From the Department of Medical Biochemistry and Biophysics

Karolinska Institutet, Stockholm, Sweden

METHODS DEVELOPMENT FOR THE INVESTIGATION OF THE MAMMALIAN GENOME RADIAL ARCHITECTURE: THE QUANTITATIVE SIDE

Gabriele Girelli

Stockholm 2021

The cover picture is a series of "*pizza-plots*" representing the localization of cell cycle phase-specific K562 Repli-seq data, radially arranged based on HAP1 GPSeq score at 1 Mb resolution. Each slice represents a chromosome. The color scale follows the viridis colormap and goes from purple for low Repli-seq values to yellow for high Repli-seq values. The plots were generated with the `ggplot2` package in R.

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by US-AB.
© Gabriele Girelli, 2021
ISBN 978-91-8016-197-8

Methods Development for the Investigation of the Mammalian Genome Radial Architecture: the quantitative side

THESIS FOR DOCTORAL DEGREE (Ph.D.)

To be defended at Karolinska Institutet,
Andreas Vesalius hall, Berzelius väg 3, Solna

On May 12th, 2021, at 9.00 AM

By

Gabriele Girelli

Principal supervisor:
Assistant Professor
Magda Bienko, Ph.D.
Karolinska Institutet
Dept. of Medical Biochem. and Biophys.

Co-supervisor:
Associate Professor
Marianne Farnebo, Ph.D.
Karolinska Institutet
Dept. of Cell and Molecular Biology

Opponent:
Group Leader
Juan Manuel Vaquerizas, Ph.D.
Max Planck Institute for Molecular Biomedicine

Examination board:
Associate Professor
Anita Göndör, Ph.D.
Karolinska Institutet
Dept. of Oncology-Pathology

Professor
Sten Linnarson, Ph.D.
Karolinska Institutet
Dept. of Medical Biochem. and Biophys.

Associate Professor
Ilaria Testa, Ph.D.
KTH Royal Institute of Technology
Dept. of Applied Physics

To Matteo, Anna, Nicolò,
and all the friends that
made this possible.

Abstract

The nucleus of mammalian cells cradles the genome, an ensemble of nucleic acid macromolecular polymers that store information in a physical form. For a cell to perform life-sustaining processes, reading and utilizing the information encoded in the genome monomers' sequence is necessary. Considerable attention has been paid to these processes since their discovery, leading to remarkable breakthroughs in our understanding of basic cell biology and the Genetics field's birth. In the past two decades, the focus has shifted from this one-dimensional approach to a more spatio-temporal perspective. It is now clear that the genome has a complex architecture, with a multitude of organizational levels at different scales. Additionally, genome architecture interplays with gene expression, and alterations to its spatial organization associate with various pathologies like cancer, premature-aging diseases, and male infertility. In this thesis, we present the development of two methods enabling the investigation of genome architecture.

In **Paper I**, we established **iFISH**, a full-stack workflow for easy DNA fluorescence *in situ* hybridization (FISH) setup and application. Specifically, iFISH includes a novel and accurately crafted database of 40 nt long oligonucleotide sequences for labeling specific human genomic loci. iFISH 40-mers provide a strikingly higher genomic coverage and shorter inter-oligo distance than other state-of-the-art databases. Moreover, the iFISH database of homologous sequences allows for the design of a 96-oligo probe in more than half of the ten kb-wide genomic regions and more than 85% of 15 kb-wide genomic regions (against a 15-30% for other databases). iFISH also includes a computational tool, easily accessible and usable via a web-based graphical user interface, for the automatic selection of optimal sets of oligos (i.e., probe design), for single-probe or homogeneous multi-probe (i.e., spotting) labeling. We applied our computational pipeline to design a total of 330 DNA FISH probes, covering all human chromosomes homogeneously, with an inter-probe distance of 10 Mb for chromosomes 1 to 16 and X and of 5 Mb for chromosomes 17 to 22.

Additionally, we systematically and individually tested most probes, whose sequences are readily available for the community to download and utilize. Furthermore, we built upon cutting-edge sequence amplification methods to provide an inexpensive and straightforward protocol for the large-scale amplification of DNA FISH probes starting from relatively low concentrated oligopools. To this end, we designed a set of novel 20-mer sequences orthogonal to the human genome and compatible with the probe-specific PCR steps of the amplification protocol. Finally, we showcased the extensive applicability and flexibility of the iFISH workflow in human IMR90 fibroblast cells, revealing the importance of a dense label sampling for correct chromatin volume estimation, and in human embryonic stem cells, uncovering overall less distinct chromosome territories, and a remarkable lack of chromosome territoriality in a subset of cells. Altogether, these results support iFISH as an empowering set of tools and resources for the research community, freely accessible online at https://www.ifish4u.org.

In **Paper II**, we presented Genomic loci Positioning by sequencing (a.k.a., **GPSeq**), a method for the genome-wide measurement of genomic loci position along the nuclear radius. GPSeq follows a straightforward protocol based on a simple and elegant concept: nuclear diffusion proceeds from the nucleus periphery towards its interior. We proved this concept by applying it to restriction enzyme diffusion, where we exploited a FISH-based method (YFISH) to visualize concentric genomic restriction signal waves generated by different digestion times. Specifically, GPSeq combines the sequencing of genomic loci restricted at different digestion time lengths into a so-called "GPSeq score," a reliable and accurate estimate of genomic loci centrality. We validated the GPSeq score against a collection of 68 DNA FISH probes, spanning 11 different chromosomes, data obtained from DamID-seq of Lamin B1, and also Hi-C chromatin contacts. Then, we utilized the radial maps drawn by GPSeq to reveal novel radial arrangements of different chromatin states and identify centrality predictors at different resolutions. Subsequently, we applied a novel 3D genome reconstruction algorithm to demonstrate how an additional centrality constraint can improve reconstructed structures' quality. Specifically, 3D genome structures generated by a GPSeq-informed algorithm showed a higher correlation with FISH-based radial measurements and an arrangement of chromosome territories and genomic compartments that better reflects the underlying biology.

Additionally, structures generated by the combination of GPSeq and Hi-C intra-chromosomal contacts allowed the recovery of the inter-chromosomal contacts, further underscoring the necessity of additional constraints provided by orthogonal methods to Hi-C for a more reliable 3D genome reconstruction. Finally, we applied GPSeq to provide insight into the so-called "bodyguard hypothesis, " speculating that heterochromatin might act as a shield from exogenous mutagens for the more internally located active chromatin. In this regard, we showed that cancer-related single-nucleotide variants (SNVs) have a strikingly different radial arrangement than germline single-nucleotide polymorphisms (SNPs), with the former more peripherally located than the latter. We then showed that genomic regions involved with gene fusions in cancer tend to locate more internally and contact other chromosomes more frequently than other regions. We combined these observations and the fact that double-strand breaks (DSBs) tend to locate more internally, further confirmed from immunofluorescence experiments, to speculate that cancer-related SNVs and germline SNPs might come to be by different underlying mechanisms. Altogether, these results highlight the importance of genome-wide high-resolution radial maps in the study of genome architecture, both as a standalone resource and as a complementary feature to chromatin contacts.

Disclaimer

The biomolecular (i.e., *wet* lab) method protocols presented here are the main work of other students and researchers (mainly Dr. Joaquin Custodio and Dr. Tomasz Kallas for GPSeq, and Eleni Gelali for iFISH). Instead, this thesis is focused on the development of the analytical and deeply quantitative side (i.e., *dry* lab) of method development. At the same time, we would like to stress that it is an impossible feat to fully separate these two sides, as novel method arise only through the interplay between experimentalists and analysts (when they are not the same person).

Popular science summary

Most cells of our body have at least a copy of an instruction manual containing all the details they need to stay alive and carry out their job. This instruction manual takes the form of deoxyribonucleic acid or DNA. The cell's instruction manual needs to be stored somewhere, just like an encyclopedia is usually held on a dedicated bookshelf. But the cell's manual (also known as its "genome") does not comprise just a few volumes. Instead, this text is more than three billion characters long and written in an alphabet of four letters (A, T, C, and G). As a comparison, the Lord of the Rings' complete trilogy includes only 2-2.5 billion characters (depending on whether one counts spaces and line breaks).

In humans' case, most cells contain not one but two copies of their instruction manual—each including 23 series, the chromosomes, made of many books and chapters. The cell stores these books in its nucleus, which acts as a library dedicated to the manual. A cell can read any part of the manual at any time, as it might simultaneously need instructions from different chapters. Moreover, like a scribe, a cell can produce new copies of its genome, as daughter cells will need their copy of the instruction manual.

Can you imagine a library where all the books' pages are always open and readable? And where flocks of readers and scribes keep roaming around to read and copy these pages?

It might help to imagine the genome not as books made of pages but as papyrus scrolls or microfilms. Libraries often use microfilms to preserve their books and store them in more compacted spaces. Indeed, the cell stores the genome in a similar form: not as paperback pages, but as a single string of letters, the DNA. Now it might be easier to imagine accessing any part of the manual at any time. One would only need to keep all microfilms unfolded all the time. The cell does something in between, where the DNA is partially unfolded (or "naked") and partially folded.

You can find most readers and scribes roaming the more internal part of this library. One can imagine that area as having most of the tables and chairs, which writers and readers need to copy and read the manual. Moreover, how can we keep the microfilm under tension to be able to read it? Of course, we would need to hold it between two places! One can imagine parts of the microfilm glued to the library walls while the readers and writers handle other portions. Indeed, there is a big difference between the center and outer parts of this imaginary library.

This thesis presents two methods to study how the cell folds and places the genome's "microfilm" in its nucleus. The first is **iFISH**, a technique able to identify the position of a specific word, section, chapter, or book in our imaginary and supposedly messy library. It achieves such a feat by flooding the whole library with colors that stain only the words, sections, chapters, or books of interest. In this way, we can immediately identify them by eye.

The second is **GPSeq** (pronounced "*gee-pea-seek*"), which tells us whether the instruction manual parts we are interested in are closer to the central area or the library walls. This method

is a bit more complicated than iFISH. The idea behind GPSeq is to overflow the library with people slowly walking from the library's walls towards its internal area. As these people walk, they keep reading aloud all the instruction manual parts that they stumble upon. A computer hears everything they read out loud and then tells us the location of a specific part of the genome based on when the roaming people read it for the first time. Of course, chapters stored close to the library walls will be read first, as it takes a shorter time to reach them. Similarly, these GPSeq people will read chapters stored in the library's interior for last, as it takes time to get to them.

Matching terminology

- library = nucleus
- instruction manual = genome
- GPSeq people = restriction enzyme
- readers = RNA polymerases
- scribes/writers = DNA polymerases
- library's walls = nuclear envelope
- colors = FISH probes
- reading aloud = sequencing

Riassunto della tesi a fini divulgativi

La maggior parte delle cellule del nostro organismo possiede almeno una copia di un manuale di istruzioni che contiene tutti i dettagli necessari per funzionare correttamente e sopravvivere. Questo manuale, che prende la forma di un acido desossiribonucleico anche noto come DNA, ha bisogno di un luogo dove essere conservato, come un'enciclopedia viene esposta su un ripiano dedicato. Diversamente da un'enciclopedia però il manuale della cellula, chiamato "genoma", non è composto da pochi volumi. Questo testo è, invece, lungo più di tre miliardi di caratteri ed è scritto in un alfabeto di quattro lettere (A, T, C, e G). Come paragone si pensi che la trilogia completa de "Il Signore degli Anelli" include soltanto tra i due e i due miliardi e mezzo di caratteri (a seconda che si contino o meno spazi e accapo).

Nel caso degli esseri umani, quasi ogni cellula contiene non una ma due copie di questo manuale d'istruzioni. Ogni copia del manuale include ventitré serie, i cromosomi, composte a loro volta di molti libri e capitoli. La cellula ha una biblioteca dedicata dove conservare questi libri: il nucleo. La cellula può leggere qualsiasi parte del manuale in ogni momento, in quanto potrebbe avere bisogno allo stesso tempo di istruzioni scritte in capitoli diversi. Inoltre, come un copista, la cellula può produrre nuove copie del suo genoma, dato che le cellule figlie avranno bisogno della propria copia del manuale di istruzioni.

È possibile immaginare una biblioteca dove tutte le pagine di tutti i libri sono sempre aperte, pronte a essere lette, in cui stormi di lettori e di copisti girovagano di continuo per leggere e copiare queste pagine?

Potrebbe essere d'aiuto immaginare il genoma non come dei libri ma come rotoli di papiro o microfilm su pellicola. Le biblioteche usano spesso il formato in microfilm per preservare i loro testi più preziosi e conservarli in spazi più ristretti. Infatti, la cellula conserva il proprio genoma in una forma molto simile: come una pellicola di lettere (il DNA) e non come libri cartacei. A questo punto, è più facile immaginare di avere accesso a una qualsiasi parte del manuale in ogni momento, sarebbe solo necessario che il microfilm fosse sempre svolto. La cellula si comporta in modo simile, in quanto il suo DNA è in parte svolto (o "nudo") e avvolto.

In questa ipotetica "biblioteca del genoma" si può trovare la maggior parte dei lettori e degli scrittori nell'area più interna: una zona attrezzata con sedie, tavoli e postazioni necessarie affinché lettori e scrittori possano leggere e copiare il manuale. Inoltre, come possiamo tenere il microfilm in tensione per poterlo leggere? Lo si potrebbe impugnare alle due estremità, oppure potremmo immaginare parti di microfilm fissate alle pareti della biblioteca, mentre i lettori e gli scrittori ne impugnano altre. In questa nostra biblioteca immaginaria, ciò che succede sulle pareti è decisamente diverso da quello che avviene al suo interno.

