UiO : University of Oslo

Tharvesh Moideen Liyakat Ali

Three-dimensional topology of the genome: A computational modeling perspective

Thesis submitted for the degree of Philosophiae Doctor

Department of Molecular Medicine
Institute of Basic Medical Science
Faculty of Medicine

University of Oslo

2020

© Tharvesh Moideen Liyakat Ali, 2020

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-751-2

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgement

The work presented in this thesis was conducted from 2017 to 2020 at the Department of Molecular Medicine, Institute of Basic Medical Sciences, and funded by The Research Council of Norway and the University of Oslo. This work was supervised by Prof. Philippe Collas, Dr. Jonas Paulsen and Dr. Annaëlle Brunet.

First, I would like to thank my supervisor Philippe Collas for giving me the opportunity to work in his scientifically vibrant lab, helping me to understand chromatin biology, supporting and motivating me throughout my PhD. Second, I would like to thank profusely my co-supervisor Jonas Paulsen for guiding me through the field of computational 3D genome biology, giving me freedom to work at my own pace and making me better in programming and statistics; thank you for all the scientific coffee sessions and non-scientific lunch and evening sessions; you have been more a friend than a supervisor. Then, I would like to thank my third co-supervisor Annaëlle Brunet for helping, supporting and motivating me during the last year of my PhD; I would definitely miss the cacao discussion sessions.

I would like to thank all the members of Collas lab for their support throughout my PhD, on both academic and non-academic matters. I am grateful to Anita and Kristin for all the wet lab works. I would like to thank Sumithra for helping me in the administration tasks. I would like to thank Annaëlle for sharing the office for a bit more than two years (I know it was painful), Nolwenn for sharing the breaks and scientific discussions, Frida for lending your ears to hear my complains about the PhD and, Dunia, Aurelie, Sarah and Thomas G for wonderful beer-brewing and board gaming sessions.

Finally, I would like to thank my family and friends for being supportive during both the euphoric and turbulent times of my PhD. Specially, I would like to thank my girlfriend for being immensely supportive and baking cakes to help me to get back during those aforementioned turbulent times.

Tharvesh Moideen Liyakat Ali
Oslo, October 2020

List of Publications

Paper I Jonas Paulsen, Tharvesh M. Liyakat Ali, Philippe Collas. **Computational 3D genome modeling using Chrom3D.** *Nature Protocols* (2018), 13, 1137-1152. DOI: 10.1038/nprot.2018.009.

Paper II Jonas Paulsen, Tharvesh M. Liyakat Ali, Maxim Nekrasov, Erwan Delbarre, Marie-Odile Baudement, Sebastian Kurscheid, David Tremethick, Philippe Collas. **Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation.** *Nature Genetics* (2019), 51, 835-843. DOI: 10.1038/s41588-019-0392-0.

Paper III Tharvesh M. Liyakat Ali, Annaël Brunet, Philippe Collas, Jonas Paulsen. **TAD cliques predict key features of chromatin organization.** *Manuscript* (2020).

Published papers not included in this thesis

Frida Forsberg, Annaël Brunet, Tharvesh M. Liyakat Ali, Philippe Collas. **Interplay of lamin A and lamin B LADs on the radial positioning of chromatin.** *Nucleus* (2019), 10, 7-20. DOI: 10.1080/19491034.2019.1570810.

Philippe Collas, Tharvesh M. Liyakat Ali, Annaël Brunet, Thomas Germier. **Finding Friends in the Crowd: Three-Dimensional Cliques of Topological Genomic Domains.** *Frontiers in Genetics* (2019), 10, 602. DOI: 10.3389/fgene.2019.00602.

List of Abbreviations

3C	chromosome conformation capture.
3D	three-dimensional.
4C	chromosome conformation capture-on-chip.
5C	chromosome conformation capture carbon copy.
^{m6}**A**	adenine-6-methylation modification.
bp	base-pair of DNA.
ChIA-PET	chromatin interaction analysis with paired-end tag sequencing.
ChIP	chromatin immunoprecipitation.
CTCF	CCCTC-binding factor.
DamID	DNA adenine methyltransferase identification.
DNA	deoxyribonucleic acid.
ESC	embryonic stem cell.
FDR	false discovery rate.
FISH	fluorescence *in situ* hybridization.
GAM	genome architecture mapping.
GUI	graphical user interface.
kb	kilo base, 1000 bp of DNA.
LAD	lamina associated domain.
Mb	mega base, 1 million bp of DNA.
MC	Monte Carlo.
me1,me2,me3	mono-methylated, di-methylated and tri-methylated.
NAD	nucleolus-associated domain.
NCHG	non-central hypergeometric.
NUP	nucleoporin.
PCA	principal component analysis.
PCR	polymerase chain reaction.
qPCR	quantitative polymerase chain reaction.
RNA	ribonucleic acid.
RNAP II	RNA polymerase II.
rRNA	ribosomal RNA.
SNP	single nucleotide polymorphism.
SPAD	speckles associated domains.
SPRITE	split-pool recognition of interactions by tag extension.
TAD	topologically associating domain.
TF	transcription factor.
TSS	transcription start site.

Contents

Chapter 1

Introduction

1.1 Organisation of the mammalian genome in space and time

The 3-dimensional (3D) organization of the genome is paramount to the proper regulation of gene expression during development and for tissue homeostasis. Much of our knowledge of the genome however still depends on a one-dimensional, linear, representation of a reference genome. It is now clear though that 3D analyses of the genome enabled by development of wet-lab and computational methods enhance the information content of genomic studies and give new insights into gene regulation and disease mechanisms [1]. For example, genomic characteristics identified and mapped onto a linear genome may be deferentially organized in the 3D nucleus space; this may result in interpretations of these features that are not visible in one dimension (Fig. 1.1). Since the first release of the human genome sequence [2], endeavors have focused on interpreting genome sequence by mapping elements regulating gene expression, histone and DNA modifications, chromatin-modifying enzymes and transcription factor binding [3]. These efforts have been accompanied by studies aiming to identify the principles of 3D chromatin folding [1] using wet-lab and computational approaches developed over the past two decades. Moreover, temporal analyses of the 3D genome aim to provide a 4th dimension to the changes in genome architecture, where the 4_{th} dimension is time [4, 5]. This thesis addresses computational approaches to model the human genome in 3D **(Paper I)** and new fundamental aspects of higher-order chromatin conformation during stem cell differentiation **(Papers II and III)**.

1.1.1 Chromosome territories

Increasing evidence over the past two decades indicates that the 3D organization of mammalian genome in the interphase nucleus is not random. In interphase, the highest level of genome architecture is in the form of chromosome territories, a view already proposed by Rabl in 1885 [6]. According to this model, each chromosome preferentially occupies a distinct space and volume in the nucleus [6, 7] (Fig. 1.2). The concept of chromosome territories relies on the view that proximity, or contact frequency, of chromosomal regions is higher within chromosomes (these are cis or intra-chromosomal interactions) than between chromosomes (trans or inter-chromosomal interactions). Individual chromosome territories can be microscopically detected using fluorescent oligonucleotide probes (sometimes called chromosome paints) coupled with fluorescent in situ hybridization (FISH; section 1.2.1) [7, 8]. Chromosome territories can also be observed using

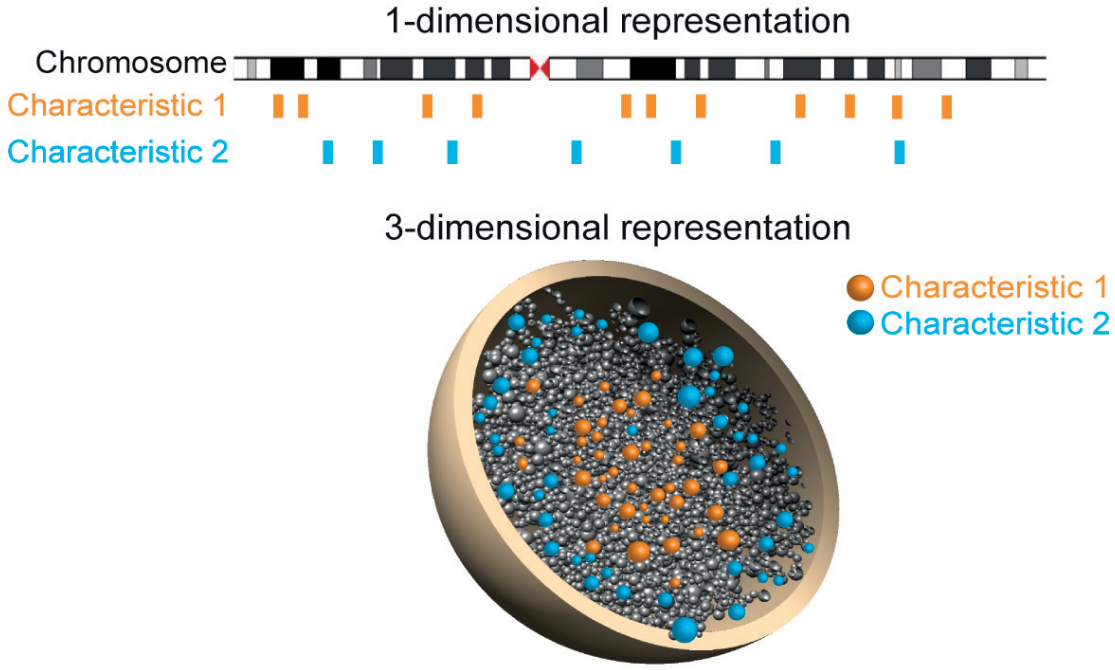

Figure 1.1: 3D analysis of the genome enhances the information content of genomic analysis. Hypothetical genome characteristics (e.g. histone modifications, or transcription factor binding sites) mapped on a one-dimensional (linear) genome only reveal distinct localizations. However, mapping the features on a 3D representation of the genome, e.g. generated by computational modeling, provides information on their radial (center-to-periphery) positioning.

chromosome conformation capture methods and by computational modeling of the 3D genome (sections 1.2.2 and 1.4) [9, 10]. The positioning of mammalian chromosomes relative to other chromosomes and to the nuclear lamina, at the nuclear periphery, is well conserved between cells and cell types, although some variation exists [11, 12]. For example, chromosome 18 is consistently more peripherally located than chromosome 19 both in mouse and human cells [13]. As addressed later, the association of specific chromosome regions with the nuclear lamina plays a role in the radial positioning of the genome [14, 15], and is likely a key factor in the overall sub-nuclear spatial placement of chromosomes.

1.1.2 Compartments

The next hierarchical level of genome organization is compartmentalization. Within chromosome territories, chromatin can be partitioned into two multi-megabase (Mb) sized 'A' and 'B' compartments [16]. This form of segmentation of the genome is detectable from the analysis of genome-wide 3C (a technique called Hi-C) and is described in section 1.3 (see also Fig. 1.3 below). Inasmuch as chromosome territories, the concept of compartments relies on the observation that genomic loci within a compartment contact each other more frequently than between compartments (Fig. 1.2).

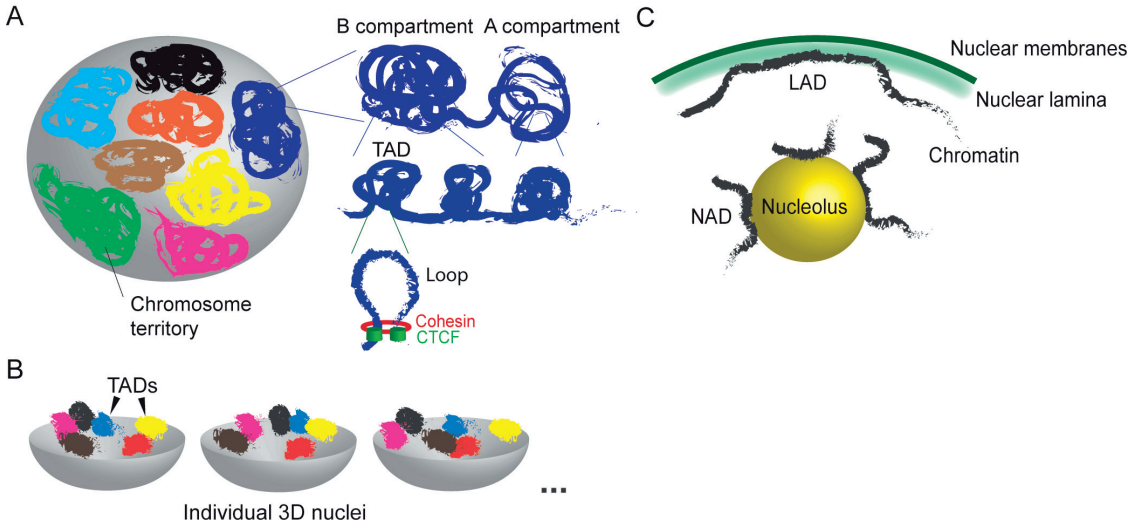

Figure 1.2: 3D chromatin organization in the nucleus. (A) Hierarchical organization of chromatin, into chromosome territories, A and B compartments and within compartments, topological domains such as TADs containing chromatin loops, bringing e.g. enhancers and promoters in proximity. Loop formation involves CTCF and cohesin proteins (see main text). (B) Topological domains such as TADs can display distinct spatial associations with other domains between cells in a population, or between time-points in a time-series experiment. (C) Interactions of chromatin with the nuclear lamina via a LAD, and with the nucleolus via NADs. Both LADs and NADs are heterochromatic, suggesting that the nuclear lamina and the nucleolus serve as a preferred anchor sites for heterochromatin domains.

A compartments contain mainly open, gene-rich and transcriptionally active parts of the genome marked by post-translational histone modifications typical for active genes, such as histone H3 lysine 4 trimethylation (H3K4me3), H3K36me3, H3K9 acetylation (H3K9ac) and H3K27ac [16]. A compartments are therefore largely euchromatic. B compartment chromatin is in contrast more compact or closed, gene-poor, transcriptionally silent and marked by repressive histone modifications such as H3K9me3 and H3K27me3 [16]; B compartments are therefore overall heterochromatic.

Based on high-resolution chromosomal contact maps derived from Hi-C data (section 1.2.3), combinations of histone modifications (chromatin states) and DNA replication profiles, A and B compartments have been further classified into six subcompartments (A1, A2, and B1, B2, B3, B4) [16]. A1 and A2 are transcriptionally active subcompartments with general characteristics of A compartments. Nonetheless, A1 and A2 differ by their DNA replication timing, where A1 finishes replicating at the beginning of S phase whereas A2 continues replicating until the middle of S phase. A2 exhibits more heterochromatin features compared to A1, such as H3K9me3 and adenine-thymine (AT)-rich DNA sequences, and accommodates longer genes. B compartments can be subcategorized to 4 subcompartments [16]. Subcompartment B1 consists of facultative

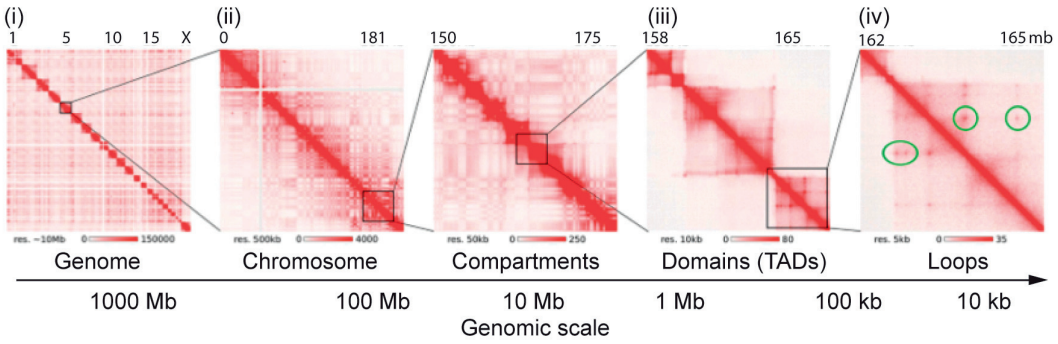

Figure 1.3: Hi-C contact matrices shown at different scales reveal the main features of hierarchical 3D genome conformation, including chromosome territories, compartments, domains or TADs, and chromatin loops within TADs. Modified with permission from [17].

heterochromatin marked by the Polycomb histone modification H3K27me3 and replicates in mid S phase. B2 contains pericentric heterochromatin associated with both the nuclear lamina and the nucleolus, whereas B3 is enriched at the nuclear lamina but not around nucleoli. This shows that chromatin in B2 compartments interact pericentric lamins and chromatin in B3 compartments interact with lamins at the periphery. Both B2 and B3 subcompartments start replication in late S-phase. B4 is a special type of subcompartment with a specific chromatin pattern observed only in chromosome 19, containing ∼47% of KRAB-ZNF superfamily genes [16].

1.1.3 Toplogically Associating Domains

Within A and B compartments, smaller domains of sub-megabase size on average harbor a high density of chromosomal contacts, more so than between such domains (Fig. 1.2A). These domains have been reported as topologically associating domains, or TADs [18–21]. Like A/B compartments, TADs are defined computationally as genomic domains along the Hi-C matrix diagonal which exhibit high interaction frequencies within them compared to the interaction frequency between them (Fig. 1.3ii and iii) [18]. Average TAD size ranges from tens of kilobases (kb) to 1-2 Mb with a median size of ∼880 kb [18] (Fig. 1.3iii). TADs have well defined borders (or boundaries) along the linear genome (Fig. 1.3), which are enriched in binding sites for cohesin and the chromatin insulator protein CCCTC-binding factor (CTCF) [18]. TAD structure is also well conserved between cell types [18, 20], as is their linear position along the genome, not only between cell types [22]. These observations suggest that TADs may represent fundamental units of 3D genome organization.

A process of chromatin 'loop extrusion' has been proposed as a likely mechanism of TAD formation [23] (Fig. 1.2A). There, a loop is formed by chromatin extrusion through a cohesin ring until the extrusion process is stopped by a 'barrier' on chromatin, namely CTCF proteins bound to CTCF motifs that

lie in a convergent orientation (facing each other inward), which stops loop extrusion [24]. I address the loop extrusion model of chromatin folding later in this Introduction. A loop extrusion mechanism of TAD formation agrees with experimental data where some TADs show boundaries enriched in cohesin and CTCF motifs in a convergent orientation [24]. Interestingly, inversion or deletion of CTCF sites in TAD boundaries weakens (i.e. tends to erase) these boundaries [25, 26]. Removal of the cohesin loading factor NIPBL [27], or depletion of the cohesin subunit RAD21 [28], also results in weak TAD boundaries, emphasizing the importance of cohesin in the formation and maintenance of at least a proportion of TADs. Of note, the concept of loop extrusion as a mechanism of TAD formation is related to, but should not be confounded by the process of chromatin loop formation inside TADs, bringing, for example, enhancers and promoters together to regulate gene expression within TADs (Fig. 1.2A).

TADs have long been perceived as key features of genome organization [29]. Today however, advancements in 3D genome analyses, and the emergence of high-resolution Hi-C data enabled by very deep sequencing, tend to challenge the TAD definition [29]. TADs now tend to be more conservatively defined as computationally defined blocks of chromosomal interactions that can be observed along the diagonal of a Hi-C matrix (Fig. 1.3ii and iii). Therefore, in this thesis, I use for sake of simplicity the term 'TADs' for genomic segments defined as TADs by TAD-calling algorithms [30].

TADs marked by convergent CTCF motifs at their boundaries and formed by loop extrusion are also called 'loop domains' [28]. Loop domains display a high density of interactions inside the loop, are relatively conserved between cell types and can be detected in high-resolution Hi-C contact maps by an interaction point at the summit of the domains in Hi-C maps [16] (Fig. 1.3iv; green circles). Other domains, which are typically not marked by CTCF at boundaries, and which interact with other domains, are referred to as 'compartmental domains' [29]; both domain types have been proposed to explain the basis of mammalian 3D chromatin organization [29].

1.1.4 Lamina Associated Domains

The mammalian nucleus is delineated by the nuclear envelope, which consists of an outer and inner nuclear membrane, nuclear pore complexes, and subjacent to the inner membrane, facing chromatin, the nuclear lamina [31] (Fig. 1.2C). The nuclear lamina is a meshwork of polymers of intermediate filaments called lamins. A-type lamins are composed of lamins A and C (also called lamin A/C), splice variants of the LMNA gene, and B-type lamins (lamins B1 and B2), encoded by the LMNB1 and LMNB2 genes respectively [31]. The nuclear lamina provides mechanical support to the nucleus and anchors chromatin at the nuclear periphery, where it has been shown to be involved in the regulation of DNA replication and transcription in space and time [32]. Specific chromatin domains are tethered to the nuclear lamina through A- and B-type lamins at the nuclear periphery via lamina-associated domains (LADs) [33] (Fig. 1.2B). LADs have been identified by chromatin immunoprecipitation (ChIP) of lamins

[34] and by DNA adenine methyltransferase identification (DamID), a proximity DNA labeling approach [35], each followed by massive parallel sequencing.

