F = 30k_BT/\sigma^2$ (with k_B being the Boltzmann constant and T the system's temperature) so that the equilibrium bond length is approximately 1σ when one superposes the FENE potential with the purely repulsive Weeks-Chandler-Andersen (WCA) potential (see below).

An alternative way to join beads together is via the harmonic potential

$$\mathcal{U}_{\mathrm{harm}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{K_h}{2}(r_{ij} - r_0)^2, \tag{3.4}$$

where r_0 is the equilibrium bond length and K_h is the spring constant. Unlike FENE bonds, harmonic bonds can be stretched without limits. It is useful during equilibration to relax the polymer from an initial conformation where bonds are too extended, for which applying the FENE potential directly will cause numerical divergences. In addition, harmonic bonds are also used in some simulations to connect non-consecutive chromatin beads.

Like many polymers, chromatin is semi-flexible. Its bending rigidity is incorporated in the simulation model by adding a three-body cosine potential (i.e., a

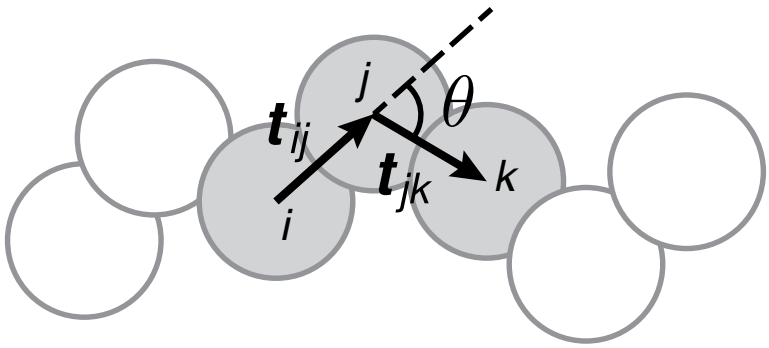

Figure 3.1: An illustration of the bending angle θ formed by two consecutive tangent vectors $\boldsymbol{t}_{ij}$ and $\boldsymbol{t}_{jk}$ which connect between beads i and j and beads j and k, respectively.

Kratky-Porod model; see Section 2.1.1)

$$\mathcal{U}_{\text{bend}}(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) = K_b \left[1 - \frac{\boldsymbol{t}_{ij} \cdot \boldsymbol{t}_{jk}}{t_{ij}\, t_{jk}}\right] = K_b(1 - \cos\theta), \tag{3.5}$$

where θ is the angle between consecutive tangent vectors $\boldsymbol{t}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$ and $\boldsymbol{t}_{jk} = \boldsymbol{r}_k - \boldsymbol{r}_j$ (Fig. 3.1). The parameter $K_b = k_B T \ell_p / \sigma$ controls the degree of bending, with ℓ_p being the persistence length of the polymer (which is half of the Kuhn length). The choice of ℓ_p for the modelled chromatin polymer depends on the level of CG; in particular, this polymer is more flexible (i.e., a lower ℓ_p) if one chooses a coarser level of description of chromatin. The specific value of ℓ_p used in my simulations will be introduced in subsequent chapters, where different levels of CG are considered. It should also be noted that the actual persistence length of chromatin is still much debated, and this is largely because the local folding (secondary structure) of the nucleosomal fibre remains poorly understood. Microscopy and chromosome conformation capture (3C) experiments have suggested a wide range of values from 30 to 200 nm for chromatin's persistence length [82–84].

3.1.3 Non-Bonded Potentials

As mentioned above, non-bonded potentials specify the interactions between non-connected beads. In general, all beads interact with one another, and the overall

non-bonded potential energy can be written as the following expansion:

$$\mathcal{U}_{\mathrm{NB}} = \sum_{i=1}^{N} \mathcal{U}_1(\boldsymbol{r}_i) + \frac{1}{2!}\sum_{i=1}^{N}\sum_{j\neq i} \mathcal{U}_2(\boldsymbol{r}_i, \boldsymbol{r}_j) + \frac{1}{3!}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k\neq j} \mathcal{U}_3(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) + \dots . \quad (3.6)$$

Here, the first term represents one-body interactions, or the internal energy of individual beads (e.g., due to an external field like gravity); the second term accounts for the energy of the two-body interactions between beads and those between beads and boundary walls (if present); the third term describes three-body interactions; and so on. In this work, I only consider one- and two-body interactions as they are sufficient to capture the essential physics of the system. Furthermore, higher order interactions are less common and introducing these terms will increase the computational cost substantially.

A classic non-bonded pairwise potential in MD simulations is the Lennard-Jones (LJ) potential

$$\mathcal{U}_{\mathrm{LJ}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = 4\epsilon \left[\left(\frac{d_{ij}}{r_{ij}}\right)^{12} - \left(\frac{d_{ij}}{r_{ij}}\right)^{6} \right], \quad (3.7)$$

where r_{ij} is the distance between beads i and j, d_{ij} (set to the bead diameter σ unless otherwise stated) is the distance at which the potential is zero, and ϵ is the minimum energy. The beads repel each other when $r_{ij} < 2^{1/6} d_{ij}$ and conversely attract to one another when beyond this point. To model the steric repulsion between beads due to excluded volume effects, I consider the Weeks-Chandler-Andersen (WCA) potential [85]

$$\mathcal{U}_{\mathrm{WCA}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = 4k_B T \left[\left(\frac{d_{ij}}{r_{ij}}\right)^{12} - \left(\frac{d_{ij}}{r_{ij}}\right)^{6} + \frac{1}{4} \right] \Theta(2^{1/6} d_{ij} - r_{ij}), \quad (3.8)$$

which is essentially a shifted LJ potential that contains only the repulsive part and ensures there is no discontinuity at the cutoff $2^{1/6} d_{ij}$ (Fig. 3.2). The Heaviside step function Θ (where $\Theta(x) = 1$ if $x > 0$ and zero otherwise) forces the energy to be zero when $r_{ij} > 2^{1/6} d_{ij}$, so there is no interaction between the particles beyond this point.

In my simulations, some chromatin beads can attract to one another and, in certain systems, to boundary walls (Chapter 5) and to protein complexes (Chapter 6), which are modelled as spherical, multivalent beads with a diameter

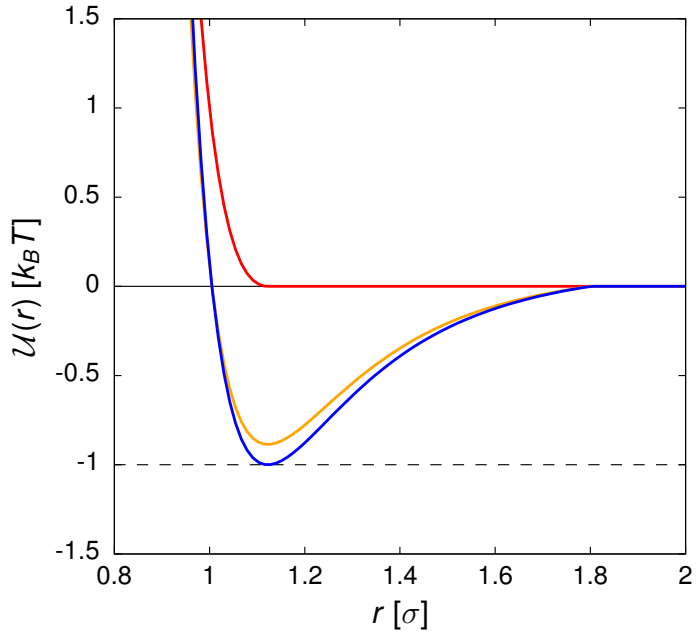

Figure 3.2: Variants of the Lennard-Jones (LJ) potential. The red curve depicts the Weeks-Chandler-Andersen (WCA) potential described in Eq. (3.8). The blue curve shows the truncated and shifted LJ potential defined in Eq. (3.9) with $\epsilon = k_BT$ and $r_c = 1.8\sigma$. The dashed line marks the minimum of this potential at $-\epsilon$. The orange curve shows the same potential but without the normalisation factor $\mathcal{N}$.

σ. These interactions are characterised using a truncated and shifted LJ potential

$$\mathcal{U}_{\mathrm{LJ/cut}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{1}{\mathcal{N}} \left[\mathcal{U}_{\mathrm{LJ}}(r_{ij}) - \mathcal{U}_{\mathrm{LJ}}(r_c)\right] \Theta(r_c - r_{ij}), \tag{3.9}$$

where the normalisation factor

$$\mathcal{N} = 1 + 4\left[\left(\frac{d_{ij}}{r_c}\right)^{12} - \left(\frac{d_{ij}}{r_c}\right)^{6}\right] \tag{3.10}$$

ensures the depth of the potential reaches $-\epsilon$ at the minimum point (Fig. 3.2). The cutoff distance $r_c \geq 2^{1/6} d_{ij}$ is set such that a portion of the potential is attractive. Note that setting $r_c = 2^{1/6} d_{ij}$ and the energy $\epsilon = k_BT$ simply results in the WCA potential. The exact values for r_c and ϵ are model-specific and depend on the types of the chromatin beads, which are informed from genomic data. These parameters will be discussed more in detail in later chapters when the simulation models for particular systems are introduced.

The LJ potential can generate numerical divergences when beads come too close together, and this can often happen during the initial equilibration of the polymer. An alternative repulsive potential that helps push apart monomers

without causing "blow-ups" is the soft potential

$$\mathcal{U}_{\text{soft}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{A}{2}\left[1 + \cos\left(\frac{\pi r_{ij}}{r_c}\right)\right]\Theta(r_c - r_{ij}), \tag{3.11}$$

where r_c is the interaction cutoff distance and A is the height of the repulsive barrier between beads. This potential is employed in this thesis during the initial relaxation of the chromatin polymer. In particular, the cutoff is set to $r_c = 2^{1/6}\sigma$, and the height A is gradually increased during the relaxation period (see subsequent chapters for the exact parameter values).

3.1.4 Brownian Dynamics

It is common that the system of interest is not situated in free space but is submerged within a solvent. In the context of simulating chromatin organisation, the chromatin fibre is located within the nucleoplasm in the cell nucleus. One could explicitly model the solvent particles and add potentials to simulate their interactions with the system; however, there are usually many more solvent particles than beads within the system, and such explicit modelling will therefore incur a high computational cost. An alternative approach is to implicitly account for the effects of the solvent by using the following Langevin equation:

$$m_i \frac{\mathrm{d}^2\boldsymbol{r}_i}{\mathrm{d}t^2} = -\boldsymbol{\nabla}_i \mathcal{U} - \gamma_i \frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t} + \sqrt{2\gamma_i k_B T}\boldsymbol{\eta}_i(t). \tag{3.12}$$

Here, the first term on the right hand side represents contributions from conservative forces. The second term accounts for the friction force on the system due to the solvent, with γ_i being the friction coefficient. The third term describes the random forces due to the bombardment of solvent particles into the system's beads, and $\boldsymbol{\eta}_i$ is a white noise vector which satisfies the statistical averages

$$\langle \eta_{i\alpha}(t) \rangle = 0 \tag{3.13}$$

$$\langle \eta_{i\alpha}(t)\eta_{j\beta}(t') \rangle = \delta_{ij}\delta_{\alpha\beta}\delta(t - t'), \tag{3.14}$$

where δ_{ij} is the Kronecker delta and $\delta(x)$ is the Dirac delta function, and the Latin indices represent the beads' indices whereas the Greek indices run over Cartesian components. This type of MD simulations is referred to as Brownian dynamics (BD) or Langevin dynamics simulations, and it neglects solvent-mediated hydrodynamic interactions between beads.

It is worth noting that at time scales much greater than the inertial time $\tau_{\text{in}} = m/\gamma$, i.e., the time over which inertial effects become insignificant, Eq. (3.12) can be approximated by an overdamped Langevin equation:

$$\frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t} = -\frac{1}{\gamma_i}\boldsymbol{\nabla}_i \mathcal{U} + \sqrt{2D_i}\,\boldsymbol{\eta}_i(t), \tag{3.15}$$

where D_i is the (translational) diffusion coefficient of bead i and is related to its friction coefficient γ_i via the Einstein relation $D_i = k_B T/\gamma_i$. Since the diffusion coefficient has units of $[\mathrm{L}]^2[\mathrm{T}]^{-1}$, one can extract another time scale $\tau_{\text{Br}} = \sigma^2/D = \gamma\sigma^2/(k_B T)$ known as the Brownian time. This is the characteristic time over which a bead diffuses across a distance of its own size.

3.2 Integration of the Equations of Motion

Integrating the equations of motion numerically is the bread and butter of MD simulations, and thus selecting an appropriate integration scheme is of paramount importance. The Verlet algorithm [86] is a standard numerical method to integrate the equations of motion which provides good numerical stability and conserves the total energy of the system [76]. It can be derived from the forward and backward Taylor expansions of a particle's position $\boldsymbol{r}(t)$ around time t:

$$\boldsymbol{r}(t+\Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\Delta t + \frac{\boldsymbol{a}(t)}{2}\Delta t^2 + \frac{\boldsymbol{b}(t)}{6}\Delta t^3 + \mathcal{O}(\Delta t^4) \tag{3.16}$$

$$\boldsymbol{r}(t-\Delta t) = \boldsymbol{r}(t) - \boldsymbol{v}(t)\Delta t + \frac{\boldsymbol{a}(t)}{2}\Delta t^2 - \frac{\boldsymbol{b}(t)}{6}\Delta t^3 + \mathcal{O}(\Delta t^4), \tag{3.17}$$

where $\boldsymbol{b}(t) = \mathrm{d}^3\boldsymbol{r}/\mathrm{d}t^3$, $\boldsymbol{a}(t)$ is the particle's acceleration, and $\boldsymbol{v}(t)$ is its velocity. From Eq. (3.2) one finds $\boldsymbol{a}(t) = \boldsymbol{F}(t)/m$. Adding these two equations together gives

$$\boldsymbol{r}(t+\Delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\Delta t) + \frac{\boldsymbol{F}(t)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4), \tag{3.18}$$

which is the standard form of the Verlet algorithm. This equation shows that the integrated position is accurate up to the fourth order in Δt. In addition to the particles' positions, one may want to acquire information about their instantaneous velocities, which are required for computing the total kinetic energy of the system and its temperature (via the equipartition theorem). A particle's

velocity can be obtained from subtracting the two expansions:

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t+\Delta t) - \boldsymbol{r}(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2), \tag{3.19}$$

which has an accuracy up to the second order in Δt. The standard form of the Verlet algorithm gives the position and velocity of a particle at different time steps. The velocity-Verlet algorithm [87] resolves this problem and is defined as follows:

$$\boldsymbol{r}(t+\Delta t) \equiv \boldsymbol{r}(t) + \boldsymbol{v}(t)\Delta t + \frac{\boldsymbol{F}(t)}{2m}\Delta t^2 \tag{3.20}$$

$$\boldsymbol{v}(t+\Delta t) \equiv \boldsymbol{v}(t) + \frac{\boldsymbol{F}(t+\Delta t) + \boldsymbol{F}(t)}{2m}\Delta t. \tag{3.21}$$

Note that in this scheme one needs to obtain the new positions, and thereby the new forces, before computing the new velocities. The velocity-Verlet scheme is merely a variant of the standard Verlet scheme and produces the same trajectories. This can be shown by considering the position and velocity at time step t:

$$\boldsymbol{r}(t) = \boldsymbol{r}(t-\Delta t) + \boldsymbol{v}(t-\Delta t)\Delta t + \frac{\boldsymbol{F}(t-\Delta t)}{2m}\Delta t^2 \tag{3.22}$$

$$\boldsymbol{v}(t) = \boldsymbol{v}(t-\Delta t) + \frac{\boldsymbol{F}(t) + \boldsymbol{F}(t-\Delta t)}{2m}\Delta t. \tag{3.23}$$

Subtracting Eq. (3.20) by Eq. (3.22) and rearranging gives

$$\begin{aligned}\boldsymbol{r}(t+\Delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\Delta t) + \left[\boldsymbol{v}(t) - \boldsymbol{v}(t-\Delta t)\right]\Delta t + \\ \frac{\boldsymbol{F}(t) - \boldsymbol{F}(t-\Delta t)}{2m}\Delta t^2.\end{aligned} \tag{3.24}$$

Finally, substituting Eq. (3.23) to this equation gives the update rule for the standard Verlet scheme.

All simulations conducted in this thesis are done using the BD framework explained above [Eq. (3.12)] and are evolved using the velocity-Verlet scheme. In practice, they are performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package (see http://lammps.sandia.gov and Ref. [88]). Specifically, integration is done in the canonical ensemble framework, where the number of particles N, the volume V, and the temperature T of the system are conserved throughout the simulation duration – this is also known

as the NVT ensemble[1]. The temperature is maintained by means of a Langevin thermostat [i.e., implementing the second and third terms of Eq. (3.12)]. The size of each time step is $\Delta t = 0.01\tau$, where τ is the simulation time unit (see the next section). For simplicity, throughout this thesis all beads have the same mass and friction coefficient (i.e., $m_i = m$ and $\gamma_i = \gamma$).

3.3 Mapping between Simulation and Physical Units

Computer simulations are typically performed in reduced (or dimensionless) units, where quantities are expressed as multiples of some fundamental entities that are intrinsic to the system. In this way, most quantities have similar orders of magnitude, and this can help limit numerical errors resulting from finite precision and floating-point arithmetic. In MD simulations, a natural way to non-dimensionalise quantities is to express them in terms of some fundamental distance, mass, energy, and the Boltzmann constant k_B (the resulting reduced units are also known as the LJ units in LAMMPS). In my simulations, distances and masses are given in units of the diameter σ and mass m, respectively, of a bead in the chromatin polymer, and energies are expressed in units of k_BT (see Table 3.1 for the mapping between physical and reduced units for common quantities). As a result, in these reduced units one has $m = \sigma = k_B = 1$. The natural time scale that can be constructed from these elementary quantities is $\tau = \sigma\sqrt{m/(k_BT)}$, and this defines the simulation time unit (though see below).

Simulation data can be converted from the reduced to physical units by fixing the values of the fundamental quantities. For the energy scale, the system's temperature is set at room temperature $T = 300$ K, so $k_BT \approx 4.1 \times 10^{-21}$ J. For the length scale, one needs to determine the physical size σ of a bead, which depends on the resolution of the simulation model (i.e., bp/bead) and the linear compaction of chromatin (i.e., bp/nm, or more often reported as the number of nucleosomes per diameter of a nucleosome – nuc/11 nm). The latter remains poorly known and depends on the local structure of chromatin (e.g., whether it exists as a 10-nm or 30-nm fibre). Another way to ascertain the bead size,

[1]As an aside, there is also the NVE ensemble, or the micro-canonical ensemble, where the number of particles N, the volume V, and the energy E of the system are conserved. Simulations which integrate the basic classical equations of motion [Eqs. (3.1) and (3.2)] sample this ensemble.

Quantity	Symbol	Mapping to Reduced (LJ) Units
Distance	r	$r^* = r/\hat{\sigma}$
Time	t	$t^* = t\left[\hat{\epsilon}/(\hat{m}\hat{\sigma}^2)\right]^{1/2}$
Velocity	v	$v^* = vt/\hat{\sigma}$
Energy	E	$E^* = E/\hat{\epsilon}$
Force	F	$F^* = F\hat{\sigma}/\hat{\epsilon}$
Temperature	T	$T^* = k_B T/\hat{\epsilon}$
Mass	M	$m^* = M/\hat{m}$

Table 3.1: Mapping between physical and reduced (LJ) units in simulations. Specifically, the fundamental distance $\hat{\sigma}$ and mass $\hat{m}$ are given by the size σ and mass m, respectively, of a chromatin bead, and the fundamental energy $\hat{\epsilon}$ is given by $k_B T$.

without relying on estimates of the linear compaction, is to defer this mapping until after completing the simulations. In this way, one can deduce the bead size by comparing the simulated structures directly with real structures observed in microscopy from a relevant biological system. Since in different chapters I will be focussing on chromatin behaviour at different length scales, I will give more details about the exact mapping of the bead size there when I present the specific simulation models.

Mapping the time scale requires a more careful analysis. From the discussions above, there are three relevant time scales in the simulations: the natural time $\tau = \sigma\sqrt{m/(k_B T)}$, the inertial time $\tau_{\text{in}} = m/\gamma$, and the Brownian time $\tau_{\text{Br}} = \gamma\sigma^2/(k_B T)$ (see Section 3.1.4). In this thesis, I am interested in time scales beyond τ_{Br}, where beads have diffused distances greater than their diameter σ, in order to sample the conformations of the chromatin fibre properly. However, these three time scales are in reality separated by orders of magnitude in the context of chromatin dynamics: assuming Stoke's law holds for a spherical chromatin bead (i.e., $\gamma = 3\pi\nu\sigma$, where ν is the viscosity of the solvent), a simple calculation using realistic values for the mass and diameter of a chromatin bead and the viscosity and temperature within the nucleus reveals that $\tau_{\text{in}} \ll \tau \ll \tau_{\text{Br}}$[2]. Owing to numerical stability, the time step Δt has to be smaller than τ. As a result, if realistic values are used for these time scales, one will have to run impractically long simulations to observe the relevant dynamics.

[2]For instance, consider a $\sigma = 30$ nm bead with 3 kbp of chromatin (around 15 nucleosomes). The mass of a nucleosome is roughly 2.6×10^5 Da $\approx 4.3 \times 10^{-22}$ kg [3], and so the mass of a bead is approximately 6.5×10^{-21} kg. Taking the viscosity ν of the nucleoplasm to be 150 cP [68], then the friction coefficient $\gamma \approx 4.2 \times 10^{-8}$ kg/s. Using $k_B T \approx 4.1 \times 10^{-21}$ J and the expressions for the various time scales, one finds $\tau_{\text{in}} \approx 2 \times 10^{-13}$ s, $\tau \approx 4 \times 10^{-8}$ s, and $\tau_{\text{Br}} \approx 9 \times 10^{-3}$ s. Note also that in reduced units the friction coefficient is $\gamma\tau/m \approx 2 \times 10^5$.

A standard workaround to this problem is to set the inverse of the inertial time γ/m to be much smaller than the realistic value (so beads have more inertia than in reality). In this thesis, I consider $\gamma/m = 1$ (or 2 in Chapter 6), compared to $\sim 10^5$ in reality[3] (in reduced units). In this way, one has $\tau_{\text{in}} \leq \tau \leq \tau_{\text{Br}}$, and the simulation dynamics can reach the Brownian time scale quickly (i.e., 1 or 2τ). As a consequence, the short time dynamics of the system are not realistic, but this is irrelevant since I am focussed on the system's behaviour at times much larger than τ_{Br}. In light of this consideration, simulation time is mapped to the physical time via the Brownian time instead of the natural time scale (i.e., setting $\tau \approx \tau_{\text{Br}}$).

In practice, the Brownian time can be determined using the aforementioned expression: $\tau_{\text{Br}} = \gamma\sigma^2/(k_B T) = 3\pi\nu\sigma^3/(k_B T)$ (assuming Stoke's law holds). This approach requires estimating the viscosity ν of the nucleoplasm. An alternative approach to obtain τ_{Br} without needing to make such an estimation is by matching the diffusive property of the simulated polymer with that of real chromatin. In particular, one can consider the mean squared displacement (MSD) of a polymer bead, which is given by

$$\text{MSD}(t) = \left\langle (\boldsymbol{r}(t_0 + t) - \boldsymbol{r}(t_0))^2 \right\rangle_{t_0}, \tag{3.25}$$

where the average $\langle \cdots \rangle$ is taken over time t_0. Theories on polymer dynamics [25] have shown that the MSD of a polymer bead exhibits the scaling

$$\text{MSD}(t) \sim t^{\zeta}, \tag{3.26}$$

where $\zeta = 1/2$ for the Rouse model (a model describing the dynamics of an ideal chain without hydrodynamic interactions), at time scales shorter than the Rouse time τ_R – the characteristic time for the whole polymer to diffuse a distance of its own size. With the mapping of the length scale known, one can use this relation to fix a value for τ_{Br} such that the MSD of a polymer bead in simulations is compatible with that of a real chromatin locus determined from single-particle tracking experiments (such as Ref. [84]).

[3] See footnote 2.

4

Simulating Chromatin with Epigenetic Modifications

In multi-cellular organisms, cells are often specialised to perform different tasks. Even though most cells carry exactly the same genome[1], they have different gene regulatory and expression patterns that give rise to their distinct phenotypes. This phenomenon, that cells can interpret and express their DNA differently, is to a large extent driven by epigenetic modifications, which are changes that do not involve alterations to the underlying DNA sequence [2, 89, 90]. At a molecular level, these modifications are typically the addition or removal of specific biochemical marks (or epigenetic marks) directly on DNA or on histone tails (i.e., post-translational modifications [PTMs] of histones [2, 90]). Examples of PTMs include acetylation, methylation, and phosphorylation (Fig. 4.1), and several classes of enzymes[2] (or modifiers) are responsible for depositing and erasing these marks.

[1] An exception to this are lymphocytes such as B and T cells, which can rearrange some parts of the DNA sequence in order to generate different antibodies to protect the host against foreign invaders, such as viruses.

[2] For example, acetylation is regulated by histone acetyltransferases and histone deacetylases, whereas methylation is mediated by histone methyltransferases and histone demethylases.

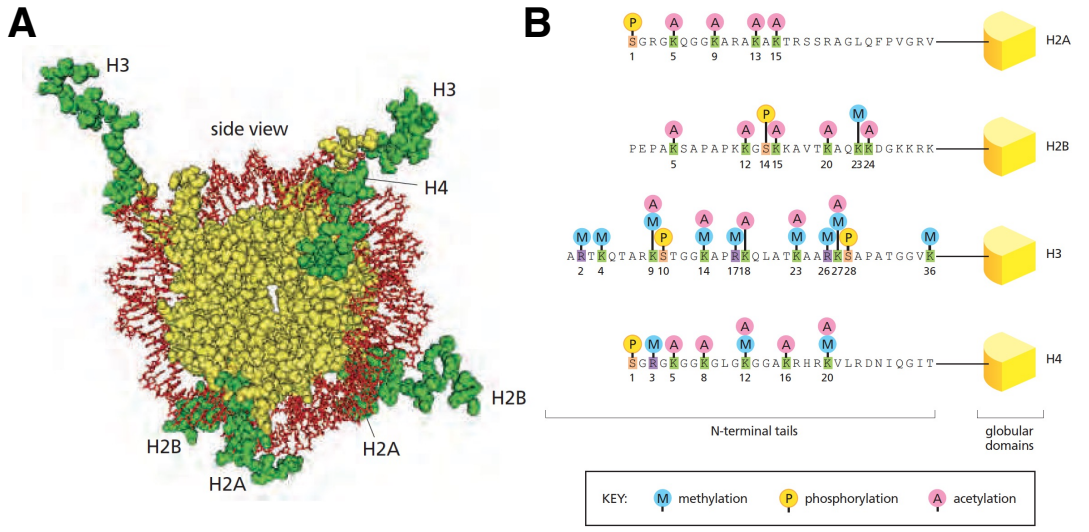

Figure 4.1: Post-translational modifications (PTMs) of histones. (A) An illustration of a histone octamer showing the various histone tails (in green) which can be subjected to modifications. (B) A schematic highlighting some common PTMs on various histone tails. Figures adapted from Ref. [2].

Cells from different tissues have their own "epigenetic landscapes", or sets of epigenetic modification patterns, along the genome. On the one hand, these landscapes are robust and heritable; they are maintained during the lifetime of cells and are faithfully re-established in their descendants [91]. On the other hand, they are dynamic and adaptable; they can be altered in response to external cues [92, 93] and are affected by disease and ageing[3] [94–96]. Furthermore, many of the epigenetic marks have a turnover time that is less than the lifetime of a cell. For instance, the average half-life of acetyl groups on histones is less than 10 minutes [97, 98], whereas methyl groups can be lost within a cell cycle [98–100]. Epigenetic modifications along chromatin can be altered due to histone replacement during DNA transcription [90, 94, 98, 101] and replication [90, 102, 103], or as a result of stochastic deposition and removal [104–106].

How cells acquire, maintain, and pass on their plastic yet reproducible epigenetic landscapes despite the rapid turnover of individual modifications remains a key topic to be understood. Biophysical modelling has helped elucidate principles of epigenetic regulations, in particular how "epigenetic domains" – stretches of chromatin modified by a similar epigenetic mark – can be formed and sustained [107]. One-dimensional (1D) models have successfully captured some aspects of the spreading and inheritance of histone PTMs [104, 105, 108–115]. In

[3]The connection between epigenetic modifications, cellular ageing (or senescence), and chromatin organisation is further explored in Chapter 5.

many of these 1D models, the spreading of histone marks is based on a positive feedback mechanism involving the modified histones and the modifiers themselves. The mechanism considers multivalent "reader" proteins which bind to histones with specific marks and "writer" proteins which deposit the marks. Crucially, the reader for a particular mark can attract the writer for the same mark. Hence, when a reader binds to a domain containing its cognate mark, it can recruit writers to deposit the same mark in nearby chromatin regions. This increases the local density of the mark, and more readers of that kind are likely to associate with the domain, which in turn attract more writers and so on, allowing the mark to propagate and the domain to grow. Simulation results from 1D models suggest that this feedback enables robust maintenance of epigenetic information even in the face of disruptions, such as DNA replication or stochastic modification [105].

Several observations support this feedback mechanism. For instance, heterochromatin protein 1 (HP1), a reader of the histone 3 lysine 9 trimethylation (H3K9me3) mark, can recruit the methyltransferase SUV39H1, a writer of this mark [98, 116, 117]. Furthermore, there are protein complexes which act both as a reader and a writer. The Polycomb repressive complex 2 (PRC2) can read the histone 3 lysine 27 trimethylation (H3K27me3) mark via its JARID2 domain, while its EZH2 domain can spread this mark [118–120]. The MLL/KMT2 enzymes responsible for depositing methylation on histone 3 lysine 4 (H3K4), a mark which correlates with gene expression, have domains that recognise the same mark [121]. Feedback for active, non-heterochromatic marks may also be driven by transcription factors and polymerases (acting as readers) if they can attract writers for these marks [98, 122].

Although 1D models have given insights on key processes mediating epigenetic domains, they cannot examine in detail the crosstalk between the three-dimensional (3D) organisation of chromatin and its epigenetic modifications. Some of the 1D models which impose effective long-range interactions have suggested that non-adjacent and looping-driven contacts are essential for the stability of epigenetic domains [105, 115]. Furthermore, previous 3D modelling work has demonstrated the epigenetic information is vital for generating chromatin structures that are compatible with data from chromosome conformation capture (3C) experiments [46, 48, 67].

A recent simulation work [68] addressed this point by explicitly coupling a bead-and-spring polymer model for the 3D folding dynamics of chromatin with a Potts-like model for the epigenetic changes along chromatin. The work showed

that this coupling can give rise to strong bistability with a first-order-like transition between two alternative states: one where the chromatin is compact and epigenetically ordered, and one where it is swollen and disordered. More importantly, there is hysteresis near the transition boundary (a typical hallmark of first-order transitions), which can naturally explain the memory and stability of epigenetic domains.

While incorporating 3D chromatin interactions gives a more realistic description of epigenetic regulation, there are issues yet to be resolved. A major issue which has not been fully addressed by feedback-based models is the principles for establishing a stable and replicable pattern of *multiple* epigenetic domains with *different* modifications. In these models, when starting from a "blank canvas" under conditions in which spreading is favourable, often a single epigenetic state will prevail and eventually cover almost the entire simulated region. However, it is clear that domains of different states need to coexist along the genome *in vivo* for there to be different expression patterns. Although in Ref. [68] long-live domains of different states can coexist by tuning the model out of equilibrium, these domains are metastable, and it is unclear how long they can be sustained.

In this chapter, I present a modelling framework which addresses this problem. In line with Ref. [68], the model considers feedback between the 3D arrangement of chromatin and its 1D epigenetic modifications. It allows *de novo* formation, spreading, and inheritance of multiple epigenetic domains. There are three key elements in this framework. First, it assumes a positive feedback between multivalent reader proteins and writer proteins, as discussed above. Second, it considers the presence of genomic bookmarking (GBM) proteins, which are transcription factors (TFs) that can bind to specific DNA sites, even during mitosis [95, 123–126]. There are several examples of these proteins: Polycomb group (PcG) proteins [127, 128] and posterior sex combs (PSCs) [129] in *Drosophila*, Esrbb [130] and Sox2 [131] in mouse, and GATA [132] and UBF [133] in human. Third, the model posits that the recruitment of specific reader and writer proteins is linked to a particular GBM protein. The framework predicts that GBM proteins enable domains of different kinds to be nucleated and sustained. After disruptive processes such as replication where a lot of the marks are lost, these proteins help recruit the appropriate readers and writers to restore the previous epigenetic landscape.

The rest of this chapter is organised as follows. In Section 4.1, I will first explain in detail the simulation model. In Section 4.2, I will discuss the possible phases

of the model when GBM proteins are absent. In Section 4.3, I will present results where these proteins are introduced. In particular, I will examine how varying the position and density of these proteins on chromatin can influence the resulting epigenetic patterns. Moreover, I will investigate how the model behaves when subjected to disruptive events, such as replication and bookmark excision. I will also show results for applying this model to simulate the whole right arm of chromosome 3 in *Drosophila*. Finally, in Section 4.4, I will draw some conclusions from the simulation results and discuss potential future work.

4.1 Simulation Model

The simulation model considers chromatin as a semi-flexible bead-and-spring polymer with N beads, in line with the framework discussed in Chapter 3. Each bead has a diameter σ and represents 3 kbp of chromatin. To model different epigenetic modifications (PTMs), beads are in one of several states (or colours), in line with previous modelling work [46, 48, 67, 68]. As a minimal model, each bead can be in one of three colours ($q = \{1, 2, 3\}$; Fig. 4.2A): red ($q = 1$; enriched in H3K27me3), grey ($q = 2$; unmarked), or blue ($q = 3$; enriched in H3K9me3). Note that this framework is general and can be used to describe any two epigenetic marks (not necessarily repressive marks) that are in direct competition, alongside an unmodified state. A more realistic model with more colours will be discussed later for chromosome-wide simulations in *Drosophila*.

In the simulations, beads of the same colour (except those in the grey, unmarked state) can attract to one another with interaction energy ϵ (in units of k_BT)[4], modelled by a truncated and shifted LJ potential [Eq. (3.9)] with an interaction cutoff $r_c = 2.5\sigma$ (Fig. 4.2D)[5]. This is to implicitly account for chromatin bridging mediated by multivalent reader proteins binding to cognate sites, as done in Ref. [68]. The interactions between beads of different states and between those in the unmarked state are purely repulsive and are described by the WCA potential [Eq. (3.8)].

[4]For simplicity, the same interaction energy is used for red-red and blue-blue interactions (i.e., $\epsilon_{\text{red}} = \epsilon_{\text{blue}} = \epsilon$). Asymmetric interaction strength is also examined, as discussed in Section 4.3.2.

[5]Note that in this chapter the potential is not rescaled by the normalisation factor $\mathcal{N}$ [see Eq. (3.10)]. With $r_c = 2.5\sigma$, $\mathcal{N} \approx 0.98$, so the rescaling is negligible.

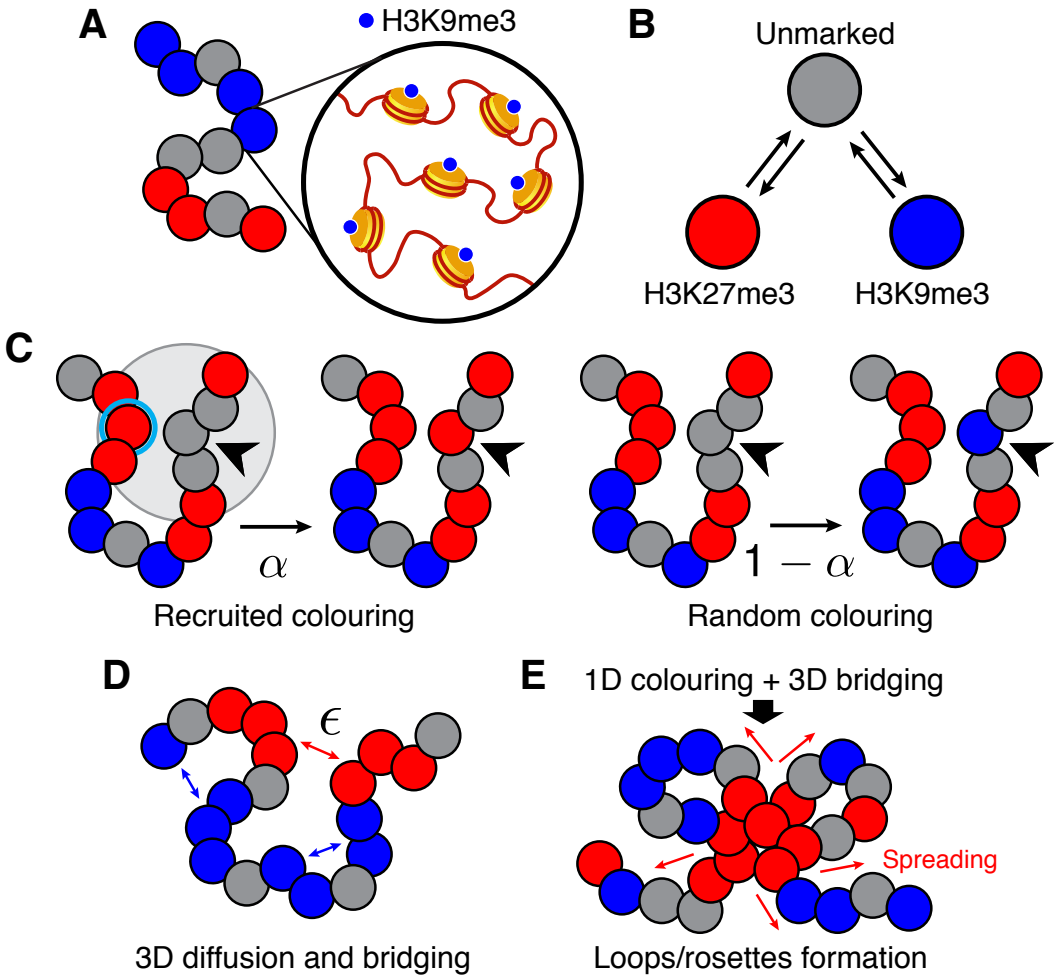

Figure 4.2: A simulation model coupling the 3D folding of chromatin and its 1D epigenetic modifications. (A) Chromatin is modelled as a bead-and-spring chain. Each bead represents 3 kbp of chromatin (roughly 15 nucleosomes) and is allowed to be in one of three epigenetic states or "colours": red (enriched in H3K27me3), grey (unmarked), and blue (enriched in H3K9me3). (B) Beads can change their state over time between red and blue via the grey state to account for the action of histone modifying and de-modifying enzymes. (C) Recolouring dynamics are governed by a Voter-like model based on the work in Ref. [105]. With probability α, a bead (indicated by the arrow) undergoes a recruited (feedback-based) colouring event, in which its colour is changed one step closer to that of a spatially proximate bead (circled in cyan, chosen randomly within a radius $r_c = 2.5\sigma$). Otherwise, with probability $1 - \alpha$, it undergoes a random colouring event, in which its state is changed stochastically while obeying the rule described in (B). (D) To implicitly model chromatin bridging mediated by multivalent epigenetic readers, beads of the same colour (except those in grey) can attract to one another with strength ϵ (in units of k_BT). (E) The coupling of 3D folding of chromatin and 1D epigenetic recolouring allows loops and rosettes to form, and these structures facilitate further spreading of epigenetic information.