Questa tesi presenta due metodi per studiare come la cellula avvolge e posiziona il "microfilm" del genoma nel suo nucleo. Il primo metodo è iFISH (pronunciato "*ai-fi-sh*"), una tecnica in grado d'identificare la posizione di specifiche parole, sezioni, capitoli o libri della nostra biblioteca immaginaria e potenzialmete caotica. Riesce a ottenere tutto ciò inondando

la biblioteca con colori che macchiano soltanto le parole, sezioni, capitoli, o libri a cui siamo interessati. In questo modo, possiamo identificarli a prima occhiata.

Il secondo metodo è GPSeq (pronunciato "*gi-pi-sik*"), che può dirci se parti del manuale si trovano più vicine all'area centrale o alle pareti della biblioteca. Questo metodo è decisamente più complesso di iFISH. L'idea alla base di GPSeq è di riempire la biblioteca di persone che camminino lentamente dall'esterno verso l'area più interna. Queste persone camminano leggendo ad alta voce tutte le parti del manuale che incontrano sul loro percorso. Nel frattempo, un computer ascolta tutto ciò che viene letto ad alta voce ed è in grado di calcolare la posizione di una parte specifica del genoma in base a quando è stata letta per la prima volta. Ovviamente, capitoli conservati vicino alle pareti della biblioteca saranno letti per primi, in quanto si impiega meno tempo per raggiungerli. Allo stesso modo, queste persone leggeranno per ultimi i capitoli conservati nelle aree più interne, in quanto impiegheranno maggior tempo per avvicinarvisi.

Corrispondenze tra vocaboli

- biblioteca = nucleo
- manuale di istruzioni = genoma
- persone di GPSeq = enzimi di restrizione
- lettori = RNA polimerasi
- scrittori/copisti = DNA polimerasi
- pareti della biblioteca = membrana nucleare
- colore = sonde per ibridazione fluorescente *in situ*
- leggere ad alta voce = sequenziare

List of Scientific Publications

1. [1]Gelali, Eleni*, Gabriele Girelli*, Masahiro Matsumoto, Erik Wernersson, Joaquin Custodio, Ana Mota, Maud Schweitzer et al. "**iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture.**" *Nature communications* 10, no. 1 (2019): 1-15.

2. [2]Girelli, Gabriele*, Joaquin Custodio*, Tomasz Kallas*, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota et al. "**GPSeq reveals the radial organization of chromatin in the cell nucleus.**" *Nature Biotechnology* 38, no. 10 (2020): 1184-1193.

*: these authors contributed equally.

Contents

List of Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BAC	Bacterial Artificial Chromosome
ChromEMT	Electron Microscopy Tomography applied on samples labeled with ChromEM
CLL	Chronic Lymphocytic Leukemia
CoM	Center of Mass
CTCF	CCCTC-binding Factor
DamID	DNA adenine methyltransferase IDentification
DNA	DeoxyriboNucleic Acid
DSB	Double-Strand Break
EMT	Electron Microscopy Tomography
ESC	Embryonic Stem Cell
FISH	Fluorescence *In Situ* Hybridization
GAM	Genome Architecture Mapping
GPSeq	Genomic loci Positioning by Sequencing
HMM	Hidden Markov Model
HOPs	Homolog-specific OligoPaints
HRP	HorseRadish Peroxidase
IF	ImmunoFluorescence
IQR	Inter-Quartile Range
IVT	*In Vitro* Transcription
LAD	Lamina-Associated Domain

MERFISH	Multiplexed Error-Robust Fluorescence *In situ* Hybridization
ML	Maximum-Likelihood
NAD	Nucleolus-Associated Domain
NGS	Next Generation Sequencing
NOR	Nucleolus Organizing Region
NSC	Neural Stem Cell
PBMC	Peripheral Blood Mononuclear Cell
PCR	Polymerase Chain Reaction
PGS	Population-based Genome Structure
PHB	Pre-Hybridization Buffer
RNA	RiboNucleic Acid
RPE	Retinal Pigment Epithelial
RRBS-Seq	Reduced-Representation BiSulphite Sequencing
RT	RetroTranscription
SAEC	Small Airway Epithelial Cell
SMC	Structural Maintenance of Chromosomes
SNIPER	Subcompartment iNference using Imputed Probabilistic ExpRessions
SNP	Single Nucleotide Polymorphism
SNR	Signal to Noise Ratio
SNV	Single Nucleotide Variant
SPRITE	Split-Pool Recognition of Interactions by Tag Extension
STORM	STochastic Optical Reconstruction Microscopy
TAD	Topologically Associating Domain
TFBS	Transcription Factor Binding Site
TSA	Tyramide Signal Amplification

List of Figures

Preface

Let us take a journey together. Imagine encountering life for the first time, in the form of a eukaryotic cell. A human retinal pigment epithelial cell (RPE) minding its own business and with no intention whatsoever to replicate.

One might be curious to see what sits at the cell's core and thus travel on highways of microtubules, the cytoskeleton, from the cell's surface towards its interior. There, they would find the nucleus: a large organelle separated from the internal cell volume, the cytoplasm, by two double-layer membranes. Inside the nucleus, what awaits us is an ensemble of long thread-like deoxyribonucleic acid (DNA) molecules, the chromosomes, that occupy some fraction of the nuclear volume like a half-empty bowl of spaghetti. These thread-like molecules are polymers, as they form through linearly assembling a set of subunits (monomers). The monomers represent the alphabet, the chromosomes are the chapters, and the genome is the manual book that contains the cell's blueprint - what we were looking for at the beginning of our journey, a journey to the center of the cell.

So, try to envision a book written not on pages but threads. Each thread is holding together a series of beads, the letters. How are these threads organized for them to be easily accessible and readable? They definitely cannot be placed on labeled shelves like in a library! In the late 1990s, such and similar questions triggered the birth of the genome architecture field, to which this Ph.D. thesis aims to contribute.

This book comprises three main parts. The Introduction (Part I) intends to give a general overview of our current knowledge of genome architecture (chapter 1) and the techniques used to investigate it (chapter 2). Part II is the actual Doctoral Thesis, focusing on the development of "Genomic loci Positioning by Sequencing" (a.k.a., GPSeq - pronounced "*gee-pea-seek*"): a novel biomolecular assay to measure the distance of genomic regions from the nuclear surface (section §4.1). This part also covers the development of iFISH: a fluorescence *in situ* hybridization (FISH) database and probe design resource instrumental for the validation of the newly developed GPSeq assay (section §4.2.1). The Final Remarks (Part III) conclude this book by providing future perspectives (chapter 7) on the presented studies.

Part I

Introduction

Chapter 1

Genome architecture

The genome of an interphase eukaryotic cell consists of multiple chromosomes: molecules of deoxyribonucleic acid usually found folded around cores of histone proteins and interacting with a number of other protein complexes and RNA (i.e., chromatin). The term genome architecture* indicates the 3D spatial organization that chromatin adopts within a cell's nucleus.

One of the first hints of the non-random nature of chromatin organization† came from the work of Dr. Emil Heitz in 1928. In a study performed on the moss *Pellia epiphylla*, Dr. Heitz revealed *heteropyknosis* of the moss' chromosomes: a different coloring of chromatin regions when treated with a chromosomal stain. In his article, he uses *heterochromatin* to indicate regions that remain heteropyknotic after late telophase and condensed during interphase and *euchromatin* for regions that become invisible after late telophase.[3–5] These heterochromatic regions appear to be preferentially located at the nuclear periphery, as further proved by electron microscopy images[6]. A recent study revealed a layer of more dense chromatin at the nuclear periphery when observing the genome of primary human Small Airway Epithelial Cells with electron microscope tomography[6].

During the past 10-15 years, the research community has seen a steady increase in the number of studies related to the genome architecture field (Figure 1.0.1). Numerous studies have revealed links between chromatin architecture and clinically relevant topics such as aging[7,8] and disease[9–13], triggering a quest for the identification of its shaping principles.

This chapter aims at providing a general overview of the topics necessary to fully grasp the aims, methods, and results presented in this doctoral thesis. It achieves this by initially presenting the genome architecture multi-scale organizational levels (section §1.1), before moving on to its radial patterns (i.e., the chromatin arrangement from the nuclear periphery towards its interior; section §1.2).

*"Genome architecture" should not be confused with "genetic architecture." The former indicates how the genome organizes in space, while the latter indicates the link between genotypic causes and phenotype variation.

†Generally, the term "chromatin architecture" is more apt to indicate short chromatin regions' organization. Here, instead, it will be used interchangeably as a synonym of "genome architecture."

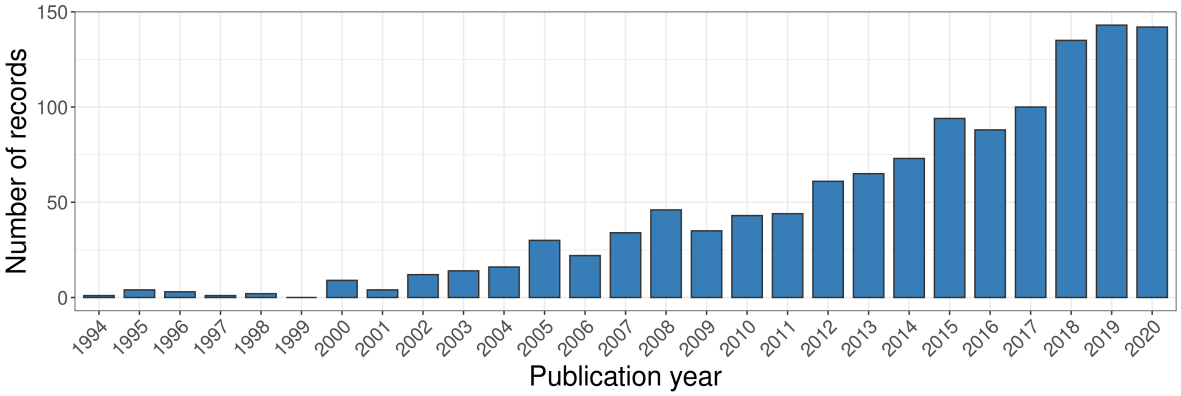

Figure 1.0.1: The number of references including the term "genome architecture" in their title, abstract, author keywords, or Keywords Plus, according to Web of Science, from 1994 to 2020. Plot generated with `ggplot2` package in R.

1.1 Multiple scales of organization

The eukaryotic genome is characterized by non-hierarchical organizational levels at multiple scales. In most human cells, the genome comprises two copies of each of 23 separate chromosome molecules, representing the highest organizational level. Instead, the lowest scale includes the linear order of nitrogenous bases and naked DNA wrapping around octameric histone protein cores to form chromatin (Figure 1.1.1). This section introduces the different levels of chromatin architecture, from chromosome- to local chromatin-scale.

1.1.1 Chromosomes and their territories

At the highest level, we find chromosomes characterized by telomeres located at their extremities and a centromere. Both centromeres and telomeres have highly repetitive sequences, which can hinder their investigation via sequencing assays.

The centromere plays a crucial role in the cell duplication process. As the genome replicates during S-phase, its copies must correctly segregate into two daughter cells. During meiosis and mitosis, centromeres are the location where a protein structure called kinetochore assembles. The kinetochore plays a fundamental role in holding the condensed sister chromatids together and attaching to the mitotic/meiotic spindle's microtubules allowing for the correct segregation of the sister chromatids to the daughter cells[15]. Moreover, chromosomes are categorized by the location of their centromere on their linear body: metacentric chromosomes have their centromere approximately located in their middle, sub-metacentric have it closer to their middle than to their telomeres, acrocentric have it closer to a telomere than to their middle, and telocentric when chromosomes have the centromere located at one of their telomeres[15].

Initially identified in 1978 by Elizabeth H. Blackburn and Joseph H. Gall, the telomeres are chromosomes' terminal regions constituted of a repetitive sequence[16]. In 1961, Hayflick and Moorhead[17] postulated a limit to a cell's replicative ability. In 1973, in line with their theory, the Russian scientist Alexei Olovnikov theorized that, due to the inability of DNA polymerase