Both methods concur in that ~1000-1500 LADs can be identified in mouse and human cells, with sizes of 10 kb to 10 Mb [32]. LADs are mainly gene-poor domains, with 2-3 genes/Mb compared to 8 genes/Mb on average in the human genome. LADs consist of molecular signatures typical of silent heterochromatin: they harbor repressive histone modifications such as H3K9me2 and H3K9me3, and H3K27me3 at their border [36, 37]. LADs overlap with late-replicating regions [38], a timing similar to the replication timing of B2 and B3 compartments [39]. Thus, LADs and associated proteins (which include transcriptional repressors) constitute a repressive compartment at the nuclear periphery [40] and are considered to be a general feature of genome organization [32, 41].

Two types of LADs have been identified across studies and irrespective of method. Constitutive LADs (cLADs) are consistently associated with the nuclear lamina across cell types; they are strongly heterochromatic, gene-poor and harbor long interspersed DNA elements (LINEs, long terminal repeats widespread along the genome) [32]. cLADs are overall conserved between mouse and human cells in their relative genomic position [41] and have been proposed to be a genomic backbone anchoring chromosomes at the nuclear periphery [32]. In contrast, facultative fLADs (also called variable vLADs) are by definition more variable between cell types and species; they also harbor a higher gene density and are less heterochromatic than cLADs [41]. Indeed, fLADs are more dynamic during cell differentiation, where entire LADs or LAD sub-domains detach from the nuclear lamina to become non-LAD (or inter-LAD) domains [38, 42–44]. Detachment from the nuclear lamina may correlate (though not always) with transcriptional activation of genes within these LADs [45]. For example, LADs containing genes important for T-cell activation are found in TADs that detach from the nuclear lamina and become expressed upon T-cell activation in vitro [46]. Similarly, work from our laboratory shows that genes bound by lamin A/C in undifferentiated human adipose stem cells (ASCs) are released from lamin interactions prior to or concomitantly with (at the time-resolution examined) their transcriptional activation during adipogenic differentiation [44].

Therefore, by anchoring chromatin at the nuclear periphery, LADs emerge as key regulators of the radial distribution of the genome [47, 48]. Based on this contention, we report in **Paper I** a computational pipeline that incorporates LAD data as a spatial constraint for chromatin in 3D structural models of the genome.

1.1.5 Other 'associated domains'

Chromatin also interacts with nuclear bodies such as nucleoli and nuclear speckles. Nucleoli are membrane-less organelle where ribosome biogenesis takes place. They form around ribosomal DNA genes and repeat elements [49]. Similar to LADs, heterochromatic interacts with the periphery of nucleoli, forming nucleolar-associated domains (NADs) [49, 50] (Fig. 1.2C). NADs are enriched in

B compartments, are gene-poor and heterochromatic, and account for up 40% of the genome [51]. Strikingly, LADs and NADs show significant overlap [52] and LAD imaging in living dividing cells reveals that LADs often redistribute to the nucleolar periphery (as NADs) in daughter cells after mitosis [53]. So, it is difficult to spatially attribute genomic regions as LADs or NADs because both the nuclear lamina and nucleoli appear to be preferred sites of heterochromatin anchoring; as such, LADs and NADs may constitute interchangeable scaffolds for heterochromatin [49].

NADs have been subdivided into type I and II NADs based on their chromatin composition and the type of LAD they correspond to [52, 54]. Type-I NADs often associate with the nucleolar periphery and the lamina and contain marks of constitutive heterochromatin. Type-II NADs are not found at the lamina, display more pronounced H3K27me3 and tend to harbor higher gene expression levels than type-I NADs. The functional significance, if any, or tethering type I or type II NADs at nucleoli and/or the nuclear lamina remains intriguing and worthy of exploration.

Nuclear speckles are membrane-less intranuclear bodies in the nuclear interior of mammalian nuclei, enriched in RNA splicing factors and associated with active genes [39, 55]. These megabase-size regions contain high RNA polymerase II, are active and have been designated speckle-associated domains (SPADs). SPADs make up 5% of the genome and provide a new feature of nuclear organization [39].

1.1.6 Temporal dynamics of chromatin domains

Associations between TADs, and between TADs and the nuclear lamina or nuclear bodies are dynamic and can vary between cell types (Fig. 1.2B) [56, 57]. As such, time plays an important role in genome conformation, and provides a 4th dimension to genome topologies. This is well illustrated during stem cell differentiation and somatic cell reprogramming to pluripotency, which both provide insights on the dynamics properties of the genome. (i) One study shows the dynamics of chromatin contacts during differentiation of human endothelial cells, through chromatin 'switches' between A and B compartments and long-range TAD-TAD interactions [58]. (ii) Another example results from single-cell Hi-C data and 3D genome modeling, and reveals the dynamics of TADs and compartments at different stages of the cell cycle [59]. Interestingly, the authors show that the average intensity of TAD borders changes during the cell cycle, where average intensity is least at mitosis, increases during G1 and plateaus during S phase [59]. (iii) High resolution in situ Hi-C analysis of neural differentiation of mouse embryonic stem cells (ESCs) shows that contacts between A compartment domains decreases while contacts between B compartment domains increase [60]. (iv) Reprogramming of B-cells to pluripotent cells has been shown to lead to a switching of domains between A and B compartments [61]. Functionally, A-to-B switching domains are enriched in immune-related genes (and correlate with gene repression) whereas B-to-A switching domains are enriched in genes related to early development that become activated [61].

These studies illustrate the importance of the temporal perspective in 3D genome architecture. The time perspective in 3D genome topologies is further explored in **Paper II**.

1.2 Molecular methods to study the 3D genome

Microscopy-based techniques, chromosome conformation capture methods combined with high-throughput sequencing techniques and genome-editing based techniques are the predominant approaches to study 3D genome architecture and dynamics. Some of the techniques relevant to my work are addressed here.

1.2.1 Microscopy imaging techniques

Visualization of specific genomic sequences in the nucleus plays an important role in the study of 3D genome organization. Several microscopy-based methods enable examinations of the co-localization or physical proximity of genomic loci and the position of loci relative to nuclear structures such as the nuclear lamina or nucleoli. Nuclear DNA is commonly visualized with a fluorescent stain such as DAPI (Fig. 1.4). Additionally, physical proximity between two (or more) genomic sites can be examined by DNA fluorescence in situ hybridization (FISH) in fixed cells [48, 62] (discussed below), or using chromatin labeling techniques (e.g. ANCHOR) in live cells [63]. Modified genome-editing tools based on the CRISPR technology are being used to monitor chromatin dynamics, but not genomic interactions, to date [64, 65]. Live-cell chromatin imaging techniques are not addressed in this thesis. We refer to a non-exhaustive list of publications on live (and fixed) cell imaging approaches applied to characterize 3D genome conformation [53, 63, 66–73].

FISH relies on the detection of specific DNA sequences in cells using fluorescently labeled complementary oligonucleotides as probes [74]. FISH has been used for chromosome and gene copy number determination, including in disease diagnosis and prognosis [75, 76]. In 3D genomics, DNA-FISH is used to study the relative position between two or multiple loci [77], or chromosome territories [7, 8] in the nuclear space (**Paper II**), and the distance between loci and nuclear components such as the nuclear lamina, nuclear speckles and nucleoli (by immuno-FISH).

FISH entails cell fixation, permeabilization and DNA denaturation and hybridization of the labeled probes to their target loci. FISH signals are detected by fluorescence microscopy (Fig. 1.4). In addition to microscope resolution, two factors are critical in the study of 3D architecture by FISH. (i) Maintenance of nuclear architecture as close as possible to the native structure. This is nearly impossible however, due to the cell fixation and DNA denaturation treatments. Yet, maintaining cells on slides or coverslips throughout the procedure ('3D FISH') arguably maintains nuclear architecture better than isolating nuclei after cell fixation and dropping the nuclei on a slide for further processing ('2D FISH'). However, a 2D versus 3D FISH comparison in our laboratory has

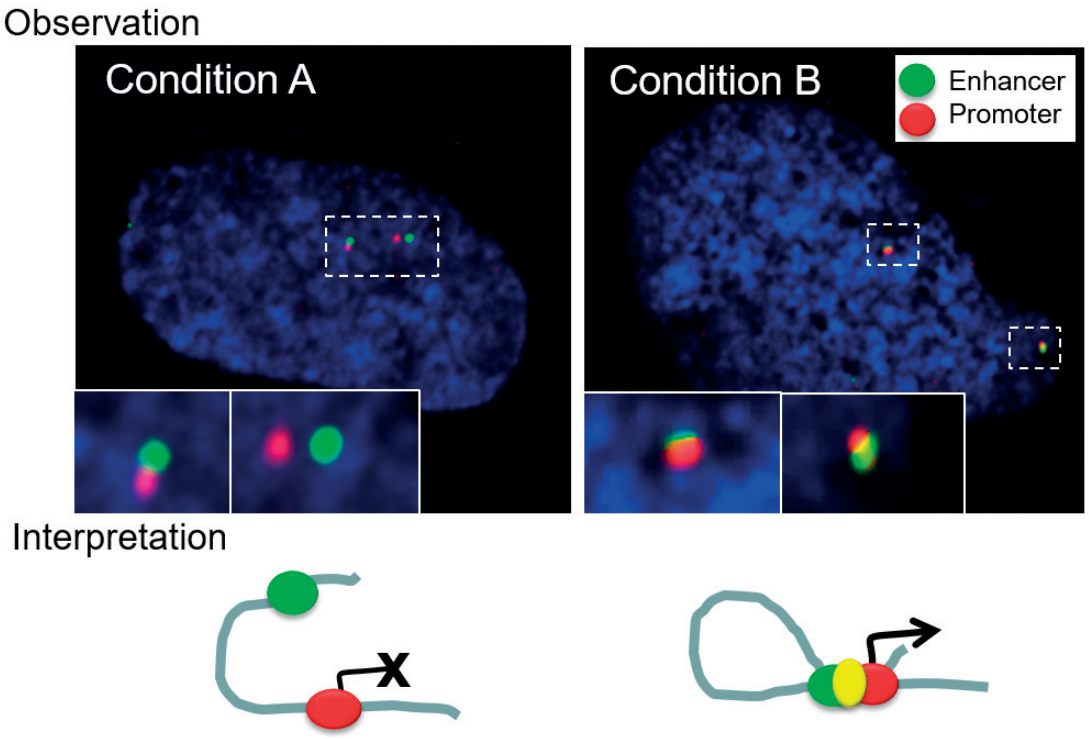

Figure 1.4: Detection of genomic sites by fluorescence in situ hybridization (FISH). Probes were designed to detect a gene enhancer (green) and promoter (red) under two experimental conditions (A, a wild-type condition and B, a nuclear lamin mutant condition) in human adipose stem cells. Note the separation of the two signals under condition A and their proximity under condition B. This is interpreted as long versus short distance of the two sites in 3D space. Promoter-enhancer proximity in B correlates here with gene activity (arrow). Nuclear DNA is stained with DAPI.

not shown significant differences in the degree of spatial association between multiple FISH probes, arguing that evidence for a more proper preservation of nuclear architecture in 3D FISH than 2D FISH is not always strong (A.L. Sørensen, T. Germier and P. Collas, unpublished data). This is presumably because formaldehyde fixation of cells or nuclei turn the latter into rigid objects that are not easily deformed through physical handling (P. Collas, unpublished observations). (ii) Length of the probes. Large probes (often bacterial artificial chromosomes covering ∼100 kb) incorporate more fluorophores than short probes and produce strong signals; they are often used to monitor chromatin compaction or the position of loci relative to nuclear landmarks. Yet long probes are suboptimal to study short distance (< 100 kb) chromatin associations, or when accurate distances between FISH signals are required; there, shorter probes (fosmids; ∼40 kb) are often used [77]. A chromatin 'contact', or proximity, is defined by a distance threshold, which mainly depends on Euclidean distance between the loci examined and microscope resolution [64].

1.2.2 Chromosome Conformation Capture methods

Recent progress in our appreciation of spatial chromatin architecture comes from high-throughput sequencing-based methods which enable the detection of genome-wide chromosomal contacts, and interactions between chromatin and intranuclear structures such as the nuclear lamina, nucleoli or nuclear speckles. Methods that can be referred to as '3D' methods, such as Hi-C, GAM and SPRITE, aim to identify chromatin contact points and lead to the view of the hierarchical organization of the genome described above [9, 56, 57]. '2D' methods such as Dam-identification (DamID)-sequencing (seq) and chromatin immunoprecipitation (ChIP)-seq are used to elucidate interactions between chromatin and nuclear structures and have been key in the identification of LADs [33] and NADs [50].

Chromosome Conformation Capture (3C)

Chromosome conformation capture, or 3C, aims to investigate pairwise chromatin interactions [78] (Fig. 1.5). Underlying 3C methods are five steps: (1) formaldehyde fixation which crosslinks interacting DNA fragments, (2) digestion of DNA with a restriction enzyme, the choice of which depends on the required frequency of base-pair digestion and the final resolution required (4- and 6-base pair cutting enzymes are frequently used [78–82]); (3) ligation of the interacting DNA fragments (the ensemble of ligation fragments is called a 3C library), (4) detection of ligation junctions (see below), and (5) analysis of these ligation products to determine contact frequencies of the regions investigated [1]. In a 3C experiment (see Fig. 1.5, primers are designed near the ends of the selected restriction fragments. The ligation frequency between non-neighboring region can then be quantified by counting the number of ligation events between the selected primer combinations. So, 3C allows quantification of contact frequencies at selected regions in 'one-versus-one' manner.

A key issue in 3C assays in general is that any two sequences that are close in linear genomic distance crosslink and ligate more frequently than sequences separated by hundreds of kb, independently of the 3D conformation of chromatin. There are many caveats in applying quantitative 3C polymerase chain reaction (PCR), and the approach requires strict controls, and careful design and data interpretation [83, 84]. Moreover, 3C experiments yield relative contact frequencies between two sites, reflecting close proximity, or 'interaction', between these sites; however additional experiments are needed to understand the functional relevance of such interactions in 3D space.

Chromosome conformation capture-on-chip (4C)

Chromosome conformation capture-on-chip (4C) is a 'one-versus-all' method (Fig. 1.5). One genomic viewpoint, or bait, is selected and proximal regions (interactors) are identified [85]. The initial steps of 4C are same as for 3C but the difference is the use of two restriction enzyme digestion steps. Once

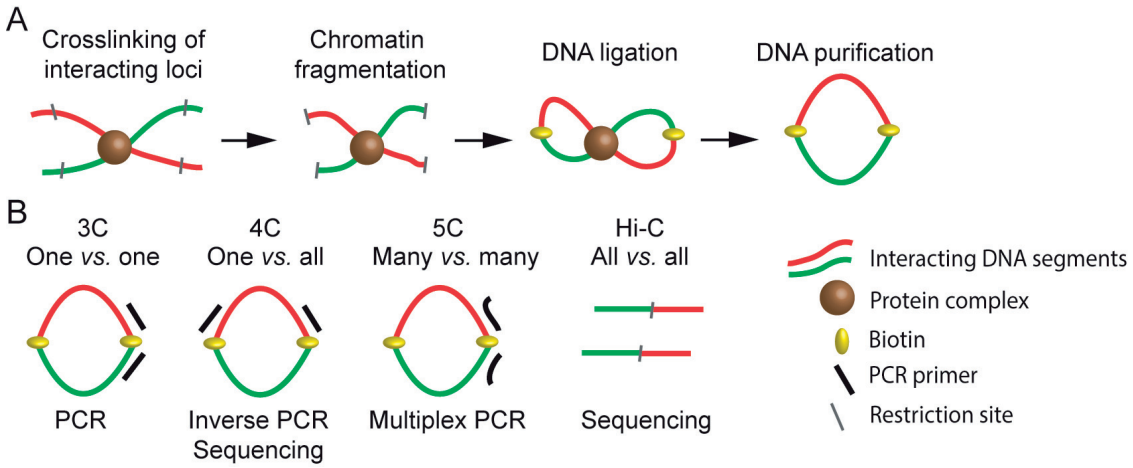

Figure 1.5: Chromosome conformation capture (3C) methods. (A) Common biochemical steps of all 'C' methods. (B) DNA sequence detection principle in 3C, 4C, 5C and Hi-C.

after the ligation step as in 3C, in 4C the ligated junctions are cut using a frequently cutting secondary restriction enzyme and self-ligated to generate DNA circles. Inverse PCR with primers specific to the first restriction junctions (viewpoint) amplifies the ligation products, which are determined and quantified using DNA microarrays or, more frequently today, high-throughput sequencing [86]. Variations of 4C, designated open-ended 3C [87], circular 3C [88] and olfactory receptor 3C [89], differ from classical 4C only in some specific steps but follow the same principle and not described here.

Chromosome conformation capture carbon copy (5C)

Chromosome conformation capture carbon copy (5C) extends 3C with the use of multiplexed ligation-mediated amplification [90]. 5C enables the capture of all interactions between several viewpoints across a selected region, and thus is a 'many-versus-many' approach (Fig. 1.5). There, a mix of 5C-specific primers is annealed onto the 3C library and ligated. These primers are designed so that the forward and reverse primers anneal across the ligated junction of the 3C products. These annealed 5C primers are ligated and the generated 5C library is amplified with universal PCR primers annealed to the 5C primers. The 5C products are quantified using microarray or sequencing. This method is limited to identifying interactions within the selected region.

1.2.3 Hi-C

The extension of 3C technologies to the determination of chromosomal interactions genome-wide ('all-versus-all') through a technique called Hi-C (Fig. 1.5) has been a breakthrough in the field of 3D genome organization [9]. Hi-C essentially follows the same initial steps as the 3C methods outlined above, including

chromatin crosslinking, fragmentation and ligation. The ligated products are purified and processed for massive parallel sequencing.

More specifically, cells are crosslinked with formaldehyde and DNA digested with a restriction enzyme. The ultimate resolution limit of Hi-C data is the restriction fragment length after DNA digestion. The original Hi-C protocol used *Hind*III and *Nco*I restriction enzymes, both recognizing and cutting a 6-base pair sequence (AAGCTT and CCATGG, respectively) [9]. Subsequently, 4-base pair cutters such as *Dpn*II have been used [16], which have more abundant target restriction sites (GATC), yielding smaller fragments and improved resolution. Other Hi-C variations rely on even shorter restriction fragments [91]. The restriction enzyme-mediated DNA overhangs are filled with biotinylated dinucleotide triphosphates. The blunt ends of biotinylated DNA segments are ligated under diluted condition to avoid self-ligation (the procedure in this case is called dilution Hi-C). Then, DNA is sheared and the ligated DNA hybrids are purified using streptavidin. The pulled-down DNA hybrids contain fragments from each of the two ligated regions and are subjected to paired-end sequencing. Hi-C demands deep sequencing to construct high-resolution genome-wide interaction maps. Details on the nature of the data generated in a Hi-C experiment and on the analysis of such data are provided in section 1.3.2.

Hi-C has been improved by introducing in situ Hi-C [10, 16]: there, the crosslinking and ligation steps are performed in (supposedly) structurally intact and permeabilized nuclei rather than in bulk suspensions. In addition to a simpler handling, the major advantage of in situ ligation is a reduction of the frequency of random ligations which are observed in dilution Hi-C. Using this method, Rao et al. [16] have achieved a resolution of up to 1 kb. In **Paper I**, we used already processed high-resolution IMR90 Hi-C data from Rao et al. [16]. In **Paper II**, we used dilution Hi-C with 25 million cells per sample, as this was the technique used in our collaborating laboratory. In **Paper III**, we used Hi-C data from four ENCODE cell lines.

Hi-C typically requires millions of cells (though more recent protocols are now adapted for much fewer cells), which makes it impossible to appreciate the variability in chromosomal interactions between cells. As addressed later in this Introduction, and in our work (**Paper II**), computational methods enable inferences on cell-to-cell variations in genome structures, but do not generate biological data. However, single-cell Hi-C now enables mapping chromosomal interactions in many individual cells. Single-cell Hi-C involves in situ crosslinking and ligation, preserving interactions in each cell. Individual nuclei are isolated to produce sequencing libraries [10, 59, 92].