To account for the turnover of epigenetic marks, beads can change colour over time according to a Voter-like model based on the 1D model described in Ref. [105]. They can be modified either by the feedback mechanism (i.e., a reader recruiting a writer of the same kind) or in a stochastic manner (i.e., modifications done by free writers). To simulate the action of modifying and de-modifying enzymes, marked beads must first be unmarked before acquiring another mark – i.e., they must transit through the grey, unmarked state (Fig. 4.2B). A recolouring step is performed after every time period of τ_{Re} within the simulations. In each recolouring step, N colour conversion attempts are done such that, on average, each bead receives one attempt. The procedure for each attempt is as follows:

1. Select a bead randomly from the fibre, say bead i. The bead is subjected to either a recruited (feedback-based) colouring event (Step 2), with probability α, or a random colouring event (Step 3), with probability $1-\alpha$.
2. Recruited colouring: Select another chromatin bead j randomly within a distance $r_c = 2.5\sigma$ from bead i. The colour of bead i is changed one step closer to that of bead j (Fig. 4.2C). In other words,
   - if $q_j = 1$, q_i undergoes the conversion $3 \to 2$ or $2 \to 1$;
   - if $q_j = 3$, q_i undergoes the conversion $1 \to 2$ or $2 \to 3$;
   - if $q_j = 2$ or $q_j = q_i$, q_i remains the same.
3. Random colouring: the colour of bead i is changed stochastically one step towards either one of the two other states with probability $1/3$, with no direct conversion between $q = 1$ and $q = 3$ permitted (Fig. 4.2C). More precisely,
   - if $q_i = 1$ or 3, it switches to 2 with a probability of $1/3$ or remains the same with a probability of $2/3$;
   - if $q_i = 2$, it has an equal probability of $1/3$ to convert to any of the three states.

   These rules ensure that there is, on average, an equal number of beads in each of the three states when the probabilities of a recruited and a random colouring event are the same (i.e., $\alpha = 1/2$).

As mentioned in Ref. [105], the overall epigenetic dynamics along the fibre are governed by the "feedback" parameter $f \equiv \alpha/(1-\alpha)$, which is ratio of the

probability of a recruited event to that of a random event. As f increases, recolouring based on feedback becomes more likely; as a result, an epigenetic mark can spread more easily along the chromatin fibre. Overall, the two key parameters of this simulation model are the epigenetic feedback strength f and the chromatin bridging strength ϵ mediated by the multivalent reader proteins.

It should be noted that the specific details of the recolouring dynamics play a less important role in determining the outcomes of this modelling framework. A Potts-like recolouring scheme (i.e., one governed by a Hamiltonian) also gives similar results and is discussed in Ref. [69]. What is critical in this framework is the coupling between the 1D epigenetic modification and the 3D folding dynamics. The synergy of these two aspects allows 3D loops and rosettes to form, and they in turn facilitate the spreading of epigenetic information to regions further away along the chromatin fibre (Fig. 4.2E).

Nevertheless, there are merits for using a Voter-like model. First, unlike the Potts-based model, this model is intrinsically out-of-equilibrium as its "rules" cannot be encoded in an effective Hamiltonian. Since PTMs are typically driven by active processes, a non-equilibrium model may be more suitable for portraying these processes. Second, in contrast to the Potts recolouring dynamics which consider one-to-many interactions along the fibre, this Voter-based model can capture more accurately the one-to-one nature in the reactions for depositing or removing histone marks.

In the following, I will specify more in detail the simulation procedure and the mapping between simulation and physical units. The simulations are conducted using the Brownian dynamics scheme discussed in Section 3.1.4. The chromatin fibre is simulated inside a periodic cube of length L. The size of the box is set to be much larger than the volume occupied by the fibre to avoid periodic boundary effects. Specifically, $L = 100\sigma$ for a chain with $N \leq 500$ beads, $L = 150\sigma$ for $500 < N \leq 2000$, and $L = 220\sigma$ for the *Drosophila* simulations presented later in this chapter.

The chromatin fibre is initialised as a random walk with each bead having a random epigenetic colour. The fibre is allowed to equilibrate for a period of $10^4\tau$, during which homotypic epigenetic interactions are switched off. More specifically, in the first 10τ of this period, non-bonded beads interact repulsively by the soft potential [Eq. (3.11)], whose height is slowly increased from 0 to $200k_BT$, such that the fibre becomes self-avoiding. This potential is replaced by

the WCA potential for the remaining time of the equilibration period. The main (production) simulation period lasts for $10^6\tau$ unless otherwise specified.

Distances in simulations are measured in units of the bead size, which is set to $\sigma = 30$ nm, in line with previous work on modelling chromatin at 3 kbp resolution [42, 46, 48]. This mapping gives a linear compaction of ∼5 nuc/11 nm, which is between the estimates obtained from fluorescence *in situ* hybridisation (FISH; 7–10 nuc/11 nm) [134] and 3C experiments (1–4 nuc/11 nm) [83]. The persistence length of the chromatin fibre is set to $\ell_p = 3\sigma = 90$ nm, which is within the range of estimates from previous experiments (30–200 nm; see Section 3.1.2). The simulation time unit τ is converted to real time via the Brownian time, or the typical time for a bead to diffuse a distance of its own size (see Section 3.3). Estimating the viscosity of the nucleoplasm to be 150 cP, the Brownian time is $\tau \approx 10$ ms [68]. The recolouring rate is set to $k_{\mathrm{Re}} \equiv \tau_{\mathrm{Re}}^{-1} = 10^{-3}\tau^{-1}$ (0.1 s^{-1}), which is compatible to the turnover timescales for histone marks [97]. In selected cases, a higher recolouring rate, $k_{\mathrm{Re}} = 10^{-1}\tau^{-1}$ (10 s^{-1}), is employed to allow faster convergence to a steady state and enable better sampling. Both rates give similar results qualitatively and no significant differences are observed.

4.2 Model Phases

To understand this model, its possible phases are first identified by varying the two key parameters (f, ϵ). Two main attributes of the chromatin fibre that can be affected by the parameters are its overall size and the order of epigenetic marks along the fibre. The former can be quantified by the radius of gyration R_g of the polymer (see Section 2.1.1) with

$$R_g^2 = \left\langle \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{r}_i - \boldsymbol{r}_{\mathrm{cm}})^2 \right\rangle, \tag{4.1}$$

where $\boldsymbol{r}_{\mathrm{cm}}$ is the polymer's centre of mass and the brackets $\langle \cdots \rangle$ denote an ensemble average over the conformations at different time steps and simulation runs. The latter can be described by an effective magnetisation

$$\widetilde{m} = \left\langle \frac{1}{N} \left[N_c(q=3) - N_c(q=1) \right] \right\rangle, \tag{4.2}$$

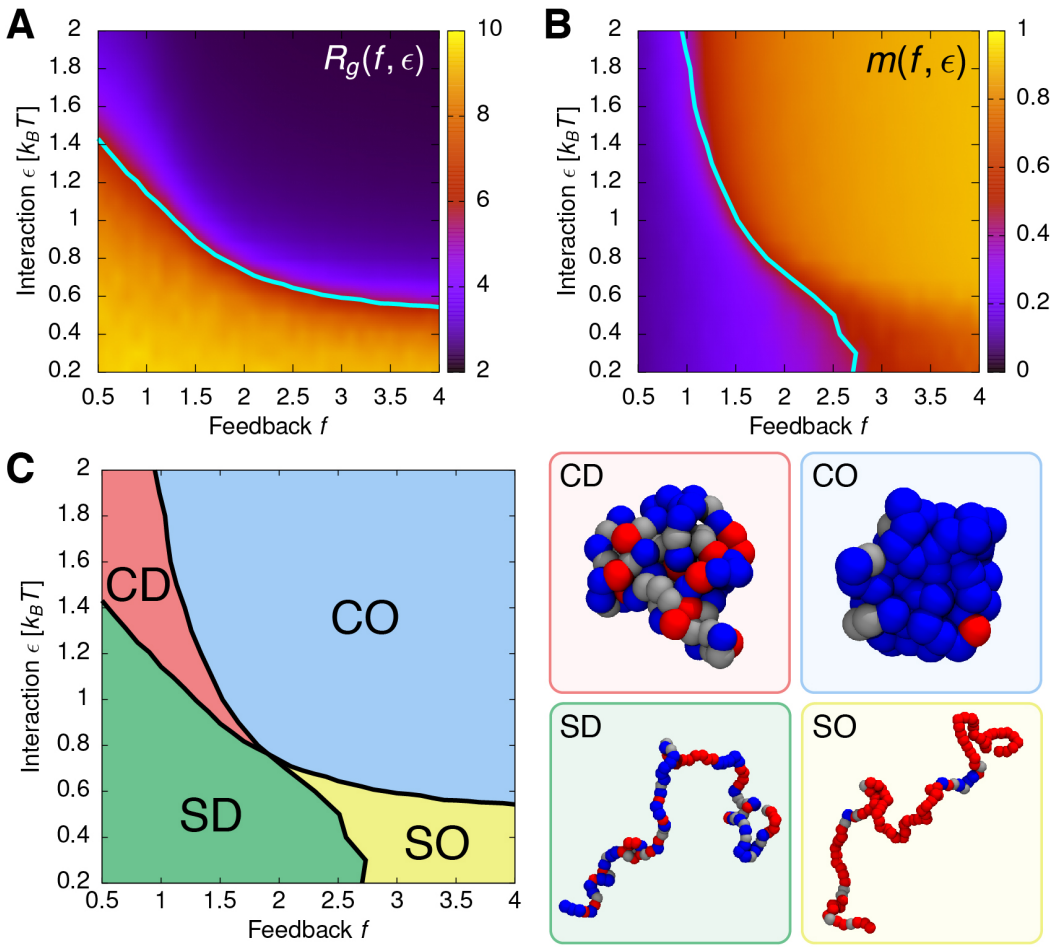

Figure 4.3: Phase diagrams of the simulation model for $N = 100$. (A) A heatmap showing the measured radius of gyration R_g in the phase space (f, ϵ), sampled in increments of 0.1 in both directions. The transition boundary (cyan line) is estimated based on the inflection points from fitting tanh curves on R_g data for each f. (B) Similar to (A), but showing the absolute effective magnetisation m of the system across the parameter space. The transition boundary is calculated using the same method as above, but from fitting tanh curves on m data for each ϵ. (C) *Left*: A phase diagram showing the four possible phases of the system by combining the transition boundaries from (A) and (B). These phases are compact-disordered (CD), compact-ordered (CO), swollen-disordered (SD), and swollen-ordered (SO). *Right*: Simulation snapshots representative of individual phases.

where $N_c(q)$ is the number of beads with colour q. The absolute value of this quantity $m \equiv |\widetilde{m}|$ is useful for distinguishing between the epigenetically disordered ($m \approx 0$) and ordered ($m \approx 1$) phases.

Measurements of these two observables reveal that there are four distinct phases in the model, characterised by the combination of a swollen or collapsed structure and an epigenetically ordered or disordered state along the chromatin fibre (Fig. 4.3). More precisely, at small ϵ, the system can either be in the swollen-

disordered (SD) phase at low f or the swollen-ordered (SO) phase at high f. When ϵ is large, the system either resides in the compact-disordered (CD) phase at low f or the compact-ordered (CO) phase at high f.

Some of these phases bear resemblance, at least qualitatively, to conformations found in real chromosomes. For instance, the CO phase is a suitable candidate for describing the structure of the inactivated X chromosome in female mammals, which is mostly coated with repressive marks and is highly condensed [2]. The CD phase is analogous to "gene deserts" (or black chromatin) in chromosomes – regions with no coherent epigenetic signal but are strongly interacting in 3D, possibly due to the action of the linker histone H1 proteins [135–137]. Moreover, the SO phase captures the arrangement of transcriptionally active and open chromatin [138].

It is intriguing to note that only the SD and CO phases are seen in the equilibrium version of the Potts-based recolouring model [68]. The remaining two phases, CD and SO, can only be obtained in a non-equilibrium model [69]. Given these "new" phases are reflective of some chromosome structures seen in physiological conditions, they further highlight the importance of capturing the non-equilibrium nature of epigenetic dynamics.

4.2.1 The Transition between the SD and CO phases

Apart from classifying the possible phases in the model, it is of interest to understand how the system transits from one phase to another. Here, I focus on the transition between the SD and CO phases. This transition captures biological processes where the spreading of an epigenetic mark occurs concomitantly with a significant change in 3D chromatin structure. A striking example of such, as noted briefly above, is the X-inactivation process during early embryonic development of female mammalian cells [2]. In this process, a repressive epigenetic mark proliferates along one copy of the X chromosomes, and that copy collapses into a highly compact structure called the Barr body. Crucially, this state is then maintained and passed on faithfully to daughter cells.

Simulation results from varying ϵ at $f = 2.0$ suggest that the transition between SD and CO phases is discontinuous or first-order-like. First and most importantly, the two phases coexist near the transition boundary, as demonstrated from plots of the joint probability density of the system's magnetisation $\widetilde{m}$ and radius of

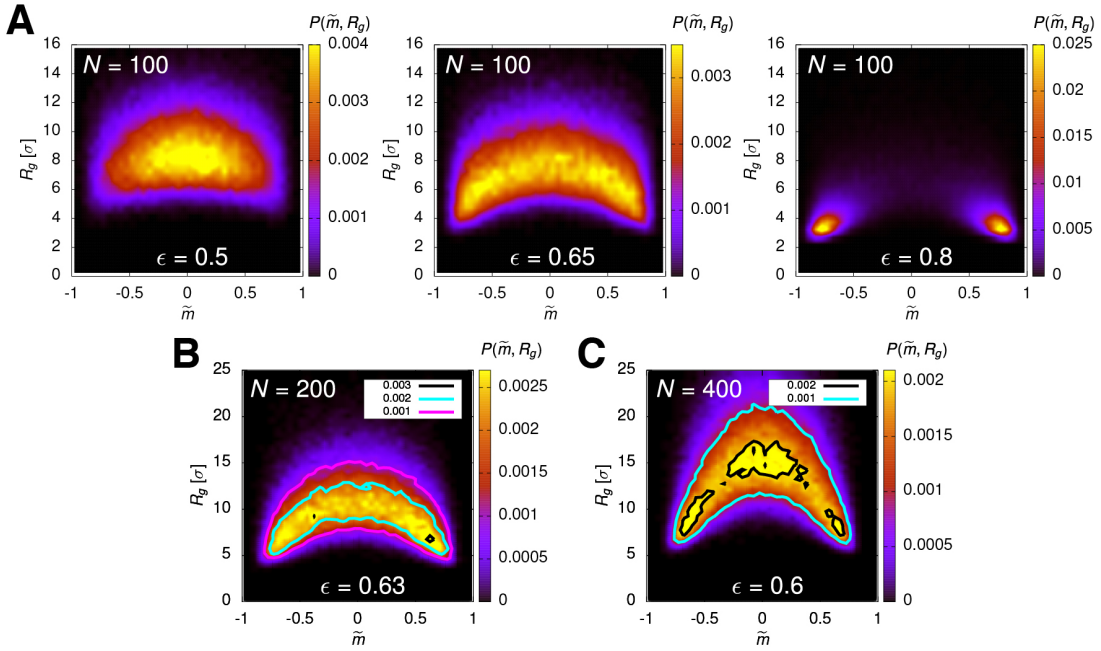

Figure 4.4: Phase coexistence in the SD-CO transition. (A) Heatmaps showing the joint probability density $P(\tilde{m}, R_g)$ for a system of $N = 100$ beads with $f = 2.0$ at three different interaction energy values: $\epsilon = 0.5$, 0.65, and 0.8. Note that the transition happens at $\epsilon_c \approx 0.65$ for this system size, and it shifts to a slightly lower ϵ as N increases due to finite-size effects. (B and C) Heatmaps displaying the joint probability density at the transition point for (B) $N = 200$ and (C) $N = 400$. Lines are drawn to highlight contour changes in the density. These plots show the emergence of three peaks as the system size increases, suggesting phase coexistence. Each plot is generated from sampling 100 simulations, except that 200 simulations are considered for $N = 400$. The bin size is 0.05 in $\tilde{m}$ and 0.5 in R_g for all plots.

gyration R_g (Fig. 4.4). In particular, at larger system sizes, three peaks become apparent in the density plots near the critical energy: the top peak corresponds to the SD phase, whereas the bottom two peaks correspond to the CO phase (one for a blue-ordered fibre and one for a red-ordered fibre). Second, there is a hysteresis cycle when moving between the phases (Fig. 4.5). When starting from a CO conformation and slowly decreasing the interaction energy ϵ, the system transits to the SD phase after some critical value of ϵ. However, when the process is reversed, with the system starting from an SD conformation and the energy increased, the transition to the CO phase does not occur at the same critical energy. In other words, the system retains memory of its previous conformation. This hysteresis effect offers a self-regulated mechanism for the system to recover its epigenetic (and polymeric) configuration in the face of disruptive processes, such as DNA transcription and replication.

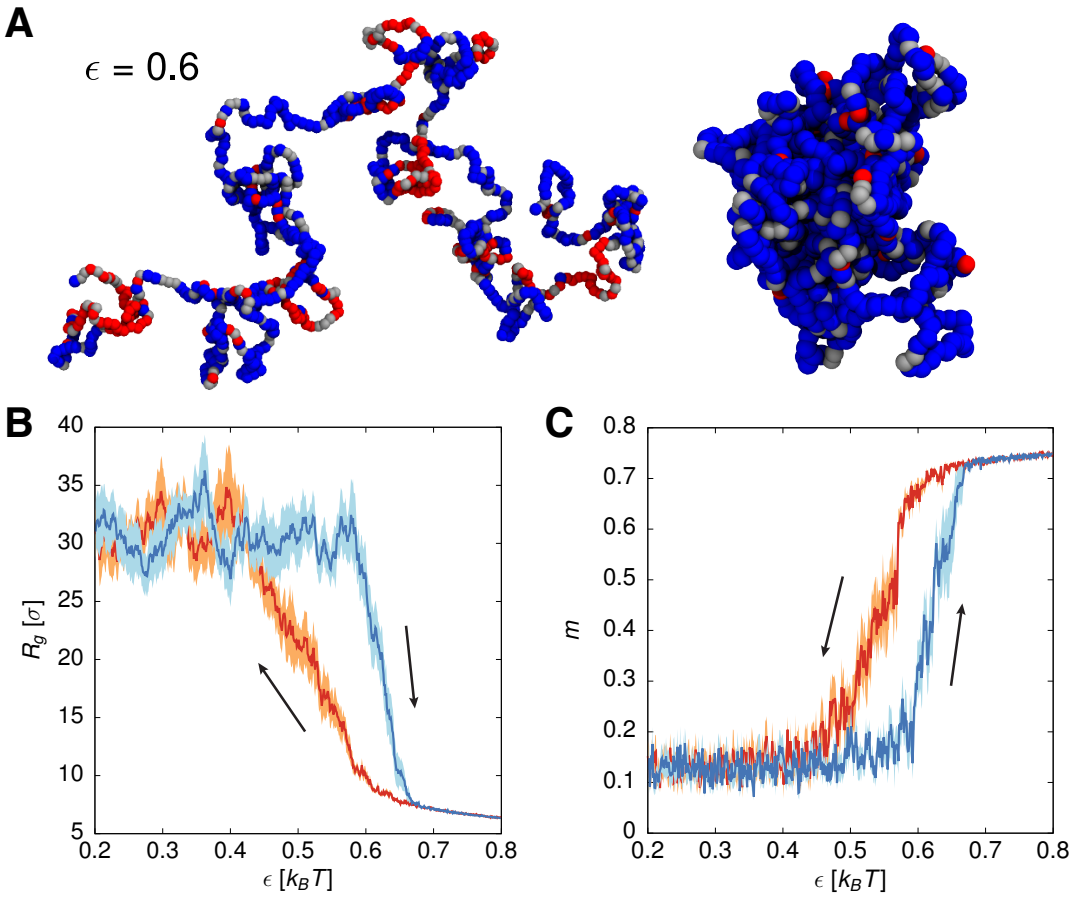

Figure 4.5: Observation of a hysteresis cycle in the SD-CO transition. Here, fixing $f = 2.0$, a chromatin fibre of $N = 1000$ beads is initialised in the CO phase at interaction energy $\epsilon = 0.8$. The energy is first decreased slowly to 0.2 (in steps of 10^{-3}) over a period of $1.5 \times 10^6\tau$ such that the system goes to the SD phase; it is then increased gradually back to 0.8 over the same duration for the system to return to the CO phase. (A) Simulation snapshots showing that, during this procedure, the system can either be in the SD or CO phase at $\epsilon = 0.6$ depending on its history. (B and C) Plots of the (B) radius of gyration R_g and (C) absolute magnetisation m of the system as ϵ varies between 0.2 and 0.8 indicate the presence of hysteresis in a small region $\epsilon \approx 0.58$–0.62. The curves in these plots are averaged over 10 simulations, with the shaded region representing the standard error of the mean.

The nature of transitions between other phases within the model is briefly outlined here. By performing a similar analysis to the one described above, the transition between the SD and CD phases is found to be smooth and continuous. Notably, there is no hysteresis cycle and coexistence is not observed. The order of other transitions can be inferred from previous work. The SO-CO transition can be mapped to the coil-to-globule transition (i.e., the Θ collapse) in polymer physics, which is known to be continuous [25, 26]. The SD-SO and CD-CO transitions only involve changes in the epigenetic order along the fibre without a change in 3D

structure, so they can be effectively described by the 1D recolouring model [105] that this work is based on, which suggests a more continuous-like transition.

Although this work focusses on the direct transition between the SD and CO phases, it is valid to ask whether indirect pathways connecting the two phases are biologically relevant. For instance, the system can move from the CO phase to the SD phase via the SO phase, thereby crossing continuous boundaries only. This pathway could be a model for processes involving a change of identity of a cell, such as reprogramming. Exploring these alternative pathways and their connections to biological processes would be an interesting avenue for future work.

4.3 Genomic Bookmarking (GBM)

Within the epigenetically ordered (CO and SO) phases, feedback between readers and writers allows a single histone mark to spread and dominate along the chromatin fibre (see, e.g., Fig. 4.3C). Though large-scale spreading of repressive marks is seen in the telomere position effect in yeast [139] and in position-effect variegation in *Drosophila* [140], it is more common to find a heterogeneous epigenetic pattern (i.e., domains of different epigenetic states) along chromatin, and the minimal model discussed above cannot account for this observation.

To address this problem and to secure faithful inheritance of epigenetic patterns, genomic bookmarking (GBM) is introduced to the model. This mechanism postulates that there are "bookmarking" proteins, typically transcription factors (TFs), that are expressed in a tissue-specific manner and can bind to specific DNA sites during interphase and mitosis [95, 123–126] (Fig. 4.6A). When bound to chromatin, these bookmarks are assumed to recruit appropriate read-write factors to spread a particular epigenetic mark. Thus, with multiple species of bookmarks bound to various sites along chromatin promoting the spreading of different marks, a specific (heterogeneous) epigenetic landscape can be established.

In the model, GBM is simulated by permanently fixing the epigenetic colour of some beads to describe the presence of chromatin-bound bookmarking proteins helping to spread a particular epigenetic mark[6]. Apart from being non-recolourable, these beads behave the same as other beads – i.e., they can associate with other like-colour beads and can influence the colour change of normal beads

[6]Note that the model does not distinguish between the binding sites of the bookmarking proteins and the proteins themselves, which are represented implicitly.

in their surrounding. In the following, bookmarks for the red state are coloured in orange, whereas those for the blue state are coloured in cyan. As an example, Fig. 4.6A depicts a bookmark, drawn as an orange square (e.g., posterior sex comb [PSC] – a TF), which binds to DNA and gathers read-write factors (e.g., Polycomb repressive complex 2 [PRC2]) to propagate the red mark (e.g., H3K27me3) to its neighbourhood.

It should be emphasised that the actual spreading of a colour along chromatin is driven by a local increase in density of that colour, even in the case with GBM. Reader-mediated bridging between like-colour beads increases the local concentration of a colour, leading to an increase in likelihood of a bead in the vicinity to switch to this colour. The choice of which colour dominates locally is determined by symmetry breaking (as red-red and blue-blue interactions have the same strength, though see Section 4.3.2 below). GBM biases the local concentration to favour a specific colour and enables it to proliferate.

4.3.1 Varying the Pattern and Density of Bookmarks

To explore how GBM affects the epigenetic and spatial configuration of the system, a fraction ϕ of beads along the chromatin fibre are designated as bookmarks, and three different bookmarking patterns are examined:

(i) *Clustered*: bookmarks are evenly spaced along the fibre, with the colour switching after every n_c consecutive bookmarks.

(ii) *Mixed*: same as clustered, but the colour alternates from one bookmark to the next (i.e., $n_c = 1$).

(iii) *Random*: bookmarks are assigned stochastically along the fibre while respecting the bookmarking density ϕ globally.

These GBM patterns are simulated using a chromatin fibre of $N = 1000$ beads, with the density of bookmarks set to $\phi = 0.1$ and $(f, \epsilon) = (2.0, 0.65)$. The polymer is initialised to a swollen-disordered (SD) configuration, and non-bookmarked regions are populated with an equal number of red, blue, and grey beads. Figs. 4.6B–D show results for these GBM patterns. Kymographs depicting the colour of each bead along the polymer over time indicate each pattern's ability to form epigenetic domains. The clustered pattern (Fig. 4.6B) yields a stable set of domains of alternating colours. In contrast, the mixed pattern (Fig. 4.6C)

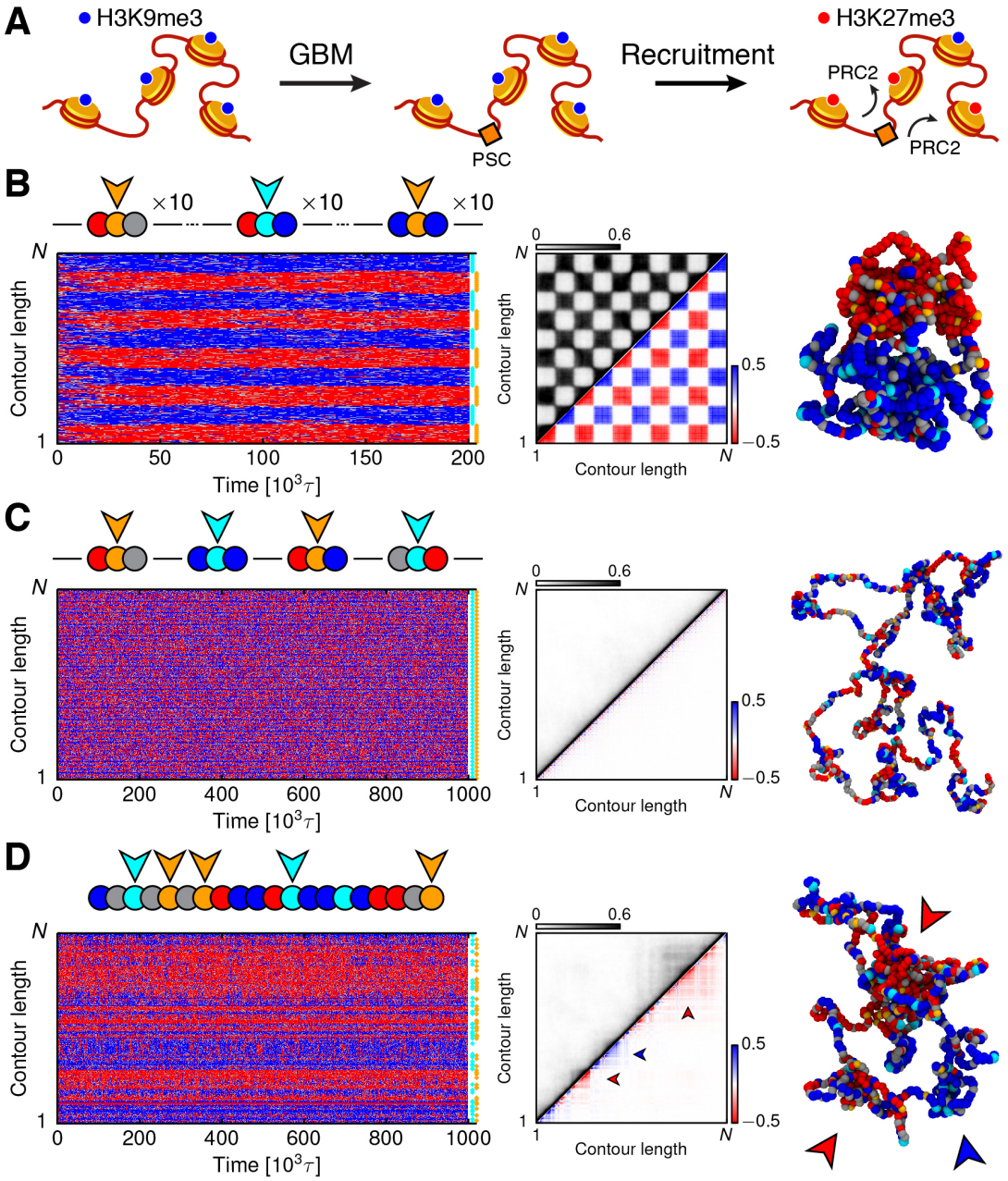

Figure 4.6: Simulating genomic bookmarking (GBM). (A) The model hypothesises that transcription factors (TFs) binding to particular DNA loci help recruit specific readers and writers to spread the respective histone marks. Here, the TF posterior sex comb (PSC) helps spread the H3K27me3 mark by recruiting the reader and writer Polycomb repressive complex 2 (PRC2). (B–D) Simulations considering different patterns of bookmarking. A chromatin fibre of $N = 1000$ beads is simulated starting from an SD configuration, with $(f, \epsilon) = (2.0, 0.65)$. GBM is modelled by fixing the colour of a fraction $\phi = 0.1$ of the beads (cyan and orange beads are bookmarks for blue and red beads, respectively). Three bookmarking patterns are considered: (B) clustered ($n_c = 10$), (C) mixed ($n_c = 1$), and (D) random (see illustrations at the top of each panel and the main text for details). (*Continued at the bottom of the next page.*)

discourages the precipitation of domains. Remarkably, the random pattern (Fig. 4.6D) manages to produce a few domains despite no apparent correlation between the positions of the bookmarks.

Snapshots and contact maps (i.e., heatmaps plotting the frequency of interaction between each pair of beads along the fibre) in the same figure reveal that these GBM patterns can alter the 3D chromatin organisation substantially. Without GBM, the parameters employed here normally drive the system into a collapsed phase[7]. Yet, for the mixed and random patterns, it is typical to observe an open or a partially condensed structure. In the random case, local collapses occur at regions where epigenetic domains are formed (see arrowheads in Fig. 4.6D), but unlike in equilibrium models, these domains are spatially demarcated and long-range interactions between domains of the same colour are rare. In contrast, the clustered pattern allows like-coloured domains to coalesce, eventually leading to phase separation[8] of the two epigenetic colours, as reflected by the chequerboard pattern shown in the contact map (Fig. 4.6B). This pattern is similar to that formed by A/B compartments in maps from Hi-C experiments [20, 22, 23]. It should be noted that the conformations described here are independent of the initial conditions of the polymer. Starting the system from a compact-disordered (CD) regime, which could be a model for folded mitotic chromosomes, gives rise to similar 1D epigenetic patterns and 3D polymeric arrangements when the system reaches a steady state (see Ref. [69]).

[7]This takes into account the fact that the transition line $\epsilon_c(f)$ between the SD and CO phases becomes lower than that shown in Fig. 4.3 as the system size N increases.

[8]The extent of this epigenetic phase separation (i.e., whether the system micro- or macro-phase separates) is studied more in detail using both a field theory approach and Brownian dynamics simulations in Refs. [141] and [142].

Figure 4.6 ***(continued)*****:** Recolouring rate is set to $k_{\mathrm{Re}} = 0.1\ \mathrm{s}^{-1}$ in (B) and $k_{\mathrm{Re}} = 10\ \mathrm{s}^{-1}$ in (C) and (D). *Left*: Kymographs showing the colour of each bead along the fibre over time. Positions of bookmarks are indicated at the right margin. *Middle*: Heatmaps showing the frequency of contact between each pair of beads. Here, two beads are said to be in contact if their spatial separation is less than $r_c = 5\sigma$. The upper triangle is a normal map showing the probability of contact between two beads. The lower triangle is an "epigenetically-weighted" map where each contact is weighted by the types of beads involved: $+1$ for blue-blue contacts, -1 for red-red contacts, and 0 otherwise. *Right*: A typical snapshot of the chromatin fibre for each bookmarking pattern. In the random case (D), GBM can give rise to locally-compacted structures (see arrowheads) without forming long-range contacts.

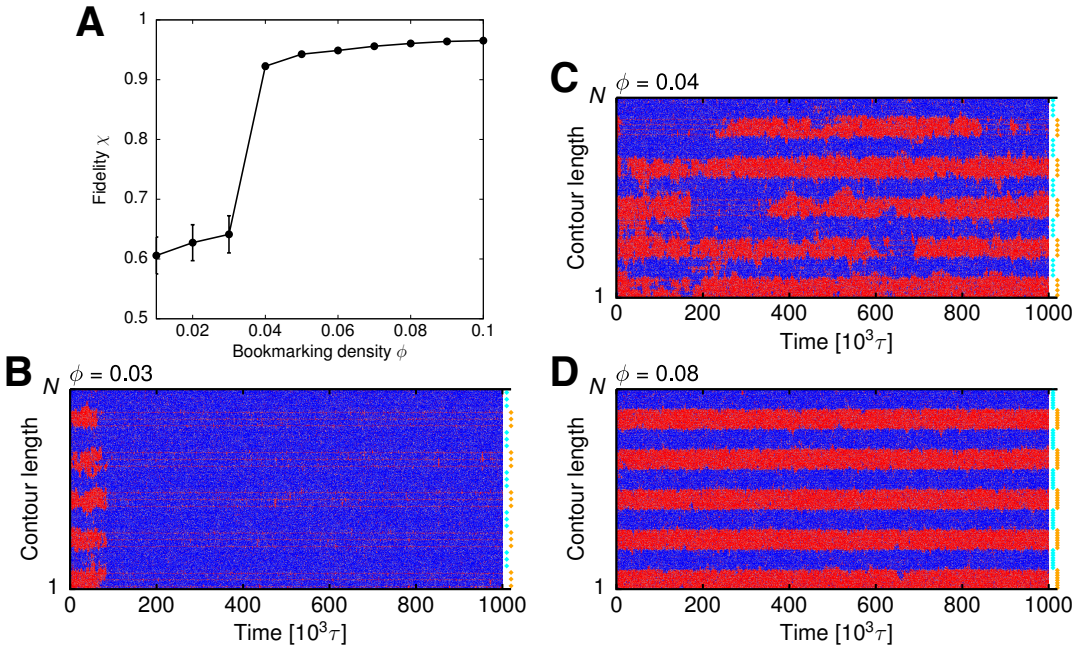

Figure 4.7: Varying the density ϕ of bookmarks bound to chromatin. Simulations are conducted in the same way as those with a clustered pattern of bookmarks in Fig. 4.6B. Each domain's size is fixed at 100 beads, and the number of bookmarks n_c within a domain is set according to ϕ. (A) A plot of the fidelity score χ, which reflects the efficiency of domain formation, as ϕ increases from 0.01 to 0.1. Error bars report the standard error of the mean from averaging over five simulation runs (these are too small to be seen for $\phi \geq 0.04$). Note that χ changes sharply from around 0.6 to near 1.0 at $\phi_c \simeq 0.04$. (B–D) Kymographs showing the epigenetic pattern of the system at (B) $\phi = 0.03$, (C) $\phi = 0.04$, and (D) $\phi = 0.08$. As before, the locations of bookmarks are indicated at the right margin.

Another way to change the GBM landscape is to vary the density ϕ of bookmarks bound to chromatin. Here, the critical density ϕ_c of bookmarks required to create stable epigenetic domains is investigated using an $N = 1000$ bead polymer. To enhance domain formation, the system is in the collapsed-ordered (CO) phase with $(f, \epsilon) = (2.0, 1.0)$. Bookmarks are assigned according to the clustered pattern, as it is the most effective strategy to generate domains among the three analysed. The density ϕ is systematically varied from 0.01 to 0.1. To simplify the comparison between results from different ϕ values, each domain's linear size is fixed at $n_d = 100$ beads (300 kbp) by adjusting the number of consecutive bookmarks n_c with the same colour accordingly. Note that this domain size is consistent with the size of a typical Hi-C domain [22, 23].

To assess the efficiency of domain formation, the probability of each bead being in the red state, $P_{\text{red}}(i)$ $(1 \leq i \leq N)$, is measured. In the ideal scenario where all domains are established (i.e., the colour of each bead matches that of its nearest

bookmark), this probability profile should conform to that of a square wave

$$\Pi(i) = \frac{1}{2}\left[\mathrm{sgn}\left[\sin\left(\frac{\pi i}{n_d}\right)\right] + 1\right], \tag{4.3}$$

where $\mathrm{sgn}(x)$ is the signum function. Based on this profile, the perfectness of domain formation is quantified by computing a "fidelity" score

$$\chi = 1 - \Delta^2, \tag{4.4}$$

where Δ^2 is the variance of the measured probability profile P_{red} from the ideal one $\Pi(i)$, i.e.,

$$\Delta^2 = \frac{1}{N}\sum_{i=1}^{N}\left[P_{\mathrm{red}}(i) - \Pi(i)\right]^2. \tag{4.5}$$

Thus, χ approaches 1 as the observed epigenetic pattern becomes more aligned with the ideal pattern.

Fig. 4.7A reports the fidelity score χ for different values of ϕ. Strikingly, the score jumps abruptly from around 0.6 to near 1.0 at $\phi_c \simeq 0.04$, signifying a transition between two phases: below ϕ_c, a single colour takes over most of the fibre (Fig. 4.7B), whereas above ϕ_c, domains alternating between red and blue states develop robustly along the fibre (Fig. 4.7D). Close to ϕ_c, domains form intermittently throughout the simulation (Fig. 4.7C). The value ϕ_c estimated here provides a prediction that around 1–10 in 400 nucleosomes[9] should be bookmarked in order to generate stable domains. Importantly, this threshold does not need to be attained throughout the genome; it is only necessary in cell-line-specific regions where it is essential to maintain domains of coherent epigenetic marks.