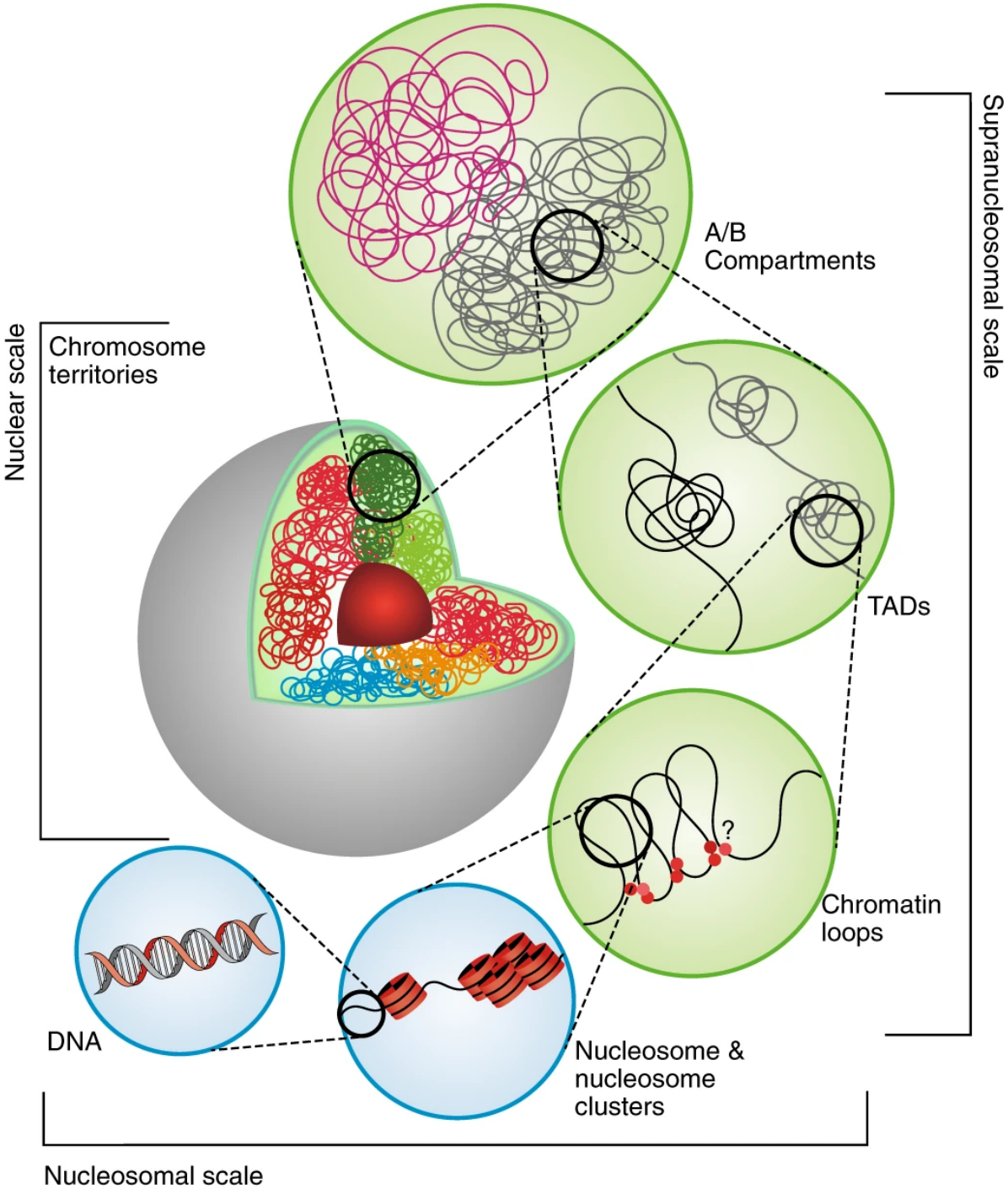

Figure 1.1.1: Multiple scales of genome organization. At the nucleosomal scale, one finds the linear sequence of bases in the naked DNA and nucleosomes. At the supranucleosomal scale, one can observe chromatin loops (see section 1.1.4), topologically associating domains (see section 1.1.3), and chromatin compartments (see section 1.1.2). At the nuclear scale, one can find chromosome territories (see section 1.1.1). Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Plants, *Three-dimensional chromatin packing and positioning of plant genomes*, Doğan E.S. & Liu C., 2018 [14].

to replicate the 3'-ends of DNA strands in the absence of a primer, chromosomes get shorter as a cell replicates[18]. In particular, Olovnikov stated that chromosome extremities might contain tandem repeats that act as a shield for the chromosome by being subjected to this shortening in their stead. Moreover, he also proposed the presence of an enzyme capable of non-template synthesis that would regenerate the terminal tandem repeats[18], today called telomerase. Later on, telomerase was identified as an active enzyme in germline and tumor cells, revealing it as a key player in long-term cellular proliferation[19].

While chromosome local intermingling events have been reported, mainly linked to concurrent expression of genes located on different chromosomes[20,21], these appear to be the exception: chromosomes are largely not intermixed. Instead, chromosomes tend to preferentially occupy separate regions, called *chromosome territories*, as initially proposed by Theodor Heinrich Boveri in his observations on blastomeres of *Ascaris megalocephala*[22]. In fact, only in the 1980s, the total genome material of hybrid cells with few chromosomes was used for DNA labeling, allowing to observe a handful of chromosomes forming territories[23,24]. Later, thanks to the establishment of chromosome sorting and DNA amplification techniques which allowed for the development of chromosome-specific labeling probes, these observations were extended to all chromosomes simultaneously[25–28].

Initial hints of the segregated nature of chromosomes during interphase came as early as the 1880s. In his observations, Carl Rabl described a peculiar arrangement of chromosomes after mitosis, which he believed to be established during telophase, where telomeres and centromeres cluster at opposite sides of the nucleus[21,29–31]. This chromosome arrangement, later referred to as *Rabl-configuration*, has been observed in various cell types from different organisms, being recently confirmed by single-cell Hi-C studies on mouse embryonic stem cells[32,33] (Figure 1.1.2).

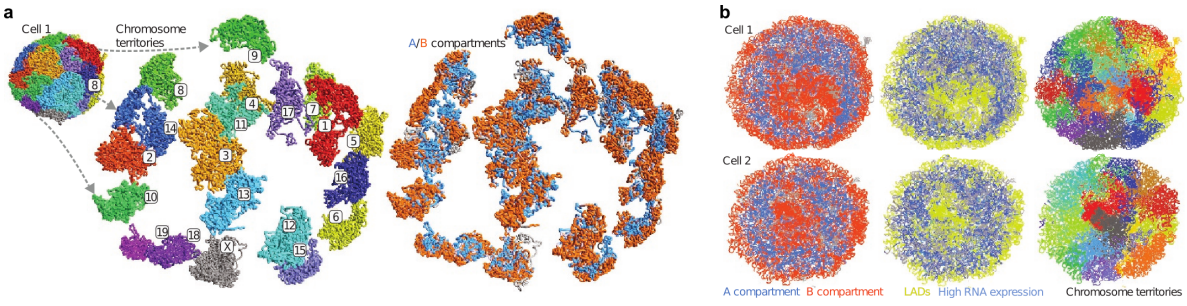

Figure 1.1.2: Chromosome territories and chromatin compartments reconstructed from single-cell Hi-C. (a) 3D genome reconstruction of haploid mouse ESCs. The expanded view shows separated chromosome territories. Left: colored by chromosome. Right: colored by A/B compartment. (b) Cross-section of super-imposed 3D genome reconstructions of haploid mouse ESCs. Chromatin beads are colored by A/B compartment (left), constitutive LADs (yellow), or highly expressed regions (blue)(middle), and chromosome (right). Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature, *3D structures of individual mammalian genomes studied by single-cell Hi-C*, Stevens T.J., et al., 2017[32].

1.1.2 Chromatin compartments

With the advent of chromosome conformation techniques (see section 2.2.1) measuring the frequency of contact between genomic regions, multiple studies revealed the presence of two major chromatin compartments – dubbed A and B compartments[34,35]. A *chromatin compartment* is defined as a set of non-consecutive genomic regions that contact preferentially other regions from the same compartment.

The first compartments were described in human cells, specifically B-lymphocytes (GM06990) and chronic myelogenous leukemia cells (K562), showing an average linear size ranging from few to tens of megabases[34]. Compartments were initially identified by eigen-decomposition of each chromosome's contact frequency matrix[‡], where the sign of the first eigenvector is used to distinguish regions belonging to the two compartments[34]. It is important to note that, on some occasions, the first eigenvector associates with the chromosome arms, while the second eigenvector correctly identifies the compartments[34].

While the eigenvector sign distinguishes the two compartments, compartment labels "A" and "B" are assigned based on the correlation between eigenvector values and epigenetic features like chromatin accessibility, replication timing, or histone marks[34,35]. Specifically, initial A compartment descriptions indicated higher gene density, gene expression, and chromatin accessibility than B compartment, alongside enrichment for active and repressed chromatin markers, respectively H3K36me3 and H3K27me3[34]. In other words, the A compartment is enriched in open and active chromatin regions, while the B compartment is enriched in more compacted and silent or repressed ones[34,35]. Notably, the A/B compartmentalization appears to be cell type-specific with loci switching compartment based on gene expression profiles[36], as the main determinant of compartmentalization appears to be transcription and not chromatin contacts (see section 1.1.4)[37–39].

Recent FISH-based studies performed on diploid human fibroblast IMR90 cells revealed that A and B compartments tend to be radially arranged, with B compartments being more peripheral than the internal A compartments[40]. Moreover, other FISH studies revealed how this polarized arrangement holds true for each chromosome territory, with A compartments at one pole (more internal) and B at the other (more peripheral)[41]. This polarized distribution of A/B compartments has also been described in structures built from single-cell Hi-C maps[32].

Recently, high-resolution Hi-C studies revealed the presence of a more refined compartment categorization. Specifically, in 2014 Rao et al. proposed the presence of at least six subcompartments, two for compartment A (namely A1 and A2) and four for compartment B (namely B1-4), based on distinct histone marker patterns[35].

Specifically, subcompartments A1 and A2 present many characteristics of A compartment, including high gene density, high gene expression, enrichment for active chromatin marks (H3K4me1, H3K27ac, H3K36me3, H3K79me2). Moreover, both subcompartments

[‡]In a contact frequency matrix, the value of cell c_{ij} is the contact frequency between genomic regions i and j. See section 2.2.1.1 for more details.

tend to localize at intermediate radial positions, being depleted at nucleolus-associated domains (NADs) and lamina-associated domains (LADs). Interestingly, while both A1 and A2 subcompartments present early replication time, they terminate replicating at different cell cycle points. Furthermore, The A2 subcompartment shows a stronger enrichment in H3K9me3 and longer genes while displaying a lower GC-content than A1[35].

Similarly, subcompartments B1, B2, and B3 present the same general characteristics as B compartments while being differentiated by several features. Indeed, subcompartment B1 appears to indicate facultative heterochromatin regions, as suggested by its enrichment in H3K27me3 and depletion in H3K36me3 chromatin marks. On the other hand, subcompartments B2 and B3 do not show these B1-specific characteristics while sharing similar late replication times. Moreover, while subcompartment B2 shows enrichment in pericentromeric heterochromatin regions, LADs, and NADs, subcompartment B3 is enriched in LADs but depleted in NADs[35].

Notably, the A and B compartments have also been annotated by methods that do not rely on eigendecomposition. In 2015, Fortin and Hansen estimated the A/B compartment eigenvector from the long-range correlation of epigenetic data (DNA methylation, DNAse hypersensitivity, and scATAC-seq) without the use of Hi-C information[42]. In 2018, Zheng and Zheng developed `CScoreTool`: a maximum-likelihood approach for estimating a C-score, closely resembling the eigenvector values[43]. Moreover, in 2019, Xiong and Ma presented Subcompartment iNference using Imputed Probabilistic ExpRessions (`SNIPER`): a machine-learning approach for subcompartment annotation from Hi-C data[44].

1.1.3 Chromatin domains

At a smaller scale, Hi-C allowed for the identification of topologically associating domains (TADs), where a region within a TAD tends to contact other regions from the same domain more frequently than regions outside it[35,45]. TADs have a variable size range, from 40 kb to 3 Mb (with a median size of 185 kb)[35] and, while genome compartmentalization shows tissue-specific features, TADs appear to be mostly conserved even across different species[45–47]. Initially thought to be the building blocks of genome architecture due to their highly conserved nature[48,49], TADs are now seen as the result of chromatin looping (see section 1.1.4, Figure 1.1.3a), as chromatin loop extrusion has been described as one of the mechanisms that drive the formation of TADs[50–52]. Two types of TADs have been described: one arising from chromatin loop extrusion, and with a characteristical CTCF enrichment and cohesin binding at its borders, and the other characterized by specific histone marks but with no CTCF enrichment[35,49].

TADs appear to restrict enhancer-promoter interactions and thus regulate gene expression[53–55] via chromatin condensation that has been suggested to be driven by transcription-associated supercoiling[56]. It follows that TAD arrangement alterations, such as domain merging or splitting, are causally linked to a variety of diseases[57], such as limb malformations[58,59]

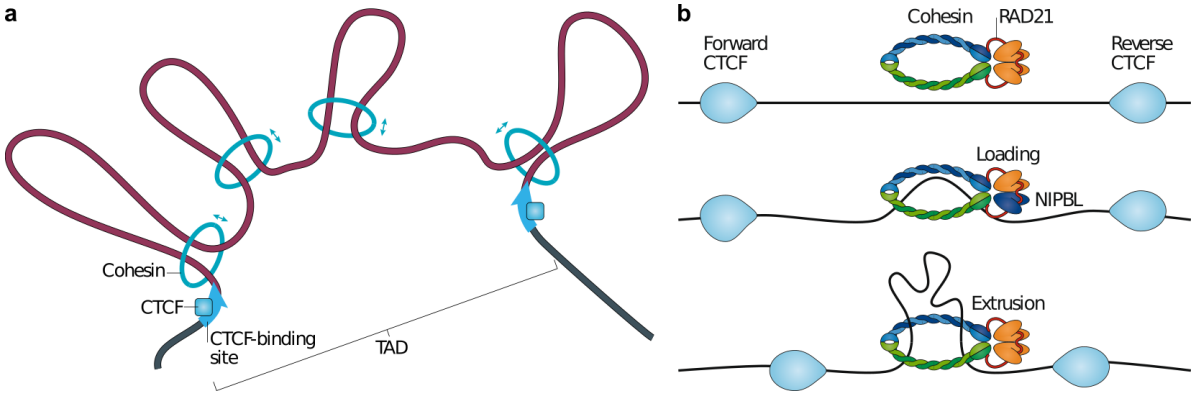

Figure 1.1.3: Chromatin loop extrusion and chromatin domains. (a) Schematic example of a topologically associating domain formed via cohesin-based loop extrusion and including multiple loops. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Reviews Molecular Cell Biology, *The role of transcription in shaping the spatial organization of the genome*, van Steensel B. & Furlong E.M.E., 2019[67]. (b) A general model of loop extrusion. Cohesin's loading onto chromatin is dependent on NIPBL. Extrusion proceeds until convergent head-to-head oriented CTCF act as a block. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Reviews Genetics, *Organizational principles of 3D genome architecture*, Rowley M.J. & Corces V.G., 2018[49].

and cancer[60]. Moreover, recent studies revealed the presence of TAD cliques, suggesting the ability of TADs to interact with each other to shape the genome architecture[61].