Single-cell Hi-C contact matrices reveal variability in the nature of contacts between cells [10] or during the cell cycle [59], which can be recapitulated by FISH analysis [69]. This cell-to-cell variation is reflected in single-cell Hi-C contact matrices that are all different from each other [10], but the union of these matrices recapitulates those generated in ensemble Hi-C. A limitation of single-cell Hi-C is the sparsity of contacts seen in each single cell [10, 59]. Typically, tens to hundreds of thousands of interactions are captured, but technical improvements claim identification of $\sim 10^6$ interactions per cell [93].

1.2.4 Hi-C and Hi-C-derived methods

Capture Hi-C

Although Hi-C captures genome-wide interactions in a presumably unbiased and unsupervised fashion in cis and trans, it requires deep sequencing to attain high-resolution contact maps necessary to query interactions between regulatory elements such as enhancers and promoters; this can be prohibitively expensive for many laboratories. To evade this issue, capture Hi-C (CHi-C) [94–96] is based on the generation of a conventional Hi-C library, and enrichment of chromosomal contacts with specific sets of genomic sites (baits; e.g. promoters, in 'promoter capture Hi-C') using an oligonucleotide-based hybridization This enriches the Hi-C library for interactions with the selected baits. CHi-C results in high resolution contact maps between the targeted regions. There are of course limitations to capture Hi-C with respect to sensitivity (identification of weak cis or trans interactions), and CHi-C was recently re-designed as 'NG ('new generation') Capture-C' with improved sensitivity [97].

HiChIP and HiChIRP

HiChIP is a combination of in situ Hi-C and ChIP. The steps include crosslinking of DNA interactions in situ and chromatin immunoprecipitation to capture DNA interactions associated with the protein of interest. Paired-end sequencing and bioinformatics analysis identifies sites of genomic enrichment in the protein of interest and interacting genomic regions at these sites. HiChIP is similar to ChIA-PET (see below) but requires much smaller cell numbers and yields over 10-fold long-range interactions-informative reads [98].

HiChIRP has been developed to target long-range interactions between chromatin and a specific RNA. The method follows the same principle as HiChIP but instead of immunoprecipitating a protein, interactions are captured by affinity-isolation of an RNA of interest followed by paired-end sequencing of the associated DNA [99].

ChIA-PET

Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) aims to identify all interactions between genomic regions bound by a protein of interest [100]. This protein is immunoprecipitated, interacting DNA segments are ligated and the ligation products are detected by paired-end sequencing. Limitations of ChIA-PET, namely the use of two different DNA linkers, has been overcome by the use of a single biotinylated DNA linker and is better suited for higher-order 3D mapping [101]. The latter ChIA-PET assay produces several types of results, such as self-ligation data (similarly to ChIP-seq), clustered inter-ligation interactions mediated by the immunoprecipitated protein, and long range interactions [101].

1.2.5 Non C based Methods

Recently, two non-ligation based methods, genome architecture mapping (GAM) and split-pool recognition of interactions by tag extension (SPRITE) have been introduced in recent years to determine chromosomal interactions.

GAM

GAM is a ligation-free method to study 3D genome architecture [56]. The principle is to measure the distance between loci by cryo-sectioning nuclei using laser microdissection, followed by massive parallel sequencing [56] (Fig. 1.6A). GAM entails slicing purified nuclei in random orientations; the DNA content of each nuclear profile is extracted, PCR-amplified and sequenced. It is expected that loci that are in close proximity in the 3D nuclear space are detected in nuclear slices more frequently than loci far apart. From slice and sequence information, a matrix is created by counting the events of co-localization of all possible pairs of loci in a large collection of nuclear slices. The matrix allows the calculation of genome-wide contact probabilities, similarly as in a Hi-C matrix. GAM notably enables inference of chromatin contacts, compartments and radial positions [56].

Unlike Hi-C, GAM claims to capture multivalent interactions involving three or more genomic regions, and to require fewer cells than Hi-C [56]. However, GAM demands time and unparalleled skills to dissect nuclei at random fashion. Moreover, the heterogeneity of 3D genome topologies between cells in a population likely requires the dissection of hundreds

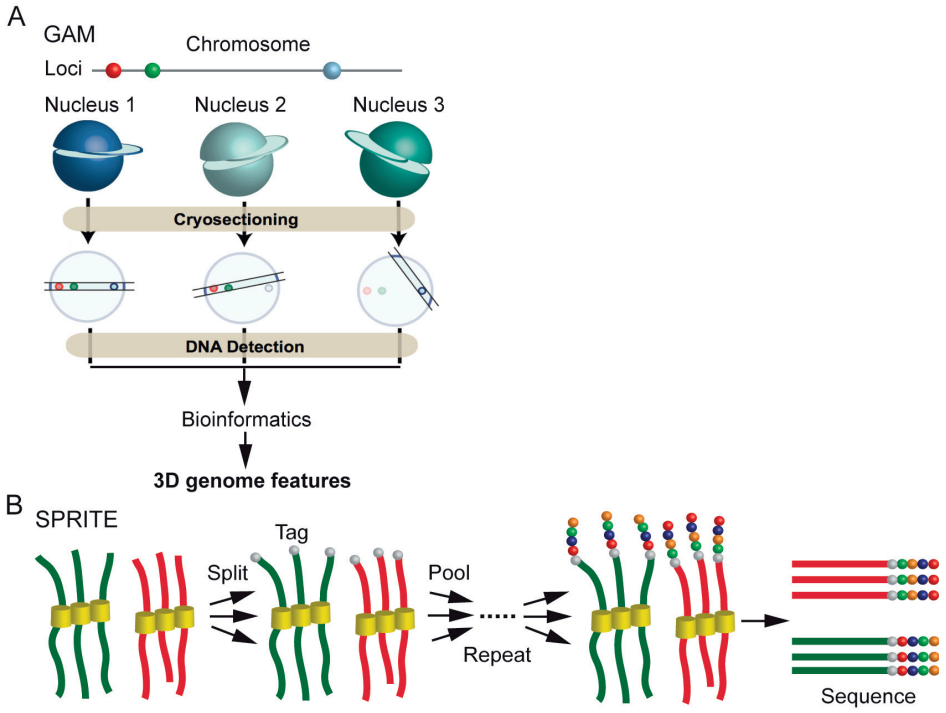

Figure 1.6: GAM and SPRITE. (A) Outline of genome architecture mapping (GAM). Loci in close proximity on a chromosome (red and green beads) are more frequently found together in thin nuclear slices than distant loci. DNA sequencing of nuclear slices and advanced computing are able to reconstitute 3D chromosomal interaction maps. In contrast to Hi-C which only detects pairwise interactions, GAM allows the detection of 3-way interactions. Reproduced with modifications and permission from [56] (Springer Nature). (B) Split-pool recognition of interactions by tag extension (SPRITE) enables detection of multi-way (here, 3-way) interactions. Redrawn and modified from [57].

or thousands of nuclei to appreciate this variation, and will influence the results [102].

SPRITE

Instead of proximity ligation, SPRITE relies on a split-pool strategy to identify genome-wide interactions between multiple regions [57]. Chromatin is crosslinked, nuclei isolated and chromatin fragmented. The chromatin fragments are split into many wells, ligated with a tag (a genetic barcode) specific to a well and re-pooled (Fig. 1.6B). The pooled fragments are again split, coupled to a new genetic barcode and re-pooled. This process is repeated multiple times, with DNA molecules accumulating unique barcodes as they are split into different wells each time. Since DNA fragments that are crosslinked (because they interact) are always

together in the same well in each split-pool round, they have identical barcodes; in contrast, distinct molecules split in different wells accumulate distinct barcodes. The DNA fragments are sequenced and accumulation of sequencing reads with identical barcodes mapped to specific genomic locations identifies sites that are assumed to interact [57].

SPRITE confirms genomic structures observed by Hi-C and GAM such as compartments and TADs. SRITE can also provide representations of chromatin regions in contact with nuclear bodies. SPRITE data notably show that active genes cluster around nuclear speckles, whereas inactive regions tend to organize around nucleoli [57].

1.2.6 Techniques to study DNA-protein interactions

DNA binding proteins play an important role in many cellular processes such as replication, splicing, transcription, DNA repair and genome organization. Nuclear lamins, transcription factors and post-translationally modified histones (e.g. by methylation or acetylation) are viewed as DNA-binding proteins. DNA-protein interactions can be studied by ChIP [103] and DamID [104].

Chromatin immunoprecipitation-sequencing (ChIP-seq)

ChIP-seq consists in the immunoprecipitation of a target chromatin-bound protein and identification of the associated DNA by sequencing [105]. DNA and proteins are crosslinked with formaldehyde, chromatin is fragmented by digestion with micrococcal nuclease or by sonication. Antibodies against the protein of interest are used to immunoprecipitate protein-DNA complexes. The crosslinks are reversed, DNA is purified and sequencing libraries are made and sequenced. ChIP has been a preferred method to profile histone modifications or transcription factor binding. ChIP-seq has been useful in mapping LADs interacting with A- or B-type lamins in various cell types or during differentiation [44, 106–110]. In **Paper II**, we used ChIP-seq to identify lamin B1 LADs and domains enriched in H3K27me3 and H3K9me3. In **Paper III**, we used publicly available CTCF ChIP-seq data to categorize TADs.

DamID-seq

DamID (Dam identification) was introduced as an alternative to ChIP to study DNA-protein interactions [104]. It is a proximity DNA labeling method where the bacterial DNA adenine methyltransferase (Dam) is

fused to a protein of interest (e.g. lamin B1, to identify LADs). Dam is targeted to DNA regions associated to the protein of interest where it methylates adenines at GATC sites to generate 6-methyl-adenine (^{m6}A), which does not normally occur in eukaryotes. Genomic regions containing ^{m6}A introduced by Dam are selectively amplified and hybridized to DNA microarray or sequenced [104, 111]. DamID has been used in several applications, for instance to probe genomic regions associated with nuclear lamins at the nuclear periphery [41]. In **Paper II**, we used mouse constitutive cLAD data generated by DamID-microarray from Peric-Hupkes et al. [38] to study association between cLADs and TAD cliques in mouse embryonic stem and differentiated cells. Of note, variations of DamID [112] allow the visualization of LADs in living cells, providing new insights on the dynamics of LADs between cells and over time [53, 113, 114].

1.3 Computational techniques to study 3D genome

1.3.1 Analysis of ChIP-seq data

The aim of ChIP-seq is to identify genomic regions enriched in the chromatin-bound factor of interest or in a specific histone modification. Significantly enriched regions are commonly called peaks or domains depending on their width. Several ChIP-seq peak calling software are available – e.g. MACS [115].

In a typical ChIP-seq data analysis, sequencing reads are mapped to a reference genome using a mapping tool such as BWA [116] or Bowtie [117]. Then, reads that are mapped to multiple location are filtered out to get uniquely mapped reads (this decreases the number of false positives). After filtering, peak callers are used to identify peaks. MACS is a widely used tool to identify peaks of transcription factors or and histone modifications and uses a dynamic local Poisson distribution to ascribe significance [115]. We have used MACS in **Paper II** to identify H3K9me3- and H3K27me3-enriched regions in human adipose stem cells.

LADs can also be mapped by ChIP-seq. In contrast to most histone modifications however, lamin ChIP-seq typically reveals broad domains of low-level enrichment. Thus analysis requires a different strategy to identify LADs, which has led to the release of Enriched domain detector (EDD) by our laboratory [108]. EDD bins the genome equally and calculates the smallest bin size that contains signal maxima by using the Agresti-Coull method. Then, each bin is scored and a gap penalty is assigned for non-informative bins. EDD detects domains using a linear algorithm by

identifying maximal scoring subsequence bins. Finally, for each domain, a P-value is assigned by Monte Carlo trial [108]. We have used EDD in **Paper II** to identify lamin B1 LADs.

1.3.2 Analysis of Hi-C data

Preprocessing and filtering

The first step, detailed in this section, is to map the paired-end sequences to a reference genome, and filter the noisy reads. Mapping can be performed using standard tools such as BWA [116] or Bowtie [117]. Ideally, the two ends of paired-end Hi-C reads correspond to two interacting loci linearly far apart along the genome; thus paired-end reads are expected to map to different locations. Some of the reads span the ligation junction, so parts of reads are from distant interacting loci; these reads are called chimeric reads. Chimeric reads require a specific strategy to be mapped, or are otherwise discarded with a full-read mapping approach. This peculiarity of Hi-C reads leads to chimeric read mapping which are implemented in many pipelines such as ICE [118], HiCUP [119], HiC-Pro [120] and TADbit [121]. Mapped reads are filtered to remove artefactual noise; to do so, reads are first removed based on filters common to all sequencing-based methods such as number of mismatches in a read, reads mapped multiple times, quality score of reads and PCR duplicates. After initial filtering, reads are assigned to the nearest restriction sites in the reference genome, so that the mapped reads are expected to be close to the restriction site; thus, reads that are mapped far from the closest restriction site are filtered out. Read-pairs from random ligation, self-ligation and dangling ends are removed. As a result, only informative read pairs, also called valid pairs, are filtered in and used for downstream analysis [118–121].

Binning

Valid read pairs are used to generate raw contact maps. Hi-C reads are not analyzed at the fragment level because of sparsity and difficulty of the analysis. Instead, reads are aggregated to genomic bins of fixed size. The result is a symmetrical matrix which contains the frequency of interactions between two genomic bins (X_{ij}) (Fig. 1.7A,C). The genome-wide interaction map contains both intra-chromosomal and inter-chromosomal data. The resolution, or bin size, is selected based on the

depth of sequencing but there is no way of *a priori* knowing an optimal bin size. Rao et al. proposed to select a minimum bin size such that 80% of all bins are covered by at least 1000 reads [16]; another approach is to use a fragment level resolution (expectedly very high) despite the computational demand. I would like to mention that there are methods developed to analyse Hi-C data without binning, for example, a method by Spill *et al.* [122].

Normalization

Raw data from Hi-C include biases inherent to the experiment which directly affect the Hi-C contact map. To eliminate these, a normalization is done mainly based on two principles: explicit factor and implicit factor correction.

In the explicit factor correction, as the name suggests, all explicit factors such as read mappability, fragment length between two restriction sites and GC content are taken into a single factor vector. This factor vector is used to normalize the contact probabilities using non-parametric step functions [123]. The major limitation of this approach is computational cost. HiCNorm is another explicit factor normalization method where a single parametric Poisson regression model is used to model reads at the bin level and much faster than non-parametric step functions [124].

The implicit correction method was implemented in the ICE pipeline [118]. ICE performs iterative correction to remove biases without any explicit assumption of sources of biases, and assumes that all regions of the genome have the same coverage, that is, equal visibility. This method is also known as a matrix-balancing algorithm; the resulting normalized matrices have equal-sum rows. An improved version of iterative correction which exploits the sparsity of high-resolution data is implemented in HiC-Pro [120]. Recently, visibility normalization by combining the advantages of the explicit and implicit methods has been implemented in HiCorr [125]. A new method called Binless normalization [122] handles normalization at the read pair level without any assumption of explicit biases or equal visibility of loci.

Hi-C matrices of cancer genomes may contain biases such as aneuploidy, copy number variation and translocations, in addition to the sequence-level biases mentioned above. These biases are addressed in CaICB, a regression-based chromosome iterative correction tool [126]. Additional software packages useful in the analysis of Hi-C data from aneuploidy cancer cells: OneD explicitly corrects regional copy number variation in

the Hi-C matrix [127]; LOIC and COIC are extensions of matrix balance algorithms and are appropriate for cancer genomic studies [128].

Identification of compartments

As discussed earlier, A/B compartments can be identified from Hi-C data (Fig. 1.3; 1.7D). Once biases are eliminated from Hi-C matrices, expected data matrices (dependent on the linear chromosomal distance) are generated from the normalized matrices for each chromosome. Then, observed/expected (O/E) matrices are generated (Fig. 1.7B), followed by the calculation of Pearson correlation values between all pairs of rows and columns in these O/E matrices. A principal component analysis (PCA) is applied on the generated Pearson correlation matrices. The resulting first principal component, PC1, defines A/B compartments. Genomic regions with positive eigenvalues are A compartments, and negative eigenvalues are B compartments [9] (Fig. 1.7D). Though, the computational definition of genomic regions as A or B compartment has to be verified by calculating the GC content of the genomic regions and swapped to ascertain that A and B reflect GC-rich and -poor regions, respectively. A similar but improved method has been implemented in the Cworld (https://github.com/dekkerlab/cworld-dekker) and in the HiTC R package [129] where instead of using O/E matrices, loess Z-score matrices are generated before calculating Pearson correlations and PCA. A new faster and efficient statistical method to identify compartments is implemented in CscoreTool; there, the score $C_i = 2P_i - 1$ is calculated to reflect that a given genomic bin P_i is in an A compartment [130].

Identification of TADs

Several algorithms have been developed to identify TADs across genomes. The first approach to identify TADs genome-wide, with a tool called DomainCaller [18], calculates one dimensional score called the directionality index (DI) to quantify the degree of upstream and downstream interaction bias of a given bin, defined as:

$$DI = \left(\frac{B-A}{|B-A|}\right)\left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E}\right) \tag{1.1}$$

where in equation 1.1, A is the number of mapped reads from a given bin to the upstream 2 Mb, E is the number of mapped reads from a given bin to the downstream 2 Mb, and E is the expected number of

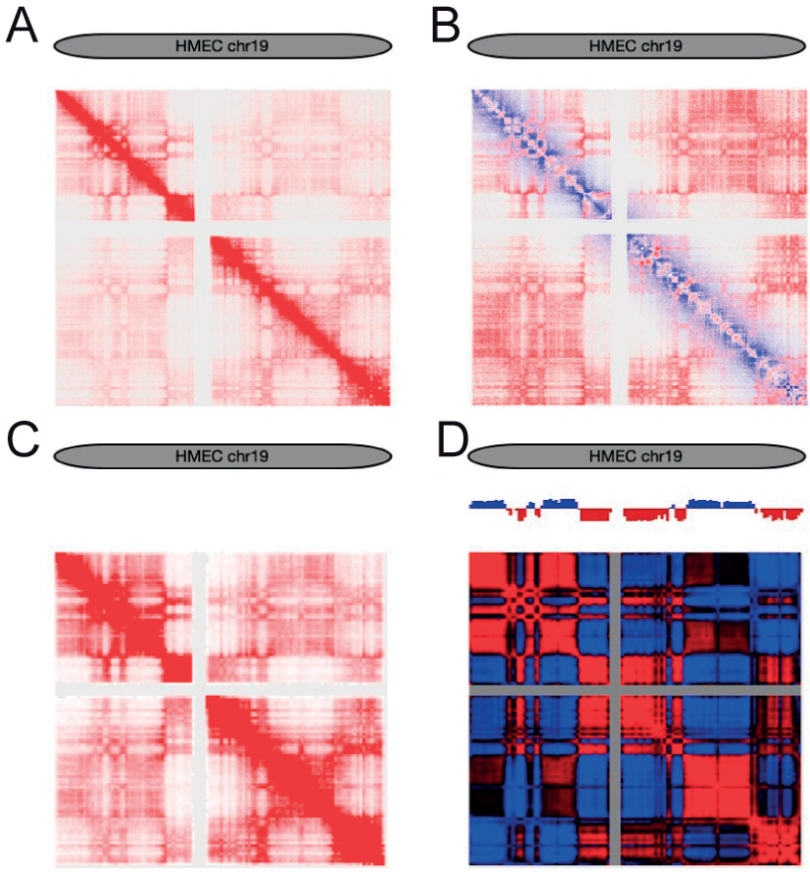

Figure 1.7: Hi-C matrices of the entire chromosome 19 in human mammary epithelial cells (HMEC) at different resolutions. (A) Observed matrix at 50 kb resolution. (B) Observed/Expected (O/E) matrix at 50 kb resolution. (C) Observed matrix at 500 kb resolution. (D) Bar graph shows the eigenvector of PC1 (from PCA) and the typical plaid pattern of a Pearson correlation matrix at 500 kb resolution. A and B compartments are shown in blue and red, respectively, in the track above the O/E matrix. T.M. Liyakat Ali, unpublished.

reads from a null distribution. The DI is based on the χ^2 distribution and is segmented with a hidden Markov Model (HMM) where a sharp transition from an upstream interaction bias to a downstream interaction bias identifies TADs.