4.3.2 Asymmetric Interactions, DNA Replication, and Bookmark Excision

All simulation results presented thus far assume symmetric interactions between like-coloured beads (i.e., $\epsilon_{\mathrm{red}} = \epsilon_{\mathrm{blue}} = \epsilon$). Since different proteins regulate individual histone modifications, a more detailed model may incorporate asymmetric,

[9]This estimate is based on the fact that each bead corresponds to roughly 15 nucleosomes and not all nucleosomes within a bead have to be bound by a bookmark.

colour-specific interaction strengths. Consider the case where the blue-blue interaction is stronger than the red-red interaction ($\epsilon_{\text{blue}} > \epsilon_{\text{red}}$), and the system is within the CO phase. Without GBM, a randomly coloured chain will eventually succumb to the invasion by the blue mark. Nevertheless, by introducing enough red bookmarks locally, a red domain can still develop. Figs. 4.8A and B show an example of this scenario. Here, a chain of $N = 2000$ beads is simulated with $\epsilon_{\text{blue}} = 1.0$, $\epsilon_{\text{red}} = 0.65$, and $f = 2.0$. At the middle of the fibre, a region of 200 beads is seeded with red bookmarks (coloured orange here) at density $\phi = 0.1 > \phi_c$ according to the clustered pattern. Starting from an SD configuration with an equal number of red, blue, and grey beads in the non-bookmarked regions, the blue mark spreads to all parts of the fibre except the central segment, where the bookmarks allow a red domain to thrive and remain stable. This result suggests that the competition between a histone mark mediated by GBM and an antagonistic, more spreadable mark can also lead to domain formation.

GBM provides a mechanism to develop heterogeneous domains; it is of interest to understand the stability of these domains under extensive perturbations. To this end, simulations are performed in which half of the non-bookmarked beads are randomly recoloured at regular intervals, mimicking the lost of histone marks during semi-conservative replication of the chromatin fibre (Fig. 4.8C). Using the same set-up as above (i.e., $\epsilon_{\text{blue}} > \epsilon_{\text{red}}$ and a red-bookmarked central region), the system undergoes "replication" every $10^5\tau$. The kymograph in Fig. 4.8D shows that the red domain recovers quickly after each round of replication and remains robust against invasion by the blue mark throughout the simulation.

Recent experiments in *Drosophila* [128, 143] investigated the influence of Polycomb response elements (PREs; chromatin loci with which bookmarking proteins for Polycomb-related histone marks associate) on epigenetic memory by analysing the change in the activity of Polycomb-silenced genes under artificial insertion or deletion of PREs. Inspired by these experiments, a similar scenario is considered in the simulations where red bookmarks are randomly excised as the system undergoes replication (Fig. 4.8E). More precisely, a quarter of the initial number of bookmarks is deleted at each replication event until no bookmark remains. Fig. 4.8F reports how the epigenetic profile evolves over time as the population of bookmarks diminishes after successive replication events. The central red domain survives the first few replication cycles when the bookmarking density ϕ remains above the threshold ϕ_c. Notably, when ϕ drops below this value, the domain

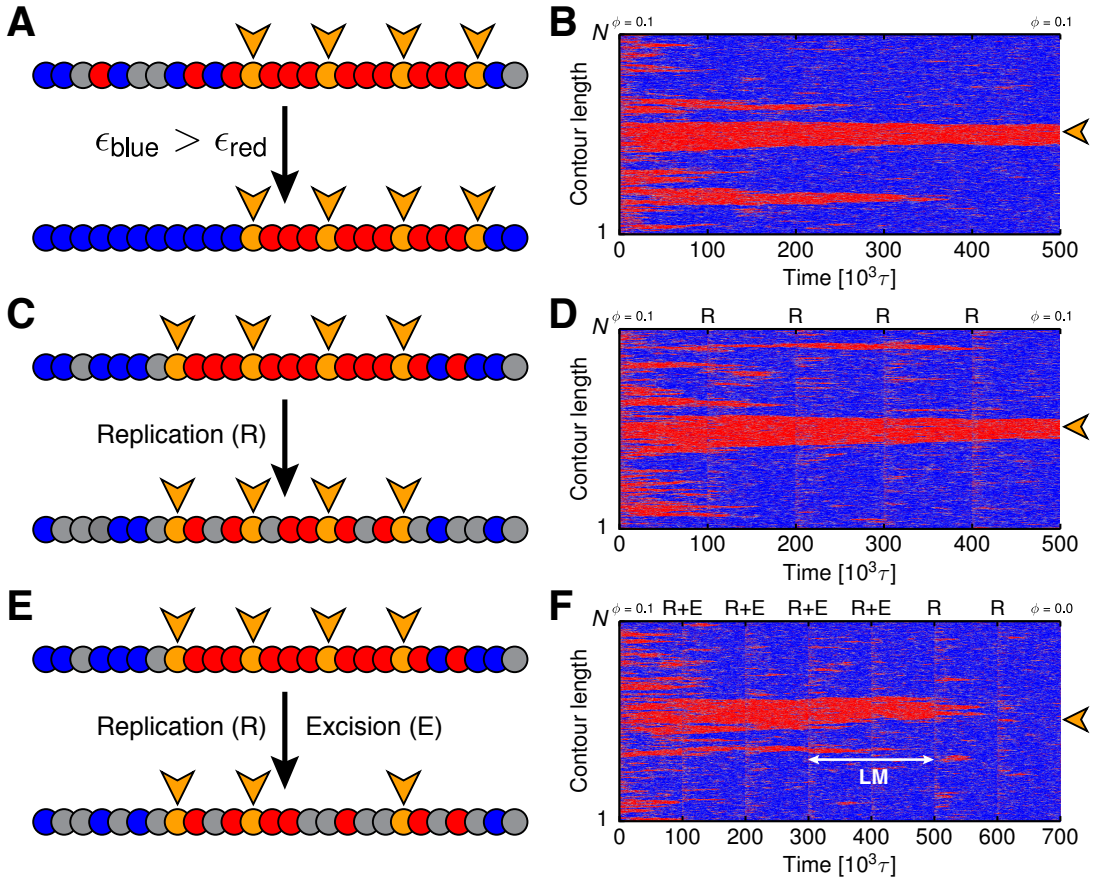

Figure 4.8: GBM simulations with asymmetric interactions, DNA replication, and bookmark excision. All simulations are done using a $N = 2000$ bead chain with $f = 2.0$. The interaction strengths between like-coloured beads are set asymmetrically, with $\epsilon_{\text{red}} = 0.65$ and $\epsilon_{\text{blue}} = 1.0$ for red-red and blue-blue interactions, respectively. The central 200-bead region of the chain is deposited with red bookmarks (orange beads) at density $\phi = 0.1 > \phi_c$ using the clustered pattern. (A and B) Illustrations and a kymograph of the epigenetic profile from a typical simulation of this case of asymmetric interactions and bookmarking at the central region (orange arrowhead in the kymograph). The blue mark takes over most parts of the fibre apart from the central segment, where bookmarks allow a red domain to be sustained. (C and D) Same as (A and B), but the chain undergoes semi-conservative replication (R) every $10^5\tau$, where half of the non-bookmarked beads are randomly recoloured. The red domain is maintained throughout the simulation despite the perturbations. (E and F) Same as (C and D), but immediately after each replication a quarter of the initial number of bookmarks (chosen randomly) are excised (R+E) until no bookmark remains. The red domain is inherited at least until $\phi < \phi_c$, after which local memory (LM) can prolong the domain's existence by a few more replication cycles.

does not disappear immediately but remains in place for a few more cycles by local memory (LM) [92, 106, 144]; the higher concentration of red marks and the positive read-write feedback prolong the domain's lifetime. These observations

are in agreement with experiments: regions enriched in H3K27me3 in *Drosophila* only fade away gradually after PRE excision [128].

It should be stressed that the findings reported in this section do not depend on the initial conditions. A chromatin fibre initialised in a compact-disordered (more mitotic-like) configuration gives rise to the same outcomes, further demonstrating the robustness of the results presented here (see Ref. [69]).

4.3.3 Predicting the Epigenetic Domains of a Full Chromosome in *Drosophila*

To test the GBM framework in a more realistic situation, the model is employed to simulate the whole right arm of chromosome 3 in *Drosophila* S2 cells at 3 kbp resolution (a total of 27.905 Mbp; $N = 9302$). The aim is to see whether the model can recapitulate the epigenetic patterns observed in experiments with minimal input data. In practice, the focus is on the profile of the Polycomb-related mark, H3K27me3, along the chromosome. It is known that posterior sex combs (PSCs), a candidate species of bookmarks, bind to PREs during interphase and mitosis and can recruit Polycomb repressive complex 2 (PRC2), which regulates the H3K27me3 mark [129]. Here, GBM simulations are conducted with bookmarks placed along the chromatin fibre at peaks identified from chromatin immunoprecipitation with sequencing (ChIP-seq) data for PSC [129]. Extending the three-colour model, non-bookmarked beads are coloured according to chromatin states from a hidden Markov model (HMM) [136]: beads corresponding to promoters/enhancers (state 1), transcriptionally active regions (states 2–4), and gene deserts (state 9) are coloured red, green, and grey, respectively, and they are non-recolourable during the simulation (Fig. 4.9A). Furthermore, there is an attractive interaction between promoter/enhancer beads to facilitate looping, and similarly between gene-desert beads to model compaction mediated by the linker histone H1 [137] (see Table 4.1 for a full list of parameters). The remaining beads of the polymer ($\sim$20%) are left unmarked (white) initially, and over time their states can be converted into that for heterochromatin (blue) or Polycomb (purple) by the recolouring mechanism of the model.

To quantify the model's ability to reproduce epigenetic domains, the probability of each bead being in the Polycomb state is measured. This profile can be viewed

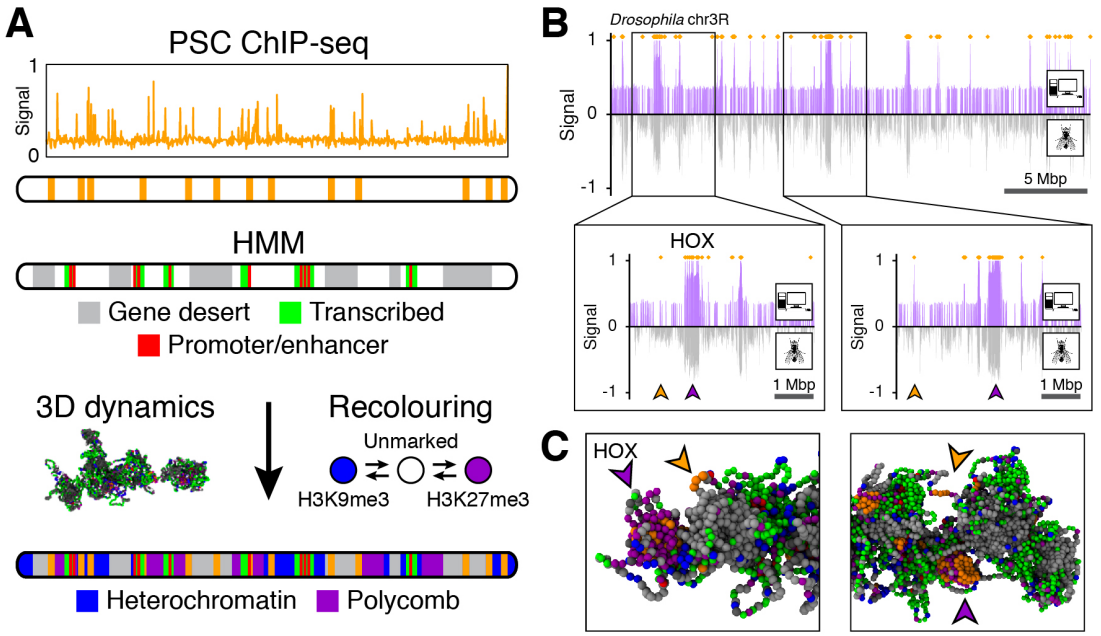

Figure 4.9: GBM simulations of chromosome 3R in *Drosophila* S2 cells. (A) Illustrations of the simulation setup. The chromosome is modelled at 3 kbp resolution corresponding to $N = 9302$ beads. Bookmarks (orange beads) are positioned according to the peaks from PSC ChIP-seq data [129]. Non-bookmarked beads are coloured using the chromatin states from a hidden Markov model (HMM) [136]. Beads covering promoters/enhancers, transcribed regions, and gene deserts are permanently coloured red, green, and grey, respectively. The remaining beads (∼20%) can switch between three colours during the simulation: unmarked (white), heterochromatin (blue), or Polycomb (purple). Preparation of PSC ChIP-seq and HMM data courtesy of D. Michieletto. (B) A comparison of the normalised H3K27me3 ChIP-seq profile from simulations (i.e., the probability of each bead being in the Polycomb state) with that from experiments [136]. Insets correspond to the regions 1–6 Mbp (*left*; covering the HOX gene cluster) and 10–15 Mbp (*right*). (C) Representative simulation snapshots of these two regions. Loci marked by the orange and purple arrowheads in (B) and (C) show that not all bookmarked beads lead to nucleation of Polycomb domains. The non-nucleating bookmarked segments (orange arrowheads) are far away from other bookmarked loci (purple arrowheads) in 3D space (despite being close together linearly) and thus are difficult to generate Polycomb domains.

as an *in silico* ChIP-seq signal for the H3K27me3 mark, and it is compared with that from *in vivo* data [136] (Fig. 4.9B). The two data sets are in good agreement (Pearson correlation coefficient $r = 0.46$; a random data set gives $r = 0.006$), further demonstrating that GBM provides a robust mechanism to generate epigenetic domains. It is interesting to note that regions annotated with bookmarks do not necessarily yield Polycomb domains (Figs. 4.9B and C). Bookmarks which have fewer spatial interactions with other bookmarks are less

Interaction	Energy $[k_BT]$
Polycomb ↔ Polycomb	0.60
Heterochromatin ↔ heterochromatin	0.72
PSC ↔ PSC	0.90
PSC ↔ Polycomb	0.90
Gene desert ↔ gene desert	0.60
Promoter/enhancer ↔ promoter/enhancer	3.00
Promoter/enhancer ↔ transcribed	0.60
Transcribed ↔ transcribed	0.60

Table 4.1: Energies for interactions between different bead types in the simulations for *Drosophila* S2 cells. These interactions are modelled by a truncated and shifted LJ potential. Other interactions not listed here are purely repulsive and are modelled by the WCA potential.

likely to cause domain formation. This observation once again highlights the importance of 3D chromatin organisation in epigenetic regulation.

4.4 Summary and Discussions

In this chapter, I have devised a modelling framework for the *de novo* formation and regulation of epigenetic modifications on chromatin. Two integral components of the framework are the positive feedback between multivalent epigenetic readers and writers and the dynamic coupling between the 3D structure of chromatin and its 1D epigenetic information. The framework is investigated by means of simulating chromatin as a recolourable bead-and-spring polymer. Each bead is assigned a colour corresponding to the local epigenetic state along chromatin. Chromatin bridging mediated by readers is modelled implicitly by incorporating an attraction between like-coloured beads. Writers are accounted for by allowing beads to change their colours over time according to non-equilibrium rules which mimic active processes in epigenetic regulation [105] (Fig. 4.2).

From varying the attraction strength between like-coloured beads and the feedback between readers and writers, there are four distinct phases in the model, characterised by the degree of polymer compaction and the order of the colouring (Fig. 4.3). Notably, at suitable feedback, an increase in attraction allows a single epigenetic colour to spread and take over the chromatin fibre, turning the fibre from a swollen-disordered (SD) to a collapsed-ordered (CO) state via a discontinuous transition (Fig. 4.4). Hysteresis accompanying this transition

provides a biophysical mechanism for epigenetic memory (Fig. 4.5). A chromatin fibre in the ordered state can recover from perturbations such as replication, where a lot of epigenetic marks are lost or replaced, by "reading off" the remaining marks on the fibre.

A major issue with this minimal model is its inability to produce a heterogeneous, multi-domain epigenetic pattern that is typical found within a chromosome. This is reconciled by introducing genomic bookmarking (GBM) to the model. This framework posits that there are bookmarking proteins, envisaged to be transcription factors (TFs), which can bind to sequence-specific loci and recruit readers and writers to spread certain epigenetic marks along the chromatin fibre.

GBM enables the model to nucleate distinct epigenetic domains in different ways. First, domains can form when there are two (or more) species of bookmarks bound to chromatin at density above a critical threshold (Fig. 4.7). The patterning of bookmarks not only affects the extent of domain formation; it also alters the 3D structure of chromatin (Fig. 4.6). Second, domains can also develop when there is a single type of bookmark competing against a more spreadable histone mark (Figs. 4.8A and B). This scenario is in line with the view that heterochromatin can spread extensively and can be stopped by actively transcribed regions (where TFs or bookmarks are bound).

Reassuringly, results from GBM simulations are consistent with experimental data in *Drosophila*. Simulating replication and bookmark excision events in the model produces trends similar to those observed in genome editing experiments [128] (Figs. 4.8C–F). The insertion of Polycomb response elements (PREs) provides binding sites for bookmarks of the H3K27me3 mark and causes domains of this mark to develop. In contrast, the deletion of PREs results in the gradual loss of these domains. The model also suggests that, despite the removal of bookmarks, domains can remain stable as long as the overall density of bookmarks is above the critical threshold. This is in agreement with work showing that Polycomb domains mediated by multiple PREs are less sensitive to the deletion of a subset of these elements [145].

To study GBM in a more realistic situation, simulations are done for the entire right arm of chromosome 3 in *Drosophila* S2 cells (Fig. 4.9). Utilising only the posterior sex combs (PSCs) ChIP-seq peaks as the binding sites for Polycomb bookmarks, these simulations can reproduce the profile of H3K27me3 signal from ChIP-seq experiments [136]. Of note, not all bookmarks give rise to domains;

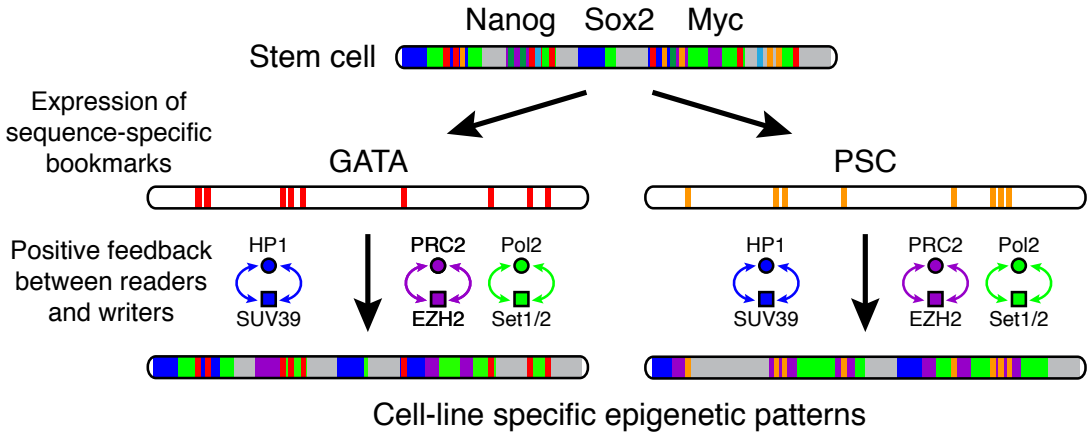

Figure 4.10: GBM and cell differentiation. The speculation here is that cell differentiation involves two stages. In the first stage, positional and environmental effects trigger the expression of cell-line-specific DNA binding factors (i.e., the bookmarks) such as GATA or PSC. In the second stage, these factors attract epigenetic readers and writers to establish epigenetic patterns tailored to each cell line via the positive feedback mechanism.

domain nucleation is dependent on the local 3D context of a bookmark, such as whether it is spatially proximate to other bookmarks. This result is compatible with research showing that 3D chromatin interactions have significant influence on epigenetic regulation [91].

The simulations conducted here offer a few predictions. First, they suggest that the deletion of a sufficient number of bookmarks (e.g., due to DNA mutation or artificial excision) can cause domains to disappear and disrupt a cell's ability to retain its epigenetic identity. Second, the expression of GBM proteins may provide a mechanism for cells to establish transient, *de novo* epigenetic domains in response to external cues [92, 93]. Third, the model predicts that GBM and the thermodynamics of epigenetic modifications are already sufficient for generating domains without the need for boundary elements, such as the CCCTC-binding factor (CTCF) and cohesin. Nevertheless, these elements may still influence domain patterning by modulating the 3D chromatin context [74].

Alongside these predictions, several questions are worth examining in the future. For instance, one may ask how GBM fits in the process of cell differentiation, where tissue cells with contrasting epigenetic patterns are derived from the same population of stem cells. It is possible that positional and environmental effects first stimulate expression of cell-line-specific bookmarks, such as GATA [125] and PSC [129], and these in turn gather readers and writers to create and maintain unique epigenetic patterns (Fig. 4.10). Similarly, one can consider the role of

GBM in reprogramming, where established epigenetic patterns become rewritten. It could be that reprogramming factors such as Nanog and Myc have the ability to override the activity of existing bookmarks and thus allow new patterns to be laid on chromatin.

It is also worth reminding that GBM is proposed here as a means to regulate the spreading of epigenetic information such that distinct domains can coexist. Yet, there may be other mechanisms which could achieve this as well. For example, the current model implicitly assumes that there is a saturating number of readers and writers such that they are always available to bind anywhere along chromatin. It could be that domains can arise naturally when these proteins are modelled explicitly with their copy numbers limited.

Overall, this work suggests that GBM is an effective mechanism for nucleating and sustaining heterogeneous epigenetic domains. GBM enables cells to robustly remember their domains after replication. Furthermore, careful modulation of GBM provides a way for cells to adjust their epigenetic landscape dynamically when adapting to physiological and environmental changes. Finally, GBM is a generic framework with only a few assumptions; it would therefore be interesting to test this framework in the future to more specific biological problems, especially those involving human cells.

5

Simulating Chromatin Reorganisation in Cellular Senescence

The three-dimensional (3D) folding of interphase chromosomes in eukaryotes is complex. Across different length scales, chromatin is often partitioned into separate structural units. In the previous chapter, it was seen that chromatin segments bearing different epigenetic marks tend to segregate to form individual domains, with interactions favoured between domains sharing the same mark over those with different marks. This phenomenon is reminiscent of the formation of active and inactive (A/B) compartments as observed in Hi-C [20, 22, 23]. Another more macroscopic segregation is the concentric, layered organisation of heterochromatin (HC) and euchromatin (EC). It is well known from microscopy that HC prefers forming a layer underneath the nuclear lamina (NL) at the periphery of the nucleus and around the nucleolus near the centre[1], whereas EC resides in the region between these two layers [146, 147]. Chromatin segments which show strong affinity with the NL are known as lamina-associated domains

[1]See Chapter 1 for a summary of the major components of the cell nucleus.

(LADs). Dam-ID experiments[2] have revealed that LADs cover more than one-third of the genome in human, and these domains, ranging from 0.1 to 10 Mbp, are typically heterochromatic and are associated with gene repression and late replication [147, 149–151]. Furthermore, these domains bind to the NL dynamically during interphase [151], and some of them also coincide with nucleolus-associated chromatin domains (NADs) [152].

Although this concentric layering of HC and EC and the formation of LADs are common across many organisms [146, 147], there are conditions in which this arrangement is disrupted, and two important pathological examples are that of cellular senescence and progeria (or Hutchinson-Gilford progeria syndrome [HGPS]; a premature ageing condition). Senescence is a state in which proliferative cells permanently exit from the cell cycle[3], and it can be triggered by different pathways, such as DNA damage and replicative stress [153]. A common way to induce senescence *in vitro* is via oncogene activation (i.e., oncogene-induced senescence [OIS]), under which the cell cycle is halted within days [154]. Cells in OIS experience a significant transformation in their nuclear architecture: there is a loss of LADs, and large HC-rich domains are formed in the nuclear interior known as senescence-associated heterochromatin foci (SAHFs) [154–156]. Moreover, senescent cells are typically associated with a loss of laminB1 [157], a protein which is abundant in the lamina, and there is evidence suggesting that nuclear pore density is higher in senescent cells compared to in healthy cells [158].

Progeria, the other example, is a rare condition caused by a mutation of the lamin A/C (*LMNA*) gene [159], which leads to changes in nuclear shape and clustering of nuclear pores [160, 161]. Similar to OIS, there is large-scale genome reorganisation in progeria, including a weakening of chromatin-lamina interactions. There are also differences between the two conditions. Notably, unlike senescent cells, progeroid cells do not develop SAHFs. Hi-C experiments have also shown that the networks of chromatin interactions are different between senescence and progeria [156, 162]. At a molecular level, there seems to be an overall reduction in the HC-associated epigenetic mark, histone 3 lysine 9 trimethylation (H3K9me3), in progeria but not in senescence [163]. Yet, other HC-related proteins, such as heterochromatin protein 1 (HP1) and core histone macroH2A, seem to contribute to SAHF formation and maintenance in senescence [155, 164].

[2]Dam-ID [148] is a technique which involves attaching the DAM protein to the loci of interest (in this case lamins) to molecularly annotate DNA segments that have been close to these loci.

[3]This state is different from quiescence, where cells are dormant but can become proliferative again in response to appropriate stimuli.

Although much work has been done on understanding the nuclear reorganisation observed in senescence and progeria, the underlying principles driving such rearrangement remain elusive; in particular, how the NL and HC contribute to this process is not well understood. It also remains unclear why SAHFs are formed in senescence but not in progeria, despite the weakening of chromatin association with the NL in both conditions. Moreover, there is a lack of modelling work on chromatin which takes into account of the NL. Previous simulation models of interphase chromosomes mostly consider intra- and inter-chromosomal interactions, and only recently have the effects of the NL on chromatin organisation and dynamics started to be examined [71, 165].

In this chapter, I construct a parsimonious model based on polymer physics principles to dissect the roles played by the NL and HC in the nuclear rearrangement observed in senescence and progeria. The model explicitly accounts for HC- and NL-mediated interactions based on known biophysical processes. By varying the strength of HC-HC and HC-NL interactions, the model yields distinct organisations that are qualitatively consistent with those found in growing (healthy), senescent, and progeroid cells. The simulated chromatin conformation for senescent cells is also in agreement with data from fluorescence microscopy experiments. Critically, this minimal model is able to quantitatively reproduce the change in the network of chromatin interactions in senescence and progeria as seen in Hi-C [156, 162]. It predicts that long-range interactions are promoted as cells become senescent and are diminished as they become progeroid. Moreover, the model recapitulates the stochasticity of LADs and their variation between cells as demonstrated in previous studies [150, 151].

Simulations also reveal that the transition between the growing and senescent conditions is rather abrupt and first-order-like. This result sheds light on a biophysical mechanism which could be important for the stability of senescence and could explain why it is difficult for senescent cells with SAHFs to re-enter the cell cycle. Furthermore, the simulated dynamics of LAD detachment from the NL during this transition suggest that this process is consistent with the kinetics of polymer desorption from a surface [166, 167].

The remaining sections of this chapter are organised as follows. In Section 5.1, I will introduce the polymer model employed here to simulate chromatin organisation in growing, senescent, and progeroid cells. In Section 5.2, I will show the phases observed within this model obtained from varying the HC and NL-mediated interactions, and how the individual phases can be mapped to nuclear

architectures found in these cell conditions. I will then present a quantitative analysis on the change in chromatin interactions between these phases. I will also discuss the heterogeneity of LADs found in different simulations. In Section 5.3, I will focus on characterising the transition in which cells turn from growing to senescent, and I will present results on the dynamics of chromatin detachment from the lamina. Lastly, in Section 5.4, I will summarise the results obtained in this chapter, discuss their implications, and suggest future research directions.

5.1 Simulation Model

The simulation model incorporates both chromatin and the NL. Using the framework discussed in Chapter 3, chromatin is coarse-grained as a flexible bead-and-spring chain of N beads. Each bead, with a diameter σ, contains 10 kbp of chromatin (roughly 50 nucleosomes) and is assigned one of two colours: red for EC-rich and blue for HC-rich segments. A bead is marked as HC if its corresponding genomic region shows enrichment in the chromatin immunoprecipitation with sequencing (ChIP-seq) signal for H3K9me3 [163] and/or LaminB1 [168] (data from the lung fibroblast cell line IMR90), and it is otherwise labelled as EC (Fig. 5.1A). The use of the LaminB1 signal to determine HC regions is motivated by the observation that LaminB1 shows strong correlation with LADs identified from Dam-ID experiments [168], and LADs are typically heterochromatic.

Rather than simulating the whole genome, I focus on a single chromosome to allow a more comprehensive sweep of the model parameters. Furthermore, nuclear structures relevant to the model, such as SAHFs, occur at a chromosomal or smaller level [169]. In the following, I perform simulations for human chromosome 20 (63.03 Mbp; $N = 6303$ beads). This chromosome has broad regions enriched in active or repressive epigenetic marks [163], and it shows moderate tendency to be close to the nuclear periphery [170]. The chromosome is simulated within a cubic box of linear size $L = 35\sigma$ in order to capture a realistic nuclear chromatin density. The box represents a subsection of the nuclear periphery; it is periodic in the x-y direction and is fixed in the z direction using a WCA repulsive wall governed by Eq. (3.8) (Figs. 5.1A and B).

The NL is modelled as a $\sim 1\sigma$ thick layer of beads ($N_L = 2000$) positioned randomly just underneath the top of the simulation box (Figs. 5.1A and B). These

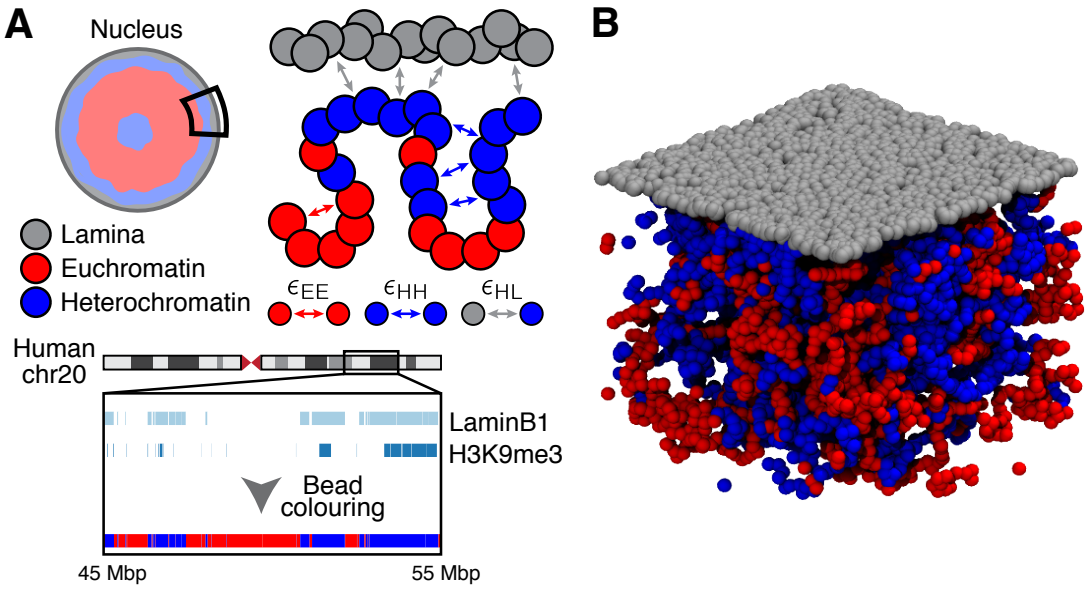

Figure 5.1: A simulation model for lamina-mediated chromosome organisation. (A) Human chromosome 20 is simulated as a flexible bead-and-spring chain within a subsection of the nuclear periphery. Chromatin beads are coloured either red or blue representing euchromatin (EC) or heterochromatin (HC), respectively. Beads are marked as HC if the corresponding chromatin segments are enriched in H3K9me3 and/or LaminB1 ChIP-seq signal (data preparation courtesy of N. A. Robertson). Beads within the centromeric region (26.4–29.4 Mbp) are also labelled as HC. The nuclear lamina (NL) is simulated as a layer of static beads (grey) located at the top of the simulation box. EC and HC beads can attract to beads of the same kind with energy ϵ_{EE} and ϵ_{HH}, respectively. HC beads can also attract to NL beads with energy ϵ_{HL}. (B) A representative snapshot of the model when the HC-HC and HC-NL interactions are weak.

beads represent lamins and other lamina-associated proteins that constitute the NL. Note that it is also possible to simulate the NL as a smooth wall as done previously [165], but the approach considered here can more accurately portray the one-to-one nature of the interactions between NL proteins and chromatin. NL beads remain static in the simulations, as it is reasonable to assume that the dynamics of NL proteins are much slower and more restricted than those of chromatin.

In the model, chromatin beads can interact with each other and with the NL beads. Previous studies have shown that various proteins mediate the interaction between HC and the NL [147, 149, 171–173]. For instance, the lamin B receptor (LBR) can tether HC segments to the lamina and can associate with HP1 [171, 174]. To account for these observations and motivated by the fact that LADs are mostly repressive, HC beads are attracted to NL beads within the simulations.

This interaction is governed by the truncated and shifted LJ potential [Eq. (3.9)] with energy ϵ_{HL} (in units of k_BT) and a cutoff $r_c = 1.8\sigma$.

HC segments are highly condensed within the nucleus. It has been thought that HP1, which can dimerise and associate with HC, facilitates such compaction [175]. To model bridging between HC segments driven by HP1 and other repressive factors, an attraction is imposed between HC beads, described by the same potential as for the HC-NL interaction, but with energy ϵ_{HH}. This procedure is consistent with the work in Chapter 4 and with previous studies which model chromatin as a block copolymer [46, 67, 69]. Typically when ϵ_{HH} is sufficiently strong, HC segments will cluster and form globules. In addition, the model incorporates a weak attraction between EC beads ($\epsilon_{\mathrm{EE}} = 0.2$) to account for promoter-enhancer interactions and cohesin-mediated looping [55, 75]. There is also no direct attraction between EC and HC beads; this is because the attractions imposed here implicitly describe the action of multivalent proteins bridging chromatin segments with similar epigenetic signature, and these proteins rarely connect those with different marks (i.e., bridging HC and EC).

The system is evolved by performing Brownian dynamics simulations as discussed in Section 3.1.4, and it is equilibrated in the following way prior to the main simulation. The chromatin fibre is first generated as a random walk in a larger box ($L = 100\sigma$) with fixed boundary conditions and the lamina absent. The fibre is allowed to equilibrate for $10^4\tau$, during which beads interact with each other via steric repulsion while ensuring chain connectivity. Within the initial $6 \times 10^3\tau$, the soft potential [Eq. (3.11)] is employed to remove overlaps in the polymer such that it becomes self-avoiding, and the height of the potential is gradually increased from 0 to $200k_BT$. The WCA potential is used for the remaining part of this equilibration period. The simulation box is then compressed slowly to $L = 35\sigma$ using indented walls in the next $5 \times 10^3\tau$. Lamina beads are then generated at the top of the box. Lastly, the chromatin chain is allowed to equilibrate with the lamina (via steric repulsion) for another $5 \times 10^3\tau$. The boundary conditions during this period are identical to those in the main simulation.

It is of interest to ascertain the physical size σ of a coarse-grained chromatin bead. This is estimated by comparing the size of HC globules found in simulations (in the parameter space mapped to the senescent condition; see the next section) with that of SAHFs found in florescence microscopy. In simulations, the radial distribution of HC beads in a globule is computed (see, e.g., Fig. 5.3E), and the globule's radius r_{HC} is estimated based on the inflection point of this distribution.

This radius is found to be $r_{\rm HC} = 9.21 \pm 0.09\sigma$ from averaging over 10 simulations. In experiments, the size of a typical SAHF is estimated from images staining H3K9me3, which is strongly enriched within SAHFs. SAHFs in focus are identified using ImageJ [176] with their areas $A_{\rm SAHF}$ determined. Averaging over 12 SAHFs gives a mean radius of $r_{\rm SAHF} = (A_{\rm SAHF}/\pi)^{1/2} = 0.64 \pm 0.04$ μm. Setting $r_{\rm HC} = r_{\rm SAHF}$ suggests that the size of each bead is $\sigma = 70 \pm 5$ nm.

5.2 Chromatin Structures in Growing, Senescent, and Progeroid Cells

To understand how alterations in HC and NL-mediated interactions contribute to chromatin reorganisation when growing cells turn senescent or progeroid, I focus on varying the parameters $\epsilon_{\rm HH}$ and $\epsilon_{\rm HL}$ while keeping $\epsilon_{\rm EE}$ fixed. Two main properties of the chromatin fibre that are affected by these parameters are the degree of chromatin adsorption to the lamina and the local compactness of the fibre. The former can be quantified by considering the distance $\overline{z}$ between the centre of mass of the chromatin fibre and the NL

$$\overline{z} = \left\langle \frac{1}{N} \sum_{i=1}^{N} z_i \right\rangle, \tag{5.1}$$

where z_i is the distance of bead i from the NL (note that all beads have the same mass), and the average $\langle \cdots \rangle$ is taken over time and simulation runs. The latter can be captured by the local density

$$\rho = \left\langle \frac{3}{4\pi N r_s^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(r_s - r_{ij}) \right\rangle, \tag{5.2}$$

where r_s is a cutoff threshold, r_{ij} is the distance between beads i and j, and $\Theta(x)$ is the Heaviside step function (i.e., $\Theta(x) = 1$ if $x > 0$ and zero otherwise). This quantity essentially measures the number of beads that are within a sphere of radius r_s centred at each bead i; thus, a more compact fibre, with more beads next to each other, has a higher ρ. The threshold r_s is set to $5\sigma \approx 350$ nm, and choosing other reasonable values (e.g., 3σ and 7σ) does not alter significantly the results presented below.

5.2.1 Model Phases

Exploring the parameter space $(\epsilon_{\mathrm{HH}}, \epsilon_{\mathrm{HL}})$ reveals that there are four phases within the model, corresponding to the combination of an adsorbed or desorbed chromatin fibre (i.e., low or high $\overline{z}$) and an extended or collapsed organisation (i.e., low or high ρ; Figs. 5.2A–C). Strikingly, three of the four phases display morphologies markedly similar to those observed in growing, senescent, and progeroid nuclei (Fig. 5.2D). At low $\overline{z}$, the adsorbed-collapsed (AC) phase ($\rho > 0$) and adsorbed-extended (AE) phase ($\rho \simeq 0$) have arrangements similar to those of growing cells, where a layer of HC harbours the NL. Yet, the AE phase shows a high degree of intermixing between EC and HC segments, which is uncommon in conventional mammalian nuclei [163, 177]. As a result, the AC phase is identified as the closest representation of the growing state. At high $\overline{z}$, the desorbed-collapsed (DC) phase ($\rho > 0$) displays organisations reminiscent to those of senescent cells. Specifically, there are large HC globules surrounded by a corona of EC, akin to SAHFs found in OIS cells [156, 163]. This similarity is also evident from comparing simulated structures from this phase with images of senescent nuclei from fluorescence microscopy (Fig. 5.3). Also at high $\overline{z}$, the desorbed-extended (DE) phase ($\rho \simeq 0$) exhibits features comparable to those of progeroid cells, such as the loss of peripheral HC [160, 161] and the large-scale intermingling of active and inactive chromatin regions [162].