Unfortunately, the definition of TADs makes them somewhat elusive, as is apparent from the number of algorithms that have been proposed and used over the years for their identification. To mention a few: the Python package `DomainCaller`, first to be proposed, identifies TADs by applying a Hidden Markov model (HMM) on a directionality index calculated on a binned genome[45]; the `Arrowhead` algorithm, implemented in the Juicer Java software[62], applies a matrix transformation to facilitate the identification of TAD boundaries[35,62]; `Armatus` combines a score function and multi-scale approach to quantify a domain's quality based on local interaction density[63]; `TADtree` is an algorithm for the identification of nested TADs, based on empirical contact frequency distributions[64]; and `TADbit` uses a Poisson regression to identify the optimal number of domains a chromosome can be divided into, based on the Bayesian Information Criterion[65].[§]

Due to their link with chromatin looping, particular interest has been shown not only for the TAD regions but also for their borders (i.e., the narrow regions between two consecutive TADs). Indeed, TAD borders show enrichment in CTCF sites, whose orientation is instrumental for chromatin looping[68]. Moreover, TAD border strength (i.e., how marked a border between two TADs is, or how easily two consecutive TADs can be distinguished) correlates with structural protein occupancy, including CTCF or cohesin, which in turn has also been linked to the ability of TADs to segregate enhancer activity to different regions[69,70].

While TADs were initially described starting from bulk Hi-C, representing a cell pop-

[§]For more details, we direct the reader to a recently published review by Forcato et al., where they compare these and other TAD calling tools[66].

ulation rather than a single-cell feature, recent FISH experiments performed on Drosophila revealed them as nanometer-sized compartments (approx. 180 nm in average diameter)[71].

1.1.4 Chromatin loops

High-resolution Hi-C maps allowed for the identification of chromatin loops as another chromatin architecture level. The fact that loops can encompass single or multiple TADs and compartments supports these three structures' interdependency[49,50].

The first chromatin loop extrusion models to be drafted include a motor protein (such as cohesin[¶]) as a requirement for chromatin loop formation[68]. Specifically, an initial loop extrusion model was proposed where a two-units DNA-binding extrusion complex would bind to a single chromatin locus, forming a small chromatin loop. Subsequently, the extrusion complex units would extrude DNA by moving in opposite directions along the genome, allowing the loop to grow without knotting[73]. More recent models, where one or two cohesin rings compose the extrusion complex, have been proposed in which the extrusion is tightly regulated by the presence of CTCF-binding sites, which could arrest the extrusion in a binding-site orientation-dependent manner[68‖] (Figure 1.1.3b). At the same time, a new motor-less model has been recently proposed, where loop extrusion is not performed by a protein but by thermal motion instead[75]. A recent study on this topic provides the first real-time imaging of an SMC complex, specifically a yeast condensin[**] complex, extruding a DNA loop, providing evidence for asymmetric (one-direction) extrusion by a single ring complex[78].

Crucial players in chromatin loops' formation are the extrusion motor protein cohesin and the transcription factor CTCF[79], which behaves as a cohesin stopper. Specifically, CTCF contains 11 zinc finger domains, with which it binds DNA, and it has been observed via atomic force microscopy to be able to circularize naked 941-bp DNA, containing 3 CTCF-binding sites, *in vitro*[80]. Moreover, the gain of CTCF-anchored chromatin loops has been proposed to mark the passage from naive pluripotent to more differentiated cells, as revealed from Hi-C data obtained from mouse ESC and NSC derivatives[81].

Recent studies focused on the role of CTCF/cohesin-driven chromatin looping in genome architecture by studying the impact of the loss of CTCF[82], the cohesin-component Rad21[83], or the cohesin-loading factor Nibpl[84] on chromatin organization. Specifically, long-range interactions, characteristic of chromatin loops-associated TADs and loop stems, vanished globally. Simultaneously, compartments appeared to be preserved and even reinforced, revealing a more refined compartment structure that reflected local chromatin features. Surprisingly, loss

[¶]For more details on cohesin's role in genome architecture and gene expression, we direct the reader to a recent review by Zhu and Wang[72].

[‖]For more details on one-sided and two-sided extrusion models, we direct the reader to a recent review by Banigan and Mirny[74].

[**]It is important to note that, while cohesin is related to chromatin loop extrusion, condensin is responsible for chromatin condensation, especially in particular cellular processes like chromatid condensation during mitosis[76] or gene repression during cell quiescence[77].

of cohesin affected the transcription only of a small subset of genes, typically located close to super-enhancers, suggesting only a modest role of chromatin loops in promoter-enhancer contact formation and a more prominent role played by chromatin features-based compartmentalization[83,85].

Regarding the role that chromatin loops play in gene regulation, they appear to be involved in gene repression during *Drosophila* development[86], and minor loops' stem contact frequency changes appear to be associated with differential gene expression[87]. Moreover, a class of cell-type unspecific super-enhancers has been recently identified as having a strong association with rapidly recovering chromatin loops[88]. Furthermore, loops play a role in establishing immunity, as looping has been associated with transcription of major histocompatibility complex loci[89] and recombination of antigen receptor genes[90].

Given the fundamental importance of chromatin loops in driving point-wise chromatin contacts (between the loop stems) and chromatin segregation[49,50], and their involvement in a large variety of cellular processes, special effort is currently dedicated to clarifying their formation mechanism.

1.2 Radial patterns

The genome of interphase cells resides in the nucleus, a cellular organelle with a double layer membrane and a protein mesh, known as nuclear lamina, coating its internal surface. This spatial constraint, represented by the nucleus membrane and lamina, is crucial in creating two different major environments: peripheral and internal.

Already from nucleic acid counter-staining experiments, it was apparent that the genome assumes a non-random radial distribution from the nuclear lamina towards the nucleus interior. Indeed, regions characterized by higher staining due to higher DNA density and compaction tend to be enriched at the periphery and are known as heterochromatin. Conversely, more relaxed, gene-enriched, active genomic regions, known as euchromatin, are enriched in the internal environment[91] (Figure 1.2.1b).

This non-random radial distribution of the genomic material seems to be tightly linked to its function. For example, rod photoreceptor cells of nocturnal animals have an inverted chromatin distribution in their nucleus, with more silenced and compact chromatin in the interior. This has been speculated to increase the ability to capture their environment's low light by acting as collecting lenses[92]. Another example is the lineage-dependent radial repositioning of large portions of the genome during granulocyte differentiation, accompanied by decompaction of the same regions[93,94].

In this section, we will describe features that characterize the nuclear peripheral (section 1.2.1) and internal environments (section 1.2.2), to then describe the radial arrangement of chromosome territories (section 1.2.3) and genes (section 1.2.4).

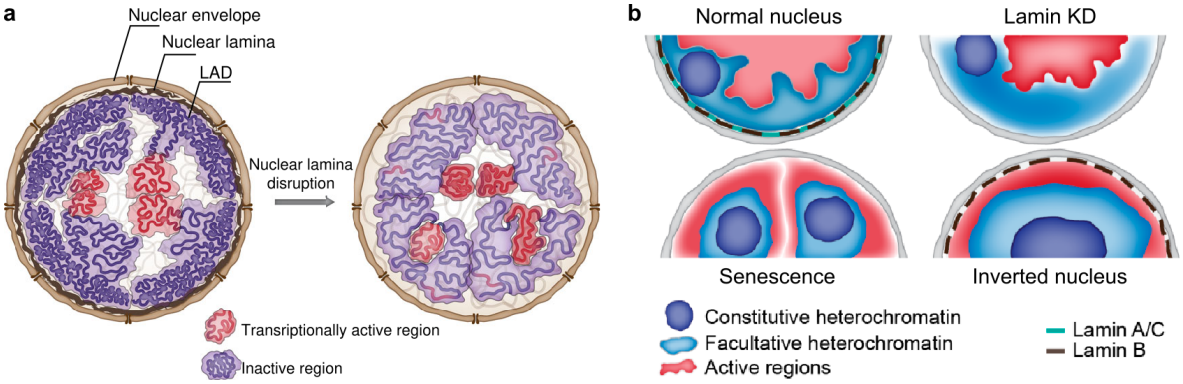

Figure 1.2.1: Nuclear lamina's role in genome architecture. (a) Schematic representation of genome architecture alterations upon nuclear lamina disruption in *Drosophila* S2 cells. Specifically, the cartoon depicts overall detachment from the nuclear envelope, LADs relaxation, and higher compaction of transcriptionally active regions. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Communications, *Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila*, Ulianov S.V., Doronin S.A., et al., 2019[95]. (b) Schematic representation of genome arrangements in cells lacking nuclear lamina-chromatin contacts. The carton depicts a normal nucleus (top left), a nucleus upon lamin knock-down (top right), the nucleus of a senescent cell (bottom left), and the nucleus of a cell with inverted heterochromatin/euchromatin arrangement (bottom right). Adapted by permission from MDPI: MDPI, Cells, *The Nuclear Lamina as an Organizer of Chromosome Architecture*, Shevelyov Y.Y. & Ulianov S.V., 2019[96].

1.2.1 The nuclear periphery

As aforementioned, the nuclear periphery is characterized by the presence of heterochromatin: i.e., mostly compacted, silenced, and gene-depleted regions. Heterochromatin appears to be held in place by interactions with the nuclear lamina, a protein mesh composed of Lamin A/C and B[97]. Thanks to the DamID method (see section 2.2.4), these interactions have been thoroughly characterized. Specifically, DamID has uncovered the so-called lamina-associated domains (LADs): large genomic regions that tend to interact with the nuclear lamina[98,99]. LADs are greatly enriched at peripheral heterochromatic regions and have characteristically low transcriptional levels, which led to postulating a silencing activity of the lamina[100]. Of note, constitutive (i.e., cell-type independent) LADs are enriched in A/T base pairs[101], consistent with their gene-depleted nature.

Recent studies focused on the nuclear envelope's role as a shaping force of the genome architecture, revealing a functional behavior on top of its intrinsic nature of geometrical constraint. Specifically, nuclear lamina disruption in *Drosophila* S2 cells induces a reduction of genome volume, caused by an increase in overall chromatin compaction. In their study, Ulianov et al.[95] hypothesized that attachment to the nuclear envelope allows for compaction of local regions and stretching of active regions towards the nuclear interior. Upon nuclear lamina disruption, *Drosophila* S2 cells show a characteristic complete detachment from the nuclear envelope, impeding proper active chromatin stretching and causing the observed genome volume reduction (Figure 1.2.1). A similar study was recently carried out on mouse lamin B knock-out embryonic stem cells[102], showing an incomplete chromatin detachment from the nu-

clear envelope, overall decompaction of LADs, and expression-linked alteration of chromatin contacts.

Interestingly, lamina-associated domains have been linked to gene regulation in development[103,104] and cancer[105], underscoring the pivotal role of lamins in shaping the genome organization. At the same time, different levels of promoter repression in LADs seem to be linked to local chromatin features, suggesting the presence of an additional mechanism in place to counteract the repressive effect of lamina-drive chromatin compaction.

1.2.2 The nuclear interior

Unlike the nuclear periphery, the internal region of the nucleus does not have a well-defined boundary. The nuclear interior appears to be gene-enriched, showing higher GC-content, more actively expressed genes, and relaxed or accessible chromatin[91]. Factually, a handful of genes have been shown to relocalize from the peripheral to the internal environment upon expression activation[106].

Interestingly, lamins' role in regulating the genome architecture appears not to be limited to the peripheral environment. Indeed, by combining data obtained by ChIP-seq, RNA-seq, ATAC-seq, and Hi-C, Pascual-reguant et al. recently described internally located euchromatin LADs, characterized by the interaction between lamins and euchromatic regions, as opposed to canonical LADs, typically constituted by peripherally located heterochromatin[107].

In particular, Forsberg et al. investigated the genome architecture regulatory role of both lamin B and A, in more detail, in HepG2 hepatocarcinoma cells[108]. In their study, chromatin interactions appeared to be more stable with lamin B than its counterpart A, but the global loss of lamin A caused an increase in chromatin mobility in the nuclear interior. Moreover, they revealed that lamin B-associated domain (B-LAD) appearance after lamin A-associated domain (A-LAD) loss results in the repositioning of the domain towards the nuclear periphery. Vice-versa, A-LAD loss in regions with locally low lamin B levels correlates with relocalization towards the nuclear interior. Altogether, these results underscore the complex interplay between lamin A and B in regulating the radial genome architecture.

1.2.3 Radial positioning of chromosome territories

Chromosome territories also appear to follow a non-random radial distribution, and, over the years, different chromosome radial positioning models have been proposed based on a variety of genetic and genomic features.

Initially, researchers approached the study of chromosome radial arrangement in the context of their localization, or the localization of centromeres, relative to the barycenter point of metaphase spreads in human cells (peripheral lymphocytes and whole blood samples). The primary rationale, in this case, was the straightforward, unambiguous identification of chromosomes in metaphase spreads. Such early-day studies hinted at a correlation between radial

positioning and chromosome size, with smaller chromosomes preferentially located closer to the metaphase spread center[109–111].

Afterward, observations on the localization of human chromosomes 18 and 19 in interphase cells put this initial model to the test. Factually, while chromosomes 18 and 19 have similar sizes (approximately 80 and 60 Mb, respectively), they differ in their radial positioning. Indeed, chromosome 18 tends to have a characteristically peripheral localization, as opposed to the internally located 19[112–114] (Figure 1.2.2a). Interestingly, these two chromosomes have strikingly different gene densities, with 18 being the most gene-depleted and 19 being the most gene-dense chromosome of the human genome[115,116]. These results seemed to point to a more gene-density-based model of chromosome radial arrangement.