The Armatus algorithm uses the Dynamic Programming approach to find domains in a Hi-C contact matrix with a tunable, single domain-length scaling parameter γ [131]. This algorithm returns non-overlapping consensus domain sets (TADs) that are consistent across multiple resolutions of the Hi-C data [131]. Rao *et al.* proposed the Arrowhead algorithm where the matrix is transformed in a way to enhance domain boundary signals; then, a heuristic algorithm identify 'corners' to determine domains [16]. A Hi-C contact map can also be considered as a 2D image used to

identify TADs as a 2D image segmentation problem [132]. Using this 2D image segmentation algorithm, blocks are detected along the diagonal of a contact matrix [132]. Another simple method called TopDom has also been proposed, where a *binSignal(i)* curve is calculated for each bin along the chromosome [133]. A *binSignal(i)* curve of i^{th} bin represents the contact frequency between the bin and the neighboring bins. Theoretically, bins around the center of TADs have high *binSignal(i)* values and low values at TAD boundaries. Domain boundaries are detected as local minima in the *binSignal(i)* series, and TADs are defined [133].

The main limitation of these methods is the assumption that TADs are not nested. However, recent substantial evidence shows that groups of sub-TADs cluster or combine to form large TADs (Fig. 1.8) [134, 135]. This limitation has been addressed in additional methods such as TADtree [136], IC-Finder [137] and PSYCHIC [138].

Several algorithms define TADs using a clustering approach, such as the Clustering-based Hi-C Domain Finder CHDF tool [139] and an unsupervised machine learning problem in ClusterTAD [140]. Graph theory approaches have also been developed to define TADs. The principle is assuming that a Hi-C matrix is the adjacency matrix of a graph where bins are represented as nodes and TADs are hubs of interacting bins in a graph. Some graph theory approaches used to define TADs include a network optimization problem in MrTADFinder [141], a Laplacian-based graph segmentation in 4D NAT [142] and an optimizing network modularity in 3DNetMod [143].

Of note, identifying TADs is still an open problem in spite of all methods released to this end, and no gold standard method that has been established at the time of this writing. In **PaperII**, we used Armatus [131] to call TADs for all Hi-C samples and replicates and identified consensus TADs conserved across replicates and samples. Our rationale for this is that TADs are well conserved across cell types, and we required an identical set of TADs across differentiation time-points to enable comparisons of TAD-TAD interactions.

Identification of interactions

The subsequent step in Hi-C data analysis is to identify 3D interactions between domains (Fig. 1.9). To this end, the genome should be segmented for example as bins or as TADs defined using one of the methods outlined explained above. Several tools are available to find 3D interactions from Hi-C contact maps. Identification of both short- and long-range interactions

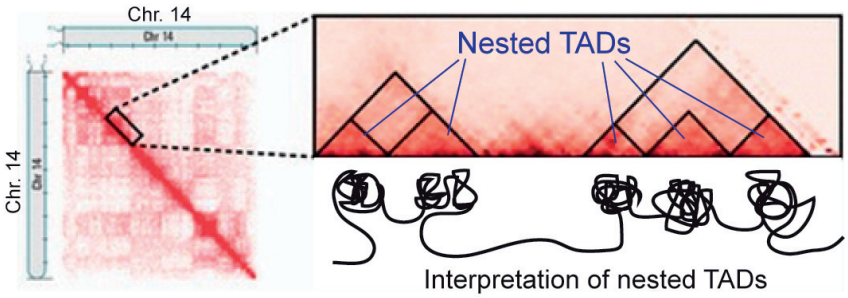

Figure 1.8: Illustration of a nested TAD structure. Left panel, Hi-C matrix for chromosome 14 in mouse embryonic stem cells (mESCs) from Dixon et al. [18]. Groups of 2 TADs (left in the enlarged matrix diagonal) and 3 TADs (right) are embedded, or nested, within large domains. A simple graphic interpretation of the nested domains is drawn. Reprinted with permission from [136].

requires a background distribution to compare interaction frequencies of the observed and expected. To date, two background models are proposed: (i) a genome-wide or chromosome-wide model where each segment pair is compared against expected interaction frequencies globally, and (ii) a local model, which compares interaction frequency of each segment pair against its surrounding segment pairs.

HICCUPS is a part of the Juicer suite developed by Durand *et al.* [144]. It employs the local model strategy to find 3D interactions by identifying contact-enriched bins relative to the neighborhood of the bins. A drawback of HICCUPS however is that it is recommended for very high resolution Hi-C maps as it finds few interactions even in high resolution data compared to another local model tool such as PSYCHIC [145]. PSYCHIC uses the local enrichment technique. The first step involves genome segmentation into domains using a unified probabilistic model and a Dynamic Programming algorithm. Then TADs are iteratively merged to a nested structure and for each TAD, interactions are modeled according to a local background model (with a power law regression) to identify significant interactions [138].

As mentioned briefly, genome-wide methods find significant chromatin interactions by comparing an interaction frequency against the expected interaction frequency derived using statistical models from the Hi-C data. Fit-Hi-C uses a genome-wide background model defined by two non-parametric splines from the input data to find the significant domain interactions [146]. It uses the iterative correction (ICE) proposed by Imakaev and colleagues [118] to remove experimental biases from the input data. GOTHiC is another example of a genome-wide background model, which uses a binomial cumulative distribution to find the significance of

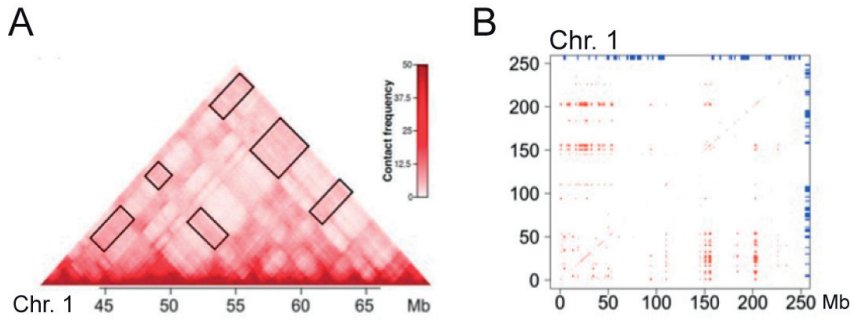

Figure 1.9: Significant TAD-TAD interactions identified in Hi-C data. (A) Hi-C matrix showing long range interactions (black boxes off the matrix diagonal, shown here as a half-diagonal flipped 90 degrees) along a segment of chromosome 1 in human adipose stem cells (used in **Paper II**). (B) Matrix plot showing significant TAD-TAD interactions within chromosome 1 in human adipose stem cells, identified using NCHG. The red pixels indicate significant TAD-TAD interactions. The blue bars on the top and right of the matrix indicate TADs interacting with the nuclear lamina (LADs).

observed interaction [147].

Throughout my PhD work, I have used the non-central hypergeometric distribution (NCHG) [148] to identify domain interactions. The algorithm was originally developed as ChiaSig to identify interactions from ChiA-PET data [148], and has been updated to identify both intra- and inter-chromosomal interactions in Hi-C data. NCHG incorporates genomic distance-dependent relationships to calculate the conditional probability of the number of interactions [148]:

$$P\left(n_{ij}|n,n_i,n_j,\omega_{ij}\right)=\frac{\binom{n_i}{n_{ij}}\binom{2n-n_i}{n_j-n_{ij}}\omega_{ij}^{n_{ij}}}{\sum_{n'_{ij}}\binom{n_i}{n'_{ij}}\binom{2n-n_i}{n_j-n'_{ij}}\omega_{ij}^{n'_{ij}}} \tag{1.2}$$

where n is the total number of interactions in a given matrix, n_i and n_j are the number of interactions involved in TADs i and j, and n_{ij} is the number of interactions between TADs i and j. ω_{ij} is the parameter that includes genomic distance and λ_{ij} is the expected interaction frequency.

$$\omega_{ij}=\frac{\lambda_{ij}\ (2\lambda-\lambda_i-\lambda_j+\lambda_{ij})}{(\lambda_i-\lambda_{ij})(\lambda_j-\lambda_{ij})} \tag{1.3}$$

NCHG is used to call TAD-TAD interactions in all the papers and an example significant TAD-TAD interaction matrix is shown in the Fig. 1.9B.

Visualization of Hi-C

Visualization of Hi-C data is important for interpretation, hypothesis generation, annotation, validation and presentation of results. Independently of any statistical analyses, chromosomal interaction patterns can sometimes be easily visually detected (Fig. 1.10), such as a genetic variation disrupting long-range interactions observed in contact maps comparing two types of data, correlation of epigenetics states and Hi-C features, changes in 3D genome topology during cell cycle or during differentiation (e.g. [62, 149]; **PaperII**). Several tools are available to visually explore and compare processed Hi-C data and integrate multi-omics datasets. Most tools help visualizing Hi-C contact maps as arc maps representing loops, circos plots, rectangular heatmaps and triangular heatmaps with basic features to pan and zoom along two dimensions; these also commonly allow integration of multi-omics data. They are either web-based or stand-alone graphical user-interfaces. Some of the widely used tools are mentioned below.

Juicebox has been developed by the lab of Leiberman-Aiden to visualize Hi-C contact maps as rectangular heatmaps [144]. It is a standalone tool written in Java and recently released as a web-based tool (https://github.com/aidenlab/Juicebox). On top of basic visualization features, it is possible to calculate and visualize Eigen values, normalized heatmaps, Pearson-correlation plaid pattern heatmaps, and others. Users can directly download published Hi-C datasets or open their own dataset in .hic format [144]. Juicebox.js has also been released (https://github.com/igvteam/juicebox.js/tree/master) for web/cloud-based visualization of Hi-C data [150]. Some of the figures in this thesis are generated using Juicebox (e.g. Fig. 1.7).

Higlass.io is another web-based interface to visualize Hi-C data as heatmaps [151]. It has many of the Juicebox features such as panning and zooming; however, a special feature of this tool is the synchronized exploration of multiple Hi-C datasets that can be arranged in a single window for better comparison [151] (Fig. 1.10). Lastly, in addition to the dynamic user-interface tools, publication-ready static plots can be generated to highlight genomic regions of interest, using Python packages such as HiCPlotter [152] and HiCExplorer [153], and R packages such as Sushi [154] and HiTC (Fig. 1.9A) [129].

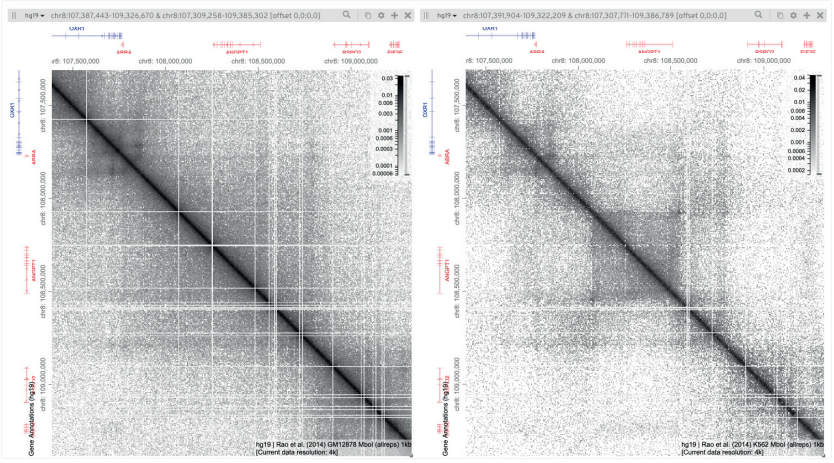

Figure 1.10: A screenshot from the Higlass.io web-based tool. This example illustrates the synchronized exploration of, here, two Hi-C datasets (left and right panes). The left pane shows chromosome 8 region of the GM1278 cell line; the right pane shows the same chromosome region in the K562 cell line. Note the distinct chromosomal interaction patterns in the two cell lines: TADs are clearly defined in K562 (right) compared to GM1278 (left).

1.4 Computational 3-dimensional genome structural modeling

The wet-lab and analysis approaches described above do not necessarily and directly provide quantitative descriptions of how chromatin interacts with other nuclear components in a 3D space (even Hi-C matrices are 2D representations of 3D interactions), how chromatin domains are repositioned in the nuclear space (e.g. during cell differentiation), variability between cells, and how chromosomes fold [1]. To fill these gaps, over the past three decades, physicists and computer scientists have developed DNA, chromatin and whole-genome modeling techniques to investigate spatial chromatin organization. Computational modeling approaches can be broadly classified into two types: (i) polymer physics-based modeling (also known as theoretical or direct modeling) and (ii) data-driven modeling (also known as restraint-based or inverse modeling) [155].

1.4.1 Polymer physics-based modeling

Polymer physics models can have the potential of bringing predictive mechanistic information of chromatin organization to a quantitative level such as chromatin compaction, interaction frequencies, position of chromosomes, end-to-end distances and other features. Polymer physics models rely on assumptions and parameters deduced from the laws of

polymer physics [156]. In polymer physics modeling, chromatin is viewed as a long semi-flexible polymer chain of N successive monomers (usually beads, cylinders or rods), with an apparent contour length s (the genomic distance). The semi-flexible polymers can adopt an infinite number of configurations but are limited by the persistence length, i.e. the length of the polymer below which it behaves as a rigid rod [156].

In the simplest model, monomers adopt a random walk without constraints, such that resulting chromatin chain configurations accommodate a confined nuclear space. Several models assume various thicknesses and compositions of the chromatin chain [157], which arguably provide a more realistic view of chromatin. For example, in the 'micelles model', mammalian chromosomes can be represented as co-polymers (polymer chains composed of two or more monomers type reflecting different biological properties) of GC-rich and GC-poor blocks modeling heterochromatin and euchromatin domains [158]. According to this model, monomers of same (chromatin) type are allowed to interact with each other (homotypic interactions) while monomers of different types repel each other [158]. This simple model predicts the observation of clusters of experimentally validated replication foci (active GC-poor regions) in mammalian nuclei [159].

Four main kinds of polymer physics-based modeling methods have been developed, with various assumptions and parameters defining how monomers in the chromatin (or DNA) polymer chain interact with each another. These are Strings and Binders Switch (SBS) models, block co-polymer models, loop extrusion models and the related Slip-link models [160] (Fig. 1.11).

Strings and Binder Switch (SBS) models

Some of the older polymer models such as the Random walk/giant-loop model [161] and the Micelles model [158] ignore the existence of chromatin-binding proteins diffused in the nucleoplasm, and which significantly affect chromatin bending and topology. In the SBS model [162], the chromatin filament is described as a self-avoiding polymer chain made of different types of monomers, or 'strings'. Monomers of different types have different binding affinities for proteins (binders) which mediate interactions of homotypic monomers at distant sites along the chain (Fig. 1.11A).

The thermodynamic state of an SBS polymer depends on two parameters: the interaction energy E_{int} and the concentration of binders C_m [163]. Based on these parameters, an SBS model can take three main

states: (i) at low E_{int} and C_m, due to its self-avoiding nature, the polymer swells and remains in a non-condensed state; (ii) at higher E_{int} and C_m values, the polymer undergoes coil-phase transitions to become folded and collapse; (iii) at even greater E_{int} and C_m, the polymer attains a thermodynamically stable and ordered globular state [163]. Changes in E_{int} and C_m values dictate the transition of folding phases, or 'switches'. These changes can be explained by relatively simple biological processes such as an increase or decrease in transcription factors concentration (binding factors) or changes in the chromatin properties (e.g. an epigenetic chromatin state) of part of the polymer (binding site) which alters the interaction energy [164].

SBS modeling can recapitulate the average contact probability $P(s)$ of interacting chromatin beads for a given chromatin chain of contour length s. Thus, SBS models can recapitulate chromatin features such as TADs and compartments found in Hi-C contact maps [162]. In more realistic SBS models, information on 'binders' and chromatin domains can be added from ChIP-seq analyses of, respectively, transcription factors and histone modifications. This information contributes to more accurately explain contact patterns along genomic regions in different organisms [165–167].

Block co-polymer models

Block copolymer modeling is a generic and minimal chromatin modeling technique also based on preferential interactions of chromatin domains with similar signatures [160, 165] (Fig. 1.11B). Genome bins of 10 kb are treated as blocks that are constrained to drive homotypic interactions. Beads in a polymer chain are connected via a harmonic potential and steric self-avoidance is provided (as in SBS models). A Gaussian-like potential energy models homotypic monomer interactions. Despite its simplicity and the exclusion of biological aspects of genome folding, block co-polymer models can recapitulate large scale Hi-C contact maps and TADs when built from epigenomic features [165].

Loop-Extrusion and slip-sling models

The loop extrusion model and related slip-sling model [167] assume that a loop extruding factor (LEF) anchors two specific points on the chromatin chain (Fig. 1.11C, D, ring) and drives progressive expansion of a loop (Fig. 1.11C,D). Loop extrusion is energy-dependent (loop extrusion model) or

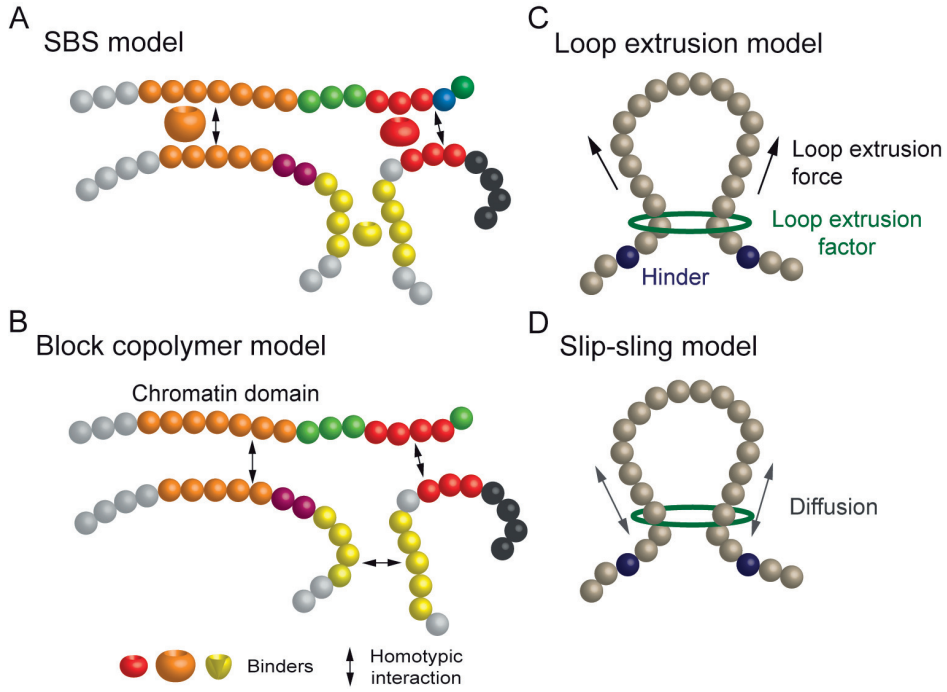

Figure 1.11: Main polymer physics-based 3D genome structural modeling methods. (A) Strings and Binders Switch model. Binders (e.g. transcription factors) mediate homotypic chromatin interactions with similar (epi)genomic properties. (B) Bock copolymer modeling also models homotypic interactions albeit without binders of SBS modeling. (C) Loop extrusion model. Chromatin loop formation is modeled by an active force acting on chromatin, extruding it through a loop extrusion factor (e.g. a cohesin complex) until it encounters a barrier, or hinder (such as a CTCF site). (D) A slip-sling model is similar to the loop extrusion model but loop formation occurs here by diffusion of a loop extrusion factor relative to chromatin.

occurs by random diffusion (slip-sling model), and stops when the LEF is hindered by a barrier, such as a protein bound to chromatin. The extrusion process therefore brings two distant regions of the chromatin polymer chain closer together at the base of the loop (Fig. 1.11C,D) [24]. The loop extrusion model can explain loops, stripes and TADs found in Hi-C data [16] (see Fig. 1.3). It is supported by studies of loop extrusion mechanisms [168] and is consistent with a view where cohesin may act as a LEF and the extrusion process is halted by CTCF bound to convergent sites [24, 169]. The loop extrusion model however cannot explain A/B compartments, implying that other mechanisms, such as those put forward in SBS and block copolymer modeling, also intervene in chromatin folding.