The association of the DC phase, which has strong HC-HC and weak HC-NL interactions, with senescence is consistent with work showing upregulation of HP1 in this condition [155]. Mass spectrometry results presented in Ref. [72] also demonstrate that HP1 and macroH2A expression are upregulated in OIS cells, and these entities have been implicated in forming SAHFs and driving heterochromatin compaction [164]. Overall, despite its simplicity with few parameters, the model is able to recapitulate the major attributes of the nuclear arrangements in various cell states. This leads one to believe that HC-HC and HC-NL interactions, the two main ingredients of the model, must be important in driving chromatin folding in these cell states and the nuclear rearrangement between physiological and pathological conditions.

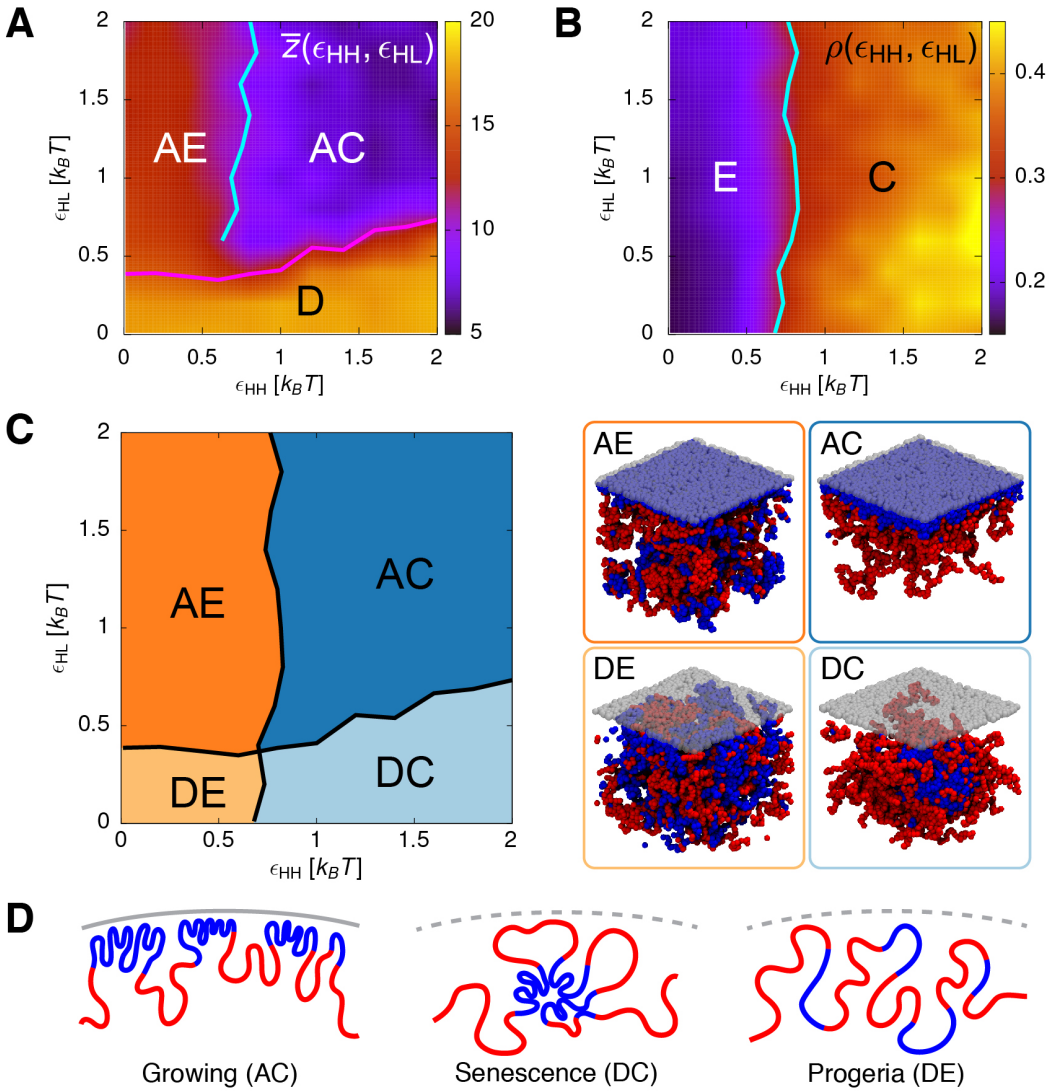

Figure 5.2: Model phases from varying HC-HC and HC-NL interactions. (A) A heatmap showing the degree of chromatin adsorption to the NL, quantified by the distance $\bar{z}$ of the chromosome's centre of mass from the NL. Across the parameter space $(\epsilon_{HH}, \epsilon_{HL})$, 10 simulations are performed at every 0.2 interval in both ϵ_{HH} and ϵ_{HL} directions. The pink line shows the boundary between the adsorbed and desorbed phases, and it is determined based on the inflection point of a tanh curve fitted to $\bar{z}(\epsilon_{HL})$ at different ϵ_{HH}. As a compact fibre tends to be closer to the NL in the adsorbed regime, $\bar{z}$ also partially captures the collapsed transition, as reported by the cyan line (computed by fitting a tanh curve as above). (B) A heatmap showing the degree of compaction of the chromatin fibre, quantified by the local density ρ. The cyan line separates the extended and collapsed phases and is estimated using the same fitting method described in (A). Its location is in agreement with that calculated from $\bar{z}$. (C) A full phase diagram combining the results from (A) and (B) and showing the four phases of the model: adsorbed-extended (AE), adsorbed-collapsed (AC), desorbed-extended (DE), and desorbed-collapsed (DC). (D) Illustrations showing typical chromatin structures in growing, senescent, and progeroid cells.

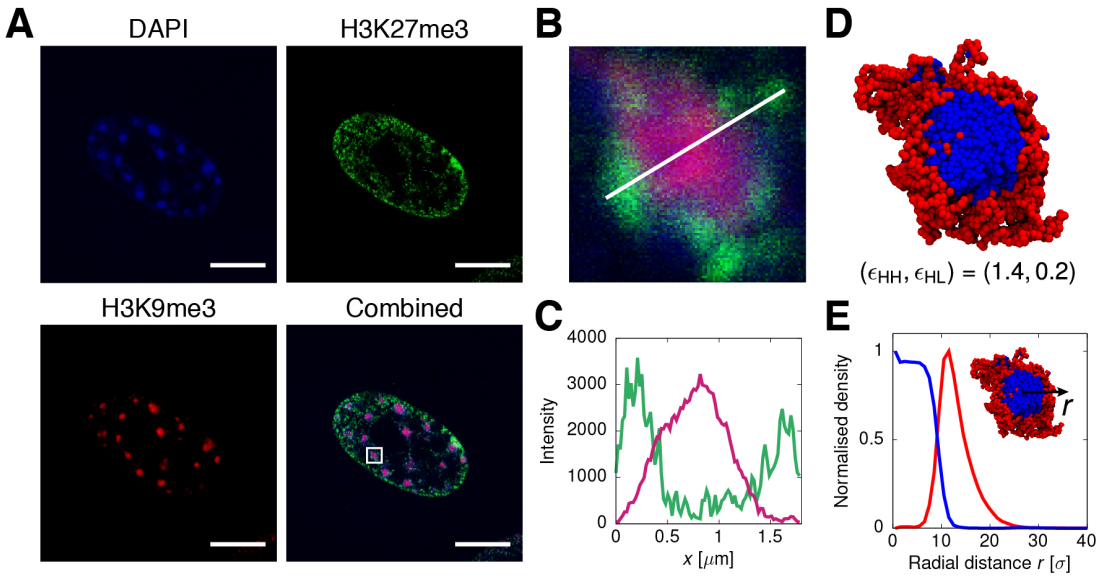

Figure 5.3: Comparing chromatin structures in senescence with those in the DC phase in simulations. (A) Confocal images of senescent nuclei with DAPI, H3K27me3, and H3K9me3 fluorescence staining. Scale bars represent 5 μm. Imaging data courtesy of N. Rattanavirotkul. (B) An enlarged view of a SAHF corresponding to the white box in the combined image in (A). (C) The intensity profiles for H3K9me3 (magenta) and H3K27me3 (green) along the white line in (B). (D) A cross-sectional view of a simulation snapshot representative of the DC phase. (E) Time-averaged radial distribution profiles of HC and EC as a function of the distance r from the centre of the globule.

5.2.2 Locality of Chromatin Interactions

Having mapped individual phases to specific cell states, I employ the model to further examine the change in chromatin interactions, or contacts, between different phases. Previous Hi-C experiments have revealed that the network of chromatin contacts in the growing phase differs substantially from those in the senescent and progeroid phases [156, 162]. A key observation is that distal contacts are enhanced in senescent cells compared to in growing cells, and the reason behind this change remains unclear. The aim here is to use simulations to understand quantitatively and mechanistically how the contact network changes between the growing (AC), senescent (DC), and progeroid (DE) phases.

To compare with Hi-C results, contact maps are constructed from the simulated structures of each phase (Fig. 5.4). These maps report the frequency (or probability after suitable normalisation) of pairwise contact between different segments along the chromosome. The simulated maps indicate that there is an increase in long-range interactions in the senescent phase relative to the growing phase. In the progeroid phase, however, distal interactions are depleted, and

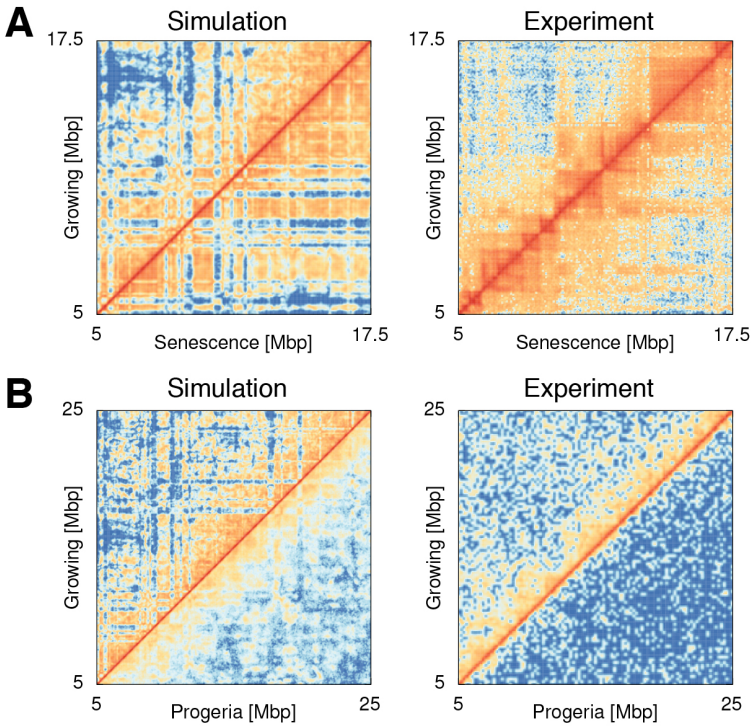

Figure 5.4: Contact maps showing chromatin interaction networks in different cell conditions. (A) Heatmaps comparing chromatin contact frequencies between growing and senescent phases in simulations and in Hi-C experiments. (B) Similar to (A), but comparing between growing and progeroid phases. Contact frequencies are shown in log scale to facilitate comparison. Contact maps are generated in the following way. In simulations, parameters are set to $(\epsilon_{\mathrm{HH}}, \epsilon_{\mathrm{HL}}) = (1.0, 1.6)$ for growing, $(1.4, 0.2)$ for senescence, and $(0.2, 0.2)$ for progeria. Two beads are considered to be in contact if their separation is less than 3σ, and the probability of contact between two beads is calculated from counting their frequency of contact over a time period of $5 \times 10^4\tau$ in each of 20 simulations. In experiments, Hi-C data from Ref. [156] are used for the comparison between growing and senescence. Maps at 50 kbp resolution are generated directly from raw sequencing data using the HiC Pro pipeline (version 2.10.0) [178]. Data from Ref. [162] are used for the comparison between growing and progeria. Maps at 200 kbp resolution are produced from the valid pairs reported in the reference, using the Age Control sample for growing and the HGPS-p19 sample for progeria. All experimental data are aligned to the GRCh37 reference genome (equivalent to the hg19 genome), and maps are normalised using the iterative corrective procedure [179].

there is a general loss of chromatin domains. These results are in line with those found in Hi-C contact maps [156, 162].

The change in the network of chromatin interactions between cell states can be quantified by the open chromatin index (OCI) as defined in Ref. [156] (Fig. 5.5A). To compute the OCI, one considers the normalised local contact signal $\mathcal{C}_\ell$ and

distal contact signal $\mathcal{C}_d$ for each chromatin bin (say bin i) in a contact map:

$$\mathcal{C}_\ell(i) = \frac{1}{N_\ell(i)} \sum_{j=1}^{N} c_{ij}\Theta(s_d - s_{ij}) \tag{5.3}$$

$$\mathcal{C}_d(i) = \frac{1}{N_d(i)} \sum_{j=1}^{N} c_{ij}\Theta(s_{ij} - s_d), \tag{5.4}$$

where c_{ij} is the contact strength between chromatin segments in bins i and j, and s_{ij} is their genomic separation. s_d is a threshold which distinguishes distal from local contacts and is set to 2 Mbp (which is close to the upper limit of the size of a topologically associating domain [TAD] [180]). $N_\ell(i)$ and $N_d(i)$ are the number of possible local and distal contact pairs, respectively, for bin i. The OCI is defined as the logarithm of the ratio of distal to local signal:

$$\mathrm{OCI}(i) = \log_2\left(\frac{\mathcal{C}_d(i)}{\mathcal{C}_\ell(i)}\right). \tag{5.5}$$

This score becomes more positive (negative) when the distal contact signal increases (decreases) with respect to the local contact signal.

Fig. 5.5 reports the OCI scores computed for the growing, senescent, and progeroid phases. Both in simulations and experiments, there is a noticeable change in the OCI (ΔOCI) when comparing the growing phase with the senescent and progeroid phases. First, the rewiring of chromatin interactions between growing and senescence is characterised by an overall positive ΔOCI (Fig. 5.5B). This upshift in the OCI is statistically significant in simulations (two-sample Kolmogorov-Smirnov test[4]: $D = 0.25$, $p < 10^{-4}$; a higher D indicates a larger separation between the samples) and in experiments ($D = 0.50$, $p < 10^{-4}$; excluding inter-chromosomal interactions). The positive ΔOCI score indicates that chromatin interactions become more distal in senescence, in line with the qualitative observations from the contact maps. Furthermore, the agreement between simulations and experiments is quantified by the Pearson correlation coefficient of the OCI values in each phase. Specifically, the correlation score is $r = 0.53$ ($p < 10^{-4}$) in growing and $r = 0.63$ ($p < 10^{-4}$) in senescence.

Second, the alteration in the chromatin network from growing to progeria gives a negative ΔOCI, reflecting a loss of long-range interactions (Fig. 5.5C). As above,

[4]The two-sample Kolmogorov-Smirnov test is a non-parametric statistical test which reports how different are the sample distributions between two sets of data. A larger D statistic indicates that the two sets are drawn from more separated distributions [181].

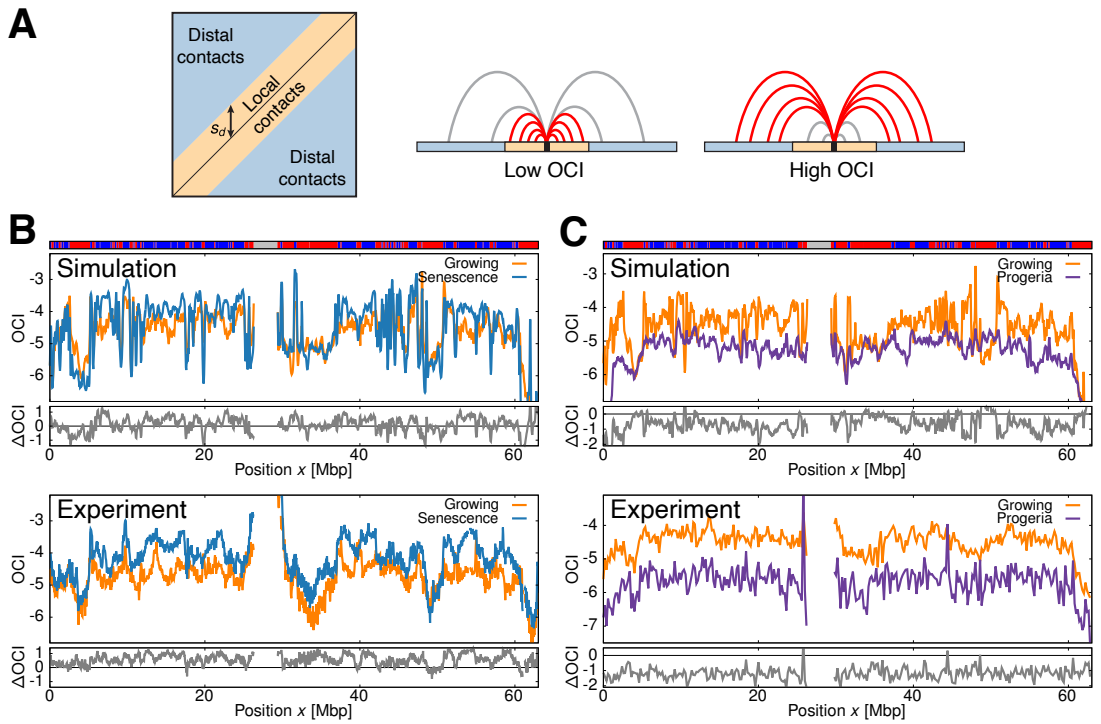

Figure 5.5: The open chromatin index (OCI) along the chromosome. (A) Illustrations explaining the OCI statistic, which measures the ratio of the distal contact strength to local contact strength along each bin in the contact map. s_d is the threshold above which contacts are defined as distal. A lower OCI indicates more local contacts relative to distal contacts. (B and C) The OCI value of each bin in the contact map for (B) growing and senescence and (C) growing and progeria, both in simulations and experiments. The difference in the OCI (ΔOCI) between cell states is reported at the bottom. The top track shows the chromatin state of each bin along the chromosome – i.e., red for EC, blue for HC, and grey for the centromeric region.

this change in the OCI is statistically significant, both in simulations ($D = 0.60$, $p < 10^{-4}$) and in experiments ($D = 0.89$, $p < 10^{-4}$). Additionally, the correlation in the OCI scores between simulations and experiments is $r = 0.63$ ($p < 10^{-4}$) in growing and $r = 0.48$ ($p < 10^{-4}$) in progeria. It should be noted that the trends reported here regarding the change in the OCI, both between growing and senescence and between growing and progeria, are unaffected by variations in the threshold s_d, as demonstrated by plotting the bin-averaged OCI as a function of s_d (Fig. 5.6).

An intriguing observation from the OCI scores is that the locality of chromatin interactions changes in the opposite direction in senescence and progeria, even though both phases suffer from a loss of chromatin-lamina interactions. A possible explanation is that large-scale HC bodies are formed in senescence but not in progeria, resulting in different contact networks. Motivated by this

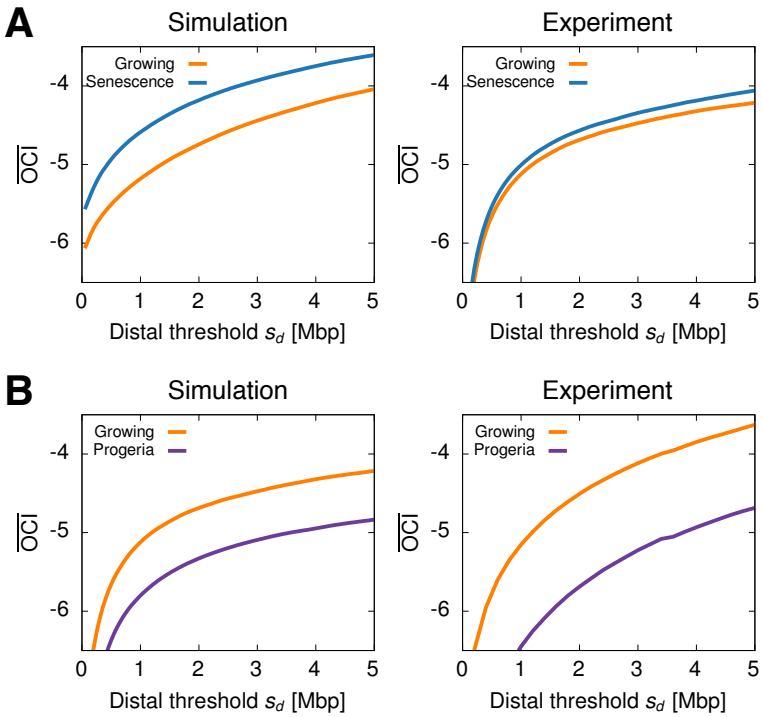

Figure 5.6: The chromosome-averaged OCI score ($\overline{\text{OCI}}$) as a function of the distal contact threshold s_d. (A) $\overline{\text{OCI}}$ in growing versus senescence. (B) $\overline{\text{OCI}}$ in growing versus progeria.

hypothesis, scatterplots are constructed displaying the OCI scores of individual chromatin bins, coloured by their states (EC or HC), to visualise whether there are differences between EC and HC segments in terms of the changes in their contact networks between cell states (Fig. 5.7). These plots indicate that HC segments generally experience a larger change in the OCI compared to EC segments (i.e., their points are further away from the diagonal) between growing and senescence. However, in the growing-progeria transition, the difference between these segments regarding the change in the OCI is less pronounced. These results are consistent with previous work showing that GC-poor chromatin regions tend to suffer a greater change in their contact network between growing and senescence [156]. They also suggest that SAHFs, the HC-rich bodies found in senescence but not in progeria, may be important in driving the formation of long-range chromatin interactions during nuclear reorganisation, possibly by allowing (polymer-polymer) phase separation between EC and HC [42, 43, 68].

One can further understand the opposite changes in distal interactions in senescent and progeroid cells by examining the decay in the contact probability $P_c(s)$ between two chromatin segments separated by genomic distance s. This decay curve can be computed from the contact maps, and theories from polymer

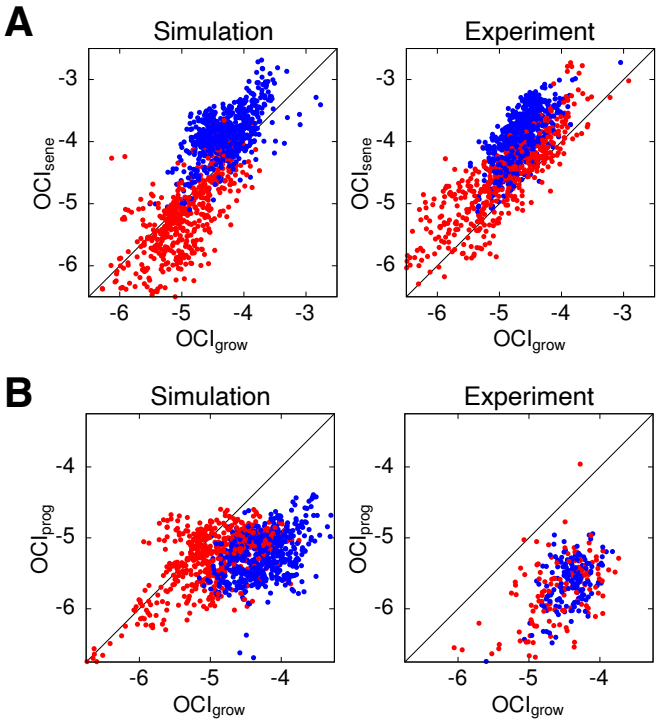

Figure 5.7: Scatterplots showing the OCI scores in different conditions for each chromatin bin, coloured according to its chromatin state (red for EC and blue for HC; the centromeric region is treated as HC). (A) OCI scores in growing versus senescence. (B) OCI scores in growing versus progeria. The diagonal line in each plot represents the case where there is no change in the OCI between conditions.

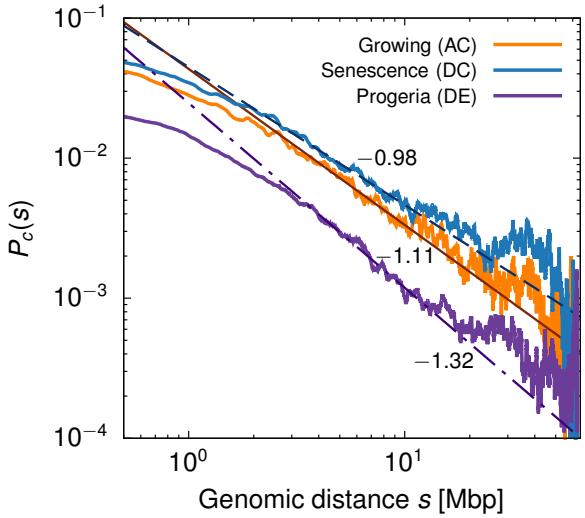

Figure 5.8: Contact probability $P_c(s)$ as a function of the genomic distance s between two chromatin segments for the simulated conformations in the growing (AC), senescent (DC), and progeroid (DE) phases. The plot is in log scale, and the contact exponent α for each phase is determined by performing a linear fit in the genomic region between 300 kbp and 1 Mbp.

physics predict that it should take the form of $P_c(s) \sim s^{-\alpha}$, where α is the contact exponent (see Section 2.1.1). Importantly, different polymeric conformations (e.g., random walk or globule-like) are characterised by different values for this exponent [26, 27, 29]. Given that the growing (AC), senescent (DC), and progeroid (DE) phases exhibit distinct polymeric organisations, one anticipates that the contact exponent α to be different between these phases. Indeed, plotting $P_c(s)$ for these phases and measuring the exponent confirm this prediction, and it is found that $\alpha_{\mathrm{DC}} < \alpha_{\mathrm{AC}} < \alpha_{\mathrm{DE}}$ (Fig. 5.8). This relation is in agreement with the changes in distal interactions reported above: since $P_c(s)$ is normalised, a smaller α implies a shallower decay in the probability and also a shift from local to non-local interactions. With $\alpha_{\mathrm{DC}} < \alpha_{\mathrm{AC}}$, $P_c(s)$ decays more gradually in the senescent phase compared to the growing phase, and thus there is an increase in likelihood of non-local contacts. In contrast, $P_c(s)$ declines more steeply in the progeroid phase as $\alpha_{\mathrm{AC}} < \alpha_{\mathrm{DE}}$, so distal contacts are less favoured. Because the simulation model focusses on the polymeric nature of chromatin, the analysis here suggests that the alteration in the chromatin interaction network between different phases is largely driven by a change in the physical folding of the chromatin fibre.

5.2.3 Cell-to-Cell Variability of Lamina-Associated Domains

Apart from examining the interaction patterns in senescence and progeria, one can use the simulation model to investigate chromatin association with the nuclear lamina (NL) in healthy, growing cells. The lamina is known to play a major role in regulating chromatin organisation and function [147]. Experiments have shown that lamina-associated domains (LADs) are highly variable between cells, and they are not conserved from mother to daughter cells [150, 151]. Remarkably, simulations within the growing (AC) phase recapitulates this stochasticity in LADs. In particular, from measuring the vertical position of each chromatin bead within the simulation box, one finds marked heterogeneity in the domains of beads adsorbed to the NL across different simulation runs, which can be associated with individual cells (Fig. 5.9A). This variability can be attributed to the stochastic nature of the Brownian dynamics scheme employed in the simulations. Nevertheless, a physical factor underpinning this phenomenon is the limited amount of surface area available at the NL with which chromatin can interact. Hence, although all HC beads have affinity with the NL, only some of

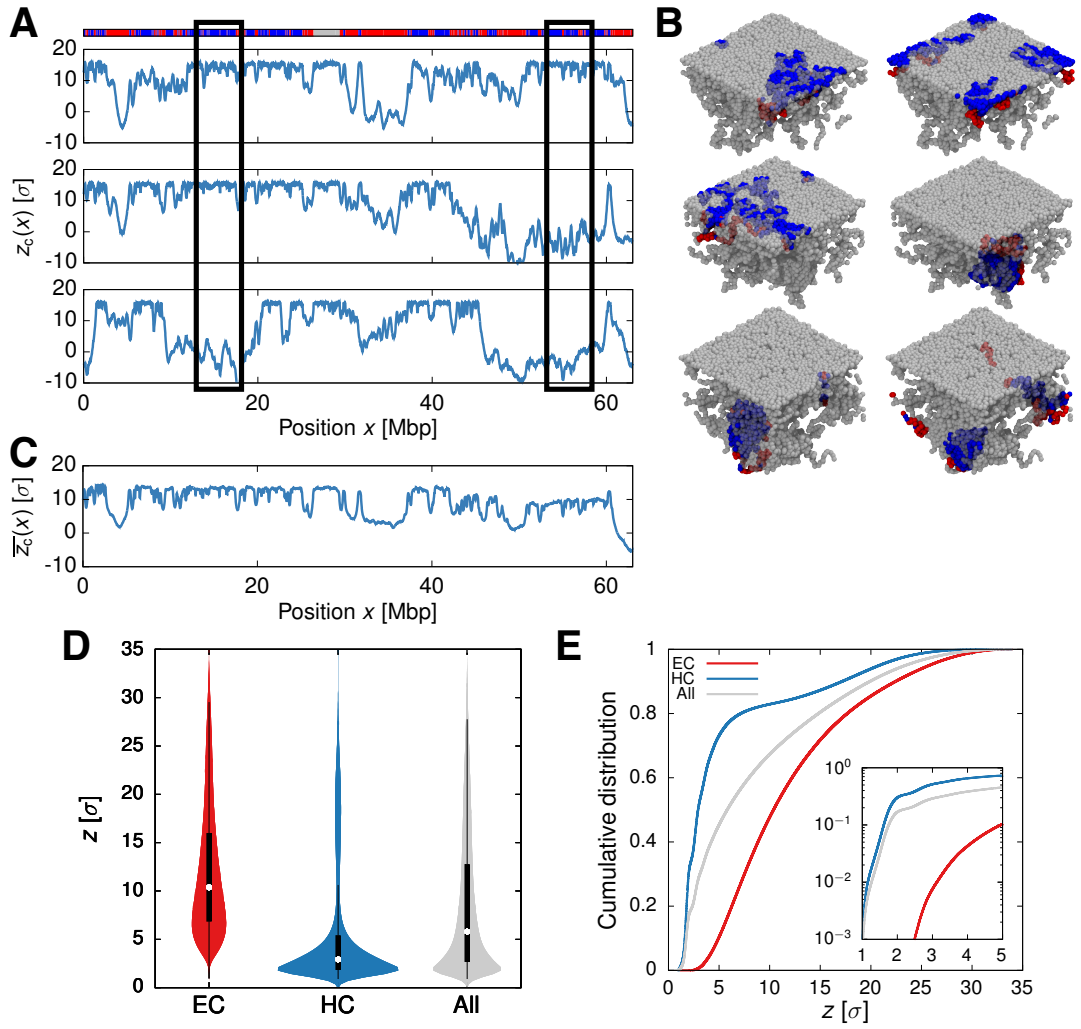

Figure 5.9: Heterogeneity of chromatin association with the nuclear lamina (NL). (A) Plots showing the vertical position z_c of each bead from the horizontal plane at the centre of the simulation box in three independent simulation runs. These simulations are conducted in the growing (AC) phase with $(\epsilon_{\mathrm{HH}}, \epsilon_{\mathrm{HL}}) = (1.0, 1.6)$. A higher z_c indicates that the bead is closer to the NL, which is located at $z_c^{\mathrm{NL}} = 17\sigma$. The top track shows the chromatin state of each bead, as in Figs. 5.5B and C. (B) Representative snapshots of these simulations colouring only the beads in the boxed regions in (A). These snapshots show that the spatial position of the same chromatin segment, particularly its location relative to the NL, can vary substantially between simulation runs. (C) A plot of the average z_c value for each bead across 20 simulation runs. (D) Violin plots displaying the distributions of the distance z from the NL for EC, HC, and all beads. (E) Cumulative distributions of z for EC, HC, and all beads. The inset shows the distributions at small z in log-linear form.

them can associate with the NL in each simulation (Fig. 5.9B). More importantly, when averaged across all simulation runs – akin to averaging over a population of cells – the profile of beads adsorbed to the NL is significantly different from that in a single simulation (Fig. 5.9C). This observation suggests that ensemble-

averaged data on chromatin folding, for example obtained from Hi-C experiments, may not fully reflect the actual organisations in individual cells [182]. This is also in line with a recent study using Hi-C and high-throughput fluorescence *in situ* hybridisation (hiFISH) to reveal that genome organisation is highly variable from cell to cell [183].

Another interesting observation from these simulations is that a small proportion of EC beads is sequestered to the NL, even though there is no direct, attractive interaction between EC and NL beads (Figs. 5.9B, D and E). This is to a large extent driven by the chromatin environment surrounding a bead. In particular, a small EC-rich region that is located within a large block of HC is prone to be co-tethered to the NL when the latter becomes attached to there. This finding highlights the importance of taking into account the epigenetic, as well as polymeric, properties of neighbouring chromatin when studying the spatial arrangement and function of a specific locus.

5.3 Chromatin Reorganisation in Senescence

The results presented thus far focus on the static, structural features of individual phases. Yet, simulations also enable one to interrogate the transitions between phases in a systematic manner that is difficult to achieve in experiments. A key transition in the phase diagram which I examine here is that from the growing (AC) to the senescent (DC) phase. The aim is to characterise this transition – i.e., whether it is first-order or continuous-like – and determine whether prominent phenotypic traits in these phases, notably SAHFs, play a role in influencing the transition behaviour. Biologically speaking, the analysis of the transition nature provides insight into the stability of individual cell conditions when subject to perturbations, such as the upregulation of HP1 proteins or sudden depletion of nuclear lamins.

5.3.1 The Transition between the Growing and Senescent Phases

The behaviour of the growing-senescence transition is studied in a similar way to that of the SD-CO transition in Section 4.2.1. Specifically, I consider whether there is hysteresis and phase coexistence near the transition boundary, which are

hallmarks of a sharp, first-order-like transition, as opposed to a continuous one. To test for hysteresis, the system is allowed to move slowly between the two phases by gradually changing the HC-NL interaction strength ϵ_{HL}. Throughout this process, the degree of chromatin adsorption to the NL is monitored by measuring the distance $\overline{z}$ between the centre of mass of the chromosome and the NL. More specifically, the system is first prepared at high ϵ_{HL} (within the growing phase); then ϵ_{HL} is slowly reduced to a low value (within the senescent phase); and finally it is increased again to its original value. Remarkably, $\overline{z}$ does not respond to the variation in ϵ_{HL} the same way in both directions (Figs. 5.10A and B). In other words, there is a hysteresis cycle in which the transition does not occur at the same critical value in each direction, and the system has memory of its previous configuration. The system also exhibits phase coexistence near the transition boundary. This is indicated by the observation of bimodality in the distribution of $\overline{z}$ near the critical point, demonstrating that the system can be either in the AC or DC phase at the same ϵ_{HL} (Fig. 5.10C).

The presence of coexistence and hysteresis suggests that the transition between the growing and senescent phases is first-order-like. With this kind of transition, there is typically a range of parameter values over which the two phases are metastable – i.e., the system may reside in a phase which does not have the lowest free energy, for example due to the presence of an energy barrier. From a biological perspective, the metastability of the senescent phase is significant, as it indicates that the transformation from growing to senescence is difficult to be reversed: to return to growing, one would need to expend energy on the system such that it escapes from the local free energy minimum associated with senescence. The spatial organisation of chromatin may provide an explanation for this metastability. In OIS cells, EC and HC segments are phase-separated by the formation of SAHFs. Notably, EC segments form a cloud of loops surrounding the HC-rich core within the foci (Figs. 5.3B and D). These loops are associated with an entropic barrier that has to be overcome when relocating HC beads to the NL for reinstating the growing condition. That is to say, the EC segments around SAHFs have to be "pushed away" in order to allow HC segments within SAHFs to reach the NL. The presence of this barrier thus stabilises the senescent phase. On a similar note, since the initial formation of SAHFs also involves a large-scale rearrangement of EC and HC, it is likely that there is another free energy barrier associated with progressing from growing to senescence, thereby securing the growing phase.

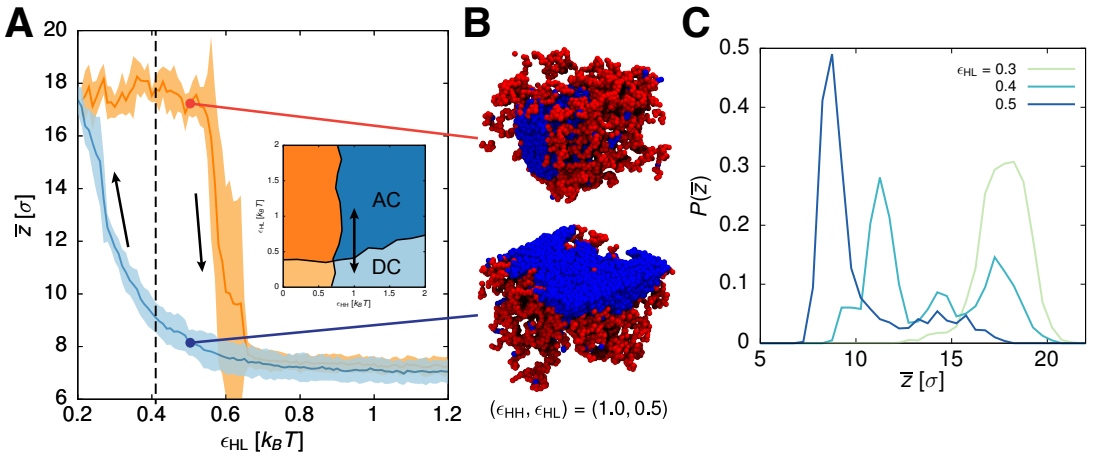

Figure 5.10: The transition between the growing (AC) and senescent (DC) phases. (A) A plot showing the distance $\overline{z}$ between the centre of mass of the chromosome with the NL as ϵ_{HL} is varied, with $\epsilon_{HH} = 1.0$. Specifically, ϵ_{HL} decreases gradually from 1.2 to 0.2 (in steps of 0.01) over a period of $10^6\tau$ (blue curve), and then it slowly increases back to 1.2 over the same amount of time (orange curve; the path of this variation within the parameter space is shown in the inset). Curves are averaged over five simulation runs, and the shaded region around each curve indicates the standard error of the mean. The dashed line shows the transition point between the AC and DC phases at $\epsilon_{HH} = 1.0$ as estimated from the phase diagram (Figs. 5.2A and C). Hysteresis is found in the region $\epsilon_{HL} \approx 0.3$–$0.6$. (B) Representative snapshots of the chromosome fibre at $\epsilon_{HL} = 0.5$, indicating that the system can either be in the AC or DC phase at this interaction strength depending on its history. (C) Probability density distributions of $\overline{z}$ at $\epsilon_{HL} = 0.3$, 0.4, and 0.5, with $\epsilon_{HH} = 1.0$. 50 simulations are sampled for constructing each distribution. Bimodality can be observed when $\epsilon_{HL} \approx 0.4$, suggesting coexistence of both the AC and DC phases.