Indeed, FISH-based experiments revealed a preferentially internal localization for GC-riched early-replicating regions, while R-riched late-replicating regions tend to localize preferentially at the nuclear periphery[117]. Interestingly the study performed by Sadoni et al. also indicated that each chromosome territory tends to follow this polarization, which hinted at the polarization of A/B compartments that was revealed only later on (see section 1.1.2,[34,35,40,41]).

A later FISH-based study focused on the correlation between chromosome radial localization and gene expression[118]. Specifically, labeling of highly or lowly transcribed genomic regions (also referred to as ridges and anti-ridges, respectively) revealed a preferentially peripheral localization of anti-ridges and internal localization of ridges. Another FISH-based study from the same year focused instead on gene-density as a potential predictor of chromosome radial localization[119]. Similarly, Küpper et al. revealed internal chromosome polarization with gene-dense and transcriptionally active regions localizing at the nuclear interior and gene-poor and transcriptionally inactive regions localizing at the nuclear periphery (Figure 1.2.2b-c). In contrast with the previous studies, Küpper et al. argued for gene density as the main predictor of chromosome radial positioning instead of replication timing[117] and transcriptional activity[118].

Interestingly, Tanabe et al. performed a comparison of evolutionarily conserved human chromosome 18 and 19 regions in chicken. Their study revealed the conservation of such regions' radial localization in interphase cells, potentially hinting at the evolutionary conservation of radial genomic localization even across highly divergent genomes[114], further underscoring the importance of the radial aspects of genome architecture.

1.2.4 Radial positioning of genes

Entire chromosome territories are not the only genomic structures to be radially distributed: the same is true for sub-chromosomal regions. Indeed, studies on a handful of genes, like the astrocytes-specific marker *GFAP*, supported a function-related differential positioning of individual genes and their alleles[121]. It follows that many studies focused on genes that tend to relocalize upon transcriptional activation[122].

For example, the functionally unrelated genes *GASZ*, *CFTR*, and *CORTBP2*, preferentially

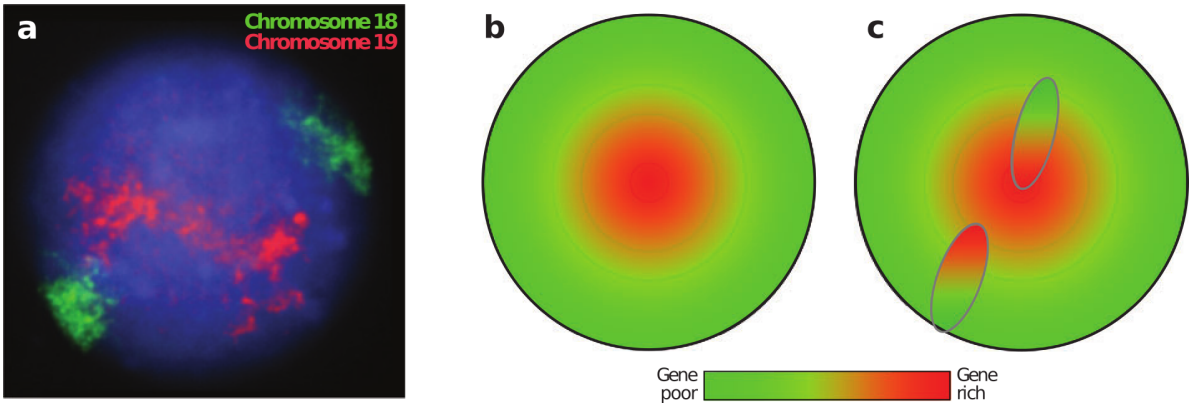

Figure 1.2.2: Radial positioning of chromosomes 18 and 19, and schema of gene density as main predictor. (a) Image of chromosome paint FISH performed on human chromosomes 18 (green) and 19 (red). Blue: DNA staining channel. (b) Schematic representation of the radial organization of gene-density in interphase cells. (c) Schematic representation of gene density radial polarization within single chromosome territories (ellipses). Adapted by permission from Annual Reviews, Inc. via Copyright Clearance Center: Annual Reviews, Annual Review of Genomics and Human Genetics, *The Spatial Organization of the Human Genome*, Bickmore W.A., 2013 [120].

located at the nuclear periphery, have been reported to relocalize to the nuclear interior upon their activation in various human cell types [106]. Moreover, in fibroblasts and B-lymphoblastoid cells, upon treatment with interferon-gamma, major histocompatibility complex genes and other regions on chromosome 6 appear to loop out from the bulk of the chromosome territory towards the nuclear interior when transcribed [89].

Studying gene radial positioning can be challenging due to the highly gene-specific nature of the results; in other words, no clear general trends seem to stand out. Moreover, in gene-dense regions, repositioning of the whole region might be driven by the activation of a subset of its genes - a known confounding factor referred to as the "neighboring effect" [122].

A recent work focused on three genes up-regulated upon embryonic stem cell differentiation [123] (*Ptn*, *Sox6*, and *Nrp1*). There, Therizols et al. tried to deconvolve the effect of radial repositioning, chromatin decondensation, and expression on replication timing. The rationale being that late-replicating regions tend to locate at the nuclear periphery. Specifically, this study shows that radial relocalization and chromatin decondensation are insufficient to cause a shift in replication time, which appeared to be a direct consequence of transcriptional activity. Nonetheless, transcription appeared to be unlinked from gene repositioning, which seems to be linked to chromatin decompaction instead [123].

Another example is the study where Tan et al. [124] focused on the radial localization of olfactory receptors by exploiting Dip-C (see section 2.2.1.2), a novel Hi-C-based technique. Specifically, this study revealed that olfactory receptor genes and enhancers locate preferentially in the nucleus interior of mouse olfactory sensory neurons while showing a preferentially peripheral localization in non-neuronal cells [124].

Altogether, these results underscore the limits of a low-resolution chromosome-scale approach to genome radial architecture and support the need for more resolved approaches that

should simultaneously measure multiple genes and non-genic regions or, ideally, the whole genome at once.

Chapter 2

Experimental methods

The genome architecture field sits at the intersection between several disciplines, ranging from biochemistry to physics. Thus, it does not come as a surprise that the methods available for its investigation are incredibly varied, with most techniques belonging to two main categories: imaging- and sequencing-based methods.

Imaging-based methods have been renowned for their high spatial and single-cell resolution while negatively impacted by a low multiplexity. Nevertheless, recent developments on this front have ameliorated this situation by improving their multiplexity[125–127]. On the other hand, sequencing-based methods provide a relatively simple and straightforward way to obtain genome-wide bulk (or, more recently, single-cell) data while most often not providing direct spatial measurements (this is not the case, for example, of spatial transcriptomics[128,129]). As one might expect, these two types of orthogonal methods can often provide insights into the same level of chromatin architecture (figure 2.0.1). Moreover, they might sometimes provide what appear as contrasting results, which require a better interpretation to be reconciled[130].

This chapter provides an overview of the techniques used in this thesis' studies*. Moreover, it also covers techniques that we later compare to the methods established in this thesis to provide a more thorough validation (see section §4.2) and rationale for their establishment (see chapter 6). Specifically, we begin by covering imaging-based techniques (section §2.1), before moving on with sequencing-based ones (section §2.2).

2.1 Imaging techniques

While imaging methods allow observing genomic regions, transcripts, or proteins directly in single cells' nuclear environment, this section focuses specifically on those techniques that include a step of genome labeling. Such techniques allow for the precise localization of genomic regions in space to distinguish between interacting and non-interacting loci based on the measured distance between them (Figure 2.1.1c).

*For details on the methods used for genome architecture investigation, we direct the reader to a recent review by Kempfer et al.[131]

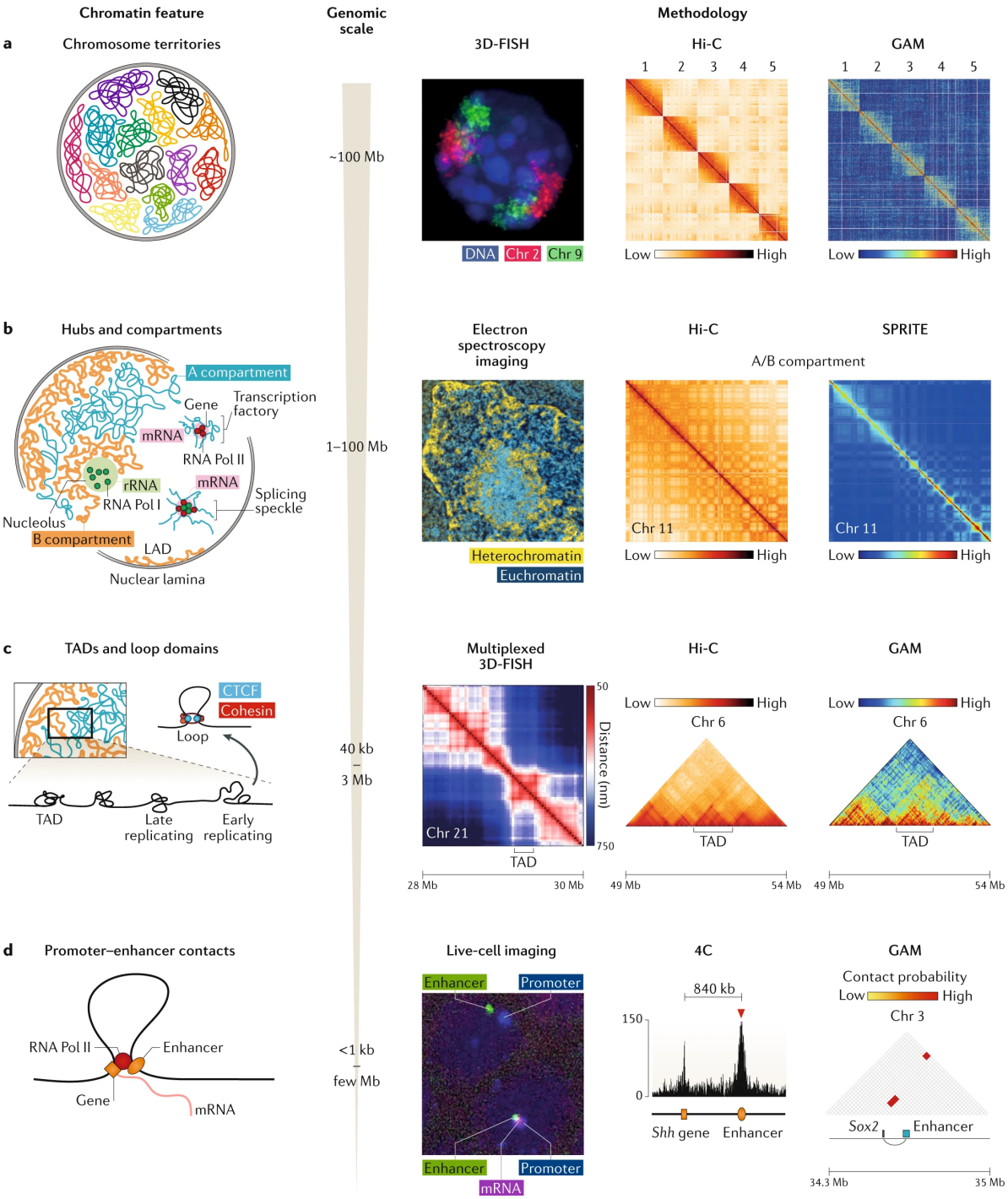

Figure 2.0.1: Methods to investigate different scales of genome architecture. (a) At the lower scales, chromosome territories (see figure 1.1.2) have been characterized by 3D-DNA FISH (see section 2.1.1), Hi-C (see section 2.2.1.1), and GAM (see section 2.2.2). (b) At sub-chromosomal scale, transcriptional hubs (not covered in this thesis) and chromatin compartments (see section 1.1.2) have been characterized by electron spectroscopy imaging[132] (not covered in this thesis), Hi-C, and SPRITE (see section 2.2.3). (c) Topologically associating domains (see section 1.1.3) and chromatin loops (see section 1.1.4) have been characterized via multiplexed 3D-DNA FISH (see section 2.1.1.1), Hi-C, and GAM. (d) Promoter-enhancer contacts have been characterized by live-cell imaging, 4C (see section 2.2.1), and GAM. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Review Genetics, *Methods for mapping 3D chromosome architecture*, Kempfer R. & Pombo A., 2020[131].

2.1.1 3D DNA Fluorescence In Situ Hybridization (3D-DNA FISH)

Fluorescence *in situ* hybridization (FISH) is one of the first visualization techniques used to study genome architecture[23,24]. FISH takes advantage of single-stranded nucleic acids' ability to hybridize to a reverse complementary single strand. By labeling the complementary single strand with a fluorophore, one can reveal the location of a specific sequence of interest in a sample[26]. Given FISH's ability to localize and, subsequently, quantify the copy number status of the targeted genomic sequences, this technique has been used in the clinical setting to identify prognostic chromosome rearrangements in cancer patients[133].

In 3D-FISH, to each target sequence, a tagged oligonucleotide is first hybridized and then visualized. It is important to note that a sufficient number of fluorophores is needed to generate a detectable FISH signal. Thus, a FISH signal can be achieved either by increasing the number of fluorescent tags per target or of proximal target sequences. Thus, in this context, the term "FISH probe" indicates the ensemble of target sequences that produce a detectable signal when labeled. While the minimum number of fluorophores and target sequences required to constitute a FISH probe is available in published protocols[125,134], particular care is needed to identify the target sequences. Specifically, all target sequences in a probe should: (1) minimize off-targets, (2) have a compatible and narrow melting temperature range, (3) avoid self-annealing, and (4) avoid signal dilution (i.e., they cannot be spatially too far apart on the target molecule)[135].