1.4.2 Restraint-based chromatin modeling using consensus methods

The availability of experimental data from imaging techniques and biochemical techniques, particularly high-throughput chromosome conformation capture techniques, prompted an alternative approach of 3D genome modeling called restrain-based, or data-driven, modeling [1]. These approaches use restraint-based methods to deduce spatial information on genomic domains directly from the data and reconstruct 3D models without assumptions on folding mechanisms. These methods are similar to methods previously used to reconstruct atomic structures of molecules from nuclear magnetic resonance data [156]. In restraint-based models, interaction frequencies derived from Hi-C or other 3C-based experiments are used to find interacting domains. 3D genome models are generated using contact probabilities or Euclidean distance between two domains obtained from contact maps as restraints. The modeling methods vary based on the resolution of representation of chromatin. Regardless, three main categories of restraint-based modeling techniques have been reported: (i) consensus methods represent the data by a single averaged 3D structure which is considered as a 'best fit' structure [170–177]; (ii) resampling methods simulate the conformation variability of structures, recapitulating structural variability between cells in a population [48, 178, 179]; (iii) deconvolution methods, also known as population-based methods deconvolute interaction maps from ensemble data [134, 180–182].

Consensus models produce a single structure from the underlying ensemble data (i.e. Hi-C data generated from a population of cells) [173, 176]. An ensemble Hi-C contact frequency matrix is converted into Euclidean distances based on the assumption that contact frequencies between two regions are inversely proportional to their genomic distance. A single 3D structure of the genome is generated, which minimizes the residual error between modeled and expected distances by a scoring function [173, 176]. The three types of scoring functions used in consensus modeling are (i) a likelihood optimization function relying on Bayesian inference [171], (ii) multi-dimensional scaling [172], and (iii) solving a generalized linear model [177]. Consensus models have been reported for the whole genome [170], a single chromosome [171, 174] or part of a chromosome [172]. In consensus modeling, physical constraints such as avoiding steric hindrance between beads in a same chromosome and preventing breakage of the bead chains can be incorporated to enhance accuracy of the resulting 3D structure [170]. One key advantage of consensus modeling is that it can rapidly generate and summarize a

structural feature from ensemble data.

However, by definition, consensus structures cannot capture the variability in genome conformations observed between cells in a population [10, 59, 69]. Additional modeling techniques have therefore been developed to produce a large number genome structures whose properties can reflect the 3D genome dynamics. Two fundamentally different approaches are used to generate ensemble 3D genome structures: resampling techniques and population-based deconvolution methods.

1.4.2.1 Restraint-based modeling using resampling methods

Resampling methods perform large number of independent optimizations, each starting from a random chromosome configuration and using the same scoring function. Depending on the statistical power required for analysis, hundreds or thousands of 3D models are generated, which are similar but not identical, and with variations that capture some of the variability in genome structures between cells in the population under study [48, 183–185]. In simulations, constraints are derived from 3C based techniques (preferably Hi-C because it contains whole-genome information for whole genome modeling), and optionally, from positional information of chromatin in the nucleus (e.g. its anchoring to the nuclear lamina, the nucleolus or speckles). Several available computational platforms and software employ resampling methods with different sources of inputs and optimization functions [48, 183–185]. Some of these derive pairwise Euclidean distance between domains from interaction frequency maps (e.g. [180]), while others use contact frequencies [178, 186] or the statistical significance of pairwise contacts [48, 182]. Examples of resampling methods to model 3D genome structures are addressed below.

Some of the earliest resampling-based models have been generated to model 3D genome structure in budding yeast [179]. In this study, modeling incorporates a bead chain restraint (bead-bead contacts), a bead chain volume restraint (accounting for chromatin thickness) and interestingly, chromatin positional constraints. Indeed, using prior knowledge on the organization of the yeast nucleus, positional constraints are added: (i) chromosomes are constrained to a confined nuclear space, (ii) telomeres are constrained to locate at the nuclear periphery (this is motivated the Rabl-like configuration in yeast where telomeres are juxtaposed near the nuclear envelope [187], (iii) centromeres are clustered and constrained to the spindle pole body at one pole of the nucleus, and (iv) rDNA repeat regions from all chromosomes are constrained to a nucleolus at the pole opposite to the spindle pole body [179]. The scoring function is defined

as a sum of all constraints, which are derived from experimental data and aims to reach a final score of zero at the end of the optimization, in order to obtain a stable genome structure [179]. Chromosomes are modeled as chains of beads. The optimization procedure using Integrative Modeling Platform (IMP) [188] and starts with a random bead configuration. This is followed by the initial optimization of the structure and simulated annealing to equilibrate the genome structure. Then, a conjugate gradient method ensures that no constraints are violated, leading to structures with a final score of zero. The optimization process is run thousands of times to generate thousands of 3D models which reflect the expected variability of genomic configuration of yeast cells in a population [179].

TADBit is a 3D genome modeling package which also includes a resampling method [121]. This framework generates ensemble of 3D genome structures also using IMP [188]. Here, chromosomes are modeled as chains of beads where each bead represents a segment of the genome (e.g. TAD). The volume of a bead is modeled based on the linear genomic size of the segment; for example, if genome segmentation is based on TADs, bead volume is proportional to the genomic size of the TAD it represents. Information for each chromosome is duplicated as an approximation to be able to model the diploid state of the genome. Contact frequencies from a 3C-based method (again, usually Hi-C) are converted into pairwise contact restraints between beads. In other words, the distance between beads are restrained if the contact frequency exceeds the cut-off value or kept apart from each other if the contact frequency is lower than the cut-off. After an initial random conformation is set, the simulation is run to minimize the IMP objective function. At the end of the simulation, a 3D genome model is generated, optimized to satisfy all restraints; yet the resulting structure is not perfect. Thus, hundreds of simulations are run parallel because each structure represents a local minimum of the IMP objective function [121, 180]. Analysis of the structures shows that properties of genome organization (e.g. chromosome territories) are preserved in the models [184].

1.4.2.2 Chrom3D – a genome modeling resampling method that incorporates positional constraints

Our laboratory has in 2017 developed and released a computational 3D genome structural modeling framework, Chrom3D, to integrate Hi-C constraints and a positional constraint based on the association of TADs with nuclear lamina (LADs) identified from lamin ChIP-seq data [48]. Of note, the initial version of Chrom3D was developed by the time I started

my thesis work, and was the starting point of my work. In Chrom3D, each chromosome is modeled as a chain of beads, where each bead represents a segment of the genome (e.g. a TAD). Size of beads are proportional to the genomic size of the segments. Input data to the Chrom3D are (i) significant pairwise interactions between segments derived from a Hi-C contact matrix and (ii) LAD information (provided e.g. by lamin A or B ChIP-seq data) as radial positional constraints of beads in the models (Fig. 1.12A). Note that adding LAD information is optional; if not available, Chrom3D can nevertheless be run. We have however in our publications systematically included LAD information in our Chrom3D modeling exercises except for local modeling (Papers I, II and III and papers not included in this thesis [43, 189]) (an example of local modeling in the Fig. 1.12D). Additionally, other constraints can be specified to e.g. position all beads within a sphere of a given radius (e.g. (5 µm) to reflect the nuclear boundary (Fig. 1.12C,E).

Modeling input information is entered into the Model Setup File (Fig. 1.13) in GTrack format [190] that can be passed as an input file. Chrom3D starts the simulation with an initial random bead chain configuration. A model is then generated using Monte Carlo optimization of a loss-score function (Fig. 1.12B). For each iteration, a random bead is selected and a random movement is imposed on the bead, from a set of pre-determined moves (e.g. bead chain translocation, rotation, wiggling or crankshaft) [48]. The loss-score for a particular iteration is calculated and the event is accepted based on the Metropolis criterion. During optimizations, the loss-score L (equation 1.4) is minimized using simulated annealing.

$$L = \sum_{i,j} k_{ij}(\parallel b_i - b_j \parallel - d_{ij})^2 \tag{1.4}$$

In the equation 1.4, k is the weight for a given interaction (to e.g. emphasize a bead interaction with the nuclear lamina or with another bead), b_i and b_j are interacting beads and d_{ij} is the distance between interacting beads.

In Chrom3D, it is possible to specify weights on each constraint based on empirical data through the parameter k of the loss-score function (equation 1.4). For example, beads containing LADs in human chromosome 18 can be given more weight to impose a stronger peripheral constraint because we know from FISH data that chromosome 18 is positioned at the nuclear periphery [191]. Finally, hundreds of simulations are run in parallel to produce an ensemble of 3D models which together recapitulate structural variability (Fig. 1.12). The radial positions of beads, which

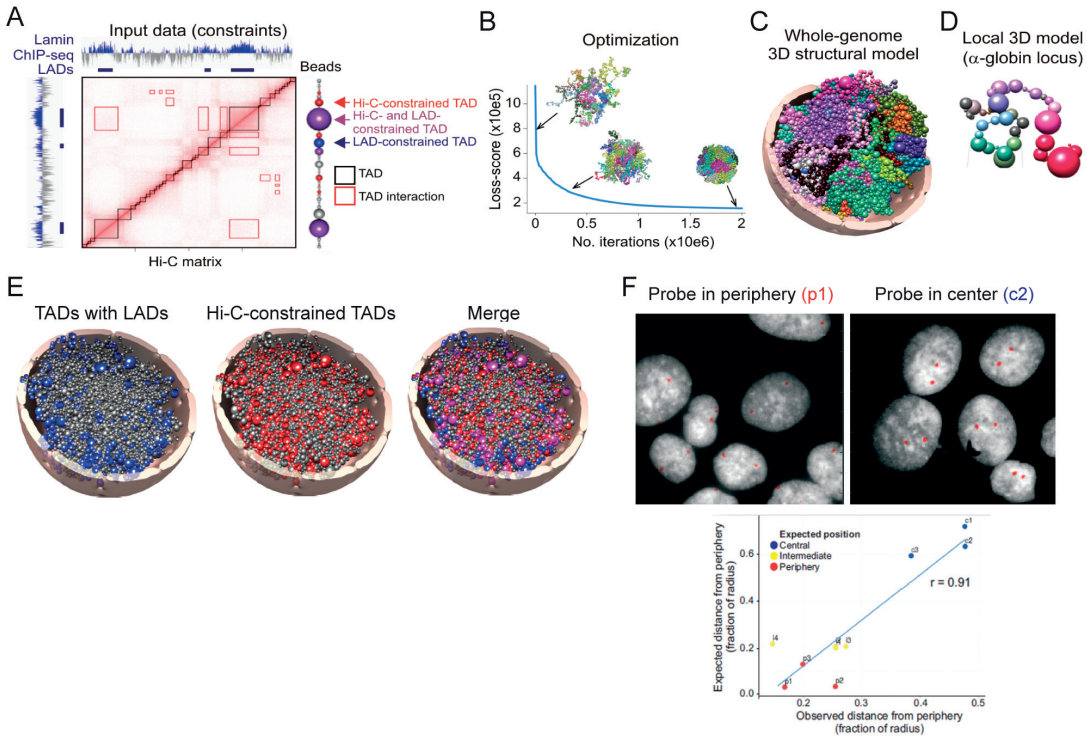

Figure 1.12: 3D genome structural modeling using Chrom3D, an ensemble modeling method. (A) Hi-C and LAD data are used as chromatin interaction constraints. Chromosomes are modeled as beads-on-a-string from domains in Hi-C data (one bead corresponds to a TAD). (B) Monte Carlo optimization starting from a randomly initialized structure. (C) Example of a Chrom3D model of a human fibroblast genome (colors represent individual chromosomes). (D) A Chrom3D model of Encode region ENm008 (500 kb) containing the α-globin gene. Hundreds of models can be made to enable statistical predictions of chromatin topologies at multiple scales. (E) Visualization of LADs, Hi-C constrained beads and both (merge) in a tomographic view of one Chrom3D structure of a HeLa cell nucleus. (F) Validation of radial positioning of loci by FISH. Panels are reproduced from [48] under CommonCreative licence.

can be determined across models, can be validated in FISH experiments [48] (Fig. 1.12F) and analysis of structures allows inference on the radial position of chromosomes, LADs [43, 48], specific loci [192] or UV-induced DNA lesions [193]. In **Paper II**, we used Chrom3D for quantitative analysis of co-localization and radial position of TADs involved in TAD cliques [62]. I have also used Chrom3D to predict the radial displacement of loci as they experimentally gain or lose LADs under external cues [43].

The algorithm behind Chrom3D is well explained with a practical application in the initial publication [48]. However, the report lacks detailed explanations of Chrom3D parameters, a step-by-step procedure to

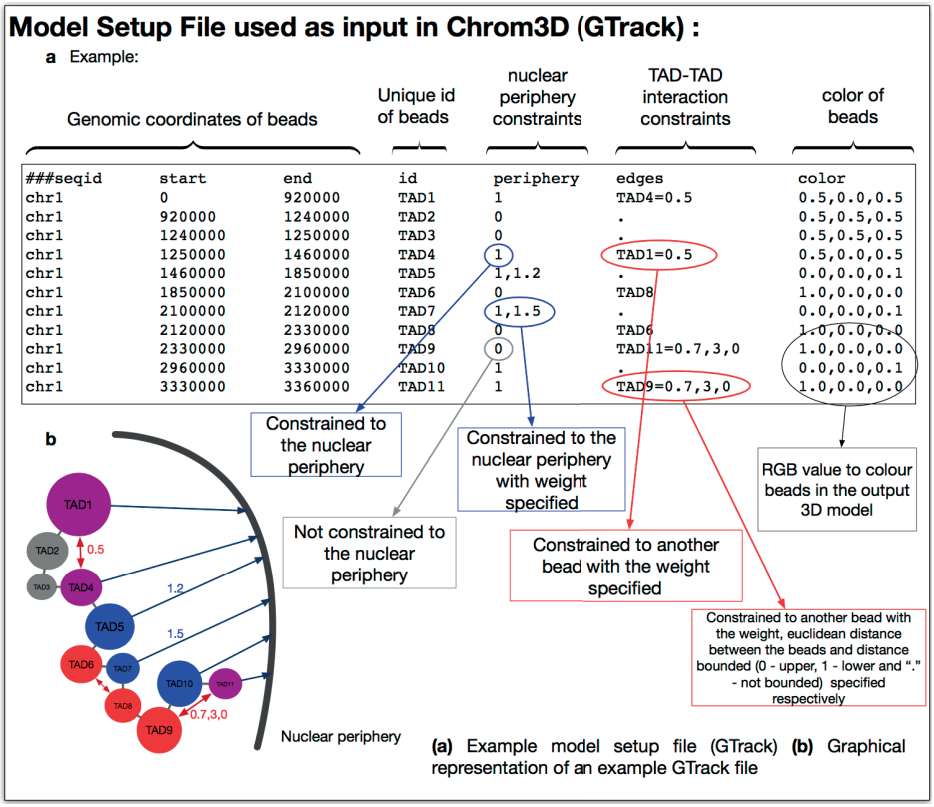

Figure 1.13: An example text and the graphical illustration of the model setup file (GTrack format). a) An example GTrack format; each field is annotated and the values for each categories of constraints are explained in the boxes. b) Graphic illustration of the example GTrack file where the black arc represents nuclear periphery, the red beads represent Hi-C constrained beads, the blue beads represent periphery constrained beads and the purple beads represent both Hi-C and periphery constrained beads.

create a model setup file (GTrack) and ways to weigh the constraints in the input file for users to fully exploit Chrom3D in their work. This led to the development of a pipeline and subsequent publication of a protocol paper [194] (**Paper I**). This paper details the process starting from processed Hi-C data and optional LAD data to create the model setup file, run Chrom3D simulations and produce publication-ready images. The protocol helps users running Chrom3D using their own data. This is maintained and supported on the github page (https://github.com/Chrom3D/pipeline).

1.4.2.3 Restraint-based modeling using deconvolution methods

The assumption behind the deconvolution methods is that ensemble Hi-C data arises from multiple chromatin structures and do not correspond to a single 3D structure shared by all cells of the cell population examined.

Multiple chromatin structures are therefore highly likely to be observed in a given population of cells [178, 181]. Deconvolution methods typically use an iterative probabilistic framework in order to demultiplex the data and construct a set of plausible structures. The generated structures recreate ensemble Hi-C data without generating physically unrealistic structures caused by violating constraints [178, 181, 195].

In deconvolution methods, rather than trying to impose ensemble Hi-C data on individual structures, structures are optimized as a group, with optimization formulated as a maximum likelihood estimation problem [182]. Moreover, the method does not need to assume any relationship between contact frequency or spatial distances, so long-range interactions can be recreated in individual structures [47, 182, 196].

Similarly to resampling methods such as Chrom3D [48], population-based modeling using deconvolution also allows for introduction of positional constraints on chromatin, for example from lamin DamID-seq (LAD) data [47]. Interestingly, the structures recapitulate the nuclear envelope and peri-nucleolar anchoring of heterochromatin domains consistent with LADs and NADs, even though no nucleolus constraint is provided [47]. Similarly to Chrom3D [48], this approach considers the structural variability between genomes and allows predictions of 3D genome properties that are not obvious in the underlying experimental data [47].

1.4.2.4 3D genome modeling and higher-order chromatin architecture in this thesis

The past two decades have witnessed an explosion of molecular- and sequencing-based approaches to study 3D genome conformation. Some of these methods also combine microscopy imaging. This development has been paralleled by the release of a plethora of computational frameworks and tools to analyze data and model chromatin in 3D. The field also witnesses the integration of physics-based modeling and restraint-based modeling to better understand genome architecture and function.

Within this context, Chrom3D was published in 2017 [48]; yet it required optimization and a computational workflow (pipeline) was needed. This is the subject of **Paper I**. In addition, there has been, up to the publication of **Paper II** in May 2019 [62], very little if no understanding of (i) how higher-order chromatin topology changes during cell differentiation, (ii) how such higher-order changes (in the form of long-range TAD-TAD interactions) would be perceived or occur in single cells. These are the topics developed in **Paper II**, whose results provide new insight into the

4D nucleome. This work was followed by a genomic characterization of TADs forming long-range associations as 'TAD cliques' in **Paper III**.

Chapter 2

Aims of the study

The 3D topology of the chromatin establishes blueprints of developmental gene expression. This is reflected at the level of the whole nucleus, on a large scale, by the radial (i.e. center-to-periphery) distribution of chromatin. Despite advancements in our understanding of 3D genome conformation, the long-lasting lack of suitable 3D genome computational modeling platform able to faithfully recapitulate and predict the positioning of genomic loci relative to each other and to the nuclear periphery has hampered our understanding of spatial genome conformation and of genome dynamics during stem cell differentiation and between cells in a population. It has also limited the ability to make testable predictions on the relationship between changes in the radial position of loci and associated gene expression changes. Lastly, it has hampered the ability to predict spatial relationships between genetic variants such as single nucleotide polymorphisms, which may provide additional clues on disease etiology.

In this context, the aims of this study were to:

- Optimize our structural 3D genome modeling platform, Chrom3D, and provide a detailed step-by-step protocol to the scientific community **(Paper I)**
- Identify and characterize higher-order changes in 3D chromatin topology during differentiation **(Paper II)**
- Develop and implement a computational approach to identify multi-TAD assemblies (TAD cliques) in single-cell Hi-C datasets **(Paper II)**
- Investigate patterns of long-range interactions between TADs across cell types **(Paper III)**
- Determine whether given genomic features define TADs in cliques versus TADs outside cliques **(Paper III)**

Own contribution:

My contributions in this thesis work have been at the conceptual and bioinformatics levels, ranging from problem identification and analysis, program conception, coding, data generation, analysis and interpretation, to the generation of figures and writing of manuscripts.

The wet-lab experiments reported in the publications have been carried by other members of the laboratory or by collaborators (Paper II). Specifically, adipose stem cell culture and cell differentiation, sample preparation for RNA-seq, ChIP-seq and Hi-C, ChIP-seq experiments (lamins and histones), FISH experiments and FISH data analysis were done by members of the Collas lab mentioned as authors and in the acknowledgement. Hi-C for Paper II was done by Dr. Maxim Nekrasov in Prof. David Tremethick's laboratory (Australian National University, Canberra, ACT, Australia).