5.3.2 Dynamics of Chromatin Detachment from the Lamina

A notable phenomenon during the transition to senescence is the detachment of LADs from the NL. The dynamics of this process are still poorly understood, as it remains challenging to probe nuclear dynamics at high temporal resolutions in experiments. Here, I perform simulations to gain further insight into how chromatin dissociates from the NL. In these simulations, the system is first initialised within the growing (AC) phase. The HC-NL interaction ϵ_{HL} is then abruptly reduced to a low level to model the loss of lamina interactions that occurs in senescence. Chromatin association with the NL is monitored over time upon the reduction in ϵ_{HL} by considering the distance z of each polymer bead from the NL and the fraction ψ of beads which remain adsorbed (Fig. 5.11). As anticipated, the bulk of the chromatin fibre migrates away from the NL

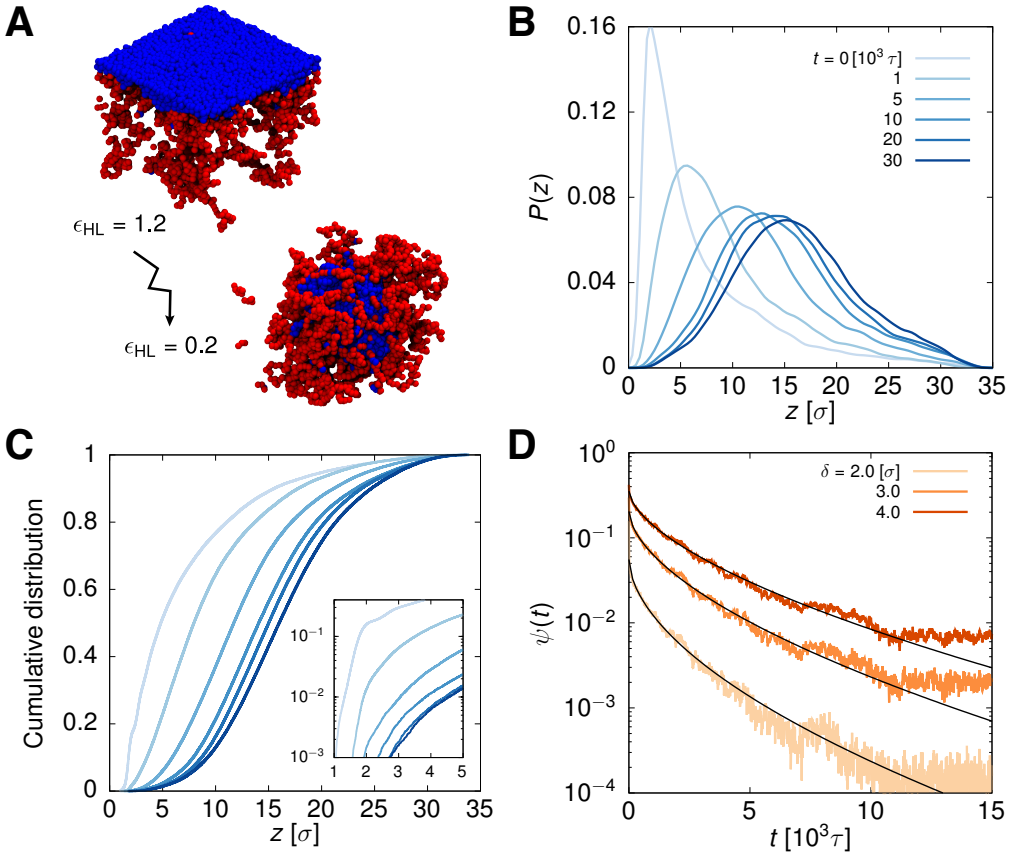

Figure 5.11: Simulated dynamics of chromatin detachment from the NL at the onset of senescence. (A) Schematics illustrating the process of detachment. The HC-NL interaction ϵ_{HL} is abruptly weakened from 1.2 to 0.2 to mimic the sudden loss of lamina-associated proteins. (B) Probability density distributions of the distance z of chromatin beads from the NL at different time points t since the reduction in ϵ_{HL}. (C) The corresponding cumulative distributions of z for the same time points. The insets display the distributions in log-linear form for small z. (D) The fraction ψ of chromatin beads which remain close to the NL (i.e., within a distance δ from the NL) as a function of the time t since the weakening in ϵ_{HL}, with different thresholds δ considered. The black lines are stretched exponential fits $f(t) = C \exp\left(-\kappa t^{\beta}\right)$ for these curves, and the fitted exponents β are 0.56, 0.58, and 0.61 for $\delta = 2$, 3, and 4σ, respectively.

over time as SAHF-like HC globules nucleate in the interior of the simulation box. Interestingly, the decay of ψ over time exhibits a non-exponential behaviour (Fig. 5.11D). Previous work in polymer physics has demonstrated that this type of decay can occur if the kinetics of the desorption process is constrained by the diffusion of polymer segments away from the surface [166, 167]. Confirmation of this trend by experiment in the future would demonstrate that the intrinsic polymer dynamics of chromatin have marked implications on the time scales in forming phenotypic features found in senescence, such as SAHFs.

5.4 Summary and Discussions

In summary, in this chapter I have developed a simple block copolymer model for the large-scale organisation of chromatin within the cell nucleus. The model explicitly incorporates the interaction between chromatin and the nuclear lamina (NL), an aspect that, to date, has been less explored in theoretical and simulation work. The model has been used to investigate the principles governing the spatial arrangement of euchromatin (EC) and heterochromatin (HC) in the context of cellular senescence and progeria, in which the nuclear architecture differs substantially from that of a healthy, growing cell. The model contains two key parameters – the HC-HC and HC-NL interaction strengths – in order to dissect the roles played by HC and the NL in driving chromatin folding.

From varying these two parameters, there are four possible phases within the model, and three of them display morphologies resembling those within growing, senescent, and progeroid nuclei as seen in microscopy experiments (Fig. 5.2). Specifically, the adsorbed-collapsed (AC) phase, with a strong HC-NL interaction, shows macroscopic features similar to those in a growing nucleus, including the aggregation of HC near the lamina and the formation of HC and EC compartments. The desorbed phases, with a weak HC-NL interaction, capture the disruption of chromatin-lamina association found in oncogene-induced senescence (OIS) and progeria. In particular, the desorbed-collapsed (DC) phase gives rise to large HC clusters within the nuclear interior, akin to senescence-associated heterochromatin foci (SAHFs) observed in OIS (Fig. 5.3). In contrast, the desorbed-extended (DE) phase exhibits marked chromatin decondensation and loss of compartmentalisation, in line with the nuclear organisation in progeria.

Upon establishing this qualitative mapping between the model phases and the biological conditions, the model is then employed to quantitatively understand the large-scale rewiring of chromatin interactions when a growing cell turns senescent or progeroid (Figs. 5.4 and 5.5). Importantly, the model recapitulates the overall trends reported in previous Hi-C experiments [156, 162]. The transition from the growing to the senescent phase is associated with a shift from local to distal interactions, whereas the transformation to the progeroid phase results in a reduction of distal contacts. The model predicts that the opposing trends in the contact patterns between senescence and progeria are partly driven by the differences in HC-mediated interactions in these two conditions (Fig. 5.7). The

strengthening of distal interaction in senescence seen here also corroborates recent Hi-C experiments and modelling work [73].

These interaction patterns are further examined from the perspective of polymer physics. According to well-established theories, the contact probability between any two chromatin loci as a function of their linear separation should decay as a power law. Simulations reveal that the decay exponent differs between the phases, with progeria having the largest exponent (decaying most steeply) and senescence having the smallest (Fig. 5.8). This result is consistent with the observed changes in the locality of chromatin contacts and provides further evidence that each phase adopts a different polymeric organisation. While previously it was suggested that progeria is a precursor to senescence [156], the results presented here offer an alternative view: namely, the nuclear arrangement of growing, senescent, and progeroid cells each corresponds to a different thermodynamic phase for interphase chromosomes.

Additionally, simulations in the growing phase demonstrate that chromatin association with the NL is highly stochastic. Even though in principle all HC domains are attracted to the lamina, only a subset of them harbour the nuclear periphery in each simulation, and the adsorption pattern varies between simulations (Fig. 5.9). This phenomenon is in agreement with Dam-ID experiments showing that lamina-associated domains (LADs) are rather heterogeneous at the single-cell level, and not all LADs are bound to the NL in each cell [150, 151]. The probabilistic nature of LADs further illustrates that chromatin organisation can vary substantially from cell to cell. Reconciling the structural information obtained from single-cell and population-wide experiments for chromatin organisation remains a key challenge to be resolved.

An important prediction from the simulations is that the progression from growing to senescence is characterised by an abrupt, first-order-like phase transition (i.e., between the AC and DC phases; Fig. 5.10). Hysteresis accompanying this kind of transition allows the system to remain metastable in a phase that is thermodynamically unfavourable. From a biophysical standpoint, this mechanism provides a natural and appealing explanation for the stability of OIS. Upon the formation of (micro)phase-separated structures such as SAHFs, a cell falls into a local free energy minimum, and thus re-entering the proliferative state becomes difficult. Functionally, the permanent exit from the cell cycle may serve as an effective way to limit the harm done to the host organism by the triggers (e.g., DNA damage and oncogene activation) that led to the senescent state.

There are several promising directions for future work in investigating chromatin interactions with the lamina. First, one can study more in detail the dynamics of chromatin desorption from the NL at the onset of senescence. The model predicts that the fraction of chromatin proximate to the NL should decay non-exponentially (Fig. 5.11). High-resolution time-course experiments can validate this finding in the future and may shed light on other spatio-temporal characteristics of LADs.

Second, the development of SAHFs following the detachment of LADs is also worth examining more closely. It is reasonable to speculate that this process should follow standard growth laws for cluster nucleation [184]. Deviation from the expected trends may uncover other principles governing SAHFs, which may in turn lead to a better understanding of the functional relevance of SAHFs.

Third, it is desirable to examine the effect of nuclear confinement on chromatin folding. The repositioning of HC from the nuclear periphery to the interior is not unique to senescence; it is also found in rod cells of nocturnal mammals [71, 172, 185]. In OIS cells there is on average one SAHF (or HC cluster) per chromosome, whereas in rod cells the formation of a single, large body of HC is more common. Notably, the size of a retinal nucleus is generally smaller than that of a senescent nucleus. It is, therefore, likely that a higher degree of confinement promotes further coalescence of HC clusters. Related to this hypothesis, recent work using a phase-field approach to simulate chromatin organisation in *Drosophila* nuclei suggests that nuclear volume remodelling can influence the kinetics of phase separation between EC and HC [186]. It would be interesting to see whether polymer models predict similar results.

Lastly, the study of cellular senescence and progeria has further demonstrated that the genome can adopt a diverse portfolio of organisations within the nucleus. It remains a fascinating challenge to unravel the link between these organisations and genome function, and to understand the benefits and drawbacks of each form of arrangement.

6

A Genome-Wide Analysis of Structure and Transcription of Regulatory Domains

The spatial organisation of the genome is intimately related to its function. Crosstalk between these two aspects can be found across all length scales within the nuclear environment, as demonstrated in previous chapters. In Chapter 4, it was observed that local chromatin folding helps regulate epigenetic patterns in order to maintain cell-type-specific expression programmes, while in Chapter 5, it was seen that large-scale genome rearrangement is involved in driving healthy cells into pathological conditions, such as cellular senescence and progeria.

A more fundamental example in which genome structure and function are strongly coupled is in gene regulation. The transcription of a gene – the copying of its DNA sequence into an RNA molecule – is a tightly regulated process which involves multiple protein complexes, such as RNA polymerases (RNAPs) and transcription factors (TFs)[1], as well as protein-binding motifs along DNA [4]. In particular, *cis*-regulatory elements (REs) – DNA sequences with which TFs associate – are

[1]Transcription factors (TFs) are protein complexes which recognise and bind to specific DNA sequences. TFs can be activators which promote gene expression or repressors which hinder it.

responsible for controlling the level of transcription, and the spatial interactions between them are often crucial to performing this task.

Two important types of REs are promoters and enhancers[2] [4, 187, 188]. Conventionally, a promoter is an RE located immediately (within a few hundred bps) upstream of the gene it regulates and is responsible for initiating transcription. An enhancer is one situated further away (a few kbps to hundreds of kbp upstream or downstream) from its target gene and serves to increase the gene's transcriptional activity. Typically, multiple enhancers can act on the same gene, while an enhancer can also influence the activity of more than one gene. It is widely believed that an enhancer modulates the expression of a gene by looping with the relevant promoter [4, 187, 188]. Promoter-enhancer interactions are ubiquitous within the genome, as demonstrated from microscopy and chromosome conformation capture (3C) experiments, and they are instrumental in gene regulation. In particular, promoters and enhancers are cell type specific, and the rewiring of their interactions is involved in development, cell differentiation, and disease [188].

While promoter-enhancer interactions often occur concomitantly with gene expression, the actual mechanisms linking these two aspects are still elusive. Work has examined various developmental or well-known regulatory loci (i.e., regions encompassing genes and their REs) to understand how these interactions are established and how they precisely control transcription [187, 188]. Modelling has played an important role in characterising the three-dimensional (3D) structures of individual loci, and this has been done using both an inverse [35, 37, 189–192] and a mechanistic approach [193–195] (see Section 2.2). In particular, from sampling a population of conformations for each target locus, simulation studies have revealed that there is large heterogeneity in chromatin folding between loci. Some loci are described by conformations that can be classified into a few distinct groups [35, 189, 193], whereas others seem to have a continuous variation in their structures that are less clusterable [194].

Deciphering the mechanistic links between the structural complexity and the transcriptional output of individual loci is challenging for several reasons. First, simulation studies have mainly investigated the former aspect, with less work focussing on the latter, making it difficult to draw connections between the two. Only recently has there been a study which attempted to predict the

[2]Other REs include silencers, which act to inhibit transcription, and insulators, which demarcate the interactions between REs. These elements are not studied here for simplicity.

transcriptional activity of REs based on simulation data [196]. Second, simulation and experimental studies have typically focussed on one or a few specific loci, so it is unclear whether principles learned from these studies can be generalised and applied across the genome. Moreover, this locus-specific approach, by definition, neglects the chromatin context of regions beyond the locus, which may also influence gene expression. Third, there is growing evidence suggesting that both the structure and transcription of individual genes can vary substantially between cells within a phenotypically homogeneous population [183, 197]. Therefore, in terms of experiments, single-cell methods are needed to verify any causal links, and it remains technically demanding to perform them genome-wide at high throughput.

Motivated by these issues, in this chapter I perform large-scale simulations to generate a pan-genomic data set which predicts the 3D structures and transcriptional activity of all activating REs (i.e., promoters and enhancers) in the human genome. In this way, universal links between these two aspects can be identified, with cell-to-cell stochasticity and the large-scale chromatin context taken into account. Here, individual chromosomes are simulated using the highly predictive heteromorphic polymer (HiP-HoP) model [194]. This recently developed mechanistic model incorporates TF binding and loop extrusion (LE) – two major principles governing chromosome organisation (see Section 2.3) – and considers the local compactness of the chromatin fibre, which is crucial for accurately capturing the structure of several gene loci [194].

This genome-wide data set provides a wealth of information about the REs. Regarding their structure, it identifies the frequently interacting chromatin targets or partners for each RE based on the ensemble of simulated conformations. It also characterises the topologies of how individual elements network with their partners and shows that there is substantial variation between them when mapping conformations to topologies. Regarding the function of REs, the data set reveals that the probability of TF binding at each RE is a strong predictor of the element's transcriptional activity. Furthermore, it suggests that the variation in this binding probability (between simulation replicas) may be related to transcriptional noise.

Importantly, this data set allows one to interrogate the relation between structure and transcription quantitatively. A prominent connection between these two aspects is that an RE with a higher number of partners typically has increased transcriptional output. Surprisingly, simulations indicate that LE activity is more

strongly correlated with transcriptional noise than the mean expression level of a gene, thus providing new insight into the functional significance of LE.

The rest of the chapter is structured as follows. In Section 6.1, I will explain the details of the HiP-HoP model, outline the simulation procedure, and describe the mapping between simulation and physical units. In Section 6.2, I will present a full-blown analysis of the structure-transcription relation for REs. Specifically, I will first examine the structural and transcriptional properties of REs separately before drawing connections between them. Finally, in Section 6.3, I will summarise the key findings and predictions of the work and highlight interesting directions for future research.

6.1 Simulation Model

The simulation model considered in this chapter is based on the HiP-HoP model introduced in Ref. [194] (Fig. 6.1A). The model is employed to simulate the entire genome of the GM12878 human lymphoblastoid cell line in a chromosome-by-chromosome manner (Fig. 6.1B). This cell line is chosen as there is a wide range of data sets available, including chromatin modifications and 3C-based data. In the following, I will discuss the various components of the model.

6.1.1 The Chromatin Fibre

In line with previous chapters, the chromatin fibre is modelled as a bead-and-spring polymer using the framework presented in Chapter 3. To fully resolve the 3D interactions between REs, each bead represents 1 kbp of chromatin and has a diameter σ. The persistence length of the fibre is fixed at $\ell_p = 4\sigma$, but this can vary due to its heteromorphic compaction (see below).

The chromatin beads are assigned different states or colours according to their local epigenetic modifications and DNA accessibility. Three epigenetic modifications are used here: histone 3 lysine 27 acetylation (H3K27ac), histone 3 lysine 27 trimethylation (H3K27me3), and histone 3 lysine 9 trimethylation (H3K9me3; Figs. 6.1A and C). These marks are typically associated with actively transcribed euchromatin, facultative heterochromatin, and constitutive heterochromatin, respectively. The chromatin immunoprecipitation with sequencing (ChIP-seq)

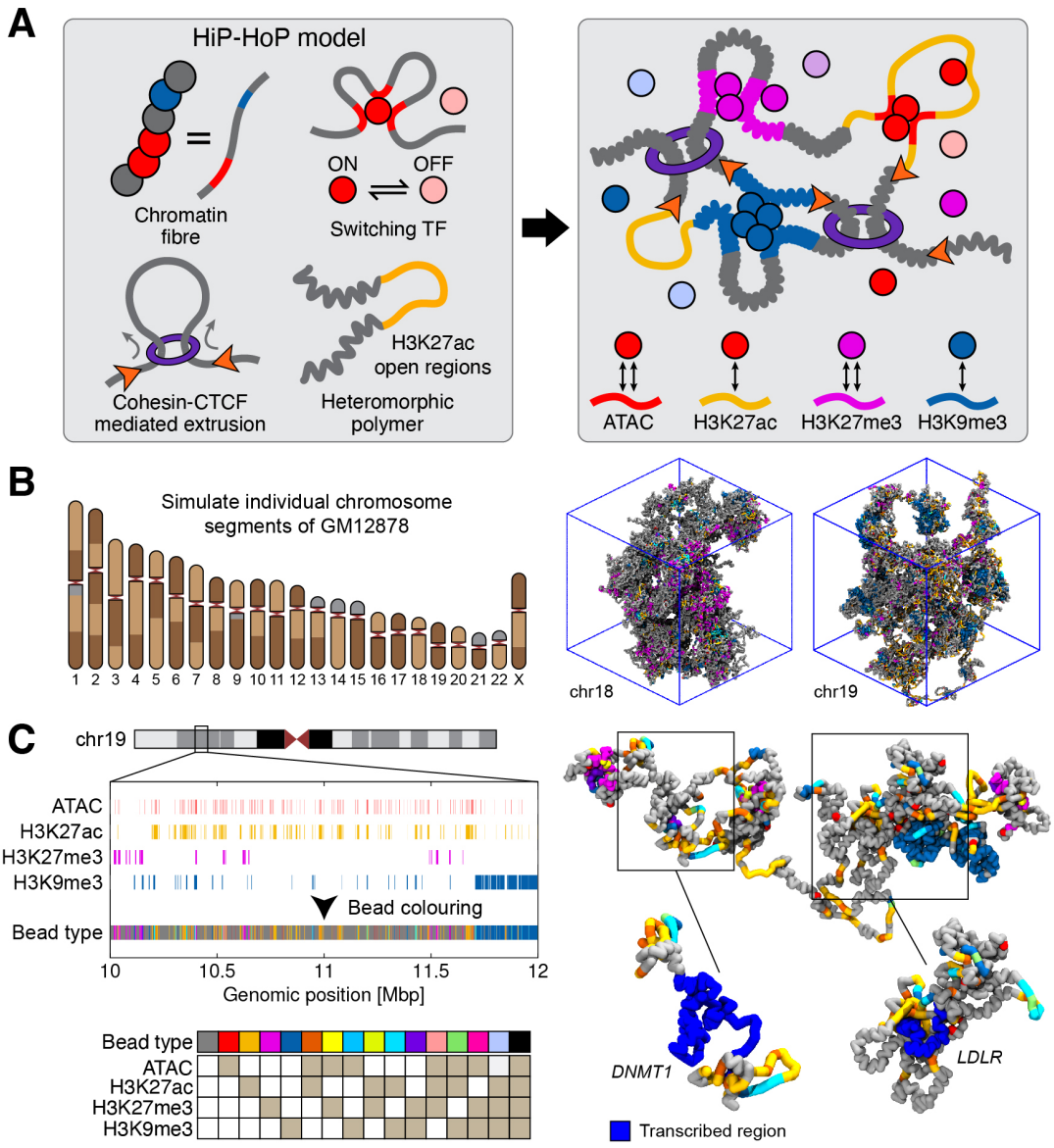

Figure 6.1: The highly predictive heteromorphic polymer (HiP-HoP) model. (A) The model represents the chromatin fibre as a bead-and-spring chain and is an amalgamation of the transcription factor (TF) model with switching [46, 47] and the loop extrusion (LE) model [54, 55]. It also considers chromatin as a heteromorphic polymer, with regions enriched in H3K27ac hypothesised to have a more open or disrupted structure [194]. The model colours the chromatin fibre according to DNA accessibility (ATAC-seq) and three other epigenetic marks: H3K27ac, H3K27me3, and H3K9me3. Additionally, there are three species of TFs: one species of active TFs (red beads) that bind strongly to ATAC and weakly to H3K27ac sites, and two species of inactive TFs, with one (Polycomb-like; magenta) binding to H3K27me3 and the other (heterochromatin-like; blue) binding to H3K9me3. (B) *Left*: The model is employed to simulate the entire genome of the GM12878 human lymphoblastoid cell line. Chromosomes are divided into segments and simulated individually, as indicated by the partitions within the ideograms shown here. *(Continued at the bottom of the next page.)*

profiles for these marks (for GM12878 cells, aligned to the hg19 reference genome) are obtained from the ENCODE database (https://www.encodeproject.org) [198, 199]. DNA accessibility is determined from an assay for transposase-accessible chromatin using sequencing (ATAC-seq)[3] data set [200]. Regions with high ATAC signals usually map to TF binding sites, such as those found at promoters and enhancers [194]. Beads are assigned particular epigenetic and/or ATAC marks if the respective chromatin regions have significant enrichment, or peaks[4], of these marks. Note that a bead can have more than one kind of mark, and its state or colour is determined based on the combination of marks it possesses (Fig. 6.1C).

A prominent feature of the HiP-HoP model is that it represents chromatin as a heteromorphic polymer, whose linear compaction varies along the contour. This feature is needed to account for a more open or disrupted conformation in acetylated regions, as observed in fluorescence *in situ* hybridisation (FISH) experiments [194]. In practice, the heteromorphic property is implemented by introducing extra harmonic springs [Eq. (3.4)] which link beads i and $i + 2$ in regions that are not annotated with the H3K27ac mark. Here, the spring constant is set to $K_h = 400k_BT/\sigma^2$ and the bond length to $r_0 = 1.1\sigma$. In this way, regions with the extra springs become crumpled and have higher linear compaction, whereas those with the mark remain open. This variability in local

[3]ATAC-seq is a technique which provides a genome-wide profile of DNA accessibility. It involves using hyperactive transposases to integrate sequencing adaptors to DNA. This process is more likely to occur in more accessible or open DNA regions, making them more likely to be sequenced and thus have a higher output signal. Alternative methods for examining DNA accessibility include DNase I hypersensitive sites with sequencing (DNase-seq), which uses restriction enzymes to probe accessibility, and micrococcal nuclease with sequencing (MNase-seq), which assesses nucleosome positioning and occupancy.

[4]Peak calling is done using `epic2` [201] for epigenetic marks with broad domains, such as H3K27ac, H3K27me3, and H3K9me3. `macs2` [202] is used for more localised marks, including ATAC, CTCF, and Rad21 (see the discussion below on modelling LE).

Figure 6.1 ***(continued)*****:** *Right*: snapshots of chromosomes 18 and 19, each simulated as a single segment. TFs are not shown for simplicity. (C) Simulation details for the region 10–12 Mbp in chromosome 19, as an example to further illustrate the model specifics. *Left*: tracks of the DNA accessibility and epigenetic data used to inform the bead type, which is based on the combination of individual marks. *Right*: a simulation snapshot of this region, with enlarged views of two example gene regulatory loci (*DNMT1* and *LDLR*; regions in black boxes) displayed at the bottom. The actual transcribed regions of both genes are highlighted in dark blue. The model can resolve the structure of individual regulatory loci at this resolution (1 kbp) across the entire human genome.

folding changes the stiffness and persistence length of the fibre: in Ref. [194], it was found that the actual persistence length in the crumpled regions is around 4.7σ, compared to the bare input value of 4σ.

6.1.2 Transcription Factor Binding

Another key ingredient of the model is the incorporation of protein complexes, referred to here simply as TFs, which bind to chromatin (i.e., the TF model; see Section 2.3.1). Similar to previous work [46, 75, 194], TFs are modelled as spherical beads with a diameter σ. These beads are multivalent and are allowed to bind to chromatin beads that are enriched in specific epigenetic modifications or have high accessibility. This attraction is modelled using the truncated and shifted LJ potential [Eq. (3.9)] with an interaction cutoff $r_c = 1.8\sigma$. To drive the formation of both transcriptionally active and inactive chromatin domains, three species of TFs are introduced: a generic active, a Polycomb-like, and a heterochromatin-like TF (Fig. 6.1A). Active TFs bind strongly ($\epsilon = 7.0$; in units of k_BT) to beads with high accessibility (ATAC peaks) and weakly ($\epsilon = 3.0$) to those enriched in H3K27ac. This aims to capture promoter-enhancer interactions and the formation of transcriptional domains [203]. For inactive TFs, Polycomb-like TFs bind to beads enriched in H3K27me3 ($\epsilon = 7.0$), representing interactions mediated by Polycomb repressive complexes (PRCs) [204, 205]. Heterochromatin-like TFs bind to beads with H3K9me3 ($\epsilon = 3.0$), modelling bridging facilitated by heterochromatin protein 1 (HP1) [175]. Importantly, TFs only interact with one another via steric repulsion, described by the WCA potential [Eq. (3.8)]. Nevertheless, thanks to their ability to bridge between multiple chromatin beads with similar marks, TFs of different species tend to (micro)phase separate and form individual clusters via bridging-induced attraction (BIA; see Section 2.3.1).

TFs also switch back and forth between a binding and a non-binding state in the model with a rate k_{sw} (Fig. 6.1A). As discussed in Section 2.3.1 and previous literature [47], this feature mimics post-translational modifications on protein complexes and accounts for the dynamical turnover of constituents within nuclear protein clusters. When TFs are non-binding, they interact with chromatin beads via steric repulsion (modelled by the WCA potential). Switching also helps regulate the size of chromatin domains and the ratio of non-local to local interactions [75].

In addition to TF-chromatin binding, there is also a weak, direct chromatin-chromatin interaction ($\epsilon = 0.4$) between beads which do not possess the active H3K27ac mark. This extra ingredient facilitates the phase separation between euchromatin and heterochromatin, and previous work has considered similar interactions when looking at compartmentalisation [74].

6.1.3 Loop Extrusion

As well as TFs, the model contains active loop extruding factors (or extruders), such as the structural maintenance of chromosome (SMC) complex cohesin [54, 55] (see the LE model in Section 2.3.2). These extruders bind to a single point on chromatin and translocate outwards to generate loops. While the actual mechanisms of how SMC complexes attach to and move along chromatin remain elusive and are subject to intense investigation, a simple extrusion model is implemented here, similar to those in Refs. [54] and [55]. Specifically, an extruder is represented as a dimer whose two ends move divergently along the chromatin fibre. For simplicity, it is modelled implicitly as a harmonic spring with short-range WCA repulsion[5]:

$$\mathcal{U}_{\text{ex}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = \mathcal{U}_{\text{WCA}}(\boldsymbol{r}_i, \boldsymbol{r}_j) + \frac{K_{\text{ex}}}{2}(r_{ij} - r_0)^2, \tag{6.1}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the positions of beads i and j, respectively, with $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$, $K_{\text{ex}} = 80k_BT/\sigma^2$ is the spring constant, and $r_0 = 1.5\sigma$ is the bond length.

In the simulations, an extruder binds randomly to any chromatin bead, say the ith one, and this is modelled by introducing a spring linking beads i and $i+3$ (as a harmonic spring may already connect beads i and $i+2$ to model the heteromorphic fibre). Once bound, its two ends translocate at speed v_{ex}, and this is done by moving the spring to the next pair of beads. Both ends of an extruder move along the fibre until colliding another extruder or reaching a CCCTC-binding factor (CTCF) bead (see below) whose orientation is opposite to the direction of travel. Note that the two ends move independently: if one end halts due to the aforementioned scenarios, the other can continue to extrude. An extruder detaches from chromatin with a rate k_{off}, upon which the spring is removed. For

[5] A harmonic bond is used here instead of a FENE bond as the latter is less tolerable to excessive stretching. The WCA potential is added to maintain volume exclusion of the bonded beads and reduce the chance of entanglement.

simplicity, it is assumed that when an extruder unbinds from the fibre, another immediately re-attaches to it (i.e., the number of extruders on chromatin remains constant).

To identify the CTCF binding sites which constrain LE, ChIP-seq data sets for CTCF and Rad21, a subunit of the cohesin complex, are obtained from ENCODE. Genomic loci which have peaks in both profiles while also containing the CTCF binding motif[6] are marked as candidate binding sites. Beads covering these sites are then labelled as CTCF beads, and the direction in which they act on the extruders is based on the orientation of the underlying motif. Motivated by the cell-to-cell variability in CTCF binding, CTCF beads are activated stochastically in each simulation run according to a probability that is linearly proportional to the score of the corresponding CTCF peak. When there are multiple peaks within the same bead, all possible outcomes are considered. For example, if a bead encompasses both a forward and backward-oriented CTCF binding site, the probabilities of the bead being a forward, backward, bidirectional, or inactive CTCF boundary are calculated based on the score of individual peaks, and an outcome is selected based on these probabilities.

6.1.4 Simulation Parameters and Setup

There are several important parameters in the model. In the TF component, there is the number of TFs of each species and the switching rate k_{sw}. In the LE component, there is the number of extruders N_{ex}, as well as the extrusion rate v_{ex} and the unbinding rate k_{off}. Although it is impractical to perform a systematic sweep of the parameter space due to the significant amount of time required for running each simulation, different regions within this space are explored. Parameters for the simulations are chosen to be consistent with previous literature, where possible, and give predictions that are broadly in agreement with Hi-C data. The procedure for varying the parameters and determining the specific set of values employed in the simulations is discussed in Section A.1.

Here, I summarise the parameter values used in the simulations. The number of TF beads N_{TF} is set to $\sim$10% of the number of chromatin beads N_{chr}, while the

[6]The consensus sequence for the CTCF binding motif is downloaded from JASPAR [206], and genome loci with this motif are identified using `fimo` [207]. When there are multiple motifs within a CTCF ChIP-seq peak, the binding orientation is taken from the motif with the highest score. If the top two motifs within a peak are oppositely oriented but have similar scores (i.e., less than 5% difference), the orientation is marked as "bidirectional".

ratio of the active, Polycomb-like, and heterochromatin-like TFs is fixed at $1/4 : 1/8 : 5/8$. Additionally, the switching rate is set to $k_{\text{sw}} = 10^{-3}\tau^{-1}$. Regarding LE, the number of extruders N_{ex} bound to chromatin is fixed at a density of 10 extruders/Mbp. The extrusion and unbinding rates are set to $v_{\text{ex}} = 4 \times 10^{-3}$ kbp/τ and $k_{\text{off}} = 2.5 \times 10^{-5}\tau^{-1}$, respectively. While the parameters are chosen to facilitate sampling and are less realistic individually, the ratio $\lambda = v_{\text{ex}}/k_{\text{ex}} = 160$ kbp (or the extruder's processivity) and the density of extruders (or the average spacing between them) are consistent with values used in the literature [55].

The chromatin fibre is simulated within a periodic cube of length L such that the density of chromatin is approximately 6.5 Mbp/μm^3. This is based on the fact that there are around 6.5 Gbp of DNA in a human diploid cell, and that the diameter of a typical cell nucleus is around 10 μm. Simulations are performed using the Brownian dynamics scheme discussed in Section 3.1.4. As done previously [194, 195], the ratio γ/m is fixed at 2 (in reduced units), which causes beads in simulations to have more inertia than in reality (see Section 3.3). This choice is needed to ensure computation time remains practically feasible. Although short-time dynamics are less realistic under this procedure, this is acceptable as the main focus is on the system's conformation at steady state.

To facilitate computation, chromosomes are simulated individually rather than as a whole (Fig. 6.1B). While this approach neglects long-range and inter-chromosomal (or *trans*-) interactions, this is tolerable as the main objective of the work is to examine the *local* chromatin structure and *cis*-interactions of REs, which typically span no longer than a few Mbps. In practice, shorter chromosomes (chromosomes 14, 15, and 17 to 22) are simulated as a single segment, whereas longer chromosomes (chromosomes 1 to 13, 16, and X)[7] are each divided into smaller segments. Breakpoints are chosen to be sites where there is low enrichment in chromatin modifications within their neighbourhood (i.e., gene deserts). A full summary of how individual chromosomes are broken down into segments is given in Section A.2.

The simulation system is initialised in the following way. In line with Refs. [39] and [195], the chromatin fibre is first generated as a mitotic-like helix conforma-

[7]The Y chromosome is not considered in this work for simplicity, as it is too short to be simulated on its own.

tion – i.e., a stack of rosettes – governed by the following set of equations

$$x(\theta) = r\cos\theta\left[(1-\xi)\cos^2(k\theta)+\xi\right] \tag{6.2}$$

$$y(\theta) = r\sin\theta\left[(1-\xi)\cos^2(k\theta)+\xi\right] \tag{6.3}$$

$$z(\theta) = \frac{p\theta}{2\pi}, \tag{6.4}$$

where r is the radius of the rosette, $r\xi$ is the radius of the gap within the interior of the rosette, $2k$ is the total number of petals per rosette, and p is the vertical distance between adjacent rosettes. These parameters are set to $r \approx L/2$, $\xi = 0.2$, $k = 6$, and $p = 1.2\sigma$. The initial simulation box is a cuboid with fixed boundaries enclosing the cylinder tightly, with dimensions $L \times L$ in the x-y directions and a height that depends on the length of the chromosome segment.

To relax the fibre from the rosette configuration, beads along the chain are initially connected together using harmonic springs (with spring constant $K_h = 200k_BT/\sigma^2$ and bond length $r_0 = 1.1\sigma$), and all non-neighbour pairwise interactions are governed by the repulsive soft potential [Eq. (3.11)], whose height is slowly increased from 0 to 10^3k_BT. After an initial relaxation simulation of duration 600τ, the springs are replaced by FENE bonds [Eq. (3.3)] and the soft potential by the WCA potential. The simulation box is then compressed slowly in the z direction for $10^4\tau$ such that it becomes a cube with the desired volume ($V = L^3$). Next, the fibre is relaxed further within the cube with fixed boundaries for $5 \times 10^3\tau$ and then with periodic boundaries for $2.5 \times 10^4\tau$. As considered in Ref. [195], to allow the fibre to quickly lose memory of the rosette-like pattern and achieve a conformation whose decay exponent in the contact probability between two loci [Eq. (2.12)] is comparable to that measured from Hi-C, extruders (with a density of $\sim$7.5 extruders/Mbp) are loaded to the fibre to perform loop extrusion without any CTCF boundaries for $2 \times 10^4\tau$. Harmonic springs for modelling fibre heteromorphicity are then added to the entire fibre, and the fibre is allowed to relax for $10^4\tau$. Next, TF beads are incorporated and are allowed to equilibrate with the fibre for $10^3\tau$, with TF-chromatin interactions being purely repulsive (modelled by the WCA potential). Finally, regions enriched in H3K27ac have their crumpled springs removed such that the fibre has different levels of local compaction. 10 independent runs are conducted using this procedure for each chromosome segment, and their final conformations are used as the starting conditions for 300 production runs (which are performed using different random seeds). From measuring structural properties such as the radius of gyration

of the fibre, it is verified that production runs starting from the same relaxed conformation do give different structures.

In the production run, the chromatin fibre is simulated for a period of $3 \times 10^5\tau$. Attractive chromatin-chromatin interactions are switched on at $10^4\tau$, and subsequently TF-chromatin interactions at $5 \times 10^4\tau$. The system is sampled every $2 \times 10^3\tau$ during the final $10^5\tau$ of the simulation period.

It is also worth appreciating the scale of computing resources required for simulating the entire human genome at 1 kbp resolution. Typically, a single simulation for a 75-Mbp (75000-bead) segment takes approximately 60 hours to complete on 18 compute cores. As a rough estimation, there are around 40 such segments in a haploid genome ($\sim$3 Gbp), and for each segment 300 simulation runs are conducted. As a result, the total amount of time required for the entire genome is $60 \times 18 \times 40 \times 300 \approx 1.3 \times 10^7$ core hours. Even with 1000 single-core computers running without interruptions, it would still require nearly 1.5 years to complete all of the simulation runs. To speed up computation, a majority of the simulations are done using the computing nodes within the Edinburgh Compute and Data Facility (ECDF) and those within the Tier 2 high performance computing facility Cirrus.

6.1.5 Mapping of Length and Time

The mapping between simulation units to physical units for the HiP-HoP model was described previously in Refs. [194] and [195]. The size of each bead σ is estimated to be around 21.8 nm from comparing distances between simulated chromatin loci to their counterparts in FISH experiments. From comparing the mean squared displacement of chromatin segments in simulations with that from microscopy for yeast chromatin [84], it was found that the simulation time unit τ maps to roughly 5 ms [195] (see also Section 3.3). This suggests that the main simulation period corresponds to approximately 25 min in real time.

6.2 3D Structure and Transcription of Regulatory Elements

The main objective of the work is to understand the connections between the spatial interactions of activating REs and their transcriptional output. To this end, I utilise the simulated 3D structures of all chromosomes to conduct a three-part analysis. First, frequent interactions are identified for each RE from the structures, and the associated topologies (or networks of interactions) are catalogued. Second, the transcriptional activity of each element is determined, and, as shown below, this can be measured both from simulations and experiments. Third, the connections between structure and transcription are explored by correlating various observables related to these two aspects.