Initially, FISH protocols did not build upon the synthesis of accurately designed target sequences. Instead, probes generation started from bacterial artificial chromosomes (BAC) with large target regions inserted[136]. This procedure has little control over any probe's sensitivity and specificity and is not compatible with labeling small genomic regions (typically targeting genomic regions of more than 100 kbp). A classical application, for example, is the so-called "chromosome painting," where entire chromosomes were labeled at the same time to identify rearrangements[26].

Nonetheless, by combining FISH with different super-resolution microscopy techniques and labeling strategies, several derivative techniques came to light over the years. Indeed, multiple recent protocols tried to meet the need for higher resolution, higher multiplexity, and better control over probe quality. Most importantly, target sequences are now being designed *in silico* starting from a reference genome and then selectively amplified to produce ready-to-use FISH probes[125,126,134,135].

We call such FISH probes "oligonucleotide-based" (i.e., with accurately designed sequences), in opposition to "BAC-based" ones. Two recently developed FISH techniques that use oligonucleotide-based probes are Oligopaints[137] and MERFISH[125,126]. Both include *ad hoc* oligonucleotide design pipelines, probe amplification, and hybridization protocols. Specifically, MERFISH and Oligopaints probes design used OligoArrayAux[138], a software suite implemented to design microarray probes[125,126,134]. More recently, Oligopaints-based studies used the `OligoMiner` pipeline instead, aiming to render the design of FISH probes

more user-friendly[139].

Grossly speaking, these two techniques differ on a few significant characteristics. First, while Oligopaints probes are generally composed of oligonucleotides of variable length[139], that is not the case for MERFISH[125,126]. Secondly, MERFISH power resides in an iterative hybridization protocol that provides each target with a specific color signature, allowing for unambiguous identification of hundreds of targets in a single sample[125,126]. On the other hand, Oligopaints designed sequences are either directly tagged with a fluorophore or combined with orthogonal sequences that hybridize to a labeled oligonucleotide in a second step[135]. Nonetheless, it is essential to note that both techniques can be combined relatively easily to answer application-specific needs, given FISH's intrinsic flexibility.

FISH experiments' spatial resolution has continuously increased over the years, thanks to advances in microscopy techniques and probe sequence design. Specifically, the development of super-resolution microscopy techniques (e.g., STORM[40]) increased FISH applicability allowing one to determine the signal location at unprecedented resolutions. Such improvements lead the analysis of FISH experiments to move from the classical 2D approach (i.e., after projecting the signals over the Z dimension) to a 3D perspective, thus providing richer data that can complement other assays.

2.1.1.1 Oligopaints

Initially published in 2012, the first Oligopaints FISH workflow introduced the computational steps to design a set of sequences (32 nt long, in the original paper) homologous to genomic regions of interest. Furthermore, it provided the protocol steps necessary to amplify such sequences from synthesized oligonucleotide libraries to produce FISH-ready probes[140].

As aforementioned, the oligonucleotide-based Oligopaints approach provided several advantages compared to BAC-based probes. First and foremost, the ability to refine a FISH probe's sequences' design allows targeting smaller genomic regions. Furthermore, by discarding any oligonucleotide able to hybridize on multiple genomic locations, the aspecific signal is greatly diminished, resulting in FISH dots with a higher signal-to-noise ratio (SNR).

The main technical limitation of the oligonucleotide-based approach is that the target sequence must be known a priori. In other words, it is not possible to design sequences if the organism of interest does not have an assembled reference genome.

Oligopaints probes amplification starts from a library of target oligonucleotides with appended probe-specific PCR adapters in the form of orthogonal 21 nt-long sequences. Notably, one of the PCR primers used during probe amplification has a fluorophore-conjugated at one of its ends. Thus, one can visualize an Oligopaints probe at the microscope right after hybridization.

Initially presented with applications on both *Drosophila* and human cells, the Oligopaints approach enabled several studies on different organisms. Interestingly, the developers of Oligopaints expanded the approach by using a different design. Specifically, they introduced

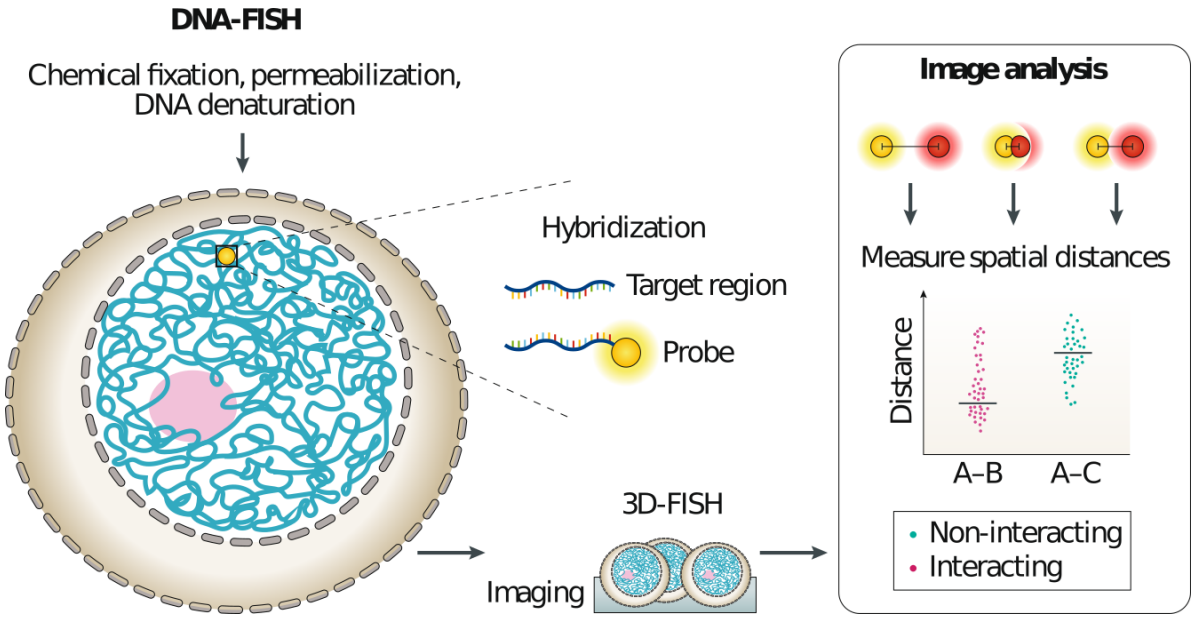

Figure 2.1.1: DNA-FISH workflow. The sample preparation process includes cell fixation, permeabilization, and DNA denaturation. Then, fluorescently labeled FISH probe hybridize to the target regions of interest. Finally, imaging of the sample allows measuring the position of the labeled regions. Panel: interaction (often referred to as "contact") is often established based on the measured spatial distance between two labeled genomic regions. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Review Genetics, *Methods for mapping 3D chromosome architecture*, Kempfer R. & Pombo A., 2020[131].

two adapters at one end of each oligonucleotide, allowing a second hybridization step targeting such primers. This secondary-hybridization approach, combined with stochastic optical reconstruction microscopy (STORM), allowed for the super-resolution imaging of the bithorax complex genomic region (approx. 316 kb)[137].

Beliveau et al. introduced homolog-specific Oligopaints (HOPs) to achieve differential allele labeling. Specifically, HOPs are Oligopaints probes whose sequences overlap with known SNPs, and the SNP base value is corrected to match the expected sequence of the targeted allele. Successful applications of HOPs include the differential allele labeling of the genomic region containing the X-inactivation center in chimeric mouse cells and the bithorax complex genomic region chimeric Drosophila cells[137]. As such, HOPs-based results support FISH's ability to distinguish alleles, potentially allowing the study of the chromatin architecture of chromosome alleles.

Furthermore, Beliveau et al. recently published `OligoMiner`[139] and `OligoMinerApp`[141] - respectively, pipeline and web-server - for the streamlined design of Oligopaints sequences.

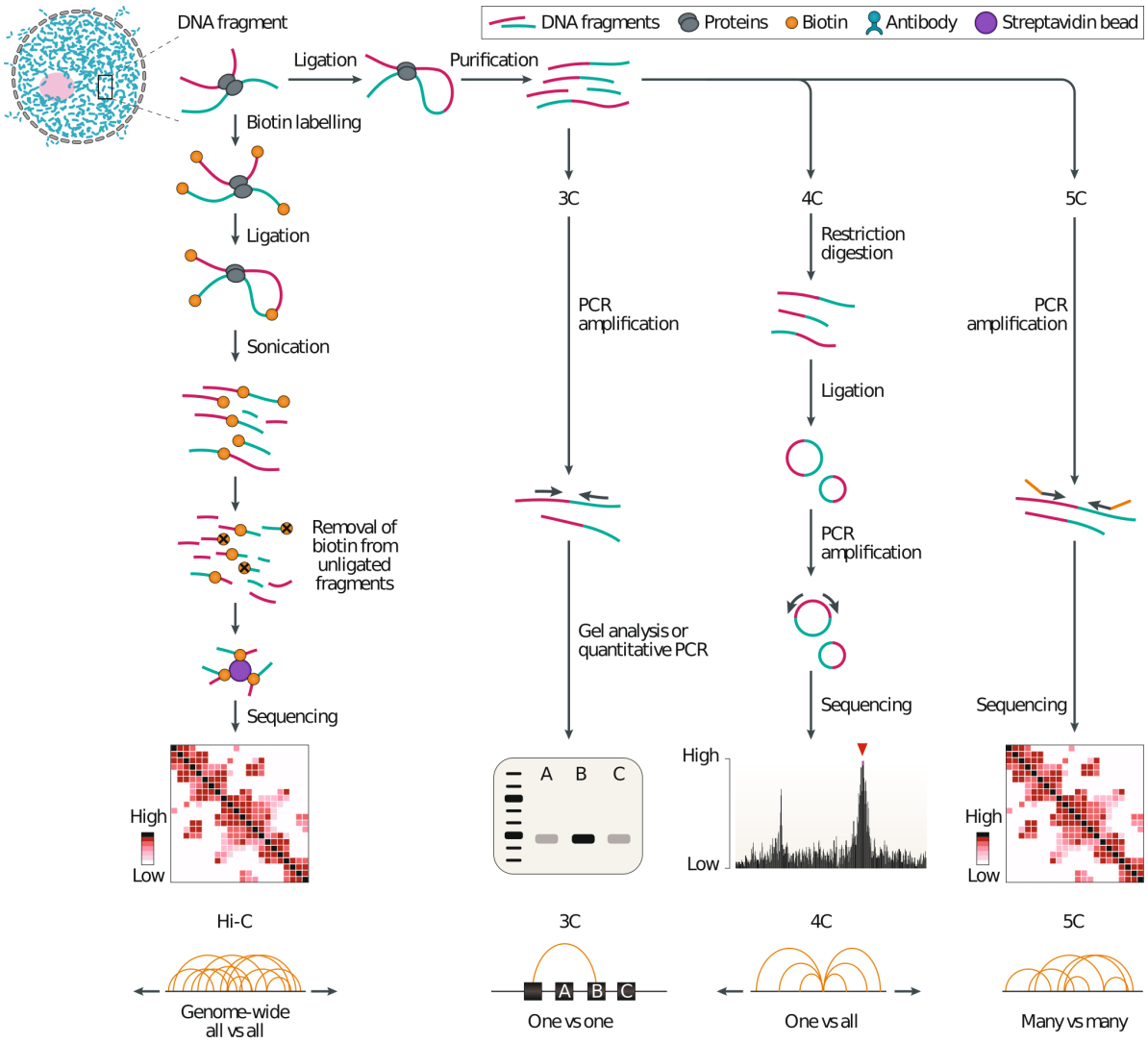

Figure 2.2.1: 3C-based methods. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Review Genetics, *Methods for mapping 3D chromosome architecture*, Kempfer R. & Pombo A., 2020[131].

2.2 Sequencing techniques

While microscopy-based techniques allow for the spatial localization of genomic loci, transcripts, or protein components, most sequencing methods used in the genome architecture field allow capturing the frequency of contacts between pairs of genomic loci.

2.2.1 Chromosome Conformation Capture techniques

Indeed, the chromosome conformation capture technique (3C) allows capturing the frequency of contacts between a genomic loci pair (i.e., one vs. one). Based on the same principles, the 4C technique allows capturing contacts between one genomic locus and the rest of the genome (i.e., one vs. all). Similarly, 5C allows capturing the frequency of contacts between two genomic loci groups (i.e., many vs. many). Finally, the Hi-C method measures the fre-

quency of contacts between all genomic loci and the rest of the genome (i.e., all vs. all)[34,142] (Figure 2.2.1).

With minor variations, all 3C-based techniques include the following steps: chromatin cross-linking, digestion, ligation, and reverse cross-linking. Briefly, chromatin is first held in place by a cross-linking agent, to be then digested and ligated to generate chimeric molecules, where genomic loci located far away in the linear sequence but close to each other in the 3D space can ligate. The final reverse cross-linking step is required to free the fragments then quantified by sequencing[142]. Later protocols skipped the cross-linking without hindering the efficiency of the assay[143].

The power of the 3C-based techniques lies in the number of cells that the assay can process simultaneously. Excluding polychimeric fragments, each digested product should ligate to up to two other fragments (one for each end) in a single cell. Thus, by collecting the ligated products of the same genomic locus of interest across many cells, a complete interaction network can be drawn[34].

Several reviews have been published, providing a thorough comparison and historical perspective on the various chromosome conformation capture techniques[142,144].

2.2.1.1 Hi-C

The development of Hi-C, published in 2009, represented a significant step towards building a genome-wide architecture model at high resolution.