My contributions were specifically as follows:

Paper I: coding to establish a seamless pipeline for Chom3D, including linking steps, coding help, and diagnostics tools, testing and optimizing the workflow; coding and testing mapping of epigenetic features onto 3D models; generation of figures; writing of the manuscript. After the publication of Paper I: I have added all the scripts to the GitHub portal (https://github.com/Chrom3D/pipeline) to make the pipeline to reach many users (open access); together with Jonas Paulsen, I am actively maintaining both Chrom3D and the pipeline providing ad hoc help to users (https://github.com/Chrom3D/Chrom3D/issues).

Paper II: concept and realization of analysis of TAD cliques during reprogramming of B cells into pluripotent cells, conceptualization and analysis of TAD cliques in single cells, including single-cell Hi-C data analyses, generation of figures, writing of the corresponding text, figure legends and methods; generation of 3D genome models and corresponding data analyses to explore the radial position of TAD cliques in adipose stem cells.

Paper III: Hi-C and TAD clique data exploration and analysis for all four cell types examined in the paper, identification and conceptualization of the questions, testing of hypotheses, data generation and analysis, generation of figures, writing of the manuscript.

In addition, I have significantly contributed to another published paper not included in this thesis (*Forsberg F, Brunet A, Liyakat Ali TM, Collas P. 2019. Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. Nucleus 10, 7-20*). I have analyzed publicly available Hi-C data for the HepG2 cell line. The 3D genome models were generated using the Hi-C and lamin A/C and lamin B ChIP-seq data from the laboratory (Forsberg and Brunet) and, wrote scripts to extract 3D coordinates of beads. I have generated 3D genome modeling parts of figures, written figure legends and methods and, approved the final manuscript.

Chapter 3

Summary of the papers

3.1 Paper I

Computational 3D genome modeling using Chrom3D
Paulsen J, Liyakat Ali TM, Collas P[$]. 2018. *Nature Protocols* 13, 1137-1152
[$]Corresponding author.

We report here Chrom3D, a computational platform for 3D structural genome modeling that simulates the spatial positioning of chromosome domains relative to each other and relative to the nuclear periphery. In Chrom3D, chromosomes are modeled as chains of contiguous beads, in which each bead represents a genomic domain. In this protocol, a bead represents a topologically associated domain (TAD) mapped from ensemble Hi-C data. Chrom3D takes as input data significant pairwise TAD–TAD interactions determined from a Hi-C contact matrix, and TAD interactions with the nuclear periphery, determined by ChIP-sequencing of nuclear lamins to define lamina-associated domains (LADs). Chrom3D is based on Monte Carlo simulations initiated from a starting random bead configuration. During the optimization process, TAD–TAD interactions constrain bead positions relative to each other, whereas LAD information constrains the corresponding bead toward the nuclear periphery. Optimization can be repeated many times to generate an ensemble of 3D genome models. Analyses of the models enable estimations of the radial positioning of genomic sites in the nucleus across cells in a population. Chrom3D provides opportunities to reveal spatial relationships between TADs and LADs. More generally, predictions from Chrom3D models can be experimentally tested in the laboratory. We describe the entire Chrom3D protocol for modeling a 3D diploid human genome, from the creation of input files to the final rendering of 3D genome structures. The procedure takes ∼18 h. Chrom3D is available on GitHub at https://github.com/Chrom3D/Chrom3D/releases/v1.0.1.

3.2 Paper II

Long-range interactions between topologically-associating domains shape the 4-dimensional genome during differentiation

Paulsen J, Liyakat Ali TM*, Nekrasov M*, Delbarre E, Baudement M-O, Kurscheid S, Tremethick D$, Collas P$. 2019. *Nature Genetics* 51, 835-843

*shared authorship. $Shared senior authorship.

Genomic information is selectively used to direct spatial and temporal gene expression during stem cell differentiation. Interactions between topologically associating domains (TADs) and between chromatin and the nuclear lamina organize and position chromosomes in the nucleus. However, how these genomic organizers together shape genome architecture is unclear. Here, using a dual-lineage differentiation system, we report long-range TAD-TAD interactions that form constitutive and variable TAD cliques. A differentiation-coupled relationship between TAD cliques and lamina-associated domains suggests that TAD cliques stabilize heterochromatin at the nuclear periphery. We also provide evidence of dynamic TAD cliques during mouse embryonic stem cell differentiation and somatic cell reprogramming and of inter-TAD associations in single-cell high-resolution chromosome conformation capture (Hi-C) data. Altogether, our findings indicate that TAD cliques represent a level of four-dimensional genome conformation that reinforces the silencing of repressed developmental genes.

3.3 Paper III

TAD cliques predict key features of chromatin organization

Liyakat Ali TM, Brunet A, Collas P$, Paulsen J$, 2020. *Manuscript.*

$Shared senior authorship.

Processes underlying genome 3D organization and domain formation in the mammalian nucleus are not completely understood. Multiple processes such as transcriptional compartmentalization, chromatin extrusion events and nuclear lamina interactions likely simultaneously and dynamically act on chromatin at multiple levels. We have explored long-range interaction patterns between topologically associated domains (TADs) across several cell types. We find that these patterns are connected to many key features of chromatin organization, including open and closed compartments,

chromatin compaction and loop extrusion processes. We find that domains that form large TAD cliques tend to be repressive across cell types when comparing gene expression, LINE/SINE repeat content and chromatin subcompartments. Further, TADs in large cliques are found to be larger in genomic size, less dense and depleted of convergent CTCF motifs, in contrast to smaller and denser 'typical' TADs explained by loop extrusion. Our results shed further light on the organizational principles that govern repressive and active domains in the human genome.

Chapter 4

Discussion

This thesis reports computational developments and analyses leading to seamless workflow of analysis of the genome in 3 dimensions (**Paper I**), new biological insights into the 3D organization of the adipose stem cell genome during differentiation (notably, TAD cliques; **Paper II**), and a genomic characterization of TAD cliques across several cell types (**Paper III**). I emphasize that my scientific contributions have been strictly computational and of bioinformatics nature, and have specifically led to the following outcomes:

- A seamless computational pipeline consisting of TAD identification and calling from Hi-C data, determination of consensus TADs between Hi-C datasets, identification of significant long-range TAD-TAD interactions, and other features. Altogether, this has led to significant improvements and, importantly, user-friendliness, of the Chrom3D framework just established in the laboratory [48] when I started my work.
- Use of Chrom3D by several laboratories around the world, as discussed below.
- The demonstration of a new level of higher-order chromatin topology in the form of TAD cliques.
- A method to identify and characterize TAD cliques in single cells.
- Application of Chrom3D genome modeling to characterize chromatin dynamics under defined experimental conditions [43] (work not reported in this thesis).

4.1 My contributions to Chrom3D

Chrom3D has been published in 2017 [48], before I started my PhD work, and I have been involved in the continuous development and support of Chrom3D. My contribution mainly includes increasing the usability and user-friendliness of this framework. For instance, prior to my involvement in the work, error messages reported by Chrom3D during data processing

```
Tharveshs-MacBook-Pro:Chrom3D-master tmali$ ./Chrom3D -o ./test_files/toy-global-example.cmm -r 3.0
-n 50000 -l 5000 ./test_files/toy-global-example.gtrack
libc++abi.dylib: terminating with uncaught exception of type std::logic_error: Error in the line #4 of the
GTrack file ./test_files/toy-global-example.gtrack
Same beadId found in the edge column, please check the edge column of the above line
Abort trap: 6
Tharveshs-MacBook-Pro:Chrom3D-master tmali$
```

Figure 4.1: Example of a user-friendly error message produced by Chrom3D. The message reports that in the 4th line, a self-interaction is defined; that is, the bead ID for a particular bead is found in the edge column (where bead-bead interactions are defined in the model setup file) of the same bead.

were highly technical; in other words, they could only be understood by skilled C++ programmers, limiting their usefulness to other users. More specifically, most error reports did not point out the source of errors. Therefore, I modified the code in order to catch errors reported by the framework and make it easier for many bioinformaticians, even with limited programming background, to understand them. For example, if a user encounters an error in the interaction information of a bead in the model setup GTrack file, the error report now points out the relevant code line (e.g. Fig. 4.1), making it a lot easier to identify for any Chrom3D user. All my contribution to Chrom3D can be found on GitHub (https://github.com/Chrom3D/Chrom3D/graphs/contributors).

4.2 Applications of Chrom3D modeling to our understanding of 3D genome architecture and dynamics

We have shown that Chrom3D can be applied to place post-translational histone modifications in the 3D models (e.g. H3K27me3, H3K27ac; **Paper I**), which, if quantified through e.g. spatial clustering analysis (unlike what we illustrate in **Paper I**), can provide predictive information on spatial domains of such modifications. Similarly, chromatin states [197], DNA methylation, TF binding or other chromatin-associated features can be mapped onto Chrom3D models with the aim of spatially characterizing their distribution.

Our laboratory has used Chrom3D to predict the radial positioning of UV-induced DNA lesions in human fibroblasts [193]. Modeling data show that UV susceptibility is enriched at the nuclear periphery relative to the nuclear center (Fig. 4.2A). The data suggest that heterochromatin at the nuclear periphery acts as a 'sink' for UV lesions, protecting the more centrally located gene-rich euchromatin, or alternatively, that DNA

lesions are more readily detected at the nuclear periphery because they are less efficiently repaired due to restricted access of the DNA repair machinery in this heterochromatic compartment [198]. Additionally, genes mutated in 5.1 to 10% of melanomas are significantly more frequently positioned at the nuclear periphery compared to genes not mutated in melanomas (Fig. 4.2B) [193]. These predictions could not be made only from mapping these DNA lesions onto a linear genome, arguing for the gain of information provided by 3D genome modeling.

I have applied Chrom3D in a recently published study (not included in this thesis) of the dynamics of interactions of chromatin with A- and/or B-type nuclear lamins, at the nuclear lamina, in HepG2 hepatocarcinoma cells used as an in vitro model of steatosis (induced by cyclosporine) [43]. ChIP-seq analysis of lamin A and lamin B reveals that a chromatin domain can interact with lamin A only (forming an 'A-LAD'), lamin B only (B-LAD) and both lamins A and B (A/B-LADs). I have produced 800 Chrom3D models of the HepG2 genome using Hi-C data for HepG2 (ENCODE, NCBI GEO accession GSE105382, sample GSM2825569) and our own LAD data for HepG2 cells before and after cyclosporine treatment. A-LADs, B-LADs and A/B-LADs (identified by ChIP-seq also in the study) were then mapped onto the Chrom3D models. Measurements of the radial positioning of these LADs highlight key features of genome organization (Fig. 4.2C): (i) A-LADs, B-LADs and A/B-LADs are more peripheral than inter-LAD regions; (ii) B-LADs and A/B LADs are closer to the nuclear periphery than A-LADs (Fig. 4.2D); (iii) a loss of lamin B (but not A) interaction correlates with displacement of loci from the nuclear periphery towards the center and, (iv) loss of lamin B from an A/B LAD or a switch from B-LAD to A-LAD also coincides with a more central position of the domain; (v) on the contrary, a gain of lamin B correlates with repositioning of the domain towards the periphery. These Chrom3D predictions were importantly validated by FISH [43]. Chrom3D models therefore enable predictions on the spatial repositioning of loci as a function of their lamin interactions.

4.3 Applications of Chrom3D modeling by other laboratories

Chrom3D is used by several groups worldwide. Our Nature Protocols publication on using Chrom3D (**Paper I**) [194] notably describes the step-by-step procedure to create the input file; this has significantly increased the number of Chrom3D users. In addition to the 'conventional' way

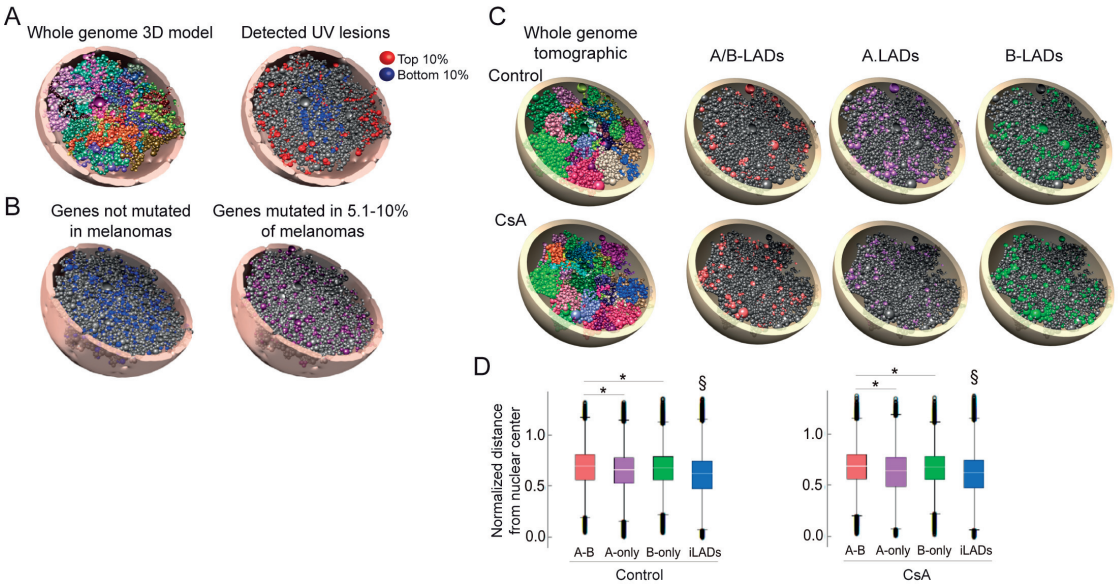

Figure 4.2: 3D positioning of genome features in Chrom3D models. (A) UV-induced DNA lesions are more readily detected at the nuclear periphery. Chrom3D model of the whole IMR90 fibroblast genome (left) and placement of the top 10% (red) and bottom 10% (blue) UV susceptibility sites in one Chrom3D model. In the models, one bead is a TAD called from Hi-C data for IMR90 fibroblasts. (B) Tomographic representation of the spatial localization of genes not mutated in melanomas (n = 1226) and genes mutated in 5-1-10% of melanomas (n = 714) (IMR90 Chrom3D models). (A, B), modified from [193] with permission. (C) Chrom3D models of the whole genome (tomographic views, left) and of A/B-LADs, A-LADs and B-LADs in control HepG2 cells and HepG2 cells treated with cyclosporine (CsA). (D) Normalized LAD distances from the nuclear center (0, center; 1, periphery) in 800 Chrom3D models of control and CsA-treated HepG2 cells. Distances for all other LADs (inter-LAD, or iLADs) are also shown. §P < 2.2x10-16 relative to each LAD class; unpaired t-tests; P < 2.2x10-16; unpaired t-tests; distances computed as absolute distances divided by the radius of the modeled nucleus (5 µm). (C, D), from [43] with permission under CreativeCommon license.

of using Chrom3D with both Hi-C and LAD constraints, other ways of running Chrom3D depend on the research question and data availability. As I have explained in the Introduction of this thesis, it is possible to run the Chrom3D with only Hi-C data (as bead-bead interaction constraints), and adding periphery interaction information is optional. Also, as demonstrated in the original Chrom3D publication [48], local modeling is possible; that is, 3D modeling of a specific locus or a genomic segment of interest. Examples of applications of Chrom3D by various research groups are discussed below.

A study by Espeso-Gil *et al.* shows the spatial colocalization of genome-wide association study (GWAS) risk sequences associated with schizophrenia and metabolic disorders using Chrom3D genome models (Fig. 4.3A). They termed the risk sequences found in confined closed proximity as 'Euclidean hot spots'. These hot spots mainly consist of risk genes with functional enrichment in lipid regulation functions, reward and addiction pathways, starvation response and regulation of food intake [199]. In addition to provide new information on spatial associations between genes enriched in specific functions, this study provides an example of application of Chrom3D using only Hi-C data without any need for lamin constraints because only chromosomal interactions are relevant here.

Other work explores the role of DNA replication timing in a 3D genome architecture context [200]. Using CRISPR-based genome editing, the authors show that early replication control elements (ERCE) play an important role in chromatin domain architecture. They also predict ERCEs along the mouse genome and show robust long-range CTCF-independent interactions between predicted ERCEs [200]. ERCEs may also be required for A/B compartmentalization [200]. In this study, they used Chrom3D to model the *Dppa2/4* locus (∼5 Mb) and show spatial proximity between ERCEs and predicted ERCE elements within the domain [200] (Fig. 4.3B).

A recent study by Tian et al. [201] exploits 3D genome modeling using Chrom3D to corroborate chromatin interaction networks. The authors developed an algorithm called MOCHI to discover cell type-specific heterogeneous interactome modules (HIMs). These HIMs represent clusters of loci which interact more frequently than expected and are regulated by the same group of TFs; in that sense, a HIM overall reflects a 'transcriptional niche' [201]. Of note however, even though the coined term of 'HIM' may be recent, the concept of spatially proximal co-regulated genes in the form of 'transcription factories' has been proposed (notably from FISH studies) as early as 1993 [202] and has been corroborated many times

since. 3D genome structural modeling, however, supplementing Hi-C data, nicely supports this concept. One of the HIMs in K562 cells is showed in Fig. 4.3C. In the 3D model, the authors show that a super-enhancer is spatially proximal to genes involved in the HIM.

4.4 Limitation of Chrom3D

Memory and time complexities

The main technical limitations of Chrom3D are memory and time complexities. The required memory and linearly increase with resolution of the data (the number of beads) and the number of iterations. So far, the maximum number of beads acceptable for each chromosome is hardcoded to 5000 but if a user wants an ultra-high-resolution models then this number can be increased in "chromosome.h" of the source code before compilation in order to increase the resolution of the model. This will, however, increase memory consumption.

Multithreading of the simulation may appear as an option to decrease runtime. However, multithreading is not supported in Chrom3D because every iteration depends on the loss-score of the prior iteration, and the move is accepted or rejected based on the Metropolis criterion. Therefore, if the iteration is multithreaded, tracking of the prior move has to be implemented, which adds more complexity. Next, the Chrom3D is implemented in C++, so it is already faster and well optimized for memory usage. Regardless, a user ideally needs to produce hundreds or thousands of models to perform statistical analysis on 3D positions; these can easily be executed in parallel to decrease runtime.

Minimizing the loss-score (i.e. finding a [local] minimum) faster might be one way to decrease runtime. In Chrom3D, simulated annealing is implemented to minimize the loss-score but is not optimally exploited. To use simulated annealing for minimizing the loss-score, the cooling rate parameter must be set between 0 and 1. The idea is that the temperature parameter gradually decreases during the simulation and is determined by the cooling rate for accepted moves. In other words, lowering the temperature during optimization speeds up loss-score minimization. Determining the optimal cooling rate is critical for faster minimization. An option to determine the cooling rate is explained in the GitHub page (https://github.com/Chrom3D/Chrom3D#temperature-parameter-and-cooling-rate).