In this analysis, to keep computation manageable while ensuring that there are enough structures, two conformations are sampled from each of the 300 independent production runs (at $2 \times 10^5\tau$ and $3 \times 10^5\tau$), giving a total of 600 structures per chromosome segment. REs along the chromatin fibre are studied by proxy based on beads corresponding to ATAC peaks (or ATAC beads), in line with previous work [196]. This procedure is motivated by the observation that chromatin segments with high accessibility typically map to these elements (active TF binding sites). In the following, I shall simply refer to REs as ATAC beads or sites. Interactions are examined for all of these beads, not just for those which map to gene promoters. This treatment provides a more comprehensive data set: in total, there are 70981 ATAC beads across all chromosome segments simulated, while only 9859 of them are mapped to (actively transcribed) promoters[8]. It should also be noted that transcription does occur at enhancers (i.e., the production of enhancer RNAs [187]), and it is of interest to compare the model predictions between these two kinds of REs.

6.2.1 Identifying Topologies of Regulatory Elements

As mentioned above, the first part of the analysis examines the structural properties of individual ATAC beads, focussing on their interaction patterns. Here, for each ATAC bead, a list of other ATAC beads with which it interacts (i.e., separated by less than 3.5σ) are detected from the 600 conformations (Fig. 6.2A).

[8]The procedure to identify promoters from ATAC beads is explained below in Section 6.2.2.

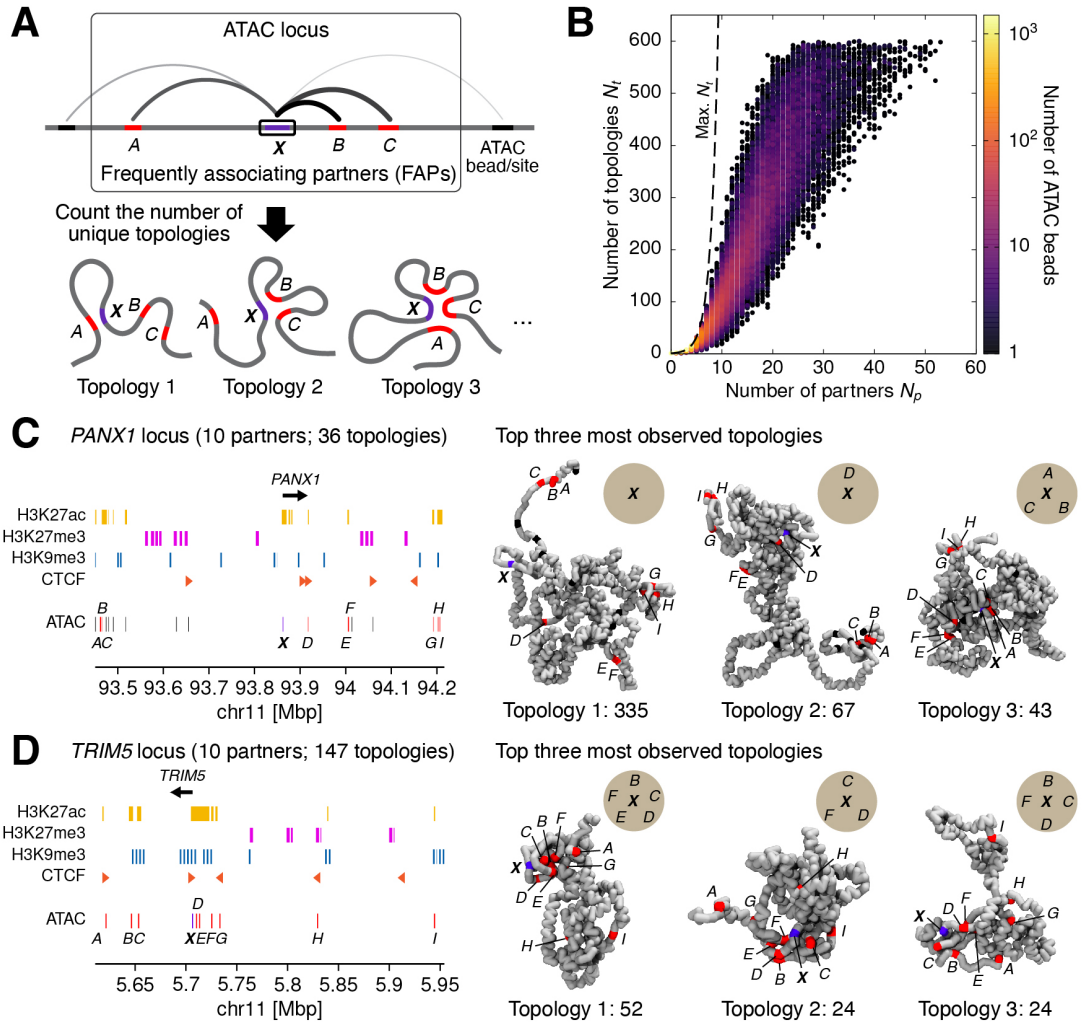

Figure 6.2: Frequently associating partners (FAPs) and interaction topologies of each ATAC bead. (A) Schematics illustrating how FAPs and topologies are determined for each ATAC bead or site, here for the one coloured in purple (also labelled as X). First, all ATAC sites (black or red) which interact with the purple site (i.e., spatially separated by less than 3.5σ) are identified, and those which do so in more than 10% of all simulated structures are considered as FAPs (red sites; labelled from A to C). The region encompassing all FAPs is defined as the structural locus of the purple site. Next, the interaction topologies, or the ways how the purple site networks with the red sites, are determined, and the frequency of observing each is calculated. (B) A scatterplot showing the number of observed topologies N_t against the number of FAPs N_p for all ATAC beads. The colour scale indicates the number of ATAC beads sharing the same coordinates, and the dashed line denotes the maximum upper bound on N_t for a given N_p (i.e., $N_{t,\max} = 2^{N_p}$). (C and D) two ATAC loci with the same number of FAPs but with very different number of topologies: the promoters for (C) *PANX1* and (D) *TRIM5*. *Left*: a panel showing tracks of the epigenetic modifications, CTCF sites, as well as the locations of the ATAC sites (the promoter in purple and FAPs in red) within each locus. *Right*: a representative snapshot for each of the top three most observed topologies of the locus. The circle at the top right of each topology indicates the specific FAPs that are interacting with the promoter.

To remove random background contacts, only the ATAC beads that interact regularly are retained, and they are referred to as frequently associating partners (FAPs). More specifically, FAPs are defined to be the ATAC beads which make contact with the target bead in more than 10% of the simulated structures; variation of this threshold value between 5–15% gives qualitatively similar trends.

The identification of FAPs allows one to define the chromatin structural unit associated with an ATAC bead (i.e., an ATAC locus) as the genomic region encompassing all of its partners (Fig. 6.2A). The median size of an ATAC locus is around 200 kbp, which is at the lower end of the scale of a topologically associating domain (TAD) [180]. This suggests that these loci are typically below the TAD level, and they may be more similar to the chromatin nanodomains (CNDs) detected recently from super-resolution microscopy [208].

With FAPs and ATAC loci defined, one can determine the interaction topologies within a locus, or the different ways that the ATAC bead of interest networks with its partners. Fig. 6.2B displays a scatterplot of the number of topologies N_t against the number of partners N_p for all ATAC beads, and it shows that N_t increases rapidly with N_p until saturating due to the limited number of conformations. This increase, to a first approximation, can be understood from a simple combinatorial calculation. If one neglects the polymeric nature of chromatin, finding the number of unique topologies for N_p partners is equivalent to counting the number of ways to create a subset within a set of N_p (distinguishable) elements. As each element can either be in the subset or not, this number is simply 2^{N_p}. Clearly, this constitutes the maximum possible number of topologies, as physical constraints of the fibre would render some not achievable. Moreover, it is not clear *a priori* whether all of the remaining permissible topologies would occur in reality, as other factors such as the local chromatin context and linear spacing between ATAC sites can also influence the interaction networks. Nevertheless, Fig. 6.2B shows that even with a modest number of partners, many topologies are detected for individual ATAC loci, suggesting that these loci display large heterogeneity in their 3D folding across the ensemble of simulated structures. This result is consistent with recent single-cell studies revealing the extensive variation in chromatin organisation within a population of cells [183].

Fig. 6.2B also shows that ATAC beads with the same number of partners can have a large difference in their number of topologies. For example, the ATAC beads corresponding to the promoters (or ATAC promoters) of *PANX1* and

TRIM5 have 10 partners each; however, the former has 36 topologies, whereas the latter has 147 (Figs. 6.2C and D). Although the epigenetic context within the neighbourhood of these loci is similar, the linear arrangement of partners is different. In particular, *PANX1* has fewer of its partners nearby compared to *TRIM5*, and the genomic distance to the nearest partner for *PANX1* and *TRIM5* is 55 kbp and 4 kbp, respectively. These differences have a clear impact on the 3D structures, as demonstrated from the top three most observed topologies for these two loci (Figs. 6.2C and D). Strikingly, for *PANX1*, a majority of conformations (335 out of 600) is associated with the topology where the ATAC promoter is alone and does not interact with any of its partners. In contrast, the top topology for *TRIM5* has multiple partners interacting with the promoter, thanks to their proximity in genomic separation. The *PANX1* locus also illustrates another interesting feature: FAPs do not need to be contiguous along the chromatin fibre (Fig. 6.2C); it is possible to skip over some ATAC beads in between, possibly as a result of the local chromatin context (e.g., epigenetic marks and CTCF boundaries).

The *PANX1* and *TRIM5* loci demonstrate that the number of conformations associated with individual topologies can vary substantially. In *PANX1*, many conformations are mapped to a single topology (Fig. 6.2C), whereas in *TRIM5*, they are spread more evenly across a larger number of topologies (Fig. 6.2D). How the population of structures is distributed among the observed topologies can be quantified by a diversity score (i.e., Shannon diversity index or Shannon entropy; Fig. 6.3A)

$$H = -\sum_{i=1}^{N_t} n_i \ln n_i, \tag{6.5}$$

where n_i is the fraction of structures in topology i. Note that this quantity is expected to scale linearly with the number of partners N_p, and this can be seen from maximising H, which is achieved when $n_i = 1/N_t$ for all i. In this case, $H = \ln N_t$, and substituting the upper bound $N_t = 2^{N_p}$ gives $H = N_p \ln 2$.

Fig. 6.3B reports the H score for all ATAC beads. Consistent with the calculation above, H scales approximately linearly with N_p initially, before saturating at high N_p again due to the limited number of conformations. Reassuringly, this statistic

[9]The Mann-Whitney U test, or the Wilcoxon-Mann-Whitney test, is a non-parametric statistical test on the null hypothesis that the two samples in consideration have the same median; in other words, when randomly drawing a value from each sample, there is an equal chance for the value from the first sample to be greater or less than that from the second [181].

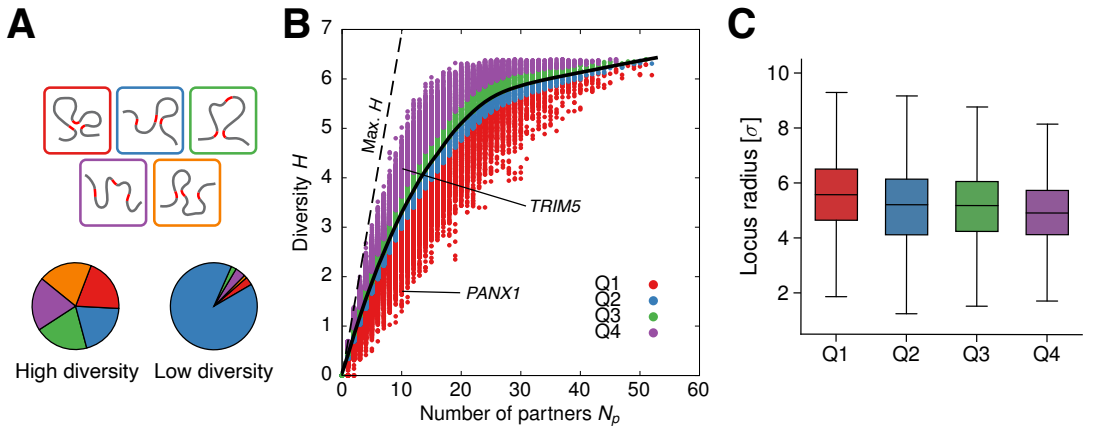

Figure 6.3: Diversity of the topologies for each ATAC bead. (A) An illustration explaining the diversity H score. Here, with a population of five topologies, a higher H is achieved when the sampled conformations are distributed more evenly among them (i.e., more equal slices in the pie chart), whereas a lower H occurs when many conformations are assigned to one of the topologies (more unequal slices). (B) A scatterplot showing the H score against the number of partners N_p for all ATAC beads. The dashed line represents the maximum possible H score for a given N_p (i.e., $H_{\max} = N_p \ln 2$). The solid black line represents the expected H score for a given N_p (i.e., $\langle H \rangle_{N_p}$), which is computed from performing a locally weighted estimated scatterplot smoothing (LOWESS) [209] fit of the data set (this is done in lieu of measuring the mean of H at each N_p in order to reduce noise at large N_p). Data points are split into quartiles (Q1 to Q4) based on the deviation $\Delta H = (H - \langle H \rangle_{N_p})/\sigma_{H,N_p}$, where σ_{H,N_p} is the sample standard deviation of H at a given N_p. (C) Boxplots showing the radius of gyration of the ATAC loci in each quartile. The difference in the median between Q1 and all other quartiles is statistically significant (Mann-Whitney U test[9]: $p < 10^{-4}$).

recapitulates the findings for *PANX1* and *TRIM5*: the former has a low diversity ($H = 1.77$) due to the uneven mapping of conformations to topologies, whereas the latter has a high diversity ($H = 4.35$), as the distribution is more balanced.

Globally, there is a large variation in the H score for ATAC beads with the same N_p. This can be quantified by measuring the deviation ΔH of each ATAC bead's H score from the expected score $\langle H \rangle_{N_p}$ at a given N_p (see the black line in Fig. 6.3B). Specifically, ΔH is defined as a z-score-like metric, i.e.,

$$\Delta H = \frac{H - \langle H \rangle_{N_p}}{\sigma_{H,N_p}}, \tag{6.6}$$

where σ_{H,N_p} is the sample standard deviation of H at a particular N_p. To help identify trends, ΔH for individual ATAC points are binned into quartiles. Interestingly, it is found that the average radius of gyration of an ATAC locus decreases as ΔH becomes larger (Fig. 6.3C). A possible explanation for this

finding is that those ATAC loci in the lower quartiles are likely to have more conformations linked to topologies with fewer interactions (e.g., similar to the scenario in *PANX1*). Structures associated with these topologies tend be more swollen due to fewer chromatin loops established, and thus a larger radius of gyration is observed overall.

6.2.2 Predicting Transcriptional Activity and Variability

The second part of the structure-transcription analysis focusses on estimating the transcriptional activity of each ATAC bead, which can be determined directly from experimental data. In particular, data from global run-on sequencing (GRO-seq)[10] [210] are used here, as they provide a genome-wide, per-base measure on this activity. Additionally, it is of interest to ask whether the model can offer predictions in this respect, as answering this question may further elucidate the mechanisms regulating transcription. Since active TFs within the model mimic complexes of RNA polymerases and activators[11] [42, 46, 196], a natural hypothesis is that the frequency of one or more TFs binding to an ATAC bead is related to the transcriptional output of the corresponding chromatin segment, with a higher frequency associated with higher output [196].

To validate this hypothesis, for each ATAC bead within a simulation run, the fraction of time where there is at least one active TF bound is recorded, and a (per-bead) distribution is obtained by combining results across all runs (Fig. 6.4A). The average of this distribution, or the TF binding probability, gives a population-wide measure of the likelihood of active TFs associating with a particular ATAC bead. Reassuringly, this probability correlates significantly with the GRO-seq signal (Spearman's $r = 0.57$, $p < 10^{-4}$)[12], as visualised by a heatmap showing the number of ATAC beads in bins according to their percentile ranks in these measures (Fig. 6.4B). This result is in agreement with the hypothesis: it indicates that the frequency of TF binding can provide a first-order approximation of the transcriptional activity at individual REs.

[10] GRO-seq is a technique which measures the density of actively transcribing RNA polymerases (RNAPs) along chromatin. In this method, these active RNAPs are allowed to "run-on" and generate nascent RNAs using brominated nucleotides. The resulting RNAs are extracted using suitable antibodies, reverse-transcribed into complementary DNAs, and then sequenced and aligned to a reference genome. Here, GRO-seq signal is binned at 1 kbp resolution to facilitate comparison with simulation data.

[11] Activators are a class of TFs – see footnote 1.

[12] Spearman correlation coefficient is used here and for other correlations presented below as it is less sensitive to outliers and more appropriate when there is a non-linear relationship.

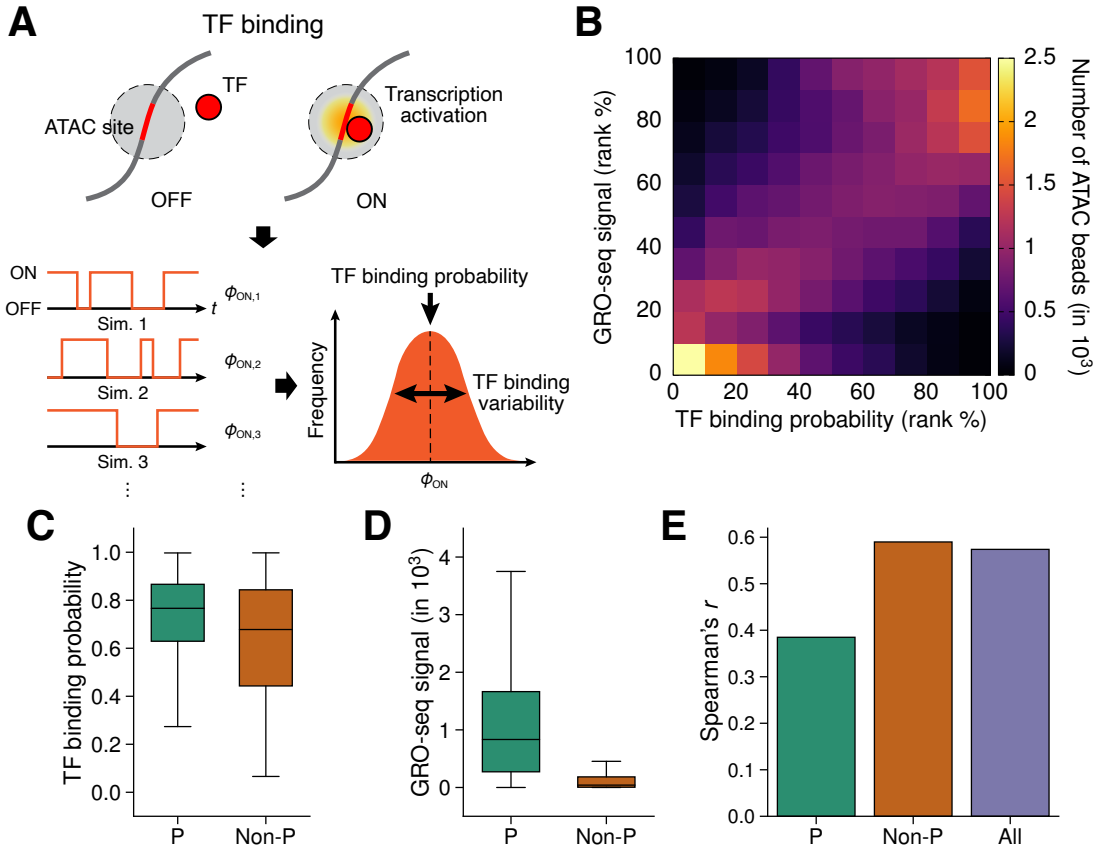

Figure 6.4: Predicting the transcriptional activity of each ATAC bead. (A) Schematics explaining how TF binding at an ATAC bead in simulations is used to predict the transcriptional output of chromatin within the bead. Here, an ATAC bead is envisaged to be in a poised state (OFF) unless a TF comes within its contact radius ($r_c = 3.5\sigma$; grey circle), upon which it becomes activated (ON). From recording a time series of TF binding activity, the fraction of time ϕ_{ON} where at least one TF associates with an ATAC bead is determined for each simulation run, and a distribution is constructed from ϕ_{ON} across different runs. The overall TF binding probability and variability are then defined respectively as the average and standard deviation of this distribution. (B) A heatmap showing the number of ATAC beads in bins according to their percentile rank in the GRO-seq signal against their rank in the TF binding probability (Spearman's $r = 0.57$, $p < 10^{-4}$). (C and D) Boxplots comparing (C) the TF binding probability and (D) the GRO-seq signal between ATAC beads that are mapped to promoters (Ps) and those that are not (non-Ps). The difference between Ps and non-Ps is statistically significant in both cases (Mann-Whitney U test, $p < 10^{-4}$). (E) A bar chart reporting the Spearman's r for the correlation between the TF binding probability and the GRO-seq signal for Ps, non-Ps, and all ATAC beads.

While the model treats all REs equally (i.e., they are all described by a single species of ATAC bead), it is worth examining whether there are differences among them regarding their level of transcription, as traditionally these elements are classified into promoters and enhancers. To make progress, ATAC

beads corresponding to gene promoters are identified by comparing the beads' genomic coordinates with those of promoters listed in the Eukaryotic Promoter Database [211]. For simplicity, beads are grouped into those that are mapped to promoters (Ps) and those that are not (i.e., non-promoters, or non-Ps), which include enhancers (and also a small proportion of other TF binding sites, such as insulators). Simulations predict that the TF binding probability is statistically higher for Ps than non-Ps (Mann-Whitney U test[13]: $p < 10^{-4}$; Fig. 6.4C), and this is matched with a higher GRO-seq signal for Ps ($p < 10^{-4}$; Fig. 6.4D). These results are consistent with the correlation discussed above, and the fact that there is a higher level of activity at Ps is reasonable given that many of them are directly responsible for initiating the transcription of protein-coding genes.

Interestingly, the correlation between the TF binding probability and GRO-seq signal strengthens if one considers non-Ps only ($r = 0.59$, $p < 10^{-4}$), whereas it deteriorates for Ps alone ($r = 0.38$, $p < 10^{-4}$; Fig. 6.4E). The improved correlation for non-Ps suggests that for many enhancers, their transcriptional activity is largely dictated by generic TF binding and is probably less regulated. On the other hand, the lower correlation for Ps indicates that, while TF binding is important, there are additional mechanisms governing transcription at these loci which are not captured by the model. In other words, these sites are more tightly regulated, for example, through biochemical pathways not represented here. This finding is also in concordance with modelling work suggesting that there is an intrinsic "activity strength" associated with each RE [212].

Another quantity examined here is the standard deviation of the distribution defined above, which measures the variation in the TF binding frequency (or TF binding variability) at an ATAC bead within the simulations conducted (Fig. 6.4A). Given that TF binding correlates with transcriptional activity, it is natural to envisage that the variability in binding is linked to fluctuations in transcriptional output, or transcriptional noise, within a population of cells. Though rapidly improving, single-cell transcriptomic experiments (e.g., single-cell RNA-seq) interrogating stochasticity in gene expression across the whole genome remain technically challenging. The simulation results presented below on TF binding variability (Section 6.2.4) provide predictions that can be validated in the future.

[13]See footnote 9.

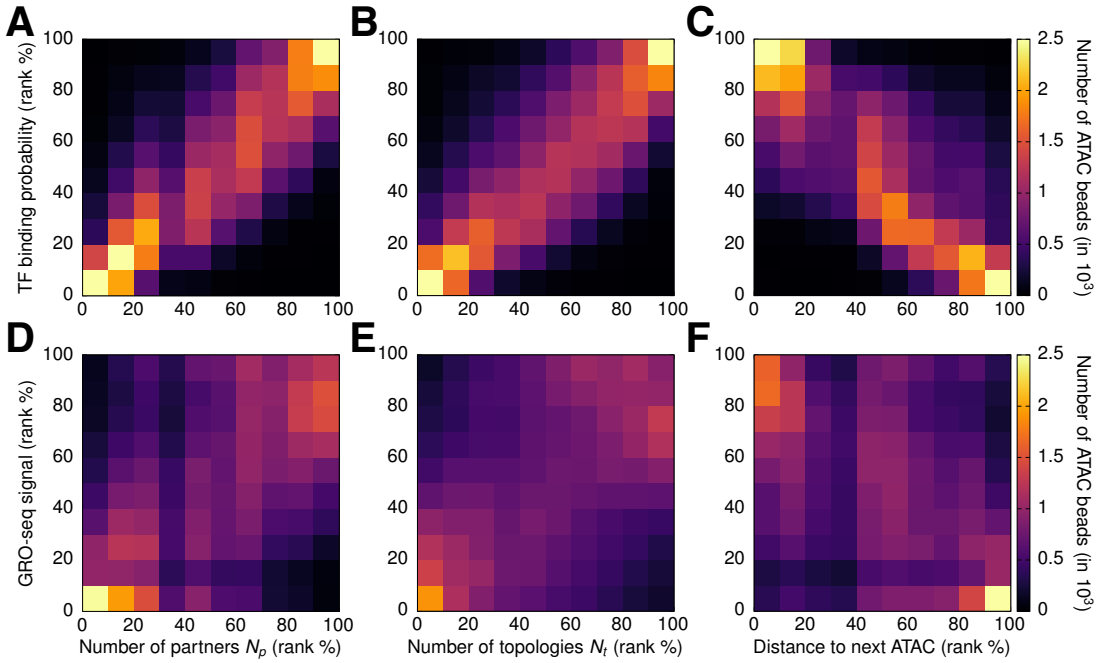

Figure 6.5: Correlating the structural properties and transcriptional activity of ATAC beads. (A–C) Heatmaps reporting the number of ATAC beads in bins according to their percentile rank in the TF binding probability against their rank in (A) the number of partners, (B) the number of topologies, or (C) the genomic distance to the nearest ATAC bead. (D–F) Similar to (A–C), but with their rank in the GRO-seq signal in place of the TF binding probability. The Spearman's r for all six heatmaps, in their display order, are 0.80, 0.77, −0.75, 0.47, 0.46, and −0.32; $p < 10^{-4}$ in all cases.

6.2.3 Linking Structure to Transcriptional Activity

Equipped with information on the structure and transcriptional output of each ATAC bead, one can turn to establish the connections between these two aspects. To identify potential links, structural observables defined above (Section 6.2.1) are correlated with measures related to transcription (Section 6.2.2), with a focus on the TF binding probability (or the mean transcriptional activity) in this section.

Notably, it is found that the number of FAPs of an ATAC bead correlates significantly with its TF binding probability (Spearman's $r = 0.80$, $p < 10^{-4}$; Fig. 6.5A) and GRO-seq signal ($r = 0.47$, $p < 10^{-4}$; Fig. 6.5D), suggesting that a bead is more frequently transcribed if it has more partners. This phenomenon can be explained by the bridging-induced attraction (BIA; see Section 2.3.1): a higher number of FAPs increases the chance of (TF-mediated) looping between the ATAC bead and its partners, thus raising the local density of binding sites. This effect, in turn, attracts more TFs to the region, reinforcing the loops and TF binding at the bead. Since the number of FAPs correlates with the number

of topologies, the latter also correlates with the binding probability ($r = 0.77$, $p < 10^{-4}$; Fig. 6.5B) and GRO-seq signal ($r = 0.46$, $p < 10^{-4}$; Fig. 6.5E).

Recent work has found that the transcriptional activity of an RE is partly determined by its genomic distance from other REs, with a higher activity when elements are closer together [196, 212, 213]. The data here corroborate this observation, showing that the distance of an ATAC bead from the next one anti-correlates with the bead's TF binding probability ($r = -0.75$, $p < 10^{-4}$; Fig. 6.5C) and GRO-seq signal ($r = -0.32$, $p < 10^{-4}$; Fig. 6.5F). This result can be explained by the decay in the contact probability $P_c(s)$ between two loci on a polymer as a function of their contour separation s (i.e., $P_c(s) \sim s^{-\alpha}$; see Section 2.1.1). When ATAC beads are near each other, their looping probability increases and thus TF binding becomes more likely.

To further illustrate the correlations discussed above, I examine chromatin regions with a hyper regulatory activity known as "super-enhancers" (SEs), which have been a topic of intense research in recent years [214, 215]. SEs are conventionally defined as regions which encompass a series of closely-spaced REs and exhibit an unusually high level of ChIP-seq enrichment in master TFs[14], the Mediator complex[15], and/or the histone mark H3K27ac [216, 217]. While these regions are functionally known for controlling cell identity and disease development, their 3D structure is less characterised, and its relation to function is still unclear. Here, the structural and transcriptional aspects of SEs are investigated more closely using the simulation data for the ATAC beads.

To proceed, ATAC beads within SEs are identified using a list of SEs (specific to the GM12878 cell line) retrieved from the dbSUPER database [218]. This procedure gives results in agreement with previous work; in particular, SE-associated ATAC (SE-ATAC) beads have a higher coverage of the H3K27ac mark in their neighbourhood, smaller genomic distance to their nearest neighbouring ATAC site, and higher transcriptional activity (quantified by GRO-seq) compared to those not belonging to SEs (non-SE-ATAC beads; Figs. 6.6A–C). More importantly, simulations predict that SE-ATAC beads have more FAPs and a higher number of topologies (Figs. 6.6D and E). Consistent with the correlations reported above, they also have a greater TF binding probability (Fig. 6.6F), which

[14]Master TFs, such as Oct4, Sox2, and Nanog, are a subset of TFs which are key to maintaining pluripotency of stem cells and to cell differentiation.

[15]Mediator is a protein complex which facilitates gene transcription by interacting with TFs and RNA polymerases.

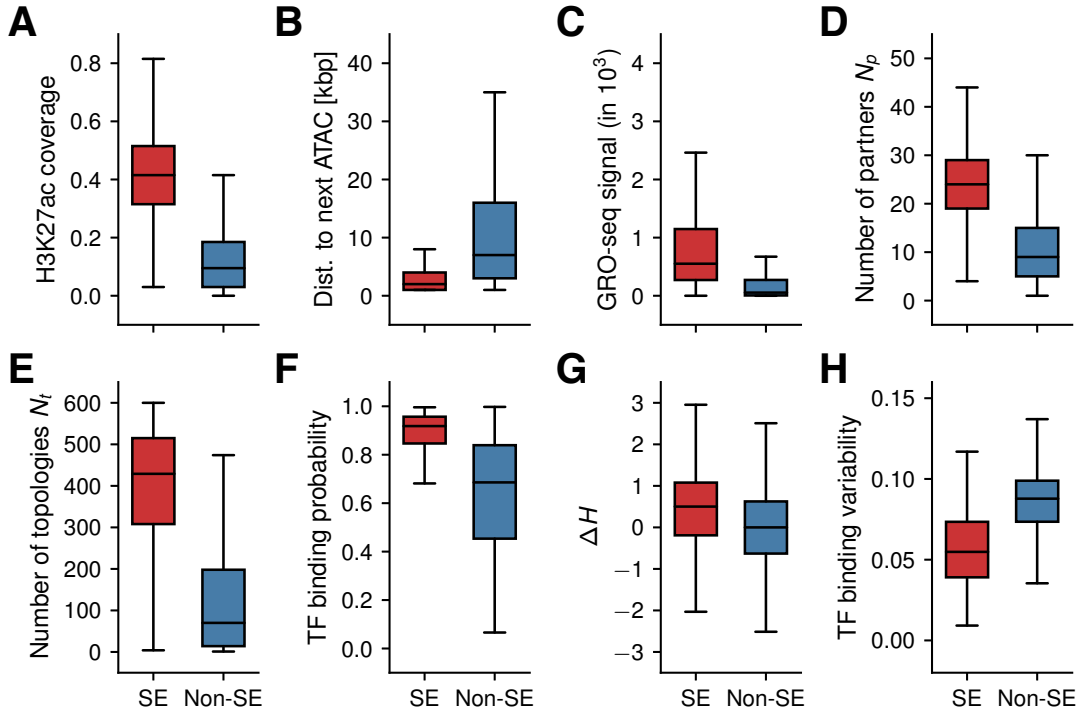

Figure 6.6: Structural and transcriptional properties of ATAC beads within super-enhancers (SEs). Here, boxplots compare the following quantities between beads within an SE (red) and those outside (Non-SE; blue): (A) the fractional coverage of H3K27ac within a 200 kbp window centred at the bead of interest; (B) the distance to the nearest ATAC bead; (C) the GRO-seq signal; (D) the number of partners; (E) the number of topologies; (F) the TF binding probability; (G) the deviation ΔH of the diversity score from the expected score; and (H) the TF binding variability. The difference between SE- and non-SE-ATAC beads in all quantities is statistically significant (Mann-Whitney U test; $p < 10^{-4}$).

is in line with the GRO-seq data. These results highlight the fundamental trend that establishing connections with more REs can elevate transcriptional output.

Interestingly, SE-ATAC beads typically have an above-average diversity score as quantified by ΔH [Eq. (6.6); Fig. 6.6G], indicating that conformations are more evenly distributed among the observed topologies for these beads compared to non-SE-ATAC beads. This finding can be explained by the fact that many partners of SE-ATAC beads are proximate in genomic distance, enabling different topologies to be sampled more easily and equally (e.g., due to a lower entropic cost of forming shorter loops). Additionally, SE-ATAC beads have markedly lower TF binding variability (Fig. 6.6H), suggesting that there is less transcriptional noise at SEs. This phenomenon can be understood as follows: since these regions typically involve many interactions with nearby regulatory partners, they are

almost always being transcribed; as a result, the fluctuation in the TF binding probability will be small.

6.2.4 Linking Structure to Transcriptional Variability

Thus far, the analysis mainly focusses on the role of TF-mediated chromatin bridging, an integral component of the HiP-HoP model, in relating the structural to the transcriptional properties of ATAC loci. However, the role played by loop extrusion (LE), another key ingredient of the model and a major player in driving chromatin folding, has not been explored. Previous experiments perturbing factors associated with LE, such as CTCF and cohesin, have not provided clear answers on the link between LE activity and transcription, with only moderate changes in gene expression observed [57–59]. Here, rather than investigating the *average* transcriptional output, I examine the impact of LE activity on transcriptional *variability* (or TF binding variability), which has been less studied.

To quantify the degree of LE activity at each ATAC bead, two observables are considered: the frequency of an extruder binding to the bead (i.e., cohesin occupancy) and the genomic distance to its nearest CTCF bead. The former provides a direct measure of LE activity, whereas the latter is related to LE since CTCFs are thought to act as barriers for extrusion. Interestingly, both observables show a statistically significant correlation with the TF binding variability (Fig. 6.7): the former with Spearman's $r = 0.26$ ($p < 10^{-4}$) and the latter with $r = -0.30$ ($p < 10^{-4}$). These correlations indicate that more LE activity can lead to an increase in transcriptional noise within ATAC loci. Mechanistically, it could be that the extruded loops arise in different places along chromatin in different simulations, thereby generating fluctuations in TF binding and transcriptional output.

To verify this trend, additional simulations are conducted with extruders removed, mimicking experiments which degrade cohesin using an auxin-inducible degron system[16] [58] (i.e., "cohesin degron" simulations; Fig. 6.8A). If the trend holds, the elimination of extruders should reduce the amount of noise in the level of

[16]The auxin-inducible degron system is a technology derived from plants that allows effective removal of a protein of interest in living cells upon the introduction of auxin molecules. In this method, the target protein is tagged with a degron (a domain or short amino acid sequence that is implicated in regulating protein degradation) that can be bound by auxin. This binding leads to ubiquitination of the protein and subsequently its removal by proteasomes.

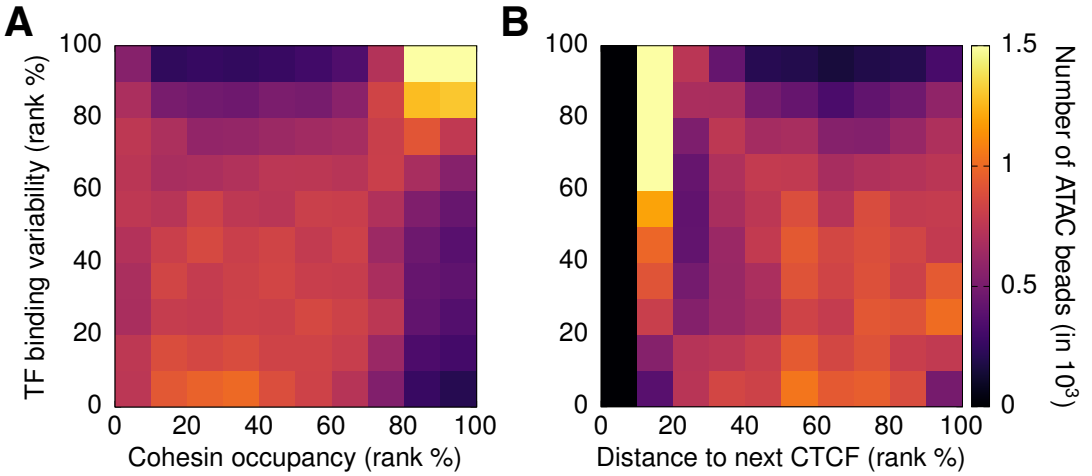

Figure 6.7: Correlating loop extrusion (LE) activity at ATAC beads with their transcriptional variability. Heatmaps showing the number of ATAC beads in bins according to their percentile rank in the TF binding variability against their rank in (A) the frequency of cohesin occupancy at the bead or (B) the distance to the nearest CTCF bead. The Spearman's r for both cases are 0.26 and -0.30, respectively, with $p < 10^{-4}$.

transcription (or TF binding). In practice, the degron simulations are performed for the region 142.5–189.0 Mbp in chromosome 1, and conformations are sampled in the same manner as before (see Section 6.1). The transcriptional properties of ATAC beads in these simulations are then compared to those in the "wild type" (WT) simulations.

Contact maps from the degron simulations are in line with those from experiments which removed cohesin [58] or knocked out its loading factor [59]. Comparing with maps from the WT simulations, these maps exhibit a loss of interactions within TADs and of LE-related architectural features, such as "dots" and "stripes", that are commonly found at TAD boundaries (Figs. 6.8B and C).

Fig. 6.8D displays a scatterplot comparing the TF binding variability at the ATAC beads in the WT and degron simulations. Strikingly, there is an overall reduction in the variability from the WT to the degron condition. This result further supports the trend mentioned above that a decrease in LE activity can reduce transcriptional noise across a population of cells. This marked reduction in the TF binding variability is in contrast to the modest change in the TF binding probability between these conditions (Figs. 6.8E and F). More specifically, the

[17]The Wilcoxon signed-rank test is similar to the Mann-Whitney U test, but for the case where the two samples are paired or related with the same sample size.