As aforementioned, Hi-C enables the measurement of all genomic loci's contact frequency between one another at a specific resolution. The result is a matrix, where each column and row corresponds to a genomic region, and the value of a cell is the contact frequency of the corresponding column/row regions. Such matrix is symmetrical and characterized by a higher signal at the diagonal, which quickly weakens as one moves away from the diagonal, as regions proximal to one another on the linear genome should also be proximal in the 3D space. Hi-C contact frequency matrices are also referred to as Hi-C maps.

The first Hi-C-based study provided the community with a coarse map at approx. 1 Mb resolution, capturing only medium-to-large scale organizational features[34]. A few years later, another study presented Hi-C contact maps with a much higher resolution, in the range of few kilobases, shedding light on more minute characteristics of the genome architecture[35].

It is essential to note that, as aforementioned, up to this point, all 3C-family techniques could only be performed on cell populations. Thus, it was unclear whether Hi-C maps represented a mere statistical property of the cell population and not the architecture of single cells[130,145].

Several computational frameworks are available for analyzing Hi-C datasets - to mention a few: `Juicer`[62], `HiCPro`[146], and `FAN-C`[147]. As Hi-C analysis tools evolve rapidly, we can only direct the reader to a recent, but already not up to date, review of different computational methods written by Forcato et al[66].

2.2.1.2 Dip-C

In 2013, to address the bulk-assay limitation of Hi-C, Nagano et al. developed single-cell Hi-C. This technique allowed for the generation of single-cell contact maps of mouse T cells, which appeared to capture a limited number of contacts[148].

Recently, Stevens et al. combined single-cell Hi-C and imaging to validate their 3D genome reconstructions of embryonic mouse ESCs[32]. Both studies showed conservation of A/B compartments and LADs, but highly variable higher resolution structure, in the analyzed cells[32,148].

Building upon the published single-cell Hi-C, a recent study by Tan et al. presented diploid-chromosome conformation capture –or Dip-C[33]. Dip-C omits the biotin pull-down step of typical Hi-C protocols and replaces it with a multiplex end-tagging amplification (META) step. Interestingly, Dip-C appears to capture two orders of magnitude more contacts than standard single-cell Hi-C.

Initially, Tan et al. generated Dip-C maps of mouse ESCs, human lymphoblasts (GM12878), and human peripheral blood mononuclear cells (PBMCs) at a maximum resolution of 20 kb. Then, they combined these contact maps with a haplotype imputation algorithm to build 3D diploid genome reconstructions. Interestingly, the centromere-telomere arrangement in the 3D reconstruction correctly segregated the three samples into three separate clusters. Moreover, applying unsupervised clustering on PCA of single-cell chromatin compartments showed the different cell types of the PBMC sample clustering separately[33].

Altogether, these results showcase the power of single-cell Hi-C for high-resolution 3D genome reconstruction.

2.2.2 GAM

Genome Architecture Mapping –or GAM– instead, differs substantially from 3C-based techniques but yields similar results. In this ligation-free method, cryosectioning and laser dissection generate several ultrathin sections from each nucleus. This process repeats for many nuclei, followed by sequencing of the material extracted from each section. Then, a probabilistic model builds contact frequency maps, similar to Hi-C ones, based on the set of genomic loci identified in each section[149] (Figure 2.2.2a).

While this technique has the advantage of measuring contact frequency based on co-occurrence in ultrathin sections, as opposed to the two-way spatially constricted contacts captured by Hi-C, GAM is a rather laborious technique that, due to basic principle design, cannot be performed on single cells.

GAM was initially applied to mouse ESCs, revealing an enrichment for long-range interactions between distant active genes and enhancers. Furthermore, thanks to its digestion- and ligation-free nature, GAM allowed capturing abundant three-way genomic contacts, which eluded previous Hi-C-based studies[149].

Furthermore, GAM was recently combined with immunoselection, allowing for its appli-

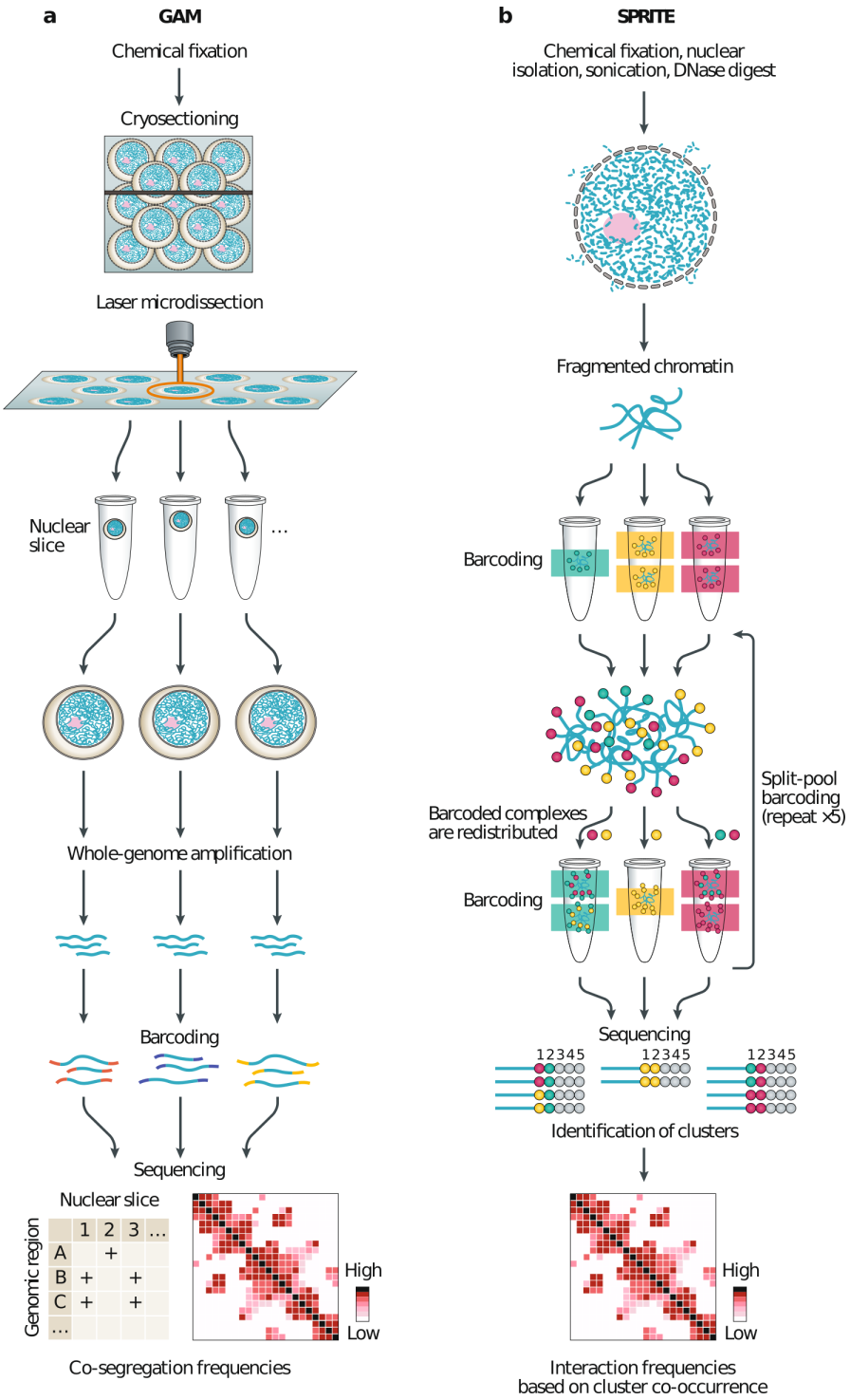

Figure 2.2.2: Schematic representation of the ligation-free GAM (a) and SPRITE (b) methods workflow. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Review Genetics, *Methods for mapping 3D chromosome architecture*, Kempfer R. & Pombo A., 2020[131].

cation in complex tissues and small numbers of specialized cells. This so-called immunoGAM allows detecting cell-type specific genome architecture features in juvenile and adult mouse brain tissue[150,151].

2.2.3 SPRITE

Another ligation-free method allowing for the detection of high-order chromatin contacts is split-pool recognition of interactions by tag extension –also known as SPRITE[152].

Specifically, SPRITE starts with relatively standard sample preparation, including a chemical fixation of the cells, followed by DNA extraction, sonication, and digestion with DNAse. The protocol proceeds by randomly separating the fragmented chromatin into separate aliquotes (split step), then labeled each with a different unique sequence (tag step). After pooling the aliquotes back together (pool step), the split-tag-pool process is repeated n times (typically three to five) (Figure 2.2.2a).

The SPRITE procedure allows to label with a unique tag sequences that belong to the same chromatin cluster, without the need for a ligation step. This crucial characteristic enables capturing long-distance chromatin contacts, usually missed by classical chromosome-conformation capture techniques.

In the study that introduced SPRITE, Quinodoz et al. applied their technique to mouse ESCs and human lymphoblasts (GM12878)[152]. This application revealed that SPRITE could capture the same results as Hi-C alongside higher-order interactions. Furthermore, by extending the method to measure DNA-RNA contacts, the researchers were able to identify two inter-chromosomal chromatin hubs that tend to contact nuclear speckles and the nucleolus, respectively[152].

2.2.4 DamID

Unlike the methods presented so far, DNA adenine methyltransferase identification (i.e., DamID) does not capture chromatin-chromatin contacts. Instead, DamID identifies genomic loci interacting with a protein of interest by exploiting eukaryotic cells' inability to methylate adenines[153].

By fusing a protein of interest with a prokaryotic DNA adenine methyltransferase (a.k.a., Dam), the fused product methylates adenines that interact with it. Then, the genome is enzymatically digested with the DpnI restriction enzyme, effectively digesting only GATC sequences with a methylated adenine. This process generates methylated DNA fragments that undergo library preparation, thus enabling identifying methylated GATC loci by sequencing[100] (Figure 2.2.3).

In its first application on human fibroblasts[153], DamID generated high-resolution maps of chromatin interaction with nuclear lamina components Lamin B1 and emerin. In this study, Guelen et al. analyzed such maps to identify lamina-associated domains (LADs), ranging in size approximately 100 kb to 10 Mb and with characteristically low transcription levels[153].

Furthermore, in 2015, Kind et al. applied DamID to single-cells[99] (scDamID). The application of scDamID on nine different human cell lines allowed the identification of constitutive LADs (cLADs, present in all cell lines) and facultative LADs (fLADs, interacting with the

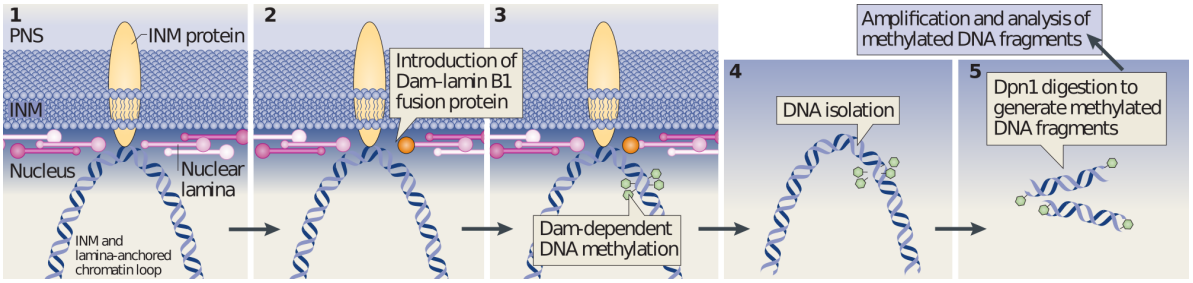

Figure 2.2.3: DamID-seq workflow. INM, inner nuclear membrane; PNS, perinuclear space. Adapted by permission from Springer Nature via Copyright Clearance Center, Inc: Springer Nature, Nature Reviews Molecular Cell Biology, *The nuclear lamins: Flexibility in function*, Burke B. & Stewart C.L., 2013[97].

lamina only in a subset of cell lines). Similarly, inter-LAD regions (iLADs) were categorized between constitutive (ciLADs) and facultative (fiLADs). Interestingly, constitutive LADs and iLAD represented 22.1% and 18.9% of the genome, respectively. Altogether, these results hint at the conservation of chromatin-lamina interactions during differentiation[99].

While being an effective technique, the need for transgenic samples expressing the protein of interest fused to Dam represents the main disadvantage of DamID. Moreover, the ability to generate methylated DNA fragments is currently limited to regions presenting the GATC restriction site of the methylation-sensitive endonuclease DpnI.

2.2.5 TSA-seq

Like DamID, tyramide signal amplification sequencing (TSA-seq) also allows identifying genomic regions interacting with a protein of interest instead of chromatin-chromatin interactions[154].

The TSA technique uses horseradish peroxidase (HRP) to generate tyramide free radicals that diffuse and react with various macromolecules, including DNA. Specifically, TSA-seq utilizes an antibody-coupled HRP to generate a biotin-tyramide free radicals concentration gradient starting from the antibody's target. Notably, the TSA-seq intensity of DNA labeling is a function of the distance between the genomic locus and the staining target. Thus, after TSA labeling, reverse cross-linking, DNA extraction, and affinity-based purification, sequencing allows to convert the obtained number of reads directly to a distance measure[154].

Chen et al. applied TSA-seq to targe the SON protein, a component of nuclear speckles, and lamin A and B, components of the nuclear lamina, on human K562 lymphoblasts. TSA-seq of lamin B1 correlated highly with DamID of lamin B1. Furthermore, Lamin A and B TSA-seq map anti-correlated with SON TSA-seq maps, indicating a more internal localization of nuclear speckles and their associated genomic loci[154].