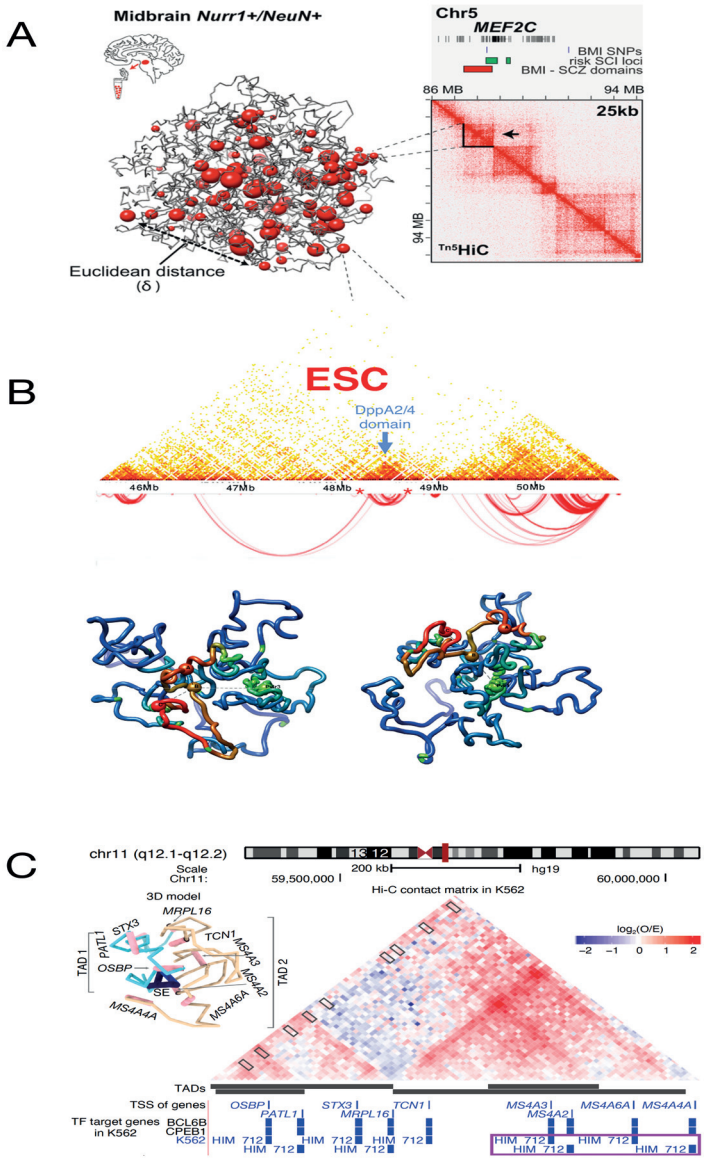

Figure 4.3: Examples of application of Chrom3D by different research groups. (A) Hi-C contact map of Chr.5 and a whole-genome Chrom3D model generated using human midbrain cell Hi-C data. The highlighted beads in red show the domains (TADs) harboring both schizophrenia and metabolic disorders risk variants. (B) Capture contact map of a ∼5 Mb region highlighting the DppA2 locus in the mouse genome and the corresponding locus in Chrom3D models, viewed from different angles. The models demonstrate the spatial proximity of early replication control elements (ERCE) and predicted ERCEs. (C) A local Chrom3D model and annotated Hi-C contact map showing contact frequencies in Chr.11 in the K562 cell line. This local Chrom3D model highlights the spatial proximity between a super-enhancer (SE) and a K562-specific heterogeneous interactome module (HIM). Panels are reproduced from multiple publications with permission [199–201]

Limitations of Chrom3D pipeline

Limitations to automate the pipeline. It would be optimal to fully automate the Chrom3D pipeline presented in **Paper I** using a simple Bash script or workflow management systems, for example, snakemake [203]. Some of the steps in the pipeline, for example *make_NCHG_input.sh* (the step that makes TAD-TAD interaction matrices for all chromosomes to identify significant interactions using NCHG), can be easily parallelized using a workflow management system. An advantage of a workflow management system is that it assesses the validity of the output from one step before passing the output from the step as an input to the next step; it also ensures that jobs are run in the correct order and all the input data are available.

Currently, identification of significant intra-chromosomal and inter-chromosomal TAD-TAD interactions from Hi-C data requires manual intervention. This step requires the P-value and effect size threshold to be decided by the user, either by visualization or through some basic tests. Here, the basic tests are comparing the NCHG output to the replicates or other datasets generated in the laboratory by calculating the Jaccard index. These threshold values could change based on the source, cell type and resolution of the Hi-C experiments. An option to avoid this manual step is using image-based algorithms such as the lasso algorithm [204]. Briefly, in this algorithm, TAD-TAD interaction matrices are treated as 2D images where each pixel represents the interaction intensity between two TADs. Then, lasso smoothens the 2D interaction matrix (an image); this step highlights or picks the TAD-TAD interactions (pixels) with the high frequency of interaction compared to the surrounding interactions (pixels). Thus, the aforementioned manual intervention step can be avoided.

There is potentially another relatively simple way to automate the pipeline. If the thresholds are known before running the pipeline based on prior knowledge, then the pipeline can be fully automated. This method requires a *config* file where the paths to the scripts, the initial input files, the parameters to functions and the thresholds are predefined. Nevertheless, this type of automation has not been tested and implemented, nor explained in **Paper I**. It is worth to note that a research group has automated this pipeline based on **Paper I**, but this automated method has to be tested in our laboratory. The automated scripts can be found in the research group's GitHub page [199] (https://github.com/sespesogil/automat_chrom3D).

```
<marker_set name="chrom3d_model">
<marker id="0" x="1.90136" y="-1.64634" z="-1.80469" radius="0.0773698" r="0.460169" g="0.627819"
b="0.28033" chrID="chr10_A" beadID="chr10_A:0-150000"/>
<marker id="1" x="1.84679" y="-1.82038" z="-1.90187" radius="0.129293" r="0.460169" g="0.627819"
b="0.28033" chrID="chr10_A" beadID="chr10_A:150000-850000"/>
<link id1="0" id2="1" r="0.460169" g="0.627819" b="0.28033" radius="0.00621167"/>
```

Figure 4.4: Example of a Chrom3D output file in CMM format. The first line contains the model name. The second and the third lines represent two beads and contain bead IDs, 3D coordinates, radii, RGB values (colors arbitrarily given to beads to highlight specific features) and chromosome IDs of beads 1 and 2. The fourth line represents a linker that connects beads 1 and 2.

Visualization and data analysis. Chrom3D generates 3D genome model outputs in the chimera marker (CMM) format which is a text file based on an XML file format to store information. The CMM format file can only be opened using the Chimera visualization software [205] and not other macromolecule visualization software such as PyMol [206], for visualization. Uses of Chimera have been explained in **Paper I** and are for visualizing structures and generating publication-ready images. Unfortunately, Chimera lacks features to add any extra information such as gene content, histone modification or gene expression data. Therefore, there is a demand to create a stand-alone or web-based graphical user interface with features to load extra information for each bead (TAD). These features would help users to frame hypotheses, enable visual comparisons of position of genes in 3D models, and communication between researchers and with broader audiences (such as students).

A Chrom3D-generated CMM file mainly contains beadID, 3D coordinates and radius of each beads (Fig. 4.4). Extracting the aforementioned information from the output CMM file requires programming skills. For instance, a user should write scripts to extract the 3D coordinates of the beads and calculate distances between beads and distances between beads and the centre of the modeled nucleus (sphere). These scripts are not included in **Paper I**.

Modeling diploid genomes. **Paper I** explains the step-by-step procedure to process haploid genome Hi-C data to create a model setup file. The pipeline contains a step to create a pseudo-diploid model setup file. This is done by duplicating the information for each chromosome from haploid Hi-C data. The resulting pseudo-diploid model setup file is used as input to Chrom3D to generate diploid 3D genome models. Nevertheless, the generated models reflect realistic difference in 3D positions between

homologs [43, 48, 62]. With advancement in biochemical techniques, there are methods to generate diploid genome Hi-C data and these data are available publicly [16, 93] for exploitation. Chrom3D is already capable of modeling real diploid genome information. Nonetheless, this requires major modification to the pipeline by adding several pre-processing steps; this would include treating two homologous chromosomes as two separate (different) chromosomes and combining interaction information in the final step to generate a single diploid model setup file. Similarly, the pipeline is not suitable to create realistic models of aneuploid cancer genome; presently, I would suggest *ad hoc* ways to pre-process aneuploid Hi-C data for modeling, depending on the nature of the actual data. This would also yield for cancer genomes containing gross genetic abnormalities such as translocations, which can be detected in Hi-C data.

4.5 TAD cliques as a novel feature of higher-order chromatin organization

TAD cliques constitute higher-order chromatin assemblies identifiable in Hi-C matrices

An important finding in this thesis work is the identification of a new higher-order level of chromatin architecture in the form of TAD cliques (**Paper II**, published in 2019 [62]). Hi-C matrices reveal interactions within TADs, between linearly consecutive TADs (along the matrix diagonal), and between linearly non-adjacent TADs (away from the matrix diagonal). Contacts between multiple TADs in Hi-C data may implicate that all TADs interact with each another or that only some of the TADs interact. To distinguish between such interaction patterns, we have opted for the use of graph theory. In graph theory, a clique is a subset of k nodes which are all connected by an edge. We have defined a TAD clique as a subset of k TADs ($k \geq 3$) that are fully connected, i.e. which all interact pairwise in the Hi-C data.

A critical step in mapping TAD cliques is the identification of statistically significant pair-wise TAD-TAD interactions, which we have done using the NCHG model [148] to calculate the probability of observing a given number of Hi-C contacts conditional on the number of interactions for the two TADs examined, the total number of interactions, and the genomic distance between the TADs (**Paper I**). We then calculate a P-value to identify statistically significant contacts, and cliques are

identified by representing all significant TAD-TAD contacts as a graph and identifying maximal cliques (**Paper II**; see also **Paper III**). Using this approach, we report, from ∼15,000 significant pairwise TAD-TAD contacts, more than 3,000 maximal cliques of 3 to 11 TADs which make up ∼50% of the genome. Thus, TAD cliques are main features of the large-scale topology of the human genome.

TAD cliques are enriched in B compartments and show all main features of heterochromatin (**Paper II**). Cliques are enriched in H3K9me3 and to a lesser extent H3K27me3; however, the linear distribution of these marks seems different, with H3K9me3 enriched over a TAD in clique, and H3K27me3 tending to be more elevated at the border of, or outside, a TAD in clique (**Paper II**; Fig. 4.5A). It remains uncertain whether H3K9me3/H3K27me3 co-enriched cliques exist or whether this is a misinterpretation of analysis of ensemble ChIP-seq data. In the case of H3K9me3/H3K27me3 co-enrichment however, TAD cliques could represent a new subtype of B compartment [16]. Our aggregation plots also cannot distinguish between distinct classes of enrichment of these marks.

I have explored this further and characterized the patterns of TAD-TAD interactions in TAD cliques identified from Hi-C data in four cell lines, and reanalyzed in **Paper III**. We notably examined the overlap of TADs in cliques with known A and B compartment sub-types, as a function of TAD clique size (**Paper III**, Fig. 5A). Overlap is strikingly low regardless of compartment sub-type. We note however that overall, A1 subcompartment overlap is reduced as clique size increases and overlap with B2 and B3 subcompartments tends to increase for larger clique sizes. These data suggest that TAD cliques are probably distinct from previously annotated subcompartments, supporting our view of these structures organizing the genome at yet another level.

Another feature of TAD cliques which has not been exploited in work shown here is that about one third are found in A compartments (**Paper II**). Intriguingly, TAD cliques in A compartments include expressed genes interspersed with H3K27me3-marked genes, but contain no LADs. In ASCs, A compartment cliques may represent associations containing genes that can be activated during differentiation; in that sense, 3-way genome architecture mapping (GAM) TAD interactions reported in mouse ES cells [56] may constitute a subset of small A compartment TAD cliques. The relationship between TAD cliques and other (non-compartment) domains is discussed below.

Interestingly, we also find that TADs in large cliques are more con-

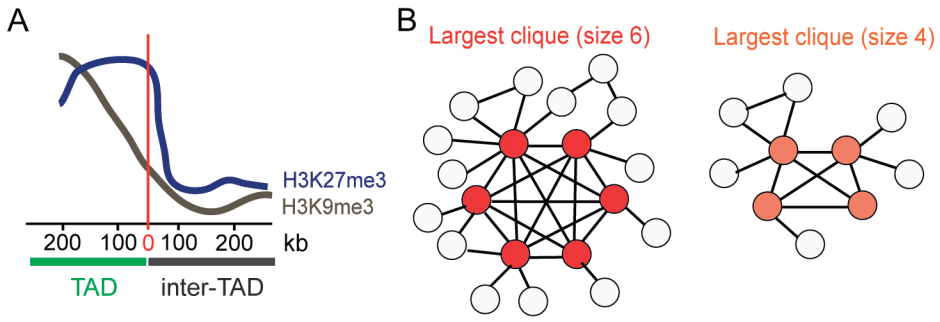

Figure 4.5: Heterochromatin state and outer-clique connectivity of TADs in cliques. (A) Schematized profiles of H3K9me3 and H3K27me3 in TADs (in clique) and inter-TADs (B) TADs in large cliques (left) show higher connectivity with TADs outside cliques than TADs in small cliques (right). Colored nodes represent TADs in the largest clique; white nodes represent TADs outside the largest clique.

nected to TADs outside the (largest) clique than TADs belonging to smaller cliques (**Paper III**). A model is depicted in Fig. 4.5B. Speculatively, that may result from heterochromatin being more compact than euchromatin and associating more readily with other heterochromatin domains, as would be predicted by homotypic interactions [13, 29, 207]. Conversely, a lower intensity of inter-TAD interactions would reflect a loser chromatin conformation which is less 'interactive' (Fig. 4.5B).

Lastly, data from **Paper III** reveal that TADs in large cliques are larger, show fewer within-TAD contacts, and are depleted of convergent CTCF motifs at their borders compared to TADs in small cliques or outside cliques. This contrasts with the smaller and more interaction-dense TADs which have been suggested to be formed by chromatin loop extrusion [208].

How do TAD cliques compare to other reported interactions between chromatin domains?

TAD cliques and SPRITE clusters. The SPRITE method described in the Introduction of this thesis detects multi-way chromosomal interactions, or "SPRITE clusters" of 3 to 14 k-mers [57]. These are interpreted as chromosomal 'hubs' arising from long-range interactions including either gene-dense, active and RNA-polymerase II-marked regions at nuclear speckles, or inactive centromere-proximal regions around the nucleolus [57]. Unlike Hi-C, SPRITE does not depend on proximity ligation and allows for detection of interactions over longer distances (Quinodoz et al.,

2018). The heterochromatic nature of TAD cliques (**Paper II**) and of NADs [50, 52] raises the possibility that a fraction of repressed SPRITE clusters could reside in TAD cliques or include several cliques at the periphery of nucleoli.

TAD cliques and 'TAD hubs'. TAD cliques show analogy to the H3K9me3-enriched so-called 'TAD hubs' reported in B compartments from long-range inter-TAD interaction in endothelial cells; like cliques, these hubs are enriched in LADs [58]. Analysis of these hubs and cliques agree in that the majority of domains fall within a pre-established conformation (such as TAD cliques or absence thereof) that is overall maintained during differentiation [58, 62] (**Paper II**). This is notwithstanding the fact that a proportion of TAD hubs and TAD cliques may grow or shrink during differentiation, by gaining or losing TADs (**Paper II**) (TAD clique dynamics is discussed below).

Nanocompartments, meta-TADs and C-walk interactions. Associations between TADs have also been reported in Drosophila cells from Hi-C and FISH data [72]. There, dynamic interactions between TADs seem to occur, arranging repressed TADs as a succession of 'nanocompartments' intercalated by active domains [72]. FISH data and inferences from 3D models of these configurations suggest that some of these nanocompartments involve linearly non-adjacent TADs – notably supporting a TAD clique idea. The nanocompartments also resemble TAD cliques identified in A compartments with H3K27me3. Our work (**Paper II**) and the Szabo study also agree that changes in inter-TAD interactions reflect discrete chromosomal contacts and not a merger or splitting of TADs.

Additional studies report TAD-TAD interactions which, however, are probably different from TAD cliques. First, the notion of 'meta-TADs' has been proposed as interactions between several neighboring TADs, but not between linearly distant TADs, which define cliques [74]. Meta-TADs are enriched in H3K27me3 and in RNA polymerase II [74] but are devoid of H3K9me3, which again segregates them from TAD cliques. Second, a variation of Hi-C using chromosome walks (C-walks) captures associations between two to four TADs, whose occurrence is enhanced by Polycomb proteins [209]. Yet, C-walks favor a view of pair-wise TAD-TAD contacts over a hub-like pattern, and random associations between active loci rather than a regulated process (this does not mean that all TAD cliques we have identified are regulated, but cliques do not involve active genes). Third, GAM reveals three-way TAD interactions that regroup active genes and enhancers [56]. These associations may constitute supra-TAD gene regulatory 'units' but are likely to be distinct from the TAD cliques we

have reported, even those found in A compartments (**Papers II and III**).

TAD cliques are dynamic topological assemblies

TAD cliques are not all static assemblies, and can expand or shrink during differentiation by gaining or losing TADs (**Paper II**). Changes in clique size do not correspond to changes in the size of B compartments or to A-B compartment switching. This again suggest that TAD cliques constitute another level of higher-order chromatin conformation different from A/B compartmentalization.

Temporal changes in inter-TAD contacts characterize not only mesenchymal and embryonic stem cell differentiation [60, 62], but also dedifferentiation, as shown during reprogramming of B cells [61]. We show a reduction in clique number during cell reprogramming, which probably reflects a loosening of chromatin structure as cells acquire pluripotency. TAD-TAD contacts also appear to be sensitive to environmental conditions. In Drosophila, heat shock response is accompanied by a decrease in intra-TAD contacts and an increase in long-range interactions [210], suggesting a 3D rearrangement of TADs that may be important for gene silencing after temperature stress. These interactions could involve a decrease in TAD border strength [210].

Relationship between TAD cliques and LADs as genome organizers for the radial positioning of chromatin

TADs and TAD cliques however are not the only features shaping genome topologies. A characteristic of TAD cliques in human and mouse cells is their enrichment in lamin interactions (LADs); yet this relationship seems to depend on TAD clique size and differentiation status (**Paper II**). The proportion of linear clique coverage by LADs increases with clique size, and differentiation correlates with an increase in the LAD content of cliques regardless of clique size. So large cliques tend to associate with the nuclear lamina and this association is exacerbated in terminally differentiated cells. Nonetheless, lamin association appears to be dispensable for the formation of cliques because many cliques are detected exist in the absence of LADs, and we found several examples of LADs appearing in pre-established cliques during differentiation. A tethering of TADs in cliques at the nuclear lamina could further compact chromatin in these TADs [211], reinforcing their repressed state.

Our Chrom3D models corroborate these features and predict a peripheral localization of TAD cliques in relation to clique size, with larger cliques more frequently found at the nuclear periphery, and differentiation (**Paper II**). One can speculate that TAD cliques strengthen a repressive state of gene expression by stabilizing peripheral heterochromatin at the nuclear lamina. Interestingly, only a subset of TADs might be sufficient to tether a clique at the nuclear lamina because in a clique containing LADs, not all TADs do harbor LADs. Extending this idea, the nuclear peripheral localization of TADs in a clique does not necessarily directly require LADs if this localization involves LADs in neighboring TADs. Our TAD clique concept further argues that these neighboring TADs need not be linearly contiguous as long as they remain spatially close in a 3D environment.

Are there TAD cliques in single cells?

We have identified TAD cliques using ensemble Hi-C data from $\sim$ 25 million cells, which makes it *a priori* impossible to infer whether cliques exist in single cells or whether they are only detectable in ensemble Hi-C data but do not exist as such. Single-cell Hi-C [10, 12, 59, 212] captures snapshots of chromosome interactions in single cells, and although contacts are sparser than in ensemble Hi-C matrices, significant pair-wise TAD-TAD contacts can be detected (**Paper II**); still, the sparsity of contacts makes identification of TAD cliques virtually impossible. To circumvent this problem, we propose an approach allowing an estimation of inter-TAD contacts in projected TAD cliques pre-identified from ensemble Hi-C data in the same cell type (here, mouse ES cells). We decompose this approach into five steps:

1. Identify significant pair-wise TAD-TAD contacts in single-cell Hi-C contact matrices
2. Identify TAD cliques in ensemble Hi-C data for the same cell type
3. Project these cliques onto individual single-cell Hi-C matrices
4. Compute TAD contact frequencies in the projected cliques and outside the cliques (using an appropriate randomization process, as described in **Paper II**
5. Compute TAD contact densities in the projected cliques

Using this strategy we have reported, from nine single-cell Hi-C datasets, an enrichment of TAD-TAD contacts in projected cliques. Furthermore, most single cells analyzed display clique-like TAD assemblies with at least 50% TAD connectivity in them (i.e. with more than 50% of TADs connected pair-wise in the projected cliques in the single-cell Hi-C data). Although this does not demonstrate the existence of TAD cliques in single cells, the subsets of TADs may display statistically significant long-range associations also in single-cell Hi-C data.

A two-color FISH analysis using probes against TADs in cliques and outside cliques supports our predictions from Chrom3D modeling, on TAD proximity in cliques versus non-cliques. The now published **Paper II** data [62] and additional experiments from our laboratory (T. Germier, A.L. Sørensen and P.Collas, unpublished data) show that TADs in cliques can form closer associations that TADs not in cliques. However, the variations in how physically close to one another TADs in a clique are, demonstrate the heterogeneity in chromatin configurations between cells and challenges the interpretation of ensemble Hi-C data on apparent 'multi-way interactions' (see Fig. 1.4).