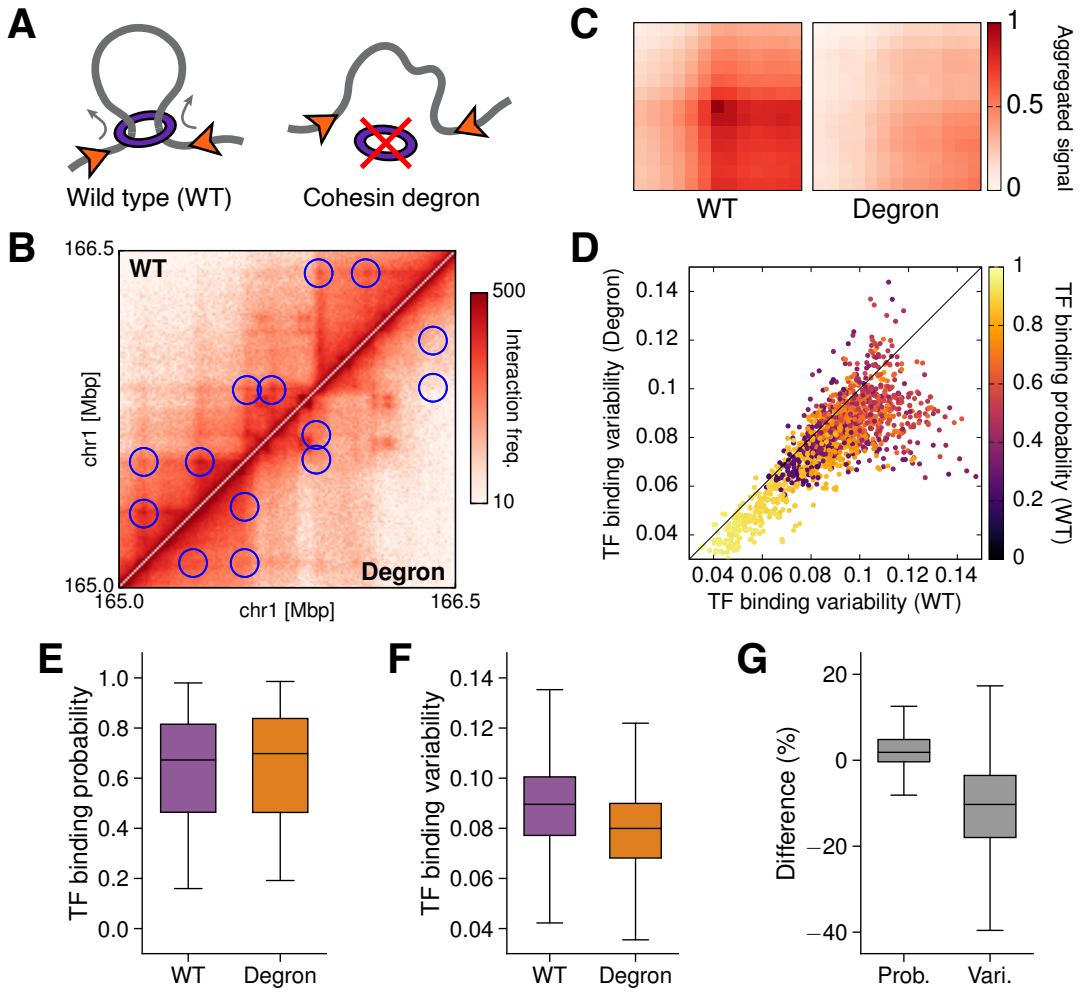

Figure 6.8: The effects of cohesin removal on transcription. (A) An illustration explaining the simulation setup. "Wild-type" (WT) simulations are done as described in Section 6.1, whereas cohesin-depleted (or cohesin degron) simulations are performed with all extrusion springs removed from the chromatin fibre. (B) A comparison of the contact maps generated from the WT (upper triangle) and degron (lower triangle) simulations. These maps are produced according to the procedure discussed in Section A.1. Interaction frequencies are shown in log scale to aid visualisation. Circles highlight the disappearance of "dots" (which typically correspond to loop anchors) at TAD boundaries upon cohesin removal. (C) Aggregated peak analysis (APA) plots comparing the relative interaction strength of these dots in the WT and degron contact maps. Dots are identified based on the possible pairs of chromatin loci with convergent CTCF motifs. (D) A scatterplot comparing the TF binding variability of each ATAC bead in the WT and degron conditions. The diagonal line represents the case where there is no change between conditions. The colour scale shows the average TF binding probability of each bead in the WT. (E and F) Boxplots comparing (E) the TF binding probability and (F) TF binding variability between the WT and degron conditions. The difference between WT and degron in these quantities is statistically significant (Wilcoxon signed-rank test[17]: $p < 10^{-4}$). (G) A boxplot showing the percentage difference in the binding probability (prob.) and variability (var.) between the two conditions.

median of the percentage change in the binding probability when extruders are removed is only around 2%, whereas it is about −10% for the binding variability (Fig. 6.8G).

6.3 Summary and Discussions

In this chapter, I have employed the highly predictive heteromorphic polymer (HiP-HoP) model [194] to simulate all chromosomes individually within the human genome (GM12878 lymphoblastoid cell line), with the aim of elucidating mechanistic links between genome structure and transcription. The HiP-HoP model incorporates two well-established mechanisms for driving chromosome organisation – transcription factor (TF) binding and loop extrusion (LE) – and accounts for the variation in the local compactness of the chromatin fibre. The focus here has been on *cis*-regulatory elements (REs; i.e., promoters and enhancers), as they are critical for controlling gene expression. For simplicity, these elements have been examined by proxy via ATAC beads, or beads with high ATAC-seq signal, since they typically have high DNA accessibility.

I have conducted a three-part analysis to understand the structure-transcription relation for the ATAC beads. In the first part, the 3D chromatin interactions related to each ATAC bead have been characterised. Simulations reveal that many ATAC beads have several frequently associating partners (FAPs) – i.e., other ATAC beads with which they often interact – and there are typically multiple ways, or topologies, regarding how they network with their partners (Fig. 6.2). Notably, the number of topologies increases with the number of partners, suggesting that beads can usually explore most of their physically possible topologies, rather than only a few, within the ensemble of simulated structures. This result is consistent with previous studies showing substantial cell-to-cell stochasticity in the local chromatin folding [183].

Intriguingly, there is also large variation in how the simulated conformations are distributed across the detected topologies. Some ATAC beads have most of their observed structures mapped to a few topologies, whereas others have their conformations spread more evenly among all topologies. A diversity score is devised to quantify this effect, and it is shown to increase with the number of partners (Fig. 6.3), a trend that can explained by a simple combinatorial analysis.

In the second part, the transcriptional properties of the ATAC beads have been examined. Apart from considering experimental data such as GRO-seq, I have investigated whether simulations can offer predictions on the level of transcription. In this respect, the frequency of active TF binding at the ATAC beads (or TF binding probability) has been found to correlate significantly with the GRO-seq signal (Fig. 6.4), indicating that it is suitable for inferring transcriptional activity. Of note, the strength of this correlation weakens if one only considers the ATAC beads mapped to gene promoters. This suggests that the transcriptional activity at these elements are more tightly controlled, and there are additional mechanisms at work which are not captured by the model. The association of TF binding probability with transcriptional activity has also led to the conjecture that the variation in TF binding across different simulation runs (or TF binding variability) can be used to estimate transcriptional variability, or transcriptional noise.

In the final part, connections have been drawn between the structural and transcriptional aspects of the ATAC beads. Here, two structural observables show a clear correlation with the TF binding probability of a bead: its number of partners and its distance from the nearest ATAC site (Fig. 6.5). The correlation with the former can be explained by the bridging-induced attraction (BIA): the higher number of partners enables more chromatin looping, which increases the local density of binding sites and thus attracts more TFs to the region. The latter can be attributed to the increase in contact probability between two beads as their genomic separation becomes smaller (due to the polymeric nature of the chromatin fibre), which in turn facilitates TF binding. These correlations have been further exemplified from the study of super-enhancers (SEs; Fig. 6.6). In particular, ATAC beads within SEs, which are highly transcribed chromatin regions, have more FAPs and topologies than those that are outside of SEs. They also have a higher diversity score and a lower TF binding variability.

Simulations also predict that LE activity is related to transcriptional variability, with higher activity associated with more stochasticity in transcription. This trend is inferred from the statistically significant correlation between the TF binding variability of an ATAC bead and the frequency of extruders (or cohesins) moving through the bead, as well as the bead's genomic distance from its nearest CTCF site (Fig. 6.7). Furthermore, "degron" simulations where extruders are removed also support this finding, as they show a marked decrease in binding variability, but only a modest change in the mean binding level (Fig. 6.8).

All in all, the analysis demonstrates that there are at least two major connections between the 3D structure of gene REs and their transcriptional output. The primary connection is concerned with the spatial interactions between REs and their mean expression level. Simulations predict that frequently transcribed genes tend to be those which interact promiscuously with many REs, as demonstrated from the example of SEs. This connection fits well with the idea of transcription hubs or factories, where genes are activated by colocalising with structural foci containing polymerases and activating factors [203]. The frequent associations between REs detected here are likely to be mediated by these foci (e.g., via BIA); thus, an increase in the number of associations helps keep a gene close to a factory, thereby promoting its likelihood of being activated.

The secondary, perhaps more surprising, connection is that between LE and transcriptional noise. While previous studies have argued that LE facilitates 3D contact between gene promoters and their distal regulatory partners [51], here simulations suggest a more complicated picture: extrusion may indeed drive the interactions between some REs, but it may also disrupt other interactions, resulting in larger variation in the interaction and expression patterns.

This connection to noise is also of interest from the perspective of evolution. Recent work has shown that "younger" genes, as determined from phylogenetic analyses, tend to have larger stochasticity in their expression level [219]. Intriguingly, CTCF-cohesin mediated LE seems to be only a recently introduced mechanism in regulating genome architecture; for instance, evidence of its regulation on TADs has mostly been confined to vertebrates [220]. These observations, along with the results presented here, lead one to believe that LE activity, transcriptional noise, and gene age are all related, and the functional significance of these connections can be examined more closely in the future.

The amount of information generated from the simulations in this chapter is immense. There are many research questions which can be studied in the future using this pan-genomic data set, of which I mention a few here. First, in Section 6.2.1, it was found that regulatory loci exhibit large variability in their structural diversity (i.e., H score). The reason behind this remains unclear and should be investigated further, for example, by looking at the relative (genomic) spacing between individual elements. In addition, motivated by the above-average H score for SEs, which are important for developmental genes, it is suspected that this structural property may have functional implications: perhaps some classes of genes prefer interacting via only a few designated topologies (for most

of the time), whereas others are less concerned with whom they interact. Such functional relevance can be examined by correlating the H score with metrics from gene ontology.

Second, while the work above considered the transcriptional activity of each RE separately, one can also explore the correlation of this activity between elements across different simulation runs and construct a transcriptional network. Previous simulation work has examined this, albeit using a simpler TF model and focussing only on a few chromosomes, and demonstrated that the resulting network is typically complex and small-world when the number of available TFs is low [196]. It would be interesting to see whether the network based on the data set here also displays this attribute. Furthermore, it would be desirable to understand how this transcriptional network relates to the structural network (i.e., the FAPs and topologies), as it may help uncover additional links between structure and function.

Finally, the simulations performed here are only for a single cell line (GM12878). To fully appreciate the connections between structure and function, one should also investigate how they alter between different cell lines, such as during reprogramming [192], and so additional genome-wide simulations will be needed in the future. Moreover, this comparative analysis can be extended to pathological conditions to explore the effect of genomic rearrangement and deletion on these connections.

7

Conclusions

In this thesis, I have employed coarse-grained polymer models and performed molecular dynamics simulations to determine some of the biophysical principles regulating genome architecture and function. The work conducted here has demonstrated that relatively simple polymer models can already provide immense details about the essential mechanisms at work in different biological situations.

In Chapter 4, I have explored the principles of how three-dimensional (3D) chromatin folding affects the dynamics and patterns of epigenetic modifications on histones along chromatin. Specifically, I have shown that coupling the spatial interactions of chromatin with histone modification dynamics, mediated respectively by epigenetic reader and writer proteins, provides an effective pathway to form an ordered epigenetic landscape along chromatin. Simulations also suggest that epigenetic memory can naturally arise as a result of this coupling due to hysteresis. Crucially, I have demonstrated that the "genomic bookmarking" mechanism, whereby proteins binding to sequence-specific sites attract readers and writers to spread a particular epigenetic mark, enables the development of heterogeneous epigenetic domains, with different chromatin regions enriched in different marks. Strikingly, this mechanism allows domain

patterns to be remembered across replication, and removal of bookmarking proteins destabilises the domains. Moreover, simulations based on the mechanism have successfully recapitulated the epigenetic patterns of the right arm of chromosome 3 in *Drosophila*.

In Chapter 5, I have dissected mechanisms governing the large-scale genome rearrangement observed in cellular senescence, where the conventional layering of euchromatin and heterochromatin is disturbed. The work has highlighted that heterochromatin- and nuclear lamina-mediated interactions are key players in this process. Varying these interactions in simulations gives rise to different polymeric phases, with conformations highly similar to those seen in a growing or senescent nucleus. I have demonstrated that the change in locality of chromatin interactions found between growing and senescent conditions can be understood by mapping the structures of these two conditions to different phases. Simulations have also captured the large variability in lamina-associated domains in the growing condition, as found in experiments. Importantly, I have shown that the structural reorganisation between growing and senescence is associated with an abrupt transition with hysteresis. This result indicates that the conformation in senescence is thermodynamically in a locally highly stable condition, thus providing a reason why cells in this state are unlikely to readopt the conformation typical to the growing state.

In Chapter 6, I have conducted a genome-wide study to decipher the mechanistic connections between chromatin folding and transcriptional activity. I have generated a collection of 3D structures of all chromosomes in the human genome at an unprecedented resolution. Utilising this rich data set, I have examined both the structural and transcriptional properties of individual gene regulatory elements to deduce how these two aspects are linked. I have found that the network of spatial interactions between elements plays a significant role in determining the level of transcription; in particular, a higher number of interactions correlates with an increase in expression. This result is consistent with the conjecture that gene expression typically occurs at structural foci known as transcriptional hubs or factories. Simulations have also provided evidence suggesting that loop extrusion activity is more strongly related to the cell-to-cell variability in transcription than to the mean transcriptional output. This outcome can be rationalised by recognising that extrusion can mediate interactions between different regulatory elements in different simulations or cells, thereby giving rise to different expression patterns.

Overall, these chapters have elucidated distinct principles for driving chromosome arrangement and function in several contexts. A key direction for future work is to understand the joint effects of these separate principles, which can be investigated by combining the various polymer models employed in this thesis. For example, it would be of interest to integrate the epigenetic read-write model from Chapter 4 with the highly predictive heteromorphic polymer (HiP-HoP) model from Chapter 6. The latter model posits that the local compaction of the chromatin fibre depends on histone modifications associated with actively transcribed regions; hence, understanding how epigenetic dynamics alter these modifications, which in turn reshape local chromatin packing, may help uncover additional mechanisms regulating gene expression. Another example would be to incorporate the effect of the nuclear lamina (Chapter 5) into the HiP-HoP model. Lamina-mediated interactions typically sequester (transcriptionally inactive) heterochromatin to the nuclear periphery while leaving (active) euchromatin within the interior. It would be informative to examine how this arrangement changes the way by which regulatory elements within euchromatic regions interact with one another and the transcriptional activity of these elements.

Finally, the results presented in this thesis are obtained purely from simulation models which only capture parts of the enormous complexity associated with eukaryotic genomes. It is my hope that this work would inspire other researchers to perform experiments to validate and further explore the principles discussed here, leading to more discoveries about genome architecture and function, and that more simulation work would follow as a result. After all, the collective effort from experimentalists and theorists has been, and will continue to be, fundamental to the advancements in genome research.

A

Further Details on the Genome-Wide Analysis of Regulatory Domains

In this appendix, I provide additional details about the parameters used in the highly predictive heteromorphic polymer (HiP-HoP) simulations for analysing the structural and transcriptional properties of gene regulatory elements (REs) across the human genome (Chapter 6). In Section A.1, I will discuss the procedure for choosing and validating the parameters used for the main (production) simulations. In Section A.2, I will provide tables listing the parameters used in the test simulations for validating the model and those in the main simulations.

A.1 Validation of Model Parameters

As discussed in Section 6.1, the HiP-HoP model involves multiple components and parameters. It is, therefore, important to validate the model to ensure that it gives results consistent with existing experimental data before employing it to simulate the entire genome. To this end, I conduct test simulations on a

smaller chromosome region to examine the effects of varying the parameters and to find a suitable set that yields predictions broadly in agreement with Hi-C experiments. The simulations are done for the region 6.5–16.5 Mbp (a 10000-bead segment) in chromosome 19, and there are two main reasons for selecting this region. First, since this work focusses on the spatial interactions between gene REs, chromosome 19 is an ideal candidate as it has a high gene density and thus provides more data per base pair than other chromosomes. Second, the specified region still contains both transcriptionally active and inactive domains (Fig. 6.1C), so it is possible to determine the effects of different TF species.

For this test region, a total of 20 sets of simulations are performed with different parameter values (see Table A.1). These explore the effects of changing the total number of TF beads N_{TF} and extruders N_{ex}, the switching time τ_{sw} (or its rate $k_{\mathrm{sw}} = \tau_{\mathrm{sw}}^{-1}$) of the TFs, as well as the extrusion τ_{ex} and unbinding time τ_{off} (or their rates $k_{\mathrm{ex}} = \tau_{\mathrm{ex}}^{-1}$ and $k_{\mathrm{off}} = \tau_{\mathrm{off}}^{-1}$) of the extruders. The ratio of the number of active N_{act}, Polycomb-like N_{poly}, and heterochromatin-like TFs N_{het} is fixed at $1/4 : 1/8 : 5/8$ unless otherwise specified. Parameters for each consecutive set of simulations are determined manually based on results from previous sets, with the aim of progressively achieving better agreement with experimental data.

More specifically, 20 replicas are generated for each set of parameters. For computational efficiency, replicas are split into groups of four, and those in the same group are simulated simultaneously by joining them together to form a longer polymer, with a 1000-bead "spacer" region demarcating individual replicas (i.e., the polymer has 45000 beads in total). This procedure has been considered before [195], and it helps maintain the chromatin density at a realistic level without significantly reducing the box size, which may lead to undesired artefacts from periodic boundaries.

To assess the ability of these simulations to predict experimental data, contact maps are generated from simulations and compared to those from Hi-C [23]. These maps are constructed in a way inspired by the crosslinking step in Hi-C for sampling pairwise chromatin interactions [194]. As mentioned in Section 6.1.4, the conformation of the chromatin fibre is sampled every $2 \times 10^3 \tau$ over a period of $10^5 \tau$. For each sampled structure, the following procedure is conducted: first, a pair of chromatin beads, say i and j, are selected randomly with their separation r_{ij} computed. Then, the pair is accepted and registered as a "read" in the contact

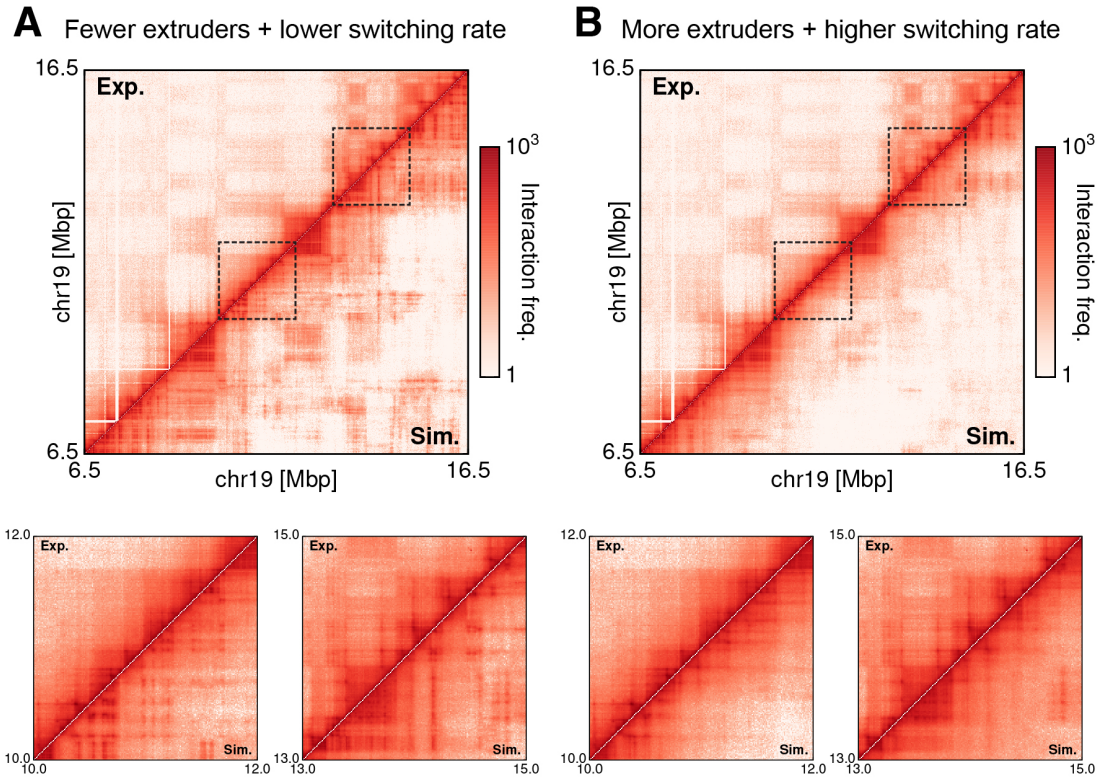

Figure A.1: Comparing simulated and Hi-C contact maps for the test region 6.5–16.5 Mbp in chromosome 19. (A) *Top*: A heatmap comparing the frequency of interaction between chromatin loci as observed from Hi-C [23] (upper triangle) and from a set of simulations (lower triangle) with fewer extruders ($N_{\mathrm{ex}} = 300$, or 6.67 extruders/Mbp) and a lower switching rate ($k_{\mathrm{sw}} = 10^{-5}\tau^{-1}$). *Bottom*: enlarged views of two specific regions, 10–12 Mbp and 13–15 Mbp, corresponding to the dashed squares in the full map. (B) Similar to (A), but for a set of simulations with more extruders ($N_{\mathrm{ex}} = 450$, or 10 extruders/Mbp) and a higher switching rate ($k_{\mathrm{sw}} = 10^{-3}\tau^{-1}$). Contact maps are displayed at 10 kbp resolution and are plotted in log scale to aid visualisation and comparison. Simulated maps are generated from sampling conformations every $2\times10^3\tau$ over a period of $10^5\tau$ in 20 runs.

matrix with probability

$$p_{\mathrm{read}}(r_{ij}) = \exp\left(-\frac{r_{ij}}{r_c}\right), \tag{A.1}$$

where $r_c = 3.5\sigma$ is a distance threshold related to the separation within which crosslinking is effective. By repeating this procedure many times and over different structures and simulation runs, reads can pile up in a way similar to that in Hi-C. By fixing the total number of reads in the simulated map to be the same as that in a Hi-C map, one can directly compare the two without performing rescaling.

Fig. A.1 compares maps created using this approach with the experimental Hi-C map from Ref. [23]. The simulated maps are for two sets of parameters: one set for the initial runs (the initial set; Fig. A.1A), with fewer extruders ($N_{\mathrm{ex}} = 300$, or 6.67 extruders/Mbp) and a lower switching rate ($k_{\mathrm{sw}} = 10^{-5}\tau^{-1}$); the other set consistent with parameters used in the main simulations (the main set; Fig. A.1B), with more extruders ($N_{\mathrm{ex}} = 450$, or 10 extruders/Mbp) and a higher switching rate ($k_{\mathrm{sw}} = 10^{-3}\tau^{-1}$). While both maps qualitatively capture interaction patterns seen in Hi-C, there are noticeable differences between them. For instance, the map from the initial set displays more long-range interactions than that from the main set (and from Hi-C), with some interactions crossing domain boundaries (see enlarged views of the maps). This can be attributed to the lower switching rate, which allows TFs to remain bound to chromatin for longer and more distal interactions to become established [75]. Additionally, the "dots" and "stripes" which often appear at the boundaries of domains mediated by loop extrusion (LE) are more subdued in the initial set, and this is likely to be due to the combined effect of having fewer extruders, which is known to suppress these features [55], and having a lower switching rate, which interferes with domain boundaries by promoting longer-range interactions. Overall, the comparison suggests that the kinetics of TF switching and LE can have a significant impact on chromatin organisation, as observed in previous studies [47, 55, 74, 75].

To quantitatively identify one of the 20 parameter sets to be used in the main simulations, I consider the similarity between the simulated map for each set and the Hi-C map; in this way, the set showing the highest agreement is selected for the production runs. A simple way to quantify the similarity between maps is to compute the Pearson correlation of the maps directly; however, since the interaction probability $P_c(s)$ between two loci is strongly dependent on their genome distance s (i.e., $P_c(s) \sim s^{-\alpha}$; see Section 2.1.1), a direct correlation is usually not informative, as the resulting score predominately reflects this dependence. In particular, correlating two maps with very different interaction patterns can still achieve a reasonable score. Hence, a robust method should factor out this distance dependence while still having the power to discern differences between maps.

A "distance-insensitive" strategy to evaluate the agreement between simulated and Hi-C maps is to compare the pattern of contacts at the scale of topologically associating domains (TADs) between them. TADs are of interest to this work as they occur at a genomic scale similar to that of the interactions between REs (no

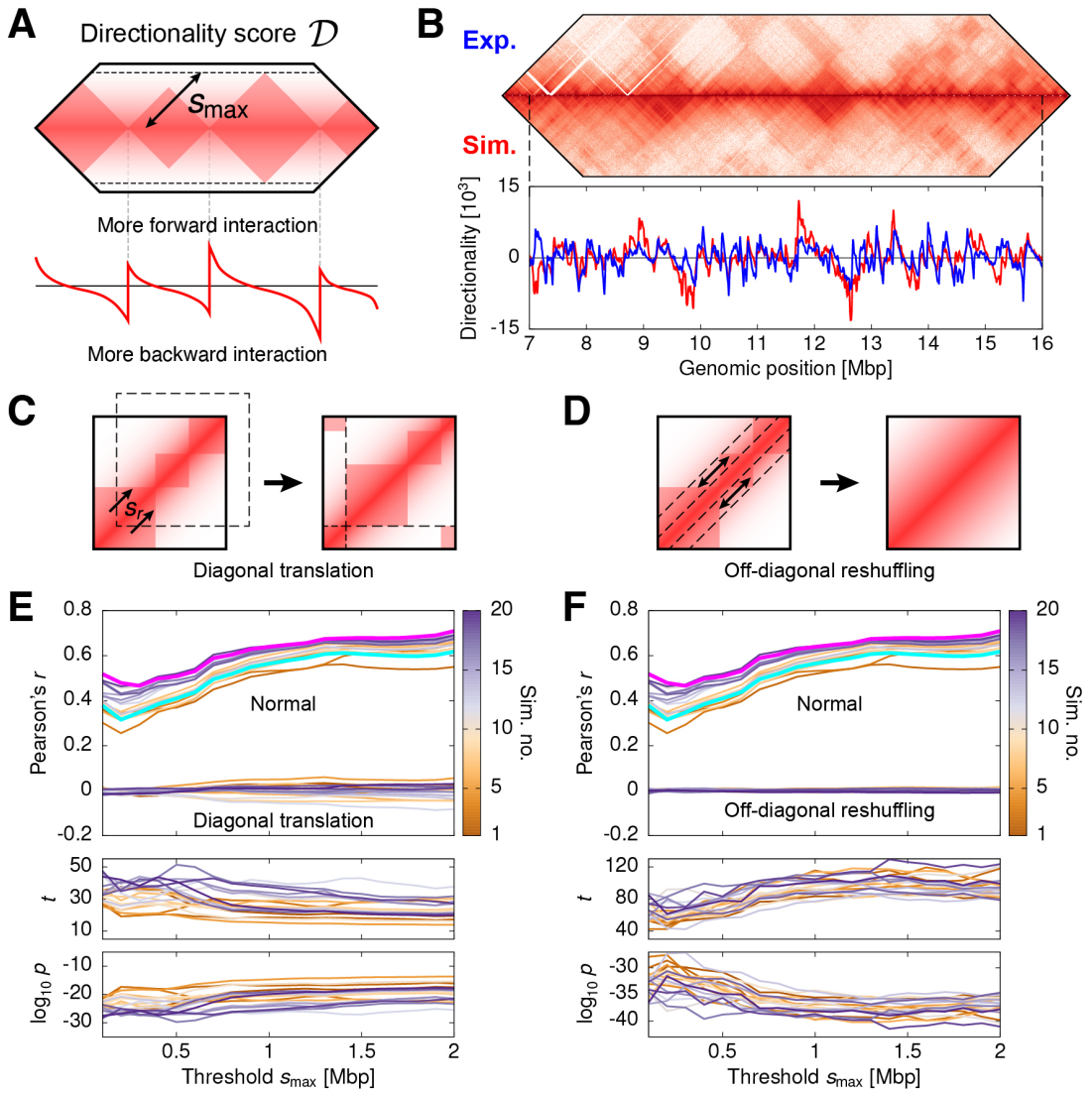

Figure A.2: Correlating simulated and Hi-C contact maps for the test region based on the directionality score $\mathcal{D}$. (A) An illustration explaining $\mathcal{D}$, which measures the aggregated frequency of interaction of a chromatin bin with other bins further ahead, subtracted by the interaction from behind (up to a genomic distance $s_{\max}$ from the diagonal). (B) Profile of $\mathcal{D}$ within the test region for a simulated map (red) and the Hi-C map (blue) from Ref. [23]. Here, $s_{\max} = 0.5$ Mbp, and the simulated map is created from the set of simulations using parameters consistent with those for the main simulations. The Pearson correlation coefficient between the two profiles is $r = 0.51$ ($p < 10^{-4}$). (C and D) Diagrams depicting the two methods for generating artificial contact maps based on the simulated maps (i.e., normal maps). (C) Diagonal translation: the map is shifted diagonally by s_r in genomic distance with periodic boundary conditions. (D) Off-diagonal reshuffling: elements within each off-diagonal stripe of the map are reshuffled randomly. (E and F) *Top*: Pearson correlation of $\mathcal{D}$ between the simulated maps (both normal and artificial) for each parameter set and the Hi-C map as a function of the threshold $s_{\max}$. Artificial maps are created from normal maps by (E) diagonal translation and (F) off-diagonal reshuffling. (*Continued at the bottom of the next page.*)

more than a few Mbps). One approach to detect TADs algorithmically from a contact map is to consider the fact that near their boundaries, interactions are highly favoured towards either the left or right of a chromatin segment [22, 46]. This bias is captured by a directionality score $\mathcal{D}$ (Fig. A.2A), which is defined for each chromatin bin (say bin i) in the map as

$$\mathcal{D}(i) = \sum_{j=i+\mathrm{bin}(s_{\min})}^{i+\mathrm{bin}(s_{\max})} c_{ij} - \sum_{j=i-\mathrm{bin}(s_{\max})}^{i-\mathrm{bin}(s_{\min})} c_{ij}, \tag{A.2}$$

where the first term on the right hand side represents the total number of interactions or reads to the right (between $s_{\min}$ and $s_{\max}$ in genomic distance), the second term represents the interactions in the corresponding region to the left, and c_{ij} is the number of interactions between bins i and j. The lower end threshold $s_{\min}$ is fixed at 20 kbp to avoid artefacts close the diagonal within the Hi-C map (here maps are at 10 kbp resolution), whereas the upper end threshold $s_{\max}$ is varied between 100 kbp and 2 Mbp. Note that because this score compares the aggregated interactions within regions either side of a bin up to the same genomic distance away, the distance dependence is effectively cancelled out. The similarity between the simulated map and the Hi-C map can then be determined based on the Pearson correlation of $\mathcal{D}$ in both maps (Fig. A.2B).

To verify the robustness of this approach, I also compute this score for artificial maps generated from rearranging the simulated maps (i.e., "normal" maps) by two different methods:

(i) Diagonal translation: the map is shifted diagonally forward by a randomly chosen genomic distance s_r, with periodic boundary conditions enforced (Fig. A.2C). In this way, most TADs are still preserved, but they are now misaligned with respect to the original map.

Figure A.2 ***(continued)*:** For each parameter set, three normal contact maps are generated, and each is employed to create 10 artificial maps (i.e., a total of 30). Average correlation scores are reported for both normal and artificial maps. The cyan line reports the average correlation (for the normal maps) for the initial parameter set, whereas the magenta line shows the correlation for the set employed in the main simulations. *Middle* and *bottom*: Welch's t-test statistic and the associated p value showing that the correlation scores for normal maps are significantly different from those for artificial maps.

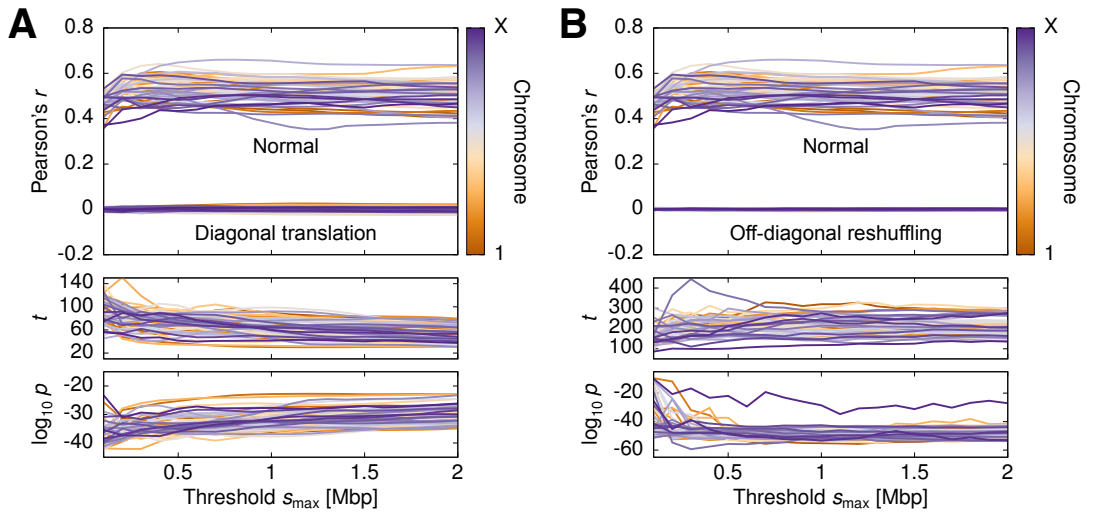

Figure A.3: Correlating the directionality score $\mathcal{D}$ between simulated and Hi-C contact maps for all simulated chromosome segments in the human genome. (A and B) *Top*: Pearson correlation of $\mathcal{D}$ between simulated (both normal and artificial) and Hi-C maps plotted as a function of the threshold $s_{\max}$. The normal simulated maps are constructed based on conformations from 300 independent simulation runs, with the same sampling frequency as before. Artificial maps are created from rearranging normal simulated maps by (A) diagonal translation or (B) off-diagonal reshuffling. *Middle* and *bottom*: Welch's t-test statistic and the associated p value showing that the correlation scores for normal maps are significantly different from those for artificial maps.

(ii) Off-diagonal reshuffling: bins within each off-diagonal stripe of the map, which covers interactions with the same genomic separation, are reshuffled randomly (Fig. A.2D). As a result, the decay in the interaction probability as a function of the genomic distance is preserved, but TADs are not.

Figs. A.2E and F show the Pearson correlation of $\mathcal{D}$ as a function of the threshold $s_{\max}$ between simulated maps (both normal and artificial ones) for each parameter set and the Hi-C map. For normal maps, the correlation score is generally higher for parameter sets whose simulations are conducted later (i.e., a higher simulation set number) than earlier sets. This reflects the fact that parameters are adjusted between sets in a way which improves how well they capture features from Hi-C maps. Importantly, parameter set 20 (magenta line in the figure) is the set with the highest correlation score, and therefore it is used for the main simulations. Moreover, artificial maps generated either by the translation or the reshuffling method do not show any correlation with the Hi-C map, with correlation scores significantly lower than those for the normal maps (Welch's t-test[1]; see Figs. A.2E

[1]The Welch's t-test is a statistical test on the null hypothesis that two data samples, which are Gaussian-distributed but with unequal variances, have the same mean. A larger t indicates that their means are more separated [181].

and F). This demonstrates that $\mathcal{D}$ is a robust statistic for comparing the similarity between contact maps.

This method of using $\mathcal{D}$ to quantify the similarity between simulated and Hi-C maps is also employed in the production runs for all of the simulated chromosome segments in the human genome (Fig. A.3). The Pearson correlation scores are mostly within the range from 0.4 to 0.6, and this is in contrast to scores of nearly zero for artificial maps created using the two methods discussed. Overall, this validation process provides support that the chosen parameters are reasonable and can yield results that are broadly in line with existing data.

A.2 Parameters for Test and Main Simulations

Two tables are presented in this section. Table A.1 lists the parameters used in the 20 sets of test simulations as discussed above. Table A.2 lists the parameters used for simulating individual chromosome segments of the human genome.

No.	$N_{\mathbf{TF}}$	$N_{\mathbf{ex}}$	$\tau_{\mathbf{sw}}$	$\tau_{\mathbf{ex}}$	$\tau_{\mathbf{off}}$	**Remarks**
1	4000	300	100000	500	40000	
2	5000	300	100000	500	40000	An extra inactive TF is introduced, and it binds to all inactive regions (i.e., H3K9me3 and H3K27me3). $N_{\mathrm{act}} = 1000$, $N_{\mathrm{pcmb}} = 500$, $N_{\mathrm{het}} = 1500$, and $N_{\mathrm{inact}} = 2000$. Direct chromatin-chromatin interactions are switched off.
3	5000	300	100000	500	40000	A single inactive TF is used, and it binds to all inactive regions. $N_{\mathrm{act}} = 1000$ and $N_{\mathrm{inact}} = 4000$.
4	4000	300	100000	500	40000	TF and chromatin interactions are switched on at the same time at $10^4\tau$.
5	4000	300	10000	500	40000	
6	4000	300	50000	500	40000	
7	4000	300	100000*	500	40000	*Switching time for active TFs is $\tau_{\mathrm{sw}}^{\mathrm{act}} = 10000$.
8	4000	300	100000	100	8000	

9	4000	300	100000	250	20000	
10	4000	450	100000	500	40000	
11	4000	450	1000	200	40000	
12	4000	300	1000	500	40000	
13	4000	300	100000	250	30000	
14	4000	300	100000	200	40000	
15	4000	300	100000	100	50000	
16	4000	200	1000	200	40000	
17	4000	450	5000	200	40000	
18	3000	450	5000	200	40000	$N_{\mathrm{act}} = 750$, $N_{\mathrm{pcmb}} = 375$, $N_{\mathrm{het}} = 1875$
19	3000	450	1000	200	40000	$N_{\mathrm{act}} = 750$, $N_{\mathrm{pcmb}} = 375$, $N_{\mathrm{het}} = 1875$
20	4000	450	1000	250	40000	

Table A.1: Parameter values explored in the test simulations for the region 6.5–16.5 Mbp in chromosome 19. Values for τ_{sw}, τ_{ex}, and τ_{off} are expressed in terms of the simulation time unit τ.

No.	**Chr.**	**Start bp**	**End bp**	N	$N_{\mathbf{ex}}$	$N_{\mathbf{act}}$	$N_{\mathbf{poly}}$	$N_{\mathbf{het}}$
1	chr1	0	73500000	73500	740	1840	920	4600
2	chr1	73500000	121500000	48000	480	1200	600	3000
3	chr1	142500000	189000000	46500	470	1160	580	2910
4	chr1	189000000	249250621	60251	600	1500	750	3770
5	chr2	0	52500000	52500	530	1310	660	3280
6	chr2	52500000	92400000	39900	400	1000	500	2490
7	chr2	95300000	147500000	52200	520	1310	650	3260
8	chr2	147500000	194500000	47000	470	1180	590	2940
9	chr2	194500000	243199373	48700	490	1220	610	3040
10	chr3	0	90600000	90600	910	2270	1130	5660
11	chr3	93400000	163500000	70100	700	1750	880	4380
12	chr3	163500000	198022430	34523	350	860	430	2160
13	chr4	0	50000000	50000	500	1250	630	3130
14	chr4	52500000	136500000	84000	840	2100	1050	5250
15	chr4	136500000	191154276	54655	550	1370	680	3420
16	chr5	0	46500000	46500	470	1160	580	2910
17	chr5	49400000	105000000	55600	560	1390	700	3470
18	chr5	105000000	180915260	75916	760	1900	950	4740
19	chr6	0	58800000	58800	590	1470	740	3680
20	chr6	61800000	103500000	41700	420	1040	520	2610
21	chr6	103500000	171115067	67616	680	1690	850	4230
22	chr7	0	58100000	58100	580	1450	730	3630
23	chr7	61000000	119000000	58000	580	1450	730	3630
24	chr7	119000000	159138663	40139	400	1000	500	2510
25	chr8	0	43900000	43900	440	1100	550	2740
26	chr8	46800000	84000000	37200	370	930	470	2330
27	chr8	84000000	146364022	62365	620	1560	780	3900
28	chr9	0	47400000	47400	470	1190	590	2960
29	chr9	65400000	141213431	75814	760	1900	950	4740
30	chr10	0	39200000	39200	390	980	490	2450
31	chr10	42300000	85000000	42700	430	1070	530	2670
32	chr10	85000000	135534747	50535	510	1260	630	3160
33	chr11	0	51600000	51600	520	1290	650	3230
34	chr11	54600000	135006516	80407	800	2010	1010	5030
35	chr12	0	34900000	34900	350	870	440	2180

36	chr12	37800000	84500000	46700	470	1170	580	2920
37	chr12	84500000	133851895	49352	490	1230	620	3080
38	chr13	19000000	56000000	37000	370	930	460	2310
39	chr13	56000000	115169878	59170	590	1480	740	3700
40	chr14	19000000	107349540	88350	880	2210	1100	5520
41	chr15	20000000	102531392	82532	830	2060	1030	5160
42	chr16	0	35300000	35300	350	880	440	2210
43	chr16	46300000	90354753	44055	440	1100	550	2750
44	chr17	0	81195210	81196	810	2030	1010	5070
45	chr18	0	78077248	78078	780	1950	980	4880
46	chr19	0	59128983	59129	590	1480	740	3700
47	chr20	0	63025520	63026	630	1580	790	3940
48	chr21	14300000	48129895	33830	340	850	420	2110
49	chr22	16000000	51304566	35305	350	880	440	2210
50	chrX	0	58600000	58600	590	1470	730	3660
51	chrX	61600000	121000000	59400	590	1490	740	3710
52	chrX	121000000	155270560	34271	340	860	430	2140

Table A.2: Parameters for simulating individual chromosome segments of the human genome. The length of each chromosome is taken from the reference genome hg19.

Bibliography

[1] C. R. Calladine, H. R. Drew, B. F. Luisi, A. A. Travers. *Understanding DNA*. Elsevier, 3rd edition (2004)

[2] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter. *Molecular Biology of the Cell*. Taylor & Francis, 6th edition (2014)

[3] P. R. Cook. *Principles of Nuclear Structure and Function*. Wiley (2001)

[4] T. D. Pollard, W. C. Earnshaw, J. Lippincott-Schwartz, G. T. Johnson. *Cell Biology*. Elsevier, 3rd edition (2017)

[5] S. Sazer, H. Schiessel. The biology and polymer physics underlying large-scale chromosome organization. *Traffic* **19**, 87 (2018)

[6] T. Misteli. The self-organizing genome: principles of genome architecture and function. *Cell* **183**, 28 (2020)

[7] B. van Steensel, E. E. M. Furlong. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* **20**, 327 (2019)

[8] H. Zheng, W. Xie. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.* **20**, 535 (2019)

[9] E. S. Doğan, C. Liu. Three-dimensional chromatin packing and positioning of plant genomes. *Nat. Plants* **4**, 521 (2018)

[10] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, T. Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157 (2005)

[11] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, C. C. O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, eaag0025 (2017)

[12] T. Cremer, M. Cremer. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010)

[13] C. O'Connor. Fluorescence in situ hybridization (FISH). *Nat. Educ.* **1**, 171 (2008)

[14] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, W. A. Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.* **145**, 1119 (1999)

[15] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, W. A. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211 (2001)

[16] H. Tanabe, S. Müller, M. Neusser, J. von Hase, E. Calcagno, M. Cremer, I. Solovei, C. Cremer, T. Cremer. Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4424 (2002)

[17] F. A. Habermann, M. Cremer, J. Walter, G. Kreth, J. von Hase, K. Bauer, J. Wienberg, C. Cremer, T. Cremer, I. Solovei. Arrangements of macro- and microchromosomes in chicken cells. *Chromosome Res.* **9**, 569 (2001)

[18] J. Dekker, K. Rippe, M. Dekker, N. Kleckner. Capturing chromosome conformation. *Science* **295**, 1306 (2002)

[19] R. Kempfer, A. Pombo. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207 (2020)

[20] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289 (2009)

[21] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, E. Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381 (2012)

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

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

[24] G. Tiana, L. Giorgetti (Editors) *Modeling the 3D Conformation of Genomes*. CRC Press (2019)

[25] A. Y. Grosberg, A. R. Khokhlov. *Statistical Physics of Macromolecules*. AIP Press (1994)

[26] M. Rubinstein, R. H. Colby. *Polymer Physics*. Oxford University Press (2003)

[27] P.-G. de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press (1979)

[28] P. J. Hagerman. Flexibility of DNA. *Ann. Rev. Biophys. Biophys. Chem.* **17**, 265 (1988)

[29] L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19**, 37 (2011)

[30] J. C. Le Guillou, J. Zinn-Justin. Critical exponents for the n-vector model in three dimensions from field theory. *Phys. Rev. Lett.* **39**, 95 (1977)

[31] A. K. Shaytan, G. A. Armeev, A. Goncearenco, V. B. Zhurkin, D. Landsman, A. R. Panchenko. Coupling between histone conformations and DNA geometry in nucleosomes on a microsecond timescale: atomistic insights into nucleosome functions. *J. Mol. Biol.* **428**, 221 (2016)

[32] F. Serra, M. Di Stefano, Y. G. Spill, Y. Cuartero, M. Goodstadt, D. Baù, M. A. Marti-Renom. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* **589**, 2987 (2015)

[33] M. Di Stefano, J. Paulsen, T. G. Lien, E. Hovig, C. Micheletti. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci. Rep.* **6**, 35985 (2016)

[34] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, M. A. Marti-Renom. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, e1005665 (2017)

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

[36] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, J. N. Onuchic. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12168 (2016)

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

[38] C. A. Brackley, D. Marenduzzo, N. Gilbert. Mechanistic modeling of chromatin folding to understand function. *Nat. Methods* **17**, 767 (2020)

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

[40] P. R. Cook, D. Marenduzzo. Entropic organization of interphase chromosomes. *J. Cell Biol.* **186**, 825 (2009)

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

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

[43] F. Erdel, K. Rippe. Formation of chromatin subcompartments by phase separation. *Biophys. J.* **114**, 2262 (2018)

[44] J.-K. Ryu, C. Bouchoux, H. W. Liu, E. Kim, M. Minamino, R. de Groot, A. J. Katan, A. Bonato, D. Marenduzzo, D. Michieletto, F. Uhlmann, C. Dekker. Bridging-induced phase separation induced by cohesin SMC protein complexes. *Sci. Adv.* **7**, eabe5905 (2021)

[45] Y. S. Mao, B. Zhang, D. L. Spector. Biogenesis and function of nuclear bodies. *Trends Genet.* **27**, 295 (2011)

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

[47] C. A. Brackley, B. Liebchen, D. Michieletto, F. Mouvet, P. R. Cook, D. Marenduzzo. Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains. *Biophys. J.* **112**, 1085 (2017)

[48] C. A. Brackley, D. Michieletto, F. Mouvet, J. Johnson, S. Kelly, P. R. Cook, D. Marenduzzo. Simulating topological domains in human chromosomes with a fitting-free model. *Nucleus* **7**, 453 (2016)

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

[50] E. J. Banigan, L. A. Mirny. Loop extrusion: theory meets single-molecule experiments. *Curr. Opin. Cell Biol.* **64**, 124 (2020)

[51] I. F. Davidson, J.-M. Peters. Genome folding through loop extrusion by SMC complexes. *Nat. Rev. Mol. Cell Biol.* **22**, 445 (2021)

[52] K. Kimura, V. V. Rybenkov, N. J. Crisona, T. Hirano, N. R. Cozzarelli. 13S condensin actively reconfigures DNA by introducing global positive writhe: implications for chromosome condensation. *Cell* **98**, 239 (1999)

[53] K. Nasmyth. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu. Rev. Genet.* **35**, 673 (2001)

[54] A. L. Sanborn, S. S. P. Rao, S.-C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, E. L. Aiden. 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 (2015)

[55] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, L. A. Mirny. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038 (2016)

[56] Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, Q. Wu. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900 (2015)

[57] E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, B. G. Bruneau. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930 (2017)

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

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

[60] J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, B. D. Rowland. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693 (2017)

[61] G. Wutz, C. Várnai, K. Nagasaka, D. A. Cisneros, R. R. Stocsits, W. Tang, S. Schoenfelder, G. Jessberger, M. Muhar, M. J. Hossain, N. Walther,

B. Koch, M. Kueblbeck, J. Ellenberg, J. Zuber, P. Fraser, J.-M. Peters. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573 (2017)

[62] T. Terakawa, S. Bisht, J. M. Eeftens, C. Dekker, C. H. Haering, E. C. Greene. The condensin complex is a mechanochemical motor that translocates along DNA. *Science* **358**, 672 (2017)

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

[64] I. F. Davidson, B. Bauer, D. Goetz, W. Tang, G. Wutz, J.-M. Peters. DNA loop extrusion by human cohesin. *Science* **366**, 1338 (2019)

[65] M. Kong, E. E. Cutts, D. Pan, F. Beuron, T. Kaliyappan, C. Xue, E. P. Morris, A. Musacchio, A. Vannini, E. C. Greene. Human condensin I and II drive extensive ATP-dependent compaction of nucleosome-bound DNA. *Mol. Cell* **79**, 99 (2020)

[66] S. Golfier, T. Quail, H. Kimura, J. Brugués. Cohesin and condensin extrude DNA loops in a cell cycle-dependent manner. *eLife* **9**, e53885 (2020)

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

[68] D. Michieletto, E. Orlandini, D. Marenduzzo. Polymer model with epigenetic recoloring reveals a pathway for the *de novo* establishment and 3D organization of chromatin domains. *Phys. Rev. X* **6**, 041047 (2016)

[69] D. Michieletto, M. Chiang, D. Colì, A. Papantonis, E. Orlandini, P. R. Cook, D. Marenduzzo. Shaping epigenetic memory via genomic bookmarking. *Nucleic Acids Res.* **46**, 83 (2018)

[70] D. Jost, C. Vaillant. Epigenomics in 3D: importance of long-range spreading and specific interactions in epigenomic maintenance. *Nucleic Acids Res.* **46**, 2252 (2018)

[71] M. Falk, Y. Feodorova, N. Naumova, M. Imakaev, B. R. Lajoie, H. Leonhardt, B. Joffe, J. Dekker, G. Fudenberg, I. Solovei, L. A. Mirny. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**, 395 (2019)

[72] M. Chiang, D. Michieletto, C. A. Brackley, N. Rattanavirotkul, H. Mohammed, D. Marenduzzo, T. Chandra. Polymer modeling predicts chromosome reorganization in senescence. *Cell Rep.* **28**, 3212 (2019)

[73] S. Sati, B. Bonev, Q. Szabo, D. Jost, P. Bensadoun, F. Serra, V. Loubiere, G. L. Papadopoulos, J. C. Rivera-Mulia, L. Fritsch, P. Bouret, D. Castillo, J. L. Gelpi, M. Orozco, C. Vaillant, F. Pellestor, F. Bantignies, M. A. Marti-Renom, D. M. Gilbert, J.-M. Lemaitre, G. Cavalli. 4D genome rewiring during oncogene-induced and replicative senescence. *Mol. Cell* **78**, 522 (2020)

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

[75] M. C. F. Pereira, C. A. Brackley, D. Michieletto, C. Annunziatella, S. Bianco, A. M. Chiariello, M. Nicodemi, D. Marenduzzo. Complementary chromosome folding by transcription factors and cohesin. *bioRxiv* (2018)

[76] D. Frenkel, B. Smit. *Understanding Molecular Simulation: from Algorithms to Applications.* Academic Press, 2nd edition (2001)

[77] B. J. Alder, T. E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208 (1957)

[78] A. Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **136**, A405 (1964)

[79] J. A. McCammon, B. R. Gelin, M. Karplus. Dynamics of folded proteins. *Nature* **267**, 585 (1977)

[80] M. Karplus, J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646 (2002)

[81] K. Kremer, G. S. Grest. Dynamics of entangled linear polymer melts: a molecular-dynamics simulation. *J. Chem. Phys.* **92**, 5057 (1990)

[82] J. Langowski. Polymer chain models of DNA and chromatin. *Eur. Phys. J. E* **19**, 241 (2006)

[83] J. Dekker. Mapping *in vivo* chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. *J. Biol. Chem.* **283**, 34532 (2008)

[84] H. Hajjoul, J. Mathon, H. Ranchon, I. Goiffon, J. Mozziconacci, B. Albert, P. Carrivain, J.-M. Victor, O. Gadal, K. Bystricky, A. Bancaud. High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome. *Genome Res.* **23**, 1829 (2013)

[85] J. D. Weeks, D. Chandler, H. C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* **54**, 5237 (1971)

[86] L. Verlet. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98 (1967)

[87] W. C. Swope, H. C. Andersen, P. H. Berens, K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. *J. Chem. Phys.* **76**, 637 (1982)

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

[89] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563 (1942)

[90] A. V. Probst, E. Dunleavy, G. Almouzni. Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.* **10**, 192 (2009)

[91] F. Ciabrelli, F. Comoglio, S. Fellous, B. Bonev, M. Ninova, Q. Szabo, A. Xuéreb, C. Klopp, A. Aravin, R. Paro, F. Bantignies, G. Cavalli. Stable Polycomb-dependent transgenerational inheritance of chromatin states in *Drosophila*. *Nat. Genet.* **49**, 876 (2017)

[92] A. Angel, J. Song, C. Dean, M. Howard. A Polycomb-based switch underlying quantitative epigenetic memory. *Nature* **476**, 105 (2011)

[93] J. Zhu, M. Adli, J. Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P. L. De Jager, D. A. Bennett, J. A. Houmard, D. M. Muoio, T. T. Onder, R. Camahort, C. A. Cowan, A. Meissner, C. B. Epstein, N. Shoresh, B. E. Bernstein. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642 (2013)

[94] P. J. Skene, S. Henikoff. Histone variants in pluripotency and disease. *Development* **140**, 2513 (2013)

[95] E. Heard, R. A. Martienssen. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157**, 95 (2014)

[96] S. Pal, J. K. Tyler. Epigenetics and aging. *Sci. Adv.* **2**, e1600584 (2016)

[97] T. K. Barth, A. Imhof. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem. Sci.* **35**, 618 (2010)

[98] G. E. Zentner, S. Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nat. Struct. Mol. Biol.* **20**, 259 (2013)

[99] T. B. Kheir, A. H. Lund. Epigenetic dynamics across the cell cycle. *Essays Biochem.* **48**, 107 (2010)

[100] C. Alabert, T. K. Barth, N. Reverón-Gómez, S. Sidoli, A. Schmidt, O. N. Jensen, A. Imhof, A. Groth. Two distinct modes for propagation of histone PTMs across the cell cycle. *Genes Dev.* **29**, 585 (2015)

[101] N. Festuccia, I. Gonzalez, P. Navarro. The epigenetic paradox of pluripotent ES cells. *J. Mol. Biol.* **429**, 1476 (2017)

[102] A. N. Scharf, T. K. Barth, A. Imhof. Establishment of histone modifications after chromatin assembly. *Nucleic Acids Res.* **37**, 5032 (2009)

[103] A. Klosin, K. Reis, C. Hidalgo-Carcedo, E. Casas, T. Vavouri, B. Lehner. Impaired DNA replication derepresses chromatin and generates a transgenerationally inherited epigenetic memory. *Sci. Adv.* **3**, e1701143 (2017)

[104] C. Arnold, P. F. Stadler, S. J. Prohaska. Chromatin computation: epigenetic inheritance as a pattern reconstruction problem. *J. Theor. Biol.* **336**, 61 (2013)

[105] I. B. Dodd, M. A. Micheelsen, K. Sneppen, G. Thon. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* **129**, 813 (2007)

[106] S. Berry, C. Dean, M. Howard. Slow chromatin dynamics allow Polycomb target genes to filter fluctuations in transcription factor activity. *Cell Syst.* **4**, 445 (2017)

[107] F. Erdel. How communication between nucleosomes enables spreading and epigenetic memory of histone modifications. *Bioessays* **39**, 1700053 (2017)

[108] K. Sneppen, M. A. Micheelsen, I. B. Dodd. Ultrasensitive gene regulation by positive feedback loops in nucleosome modification. *Mol. Syst. Biol.* **4**, 182 (2008)

[109] M. A. Micheelsen, N. Mitarai, K. Sneppen, I. B. Dodd. Theory for the stability and regulation of epigenetic landscapes. *Phys. Biol.* **7**, 026010 (2010)

[110] I. B. Dodd, K. Sneppen. Barriers and silencers: a theoretical toolkit for control and containment of nucleosome-based epigenetic states. *J. Mol. Biol.* **414**, 624 (2011)

[111] K. Müller-Ott, F. Erdel, A. Matveeva, J.-P. Mallm, A. Rademacher, M. Hahn, C. Bauer, Q. Zhang, S. Kaltofen, G. Schotta, T. Höfer, K. Rippe. Specificity, propagation, and memory of pericentric heterochromatin. *Mol. Sys. Biol.* **10**, 746 (2014)

[112] H. Zhang, X.-J. Tian, A. Mukhopadhyay, K. S. Kim, J. Xing. Statistical mechanics model for the dynamics of collective epigenetic histone modification. *Phys. Rev. Lett.* **112**, 068101 (2014)

[113] L. C. M. Anink-Groenen, T. R. Maarleveld, P. J. Verschure, F. J. Bruggeman. Mechanistic stochastic model of histone modification pattern formation. *Epigenetics Chromatin* **7**, 30 (2014)

[114] M. J. Obersriebnig, E. M. H. Pallesen, K. Sneppen, A. Trusina, G. Thon. Nucleation and spreading of a heterochromatic domain in fission yeast. *Nat. Commun.* **7**, 11518 (2016)

[115] F. Erdel, E. C. Greene. Generalized nucleation and looping model for epigenetic memory of histone modifications. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4180 (2016)

[116] L. Fritsch, P. Robin, J. R. R. Mathieu, M. Souidi, H. Hinaux, C. Rougeulle, A. Harel-Bellan, M. Ameyar-Zazoua, S. Ait-Si-Ali. A subset of the histone H3 lysine 9 methyltransferases Suv39h1, G9a, GLP, and SETDB1 participate in a multimeric complex. *Mol. Cell* **37**, 46 (2010)

[117] N. A. Hathaway, O. Bell, C. Hodges, E. L. Miller, D. S. Neel, G. R. Crabtree. Dynamics and memory of heterochromatin in living cells. *Cell* **149**, 1447 (2012)

[118] A. Kuzmichev, K. Nishioka, H. Erdjument-Bromage, P. Tempst, D. Reinberg. Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of Zeste protein. *Genes Dev.* **16**, 2893 (2002)

[119] G. Li, R. Margueron, M. Ku, P. Chambon, B. E. Bernstein, D. Reinberg. Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev.* **24**, 368 (2010)

[120] S. Aranda, G. Mas, L. Di Croce. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* **1**, e1500737 (2015)

[121] M. Ali, R. A. Hom, W. Blakeslee, L. Ikenouye, T. G. Kutateladze. Diverse functions of PHD fingers of the MLL/KMT2 subfamily. *Biochim. Biophys. Acta* **1843**, 366 (2014)

[122] H. H. Ng, F. Robert, R. A. Young, K. Struhl. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell* **11**, 709 (2003)

[123] K. D. Sarge, O.-K. Park-Sarge. Gene bookmarking: keeping the pages open. *Trends Biochem. Sci.* **30**, 605 (2005)

[124] S. K. Zaidi, D. W. Young, M. A. Montecino, J. B. Lian, A. J. van Wijnen, J. L. Stein, G. S. Stein. Mitotic bookmarking of genes: a novel dimension to epigenetic control. *Nat. Rev. Genet.* **11**, 583 (2010)

[125] S. Kadauke, G. A. Blobel. Mitotic bookmarking by transcription factors. *Epigenetics Chromatin* **6**, 6 (2013)

[126] S. S. Teves, L. An, A. S. Hansen, L. Xie, X. Darzacq, R. Tjian. A dynamic mode of mitotic bookmarking by transcription factors. *eLife* **5**, e22280 (2016)

[127] B. Schuettengruber, N. Oded Elkayam, T. Sexton, M. Entrevan, S. Stern, A. Thomas, E. Yaffe, H. Parrinello, A. Tanay, G. Cavalli. Cooperativity, specificity, and evolutionary stability of Polycomb targeting in *Drosophila*. *Cell Rep.* **9**, 219 (2014)

[128] F. Laprell, K. Finkl, J. Müller. Propagation of Polycomb-repressed chromatin requires sequence-specific recruitment to DNA. *Science* **356**, 85 (2017)

[129] N. E. Follmer, A. H. Wani, N. J. Francis. A Polycomb group protein is retained at specific sites on chromatin in mitosis. *PLoS Genet.* **8**, e1003135 (2012)

[130] N. Festuccia, A. Dubois, S. Vandormael-Pournin, E. Gallego Tejeda, A. Mouren, S. Bessonnard, F. Mueller, C. Proux, M. Cohen-Tannoudji, P. Navarro. Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network. *Nat. Cell Biol.* **18**, 1139 (2016)

[131] C. Deluz, E. T. Friman, D. Strebinger, A. Benke, M. Raccaud, A. Callegari, M. Leleu, S. Manley, D. M. Suter. A role for mitotic bookmarking of SOX2 in pluripotency and differentiation. *Genes Dev.* **30**, 2538 (2016)

[132] S. Kadauke, M. I. Udugama, J. M. Pawlicki, J. C. Achtman, D. P. Jain, Y. Cheng, R. C. Hardison, G. A. Blobel. Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. *Cell* **150**, 725 (2012)

[133] A. Grob, C. Colleran, B. McStay. Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. *Genes Dev.* **28**, 220 (2014)

[134] K. Bystricky, P. Heun, L. Gehlen, J. Langowski, S. M. Gasser. 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 (2004)

[135] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212 (2010)

[136] P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu, D. Linder-Basso, A. Plachetka, G. Shanower, M. Y. Tolstorukov, L. J. Luquette, R. Xi, Y. L. Jung, R. W. Park, E. P. Bishop, T. K. Canfield, R. Sandstrom, R. E. Thurman, D. M. MacAlpine, J. A. Stamatoyannopoulos, M. Kellis, S. C. Elgin, M. I. Kuroda, V. Pirrotta, G. H. Karpen, P. J. Park. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480 (2011)

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

[138] N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter, W. A. Bickmore. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555 (2004)

[139] P. B. Talbert, S. Henikoff. Spreading of silent chromatin: inaction at a distance. *Nat. Rev. Genet.* **7**, 793 (2006)

[140] G. Schotta, A. Ebert, V. Krauss, A. Fischer, J. Hoffmann, S. Rea, T. Jenuwein, R. Dorn, G. Reuter. Central role of *Drosophila* SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J.* **21**, 1121 (2002)

[141] D. Michieletto, D. Colì, D. Marenduzzo, E. Orlandini. Nonequilibrium theory of epigenomic microphase separation in the cell nucleus. *Phys. Rev. Lett.* **123**, 228101 (2019)

[142] D. Colì, E. Orlandini, D. Michieletto, D. Marenduzzo. Magnetic polymer models for epigenetics-driven chromosome folding. *Phys. Rev. E* **100**, 052410 (2019)

[143] R. T. Coleman, G. Struhl. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a *Drosophila* HOX gene. *Science* **356**, eaai8236 (2017)

[144] S. Berry, M. Hartley, T. S. G. Olsson, C. Dean, M. Howard. Local chromatin environment of a Polycomb target gene instructs its own epigenetic inheritance. *eLife* **4**, e07205 (2015)

[145] S. De, A. Mitra, Y. Cheng, K. Pfeifer, J. A. Kassis. Formation of a Polycomb-domain in the absence of strong Polycomb response elements. *PLoS Genet.* **12**, e1006200 (2016)

[146] I. Solovei, K. Thanisch, Y. Feodorova. How to rule the nucleus: *divide et impera*. *Curr. Opin. Cell Biol.* **40**, 47 (2016)

[147] B. van Steensel, A. S. Belmont. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780 (2017)

[148] B. van Steensel, S. Henikoff. Identification of *in vivo* DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* **18**, 424 (2000)

[149] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, B. van Steensel. Domain

organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948 (2008)

[150] J. Kind, L. Pagie, H. Ortabozkoyun, S. Boyle, S. S. de Vries, H. Janssen, M. Amendola, L. D. Nolen, W. A. Bickmore, B. van Steensel. Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178 (2013)

[151] J. Kind, L. Pagie, S. S. de Vries, L. Nahidiazar, S. S. Dey, M. Bienko, Y. Zhan, B. Lajoie, C. A. de Graaf, M. Amendola, G. Fudenberg, M. Imakaev, L. A. Mirny, K. Jalink, J. Dekker, A. van Oudenaarden, B. van Steensel. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134 (2015)

[152] A. Németh, A. Conesa, J. Santoyo-Lopez, I. Medina, D. Montaner, B. Péterfia, I. Solovei, T. Cremer, J. Dopazo, G. Längst. Initial genomics of the human nucleolus. *PLoS Genet.* **6**, e1000889 (2010)

[153] J. Campisi, F. d'Adda di Fagagna. Cellular senescence: when bad things happen to good cells. *Nat. Rev. Mol. Cell Biol.* **8**, 729 (2007)

[154] T. Chandra, K. Kirschner. Chromosome organisation during ageing and senescence. *Curr. Opin. Cell Biol.* **40**, 161 (2016)

[155] M. Narita, S. Nũnez, E. Heard, M. Narita, A. W. Lin, S. A. Hearn, D. L. Spector, G. J. Hannon, S. W. Lowe. Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703 (2003)

[156] T. Chandra, P. A. Ewels, S. Schoenfelder, M. Furlan-Magaril, S. W. Wingett, K. Kirschner, J.-Y. Thuret, S. Andrews, P. Fraser, W. Reik. Global reorganization of the nuclear landscape in senescent cells. *Cell Rep.* **10**, 471 (2015)

[157] P. P. Shah, G. Donahue, G. L. Otte, B. C. Capell, D. M. Nelson, K. Cao, V. Aggarwala, H. A. Cruickshanks, T. S. Rai, T. McBryan, B. D. Gregory, P. D. Adams, S. L. Berger. Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes Dev.* **27**, 1787 (2013)

[158] C. Boumendil, P. Hari, K. C. F. Olsen, J. C. Acosta, W. A. Bickmore. Nuclear pore density controls heterochromatin reorganization during senescence. *Genes Dev.* **33**, 144 (2019)

[159] M. Eriksson, W. T. Brown, L. B. Gordon, M. W. Glynn, J. Singer, L. Scott, M. R. Erdos, C. M. Robbins, T. Y. Moses, P. Berglund, A. Dutra, E. Pak, S. Durkin, A. B. Csoka, M. Boehnke, T. W. Glover, F. S. Collins. Recurrent *de novo* point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293 (2003)

[160] R. D. Goldman, D. K. Shumaker, M. R. Erdos, M. Eriksson, A. E. Goldman, L. B. Gordon, Y. Gruenbaum, S. Khuon, M. Mendez, R. Varga, F. S. Collins. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8963 (2004)

[161] D. K. Shumaker, T. Dechat, A. Kohlmaier, S. A. Adam, M. R. Bozovsky, M. R. Erdos, M. Eriksson, A. E. Goldman, S. Khuon, F. S. Collins, T. Jenuwein, R. D. Goldman. Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8703 (2006)

[162] R. P. McCord, A. Nazario-Toole, H. Zhang, P. S. Chines, Y. Zhan, M. R. Erdos, F. S. Collins, J. Dekker, K. Cao. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res.* **23**, 260 (2013)

[163] T. Chandra, K. Kirschner, J.-Y. Thuret, B. D. Pope, T. Ryba, S. Newman, K. Ahmed, S. A. Samarajiwa, R. Salama, T. Carroll, R. Stark, R. Janky, M. Narita, L. Xue, A. Chicas, S. Nũnez, R. Janknecht, Y. Hayashi-Takanaka, M. D. Wilson, A. Marshall, D. T. Odom, M. M. Babu, D. P. Bazett-Jones, S. Tavaré, P. A. W. Edwards, S. W. Lowe, H. Kimura, D. M. Gilbert, M. Narita. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Mol. Cell* **47**, 203 (2012)

[164] R. Zhang, M. V. Poustovoitov, X. Ye, H. A. Santos, W. Chen, S. M. Daganzo, J. P. Erzberger, I. G. Serebriiskii, A. A. Canutescu, R. L. Dunbrack, J. R. Pehrson, J. M. Berger, P. D. Kaufman, P. D. Adams. Formation of macroH2A-containing senescence-associated heterochromatin foci and senescence driven by ASF1a and HIRA. *Dev. Cell* **8**, 19 (2005)

[165] N. A. Kinney, I. V. Sharakhov, A. V. Onufriev. Chromosome-nuclear envelope attachments affect interphase chromosome territories and entanglement. *Epigenetics Chromatin* **11**, 3 (2018)

[166] H. E. Johnson, S. Granick. New mechanism of nonequilibrium polymer adsorption. *Science* **255**, 966 (1992)

[167] J. F. Douglas, H. E. Johnson, S. Granick. A simple kinetic model of polymer adsorption and desorption. *Science* **262**, 2010 (1993)

[168] M. Sadaie, R. Salama, T. Carroll, K. Tomimatsu, T. Chandra, A. R. J. Young, M. Narita, P. A. Pérez-Mancera, D. C. Bennett, H. Chong, H. Kimura, M. Narita. Redistribution of the Lamin B1 genomic binding profile affects rearrangement of heterochromatic domains and SAHF formation during senescence. *Genes Dev.* **27**, 1800 (2013)

[169] R. Zhang, W. Chen, P. D. Adams. Molecular dissection of formation of senescence-associated heterochromatin foci. *Mol. Cell. Biol.* **27**, 2343 (2007)

[170] M. R. Branco, T. Branco, F. Ramirez, A. Pombo. Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. *Chromosome Res.* **16**, 413 (2008)

[171] A. L. Olins, G. Rhodes, D. B. M. Welch, M. Zwerger, D. E. Olins. Lamin B receptor: multi-tasking at the nuclear envelope. *Nucleus* **1**, 53 (2010)

[172] I. Solovei, A. S. Wang, K. Thanisch, C. S. Schmidt, S. Krebs, M. Zwerger, T. V. Cohen, D. Devys, R. Foisner, L. Peichl, H. Herrmann, H. Blum, D. Engelkamp, C. L. Stewart, H. Leonhardt, B. Joffe. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584 (2013)

[173] A. Poleshko, K. M. Mansfield, C. C. Burlingame, M. D. Andrake, N. R. Shah, R. A. Katz. The human protein PRR14 tethers heterochromatin to the nuclear lamina during interphase and mitotic exit. *Cell Rep.* **5**, 292 (2013)

[174] Q. Ye, I. Callebaut, A. Pezhman, J. C. Courvalin, H. J. Worman. Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *J. Biol. Chem.* **272**, 14983 (1997)

[175] A. R. Strom, A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq, G. H. Karpen. Phase separation drives heterochromatin domain formation. *Nature* **547**, 241 (2017)

[176] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676 (2012)

[177] R. Wu, A. V. Terry, P. B. Singh, D. M. Gilbert. Differential subnuclear localization and replication timing of histone H3 lysine 9 methylation states. *Mol. Biol. Cell* **16**, 2872 (2005)

[178] N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C.-J. Chen, J.-P. Vert, E. Heard, J. Dekker, E. Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015)

[179] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999 (2012)

[180] J. Dekker, E. Heard. Structural and functional diversity of topologically associating domains. *FEBS Lett.* **589**, 2877 (2015)

[181] M. H. DeGroot, M. J. Schervish. *Probability and Statistics*. Pearson, 4th edition (2012)

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

[183] E. H. Finn, G. Pegoraro, H. B. Brandão, A.-L. Valton, M. E. Oomen, J. Dekker, L. Mirny, T. Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176**, 1502 (2019)

[184] A. J. Bray. Theory of phase-ordering kinetics. *Adv. Phys.* **51**, 481 (2002)

[185] I. Solovei, M. Kreysing, C. Lanctôt, S. Kösem, L. Peichl, T. Cremer, J. Guck, B. Joffe. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* **137**, 356 (2009)

[186] R. Laghmach, M. Di Pierro, D. A. Potoyan. Mesoscale liquid model of chromatin recapitulates nuclear order of eukaryotes. *Biophys. J.* **118**, 2130 (2020)

[187] D. Vernimmen, W. A. Bickmore. The hierarchy of transcriptional activation: from enhancer to promoter. *Trends Genet.* **31**, 696 (2015)

[188] S. Schoenfelder, P. Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437 (2019)

[189] B. K. Kragesteen, M. Spielmann, C. Paliou, V. Heinrich, R. Schöpflin, A. Esposito, C. Annunziatella, S. Bianco, A. M. Chiariello, I. Jerković, I. Harabula, P. Guckelberger, M. Pechstein, L. Wittler, W.-L. Chan, M. Franke, D. G. Lupiáñez, K. Kraft, B. Timmermann, M. Vingron, A. Visel, M. Nicodemi, S. Mundlos, G. Andrey. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* **50**, 1463 (2018)

[190] S. Bianco, C. Annunziatella, G. Andrey, A. M. Chiariello, A. Esposito, L. Fiorillo, A. Prisco, M. Conte, R. Campanile, M. Nicodemi. Modeling single-molecule conformations of the HoxD region in mouse embryonic stem and cortical neuronal cells. *Cell Rep.* **28**, 1574 (2019)

[191] A. M. Chiariello, S. Bianco, A. M. Oudelaar, A. Esposito, C. Annunziatella, L. Fiorillo, M. Conte, A. Corrado, A. Prisco, M. S. C. Larke, J. M. Telenius, R. Sciarretta, F. Musella, V. J. Buckle, D. R. Higgs, J. R. Hughes, M. Nicodemi. A dynamic folded hairpin conformation is associated with α-globin activation in erythroid cells. *Cell Rep.* **30**, 2125 (2020)

[192] M. Di Stefano, R. Stadhouders, I. Farabella, D. Castillo, F. Serra, T. Graf, M. A. Marti-Renom. Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. *Nat. Commun.* **11**, 2564 (2020)

[193] C. A. Brackley, J. M. Brown, D. Waithe, C. Babbs, J. Davies, J. R. Hughes, V. J. Buckle, D. Marenduzzo. Predicting the three-dimensional folding of

cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.* **17**, 59 (2016)

[194] A. Buckle, C. A. Brackley, S. Boyle, D. Marenduzzo, N. Gilbert. Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic loci. *Mol. Cell* **72**, 786 (2018)

[195] D. Rico, D. Kent, A. Mikulasova, N. Karataraki, R. Berlinguer-Palmini, B. A. Walker, B. M. Javierre, L. J. Russell, C. A. Brackley. High-resolution simulations of chromatin folding at genomic rearrangements in malignant B-cells provide mechanistic insights on proto-oncogene deregulation. *bioRxiv* (2021)

[196] C. A. Brackley, N. Gilbert, D. Michieletto, A. Papantonis, M. C. F. Pereira, P. R. Cook, D. Marenduzzo. Complex small-world regulatory networks emerge from the 3D organisation of the human genome. *Nat. Commun.* **12**, 5756 (2021)

[197] G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, B. J. Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496 (2014)

[198] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012)

[199] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, J. M. Cherry. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794 (2018)

[200] M. R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, H. Y. Chang. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959 (2017)

[201] E. B. Stovner, P. Sætrom. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392 (2019)

[202] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008)

[203] P. R. Cook, D. Marenduzzo. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Res.* **46**, 9895 (2018)

[204] E. C. Chittock, S. Latwiel, T. C. R. Miller, C. W. Müller. Molecular architecture of polycomb repressive complexes. *Biochem. Soc. Trans.* **45**, 193 (2017)

[205] S. Boyle, I. M. Flyamer, I. Williamson, D. Sengupta, W. A. Bickmore, R. S. Illingworth. A central role for canonical PRC1 in shaping the 3D nuclear landscape. *Genes Dev.* **34**, 931 (2020)

[206] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, A. Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87 (2020)

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

[208] Q. Szabo, A. Donjon, I. Jerković, G. L. Papadopoulos, T. Cheutin, B. Bonev, E. P. Nora, B. G. Bruneau, F. Bantignies, G. Cavalli. Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat. Genet.* **52**, 1151 (2020)

[209] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829 (1979)

[210] L. J. Core, A. L. Martins, C. G. Danko, C. T. Waters, A. Siepel, J. T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311 (2014)

[211] R. Dreos, G. Ambrosini, R. Groux, R. C. Périer, P. Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.* **45**, D51 (2017)

[212] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, J. M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664 (2019)

[213] D. J. Downes, R. A. Beagrie, M. E. Gosden, J. Telenius, S. J. Carpenter, L. Nussbaum, S. De Ornellas, M. Sergeant, C. Q. Eijsbouts, R. Schwessinger, J. Kerry, N. Roberts, A. Shivalingam, A. El-Sagheer, A. M. Oudelaar, T. Brown, V. J. Buckle, J. O. J. Davies, J. R. Hughes. High-resolution targeted 3C interrogation of *cis*-regulatory element organization at genome-wide scale. *Nat. Commun.* **12**, 531 (2021)

[214] S. Pott, J. D. Lieb. What are super-enhancers? *Nat. Genet.* **47**, 8 (2015)

[215] X. Wang, M. J. Cairns, J. Yan. Super-enhancers in transcriptional regulation and genome organization. *Nucleic Acids Res.* **47**, 11481 (2019)

[216] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, R. A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307 (2013)

[217] D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. A. Sigova, H. A. Hoke, R. A. Young. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934 (2013)

[218] A. Khan, X. Zhang. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164 (2016)

[219] G. V. Barroso, N. Puzovic, J. Y. Dutheil. The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics* **208**, 173 (2018)

[220] Q. Szabo, F. Bantignies, G. Cavalli. Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**, eaaw1668 (2019)