4.6 Perspective

The parallel rapid developments of wet-lab techniques and computational methods to analyze genomes in three dimensions, and the combination of such approaches, is expected to lead to further leaps in our understanding of genome dynamics in multiple dimensions. Temporal structural changes in genome conformations, and parallel explorations of multiple cell types or cell differentiation lineages, will increasingly be enabled. Analysis of aneuploidy genomes, such as cancer genomes, and of heterogeneity of genome topologies between cells (e.g. in a tumor) will also expectedly be increasingly possible. Additionally, analyses of single-cell 3D genomes will also increasingly be enabled and refined to lead to unparalleled descriptions and understanding of principles of 3D genome organization, and of the variability thereof. The field of computational genome biology is exploding and is expected to continue expanding - not only for computational experts, but also for non-computational biologists through development of user-friendly pipelines – in the years to come.

Bibliography

1. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14,** 390 (2013).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome (2001).
3. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome (2012).
4. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549,** 219–226 (2017).
5. Marti-Renom, M. A. *et al.* Challenges and guidelines toward 4D nucleome data and model standards. *Nature genetics* **50,** 1352–1358 (2018).
6. Misteli, T. Chromosome territories: The arrangement of chromosomes in the nucleus. *Nature Education* **1,** 167 (2008).
7. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics* **2,** 292–301 (2001).
8. Cremer, M. *et al.* Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome research* **9,** 541–567 (2001).
9. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).
10. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502,** 59 (2013).
11. Meaburn, K. J. & Misteli, T. Chromosome territories. *Nature* **445,** 379–381 (2007).
12. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544,** 59–64 (2017).
13. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics* **10,** 211–220 (2001).
14. Kolbl, A. C. *et al.* The radial nuclear positioning of genes correlates with features of megabase-sized chromatin domains. *Chromosome research* **20,** 735–752 (2012).
15. Bickmore, W. A. The spatial organization of the human genome. *Annual review of genomics and human genetics* **14,** 67–84 (2013).

16. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).

17. Szalaj, P. & Plewczynski, D. Three-dimensional organization and dynamics of the genome. *Cell biology and toxicology* **34,** 381–404 (2018).

18. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

19. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385 (2012).

20. Sexton, T. *et al.* Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148,** 458–472 (2012).

21. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular cell* **48,** 471–484 (2012).

22. Rudan, M. V. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell reports* **10,** 1297–1309 (2015).

23. Fudenberg, G. *et al.* Formation of chromosomal domains by loop extrusion. *Cell reports* **15,** 2038–2049 (2016).

24. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A. & Mirny, L. A. *Emerging evidence of chromosome folding by loop extrusion* in *Cold Spring Harbor symposia on quantitative biology* **82** (2017), 45–55.

25. Guo, Y. *et al.* CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162,** 900–910 (2015).

26. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* **112,** E6456–E6465 (2015).

27. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551,** 51–56 (2017).

28. Rao, S. S. *et al.* Cohesin loss eliminates all loop domains. *Cell* **171,** 305–320 (2017).

29. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* **19,** 789–800 (2018).

30. De Wit, E. TADs as the caller calls them. *Journal of molecular biology* (2019).

31. Burke, B. & Stewart, C. L. The nuclear lamins: flexibility in function. *Nature reviews Molecular cell biology* **14,** 13–24 (2013).

32. Van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169,** 780–791 (2017).

33. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

34. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics* **43,** 630 (2011).

35. Pickersgill, H. *et al.* Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature genetics* **38,** 1005–1014 (2006).

36. Chen, S. *et al.* A lamina-associated domain border governs nuclear lamina interactions, transcription, and recombination of the Tcrb locus. *Cell reports* **25,** 1729–1740 (2018).

37. Harr, J. C. *et al.* Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. *Journal of Cell Biology* **208,** 33–52 (2015).

38. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell* **38,** 603–613 (2010).

39. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology* **217,** 4025–4048 (2018).

40. Reddy, K. L., Zullo, J., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452,** 243–247 (2008).

41. Meuleman, W. *et al.* Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence. *Genome research* **23,** 270–280 (2013).

42. Brunet, A., Forsberg, F., Fan, Q., Sæther, T. & Collas, P. Nuclear Lamin B1 Interactions With Chromatin During the Circadian Cycle Are Uncoupled From Periodic Gene Expression. *Frontiers in Genetics* **10,** 917 (2019).

43. Forsberg, F., Brunet, A., Ali, T. M. L. & Collas, P. Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. *Nucleus* **10,** 7–20 (2019).

44. Rønningen, T. *et al.* Prepatterning of differentiation-driven nuclear lamin A/C-associated chromatin domains by GlcNAcylated histone H2B. *Genome research* **25,** 1825–1835 (2015).

45. Robson, M. I. *et al.* Tissue-specific gene repositioning by muscle nuclear membrane proteins enhances repression of critical developmental genes during myogenesis. *Molecular cell* **62,** 834–847 (2016).

46. Robson, M. I. *et al.* Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome research* **27,** 1126–1138 (2017).

47. Li, Q. *et al.* The three-dimensional genome organization of Drosophila melanogaster through data integration. *Genome biology* **18,** 145 (2017).

48. Paulsen, J. *et al.* Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome biology* **18,** 21 (2017).

49. Bersaglieri, C. & Santoro, R. Genome Organization in and around the Nucleolus. *Cells* **8,** 579 (2019).

50. Németh, A. *et al.* Initial genomics of the human nucleolus. *PLoS genetics* **6** (2010).

51. Dillinger, S., Straub, T. & Nemeth, A. Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One* **12** (2017).

52. Vertii, A. *et al.* Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. *Genome research* **29,** 1235–1249 (2019).

53. Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153,** 178–192 (2013).

54. Bizhanova, A., Yan, A., Yu, J., Zhu, L. J. & Kaufman, P. D. Distinct features of nucleolus-associated domains in mouse embryonic stem cells. *bioRxiv,* 740480 (2019).

55. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harbor perspectives in biology* **3,** a000646 (2011).

56. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543,** 519 (2017).

57. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174,** 744–757 (2018).

58. Niskanen, H. *et al.* Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions. *Nucleic acids research* **46,** 1724–1740 (2018).

59. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547,** 61–67 (2017).

60. Bonev, B. *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* **171,** 557–572 (2017).

61. Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature genetics* **50,** 238–249 (2018).

62. Paulsen, J. *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nature Genetics* **51,** 835–843 (2019).

63. Bystricky, K. Chromosome dynamics and folding in eukaryotes: Insights from live cell microscopy. *FEBS letters* **589,** 3014–3022 (2015).

64. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics,* 1–20 (2019).

65. Wang, W., Zhang, L., Wang, X. & Zeng, Y. The advances in CRISPR technology and 3D genome. *Seminars in Cell & Developmental Biology* **90,** 54–61 (2019).

66. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362,** eaau1783 (2018).

67. Boettiger, A. N. *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529,** 418–422 (2016).

68. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology* **2,** a003889 (2010).

69. Finn, E. H. *et al.* Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176,** 1502–1515 (2019).

70. Germier, T. *et al.* Real-time imaging of a single gene reveals transcription-initiated local confinement. *Biophysical journal* **113,** 1383–1394 (2017).

71. Maass, P. G. *et al.* Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING). *Nature structural & molecular biology* **25,** 176–184 (2018).

72. Szabo, Q. *et al.* TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science advances* **4,** eaar8082 (2018).

73. Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353,** 598–602 (2016).

74. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.* **79,** 347–372 (2015).

75. Landstrom, A. P. & Tefferi, A. Fluorescent in situ hybridization in the diagnosis, prognosis, and treatment monitoring of chronic myeloid leukemia. *Leukemia & lymphoma* **47,** 397–402 (2006).

76. Sarrate, Z., Vidal, F. & Blanco, J. Role of sperm fluorescent in situ hybridization studies in infertile patients: indications, study approach, and clinical relevance. *Fertility and sterility* **93,** 1892–1902 (2010).

77. Oldenburg, A. *et al.* A lipodystrophy-causing lamin A mutant alters conformation and epigenetic regulation of the anti-adipogenic MIR335 locus. *Journal of Cell Biology* **216,** 2731–2743 (2017).

78. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).

79. Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & de Laat, W. Looping and Interaction between Hypersensitive Sites in the Active beta-globin Locus. *Molecular Cell* **10,** 1453–1465 (2002).

80. Palstra, R.-J. *et al.* The beta-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics* **35,** 190–194 (2003).

81. Tan-Wong, S. M., French, J. D., Proudfoot, N. J. & Brown, M. A. Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proceedings of the National Academy of Sciences* **105,** 5160–5165 (2008).

82. Comet, I., Schuettengruber, B., Sexton, T. & Cavalli, G. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proceedings of the National Academy of Sciences* **108,** 2294–2299 (2011).

83. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nature Methods* **3,** 17–21 (2006).

84. Hagège, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nature Protocols* **2,** 1722 (2007).

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

86. Van de Werken, H. J. *et al.* in *Nucleosomes, Histones and Chromatin Part B* (eds Wu, C. & Allis, C. D.) 89–112 (Academic Press, 2012).

87. Würtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Research* **14,** 477–495 (2006).

88. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics* **38,** 1341–1347 (2006).

89. Lomvardas, S. *et al.* Interchromosomal Interactions and Olfactory Receptor Choice. *Cell* **126,** 403–413 (2006).

90. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16,** 1299–1309 (2006).

91. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophysical reviews* **11,** 67–78 (2019).

92. Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome biology* **16,** 175 (2015).

93. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361,** 924–928 (2018).

94. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome research* **24,** 1854–1868 (2014).

95. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications* **6,** 6178 (2015).

96. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47,** 598 (2015).

97. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* **13,** 74 (2015).

98. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13,** 919 (2016).

99. Mumbach, M. R. *et al.* HiChIRP reveals RNA-associated chromosome conformation. *Nature Methods* **16,** 489–492 (2019).

100. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58 (2009).

101. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–1627 (2015).

102. Mishra, A. & Hawkins, R. D. Three-dimensional genome architecture and emerging technologies: looping in disease. *Genome Medicine* **9,** 87 (2017).

103. Kim, T. H. & Ren, B. Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.* **7,** 81–102 (2006).

104. Van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology* **18,** 424–428 (2000).

105. Collas, P. in *Chromatin Immunoprecipitation Assays* 1–25 (Springer, 2009).

106. Gesson, K. *et al.* A-type lamins bind both hetero-and euchromatin, the latter being regulated by lamina-associated polypeptide 2 alpha. *Genome research* **26,** 462–473 (2016).

107. Lund, E. *et al.* Lamin A/C-promoter interactions specify chromatin state–dependent transcription outcomes. *Genome research* **23,** 1580–1589 (2013).

108. Lund, E., Oldenburg, A. R. & Collas, P. Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations. *Nucleic acids research* **42,** e92–e92 (2014).

109. Sadaie, M. *et al.* Redistribution of the Lamin B1 genomic binding profile affects rearrangement of heterochromatic domains and SAHF formation during senescence. *Genes & development* **27,** 1800–1808 (2013).

110. Shah, P. P. *et al.* Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes & development* **27,** 1787–1799 (2013).

111. Aughey, G. N. & Southall, T. D. Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdisciplinary Reviews: Developmental Biology* **5,** 25–37 (2016).

112. Aughey, G. N., Cheetham, S. W. & Southall, T. D. DamID as a versatile tool for understanding gene regulation. *Development* **146,** dev173666 (2019).

113. Luperchio, T. R. *et al.* Chromosome conformation paints reveal the role of lamina association in genome organization and regulation. *BioRxiv,* 122226 (2017).

114. Luperchio, T. R. *et al.* The repressive genome compartment is established early in the cell cycle before forming the lamina associated domains. *bioRxiv,* 481598 (2018).

115. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9,** R137 (2008).

116. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

117. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

118. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* **9,** 999–1003 (2012).

119. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Research* **4** (2015).

120. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16,** 259 (2015).

121. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Computational Biology* **13,** 1–17 (2017).

122. Spill, Y. G., Castillo, D., Vidal, E. & Marti-Renom, M. A. Binless normalization of Hi-C data provides significant interaction and difference detection independent of resolution. *Nature Communications* **10,** 1938 (2019).

123. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43,** 1059–1065 (2011).

124. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28,** 3131–3133 (2012).

125. Lu, L. *et al.* Robust Hi-C chromatin loop maps in human neurogenesis and brain tissues at high-resolution. *bioRxiv* (2019).

126. Wu, H.-J. & Michor, F. A computational strategy to adjust for copy number in tumor Hi-C data. *Bioinformatics* **32,** 3695–3701 (2016).

127. Vidal, E. *et al.* OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Research* **46,** e49–e49 (2018).

128. Servant, N., Varoquaux, N., Heard, E., Barillot, E. & Vert, J.-P. Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics* **19,** 313 (2018).

129. Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* **28,** 2843–2844 (2012).

130. Zheng, X. & Zheng, Y. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics* **34,** 1568–1570 (2017).

131. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9,** 14 (2014).

132. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30,** 386–392 (2014).

133. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research* **44,** 70–70 (2015).

134. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research* **45,** 2994–3005 (2017).

135. Malik, L. & Patro, R. Rich chromatin structure prediction from Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16,** 1448–1458 (2019).

136. Weinreb, C. & Raphael, B. J. Identification of hierarchical chromatin domains. *Bioinformatics* **32,** 1601–1609 (2016).

137. Haddad, N., Vaillant, C. & Jost, D. IC-Finder: Inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Research* **45** (2017).

138. Ron, G., Globerson, Y., Moran, D. & Kaplan, T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nature Communications* **8,** 2237 (2017).

139. Wang, Y., Li, Y., Gao, J. & Zhang, M. Q. A novel method to identify topological domains using Hi-C data. *Quantitative Biology* **3,** 81–89 (2015).

140. Oluwadare, O. & Cheng, J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. *BMC bioinformatics* **18,** 480 (2017).

141. Yan, K.-K., Lou, S. & Gerstein, M. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLOS Computational Biology* **13,** 1–22 (2017).

142. Chen, J., Hero, A. O. & Rajapakse, I. Spectral identification of topological domains. *Bioinformatics* **32,** 2151–2158 (2016).

143. Norton, H. K. *et al.* Detecting hierarchical genome folding with network modularity. *Nature Methods* **15,** 119–122 (2018).

144. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3,** 99–101 (2016).

145. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nature methods* **14,** 679 (2017).

146. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research* **24,** 999–1011 (2014).

147. Mifsud, B. *et al.* GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS ONE* **12,** 1–15 (2017).

148. Paulsen, J., Rødland, E. A., Holden, L., Holden, M. & Hovig, E. A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic acids research* **42,** e143–e143 (2014).

149. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342,** 948–953 (2013).

150. Robinson, J. T. *et al.* Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell systems* **6,** 256–258 (2018).

151. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology* **19,** 125 (2018).

152. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome biology* **16,** 198 (2015).

153. Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications* **9,** 1–15 (2018).

154. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi. R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30,** 2808–2810 (2014).

155. Marti-Renom, M. A. & Mirny, L. A. Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. *PLOS Computational Biology* **7,** 1–6 (2011).

156. Rosa, A. & Zimmer, C. *Computational models of large-scale genome architecture* 1st ed., 275–349 (Elsevier Inc., 2014).

157. Woodcock, C. L. & Ghosh, R. P. Chromatin higher-order structure and dynamics. *Cold Spring Harbor perspectives in biology* **2,** a000596 (2010).

158. Ostashevsky, J. A polymer model for the structural organization of chromatin loops and minibands in interphase chromosomes. *Molecular biology of the cell* **9,** 3031–3040 (1998).

159. Jackson, D. A. & Pombo, A. Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *The Journal of cell biology* **140,** 1285–1295 (1998).

160. Fiorillo, L. *et al.* A modern challenge of polymer physics: Novel ways to study, interpret, and reconstruct chromatin structure. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* e1454 (2019).

161. Sachs, R. K., van den Engh, G., Trask, B., Yokota, H. & Hearst, J. E. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences* **92,** 2710–2714 (1995).

162. Nicodemi, M. & Prisco, A. Thermodynamic pathways to genome spatial organization in the cell nucleus. *Biophysical journal* **96,** 2168–2177 (2009).

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

164. Barbieri, M. *et al.* Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences* **109,** 16173–16178 (2012).

165. Jost, D., Carrivain, P., Cavalli, G. & Vaillant, C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic acids research* **42,** 9553–9561 (2014).

166. Cheng, T. M. *et al.* A simple biophysical model emulates budding yeast chromosome condensation. *Elife* **4,** e05565 (2015).

167. Brackley, C. A. *et al.* Nonequilibrium chromosome looping via molecular slip links. *Physical review letters* **119,** 138101 (2017).

168. Vian, L. *et al.* The energetics and physiological impact of cohesin extrusion. *Cell* **173,** 1165–1178 (2018).

169. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences* **115,** E6697–E6706 (2018).

170. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465,** 363–367 (2010).

171. Hu, M. *et al.* Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology* **9** (2013).

172. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nature methods* **11,** 1141 (2014).

173. Peng, C. *et al.* The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic acids research* **41,** e183–e183 (2013).

174. Szałaj, P. *et al.* An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome research* **26,** 1697–1709 (2016).

175. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J.-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30,** i26–i33 (2014).

176. Zhang, Z., Li, G., Toh, K.-C. & Sung, W.-K. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology* **20,** 831–846 (2013).

177. Zou, C., Zhang, Y. & Ouyang, Z. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome biology* **17,** 40 (2016).

178. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* **30,** 90 (2012).

179. Tjong, H., Gong, K., Chen, L. & Alber, F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome research* **22,** 1295–1305 (2012).

180. Baù, D. & Marti-Renom, M. A. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods* **58,** 300–306 (2012).

181. Giorgetti, L. *et al.* Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157,** 950–963 (2014).

182. Tjong, H. *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences* **113,** E1663–E1672 (2016).

183. Bau, D. & Marti-Renom, M. A. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome research* **19,** 25–35 (2011).

184. Serra, F. *et al.* Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS letters* **589,** 2987–2995 (2015).

185. Le Dily, F. *et al.* Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & development* **28,** 2151–2162 (2014).

186. Meluzzi, D. & Arya, G. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic acids research* **41,** 63–75 (2013).

187. Pouokam, M. *et al.* The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific reports* **9,** 1–10 (2019).

188. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS biology* **10** (2012).

189. Collas, P., Ali, T. M. L., Brunet, A. & Germier, T. Finding Friends in the Crowd: Three-Dimensional Cliques of Topological Genomic Domains. *Frontiers in genetics* **10** (2019).

190. Gundersen, S. *et al.* Identifying elemental genomic track types and representing them uniformly. *BMC bioinformatics* **12,** 494 (2011).

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

192. Briand, N. *et al.* The lipodystrophic hotspot lamin A p. R482W mutation deregulates the mesodermal inducer T/Brachyury and early vascular differentiation gene networks. *Human molecular genetics* **27,** 1447–1459 (2018).

193. Garcia-Nieto, P. E. *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *The EMBO journal* **36,** 2829–2843 (2017).

194. Paulsen, J., Liyakat Ali, T. M. & Collas, P. Computational 3D genome modeling using Chrom3D. *Nature Protocols* **13,** 1137 (2018).

195. Sekelja, M., Paulsen, J. & Collas, P. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? *Genome biology* **17,** 54 (2016).

196. Dai, C. *et al.* Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities. *Nature communications* **7,** 1–11 (2016).

197. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43 (2011).

198. Adar, S., Hu, J., Lieb, J. D. & Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences* **113,** E2124–E2133 (2016).

199. Espeso-Gil, S. *et al.* A chromosomal connectome for psychiatric and metabolic risk variants in adult dopaminergic neurons. *Genome medicine* **12,** 1–19 (2020).

200. Sima, J. *et al.* Identifying cis elements for spatiotemporal control of mammalian DNA replication. *Cell* **176,** 816–830 (2019).

201. Tian, D., Zhang, R., Zhang, Y., Zhu, X. & Ma, J. MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome. *Genome Research,* gr–250316 (2020).

202. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *The EMBO journal* **12,** 1059–1065 (1993).

203. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

204. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1996).

205. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25,** 1605–1612 (2004).

206. DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* **40,** 82–92 (2002).

207. Gonzalez-Sandoval, A. & Gasser, S. M. On TADs and LADs: spatial control over gene expression. *Trends in Genetics* **32,** 485–495 (2016).

208. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nature Genetics,* 1–9 (2020).

209. Olivares-Chauvet, P. *et al.* Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature* **540,** 296–300 (2016).

210. Li, L. *et al.* Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Molecular cell* **58,** 216–231 (2015).

211. Ulianov, S. V. *et al.* Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nature communications* **10,** 1–11 (2019).

212. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544,** 110–114 (2017).