Interestingly, a recent application of TSA-seq on four different human cell lines (H1, HFF, HCT116, K562) revealed a conserved genome organization around nuclear speckles. Specifically, only 10% of the genome showed significant changes in its position relative to nuclear

speckles, and such movements were consistently associated with changes in cell-type-specific transcriptional activity. Furthermore, expression amplification upon heat-shock protein H1 induction is associated with chromatin condensation, concomitant with speckle-association. Altogether, these results support the view of a conserved genome arrangement around nuclear speckles [155].

To be noted, the laboratory that introduced TSA-seq recently presented a combination of TSA and mass spectrometry (TSA-MS) to study the protein composition of nuclear speckles [156].

Part II

Doctoral thesis

Chapter 3

Research Aims

Briefly, this thesis aims to develop novel biomolecular methods for investigating the mammalian genome radial architecture from a quantitative and analytical perspective.

The specific aims of the constituent papers are:

Paper I

- To develop iFISH: a publicly available resource for the design of single and spotting DNA FISH probes

Paper II

- To develop GPSeq: a novel biomolecular assay for quantifying the radial localization of genomic loci in a genome-wide manner.

Chapter 4

Results

In this chapter, we discuss the development of GPSeq, an assay for the genome-wide measurement of genomic loci localization relative to the nuclear lamina, and iFISH, a computational and methodological resource for straightforward design and in-lab amplification of FISH probes. First, we describe the first steps of GPSeq development (section §4.1), specifically establishing a genome's concentric labeling protocol and analyzing the resulting microscopy images. Then, we review the process of designing a GPSeq-based score to estimate the radial localization of a genomic region. Subsequently, we describe GPSeq's validation (section §4.2) via DNA FISH, DamID-seq, and Hi-C. To this end, we also describe the development of iFISH (section 4.2.1.1): a crucial resource for designing the large number of DNA FISH probes used for GPSeq's validation. Afterward, we recount the genome architecture aspects revealed by GPSeq (section §4.3), both in terms of genetic and epigenetic patterns, and a simple approach towards predicting radial genome organization. We continue by exploring how GPSeq can improve the construction of Hi-C-based 3D genome structures (section §4.4), which we built with a novel computational tool, dubbed `chromflock`. Finally, we conclude this chapter with insights provided by GPSeq on the radial distribution of genomic single-point mutations and double-strand breaks (section §4.5).

4.1 GPSeq development

Genomic loci Positioning by Sequencing is a straightforward assay based on a simple concept: diffusion of particles from the outside of a closed environment should proceed towards its interior. Specifically, we hypothesized that subjecting cells to enzymatic restriction for different time lengths would result in the initial digestion of peripheral chromatin, with internal chromatin restricted only at longer time points.

In other words, this protocol would allow generating growing concentric waves of restriction (Figure 4.1.1a). Moreover, we argued that we could use sequencing-derived information to assign the restricted sites to each restriction wave (corresponding to a specific digestion duration and thus radial location). Our final aim was to combine the sequencing results into a

score function, which would ideally reflect the genomic locus's radial location.

For the development of GPSeq, we selected HAP1 chronic myeloid leukemia cells for their relatively rounded nucleus and haploid genome. Here, the rationale was that a diploid non-chimeric sample would not allow for allele annotation, causing any centrality estimation to be an average between the two alleles' locations. Also, a rounded nucleus is ideal for establishing the technique, as a more complex nuclear shape can result in less easily interpretable enzyme diffusion patterns.

To prove this concentric wave concept, we initially embarked on developing a method to visualize the digested genomic loci at different enzymatic restriction time lengths and characterize the resulting images after their imaging by fluorescence microscope.

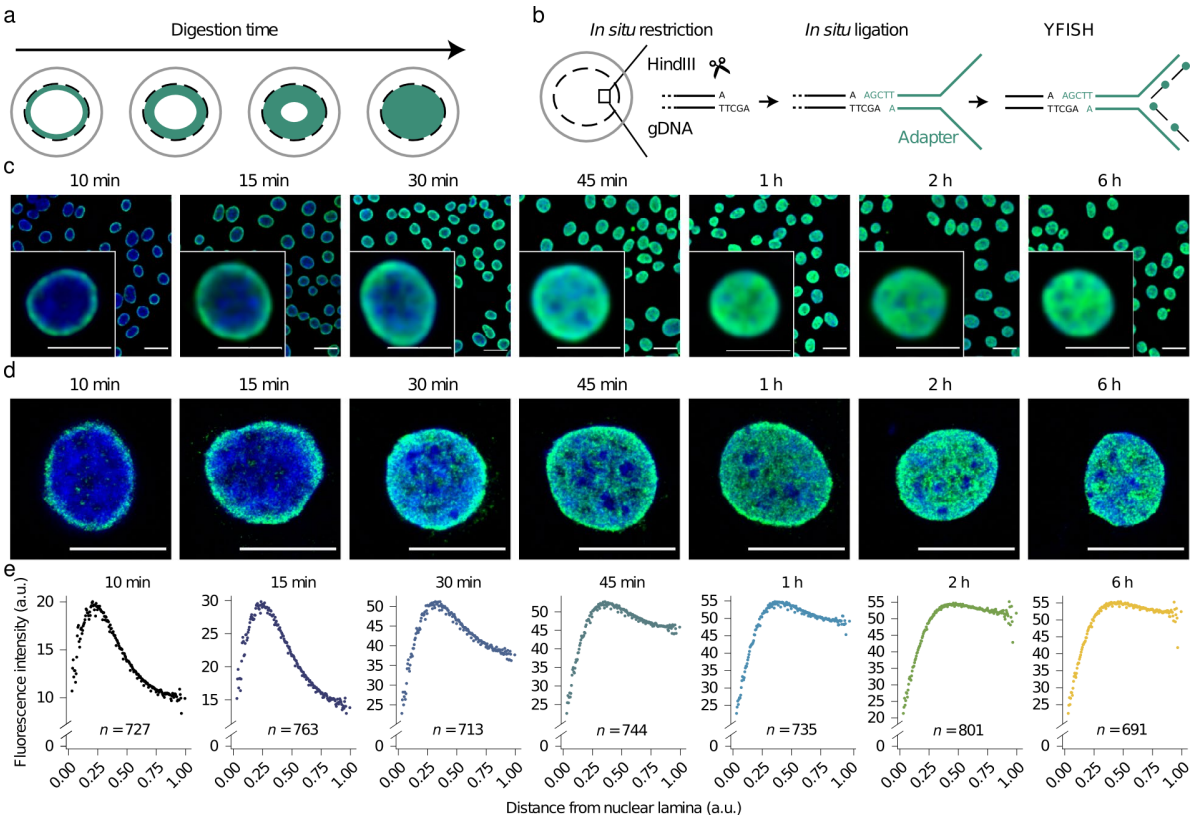

Figure 4.1.1: GPSeq YFISH proof of concept. (a) GPSeq concept of concentric waves of restrictions resulting from increasing time of digestion. (b) YFISH workflow, consisting of *in situ* adapter ligation and FISH. (c) Midsection of selected field of views acquired by widefield microscopy. YFISH signal and DNA staining by Hoechst 33342 correspond, respectively, to the green and blue channels. Inset show zoomed-in single nuclei from each respective field of view. Scale bars correspond to 20 µm and 10 µm for the fields and insets, respectively. Dynamic range was separately set for each condition to allow better visualization of the corresponding wave signal. (d) Same as c but acquired by STED microscopy. (e) Median radial YFISH signal profiles calculated with `pygpseq` for each digestion time (section 5.2.4). The number of nuclei analyzed is reported as n. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

4.1.1 YFISH and image analysis

To visualize restricted genome sites, we implemented YFISH: a novel fluorescence *in situ* hybridization method based on a Y-shaped double-stranded DNA construct (YFISH adapter). The YFISH adapter comprises three parts: (1) sticky-hands complementary with the restricted sites, allowing for adapter ligation; (2) a double-stranded DNA central region, essential for assembling the adapter; and (3) two non-complementary single-stranded flaps (one with 3'-5' and one with 5'-3' orientation), which constitute the acceptors for fluorescently-labeled oligonucleotides (Figure 4.1.1b).

We prepared the cell samples based on an adapted version of the 3D-FISH protocol (section 5.1.2), followed by genome restriction, and YFISH adapter ligation (section 5.1.4.1). Afterward, we hybridized fluorescently labeled oligonucleotides to the single-stranded portion of the YFISH adapters via FISH. Finally, we visualized and imaged the samples at the microscope (section 5.1.4.2).

The first GPSeq-YFISH experiment was performed with the 6-base cutter HindIII and consisted of six conditions, corresponding to the following restriction time lengths: 10 min, 15 min, 30 min, 45 min, 1 h, 2 h, and 6 h. It is important to note that we always performed image acquisition with the same optical configuration and with a dynamic range set on the longest restriction condition to obtain images with comparable fluorescence intensity values across different conditions.

Visual inspection confirmed YFISH signal waves' presence, growing from the nuclear periphery towards its interior with growing digestion time lengths (Figure 4.1.1c-d). Specifically, a sharp and bright crown-like signal enriched at the nuclear periphery was already visible at 10 min of digestion. After two hours, saturation appeared to be reached, with longer digestion not contributing to the signal wave's further growth.

To provide an appropriate quantitative analysis of the YFISH signal waves, we developed a Python 3 package called `pygpseq`. Specifically, `pygpseq` provides full-stack tools for image analysis, including conversion from microscopy-proprietary format to TIFF, identification, and removal of out-of-focus fields of view, calculation of cell population median radial fluorescence intensity profiles, and generation of a summary report in PDF format (see section 5.2.4 for more details). Quantification of YFISH signal radial profile, calculated over the whole sampled cell population, confirmed our observations (Figure 4.1.1e), with the profile's peak growing in size until reaching saturation after around 2 h of digestion.

Furthermore, we expanded the `pygpseq` package with *ad-hoc* scripts to perform single-cell and single-radii characterization of the YFISH signal profiles (section 5.2.4). Briefly, for each profile, we calculated the intensity and the position relative to the lamina of the profile peak and inflection point. We interpreted the profile peak and inflection point positions as indicators of enzyme penetration within the nuclear space. We also defined a "contrast" measure as the ratio of the intensity at peak and inflection point, respectively, which we expected to decrease as saturation is approached, with increasing digestion times (Figure 4.1.2a). All three

profile features appeared to be relatively consistent across nuclei for each digestion condition, as revealed by narrow interquartile ranges (Figure 4.1.2b-d). Moreover, the contrast decreased with increasing digestion times, as expected, reaching saturation at 2 h (Figure 4.1.2d). Furthermore, the similarity between the 45 min and 1 h conditions suggested discarding either to achieve a more straightforward experimental setup including only five conditions, specifically: 10 min, 15 min, 30 min, 1 h, and 2 h.

In the case of single-radii characterization, we extracted each nucleus and randomly drew 200 radii that distributed them homogeneously in the 3D space from the nuclear volume centroid to the nuclear-fitted surface. Then, we interpolated each radius's intensity profile from 100 points homogeneously distributed along the radius itself (Figure 4.1.2e). For each nucleus, the calculated standard deviation of the three profile features appeared to be relatively

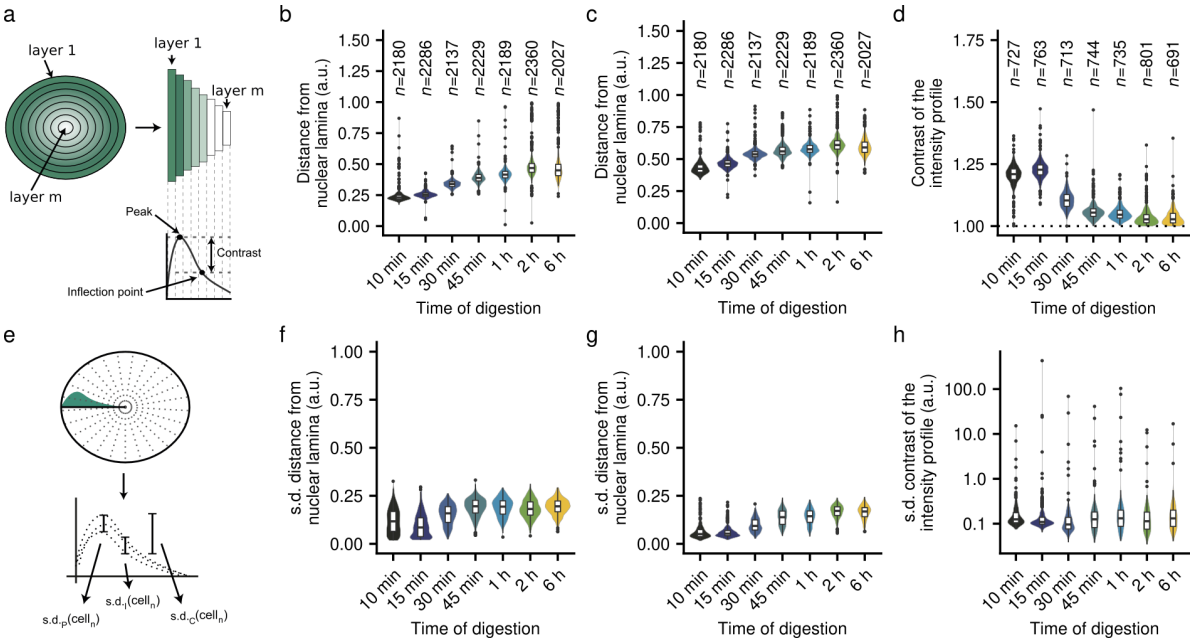

Figure 4.1.2: GPSeq YFISH characterization at single-cell and single-radius level. (a) Single-nucleus analysis schema. Each nucleus is divided into m layers (typically, m=200). For each layer, all voxels are identified, and their median intensity is calculated. The median values along the nuclear radius are then smoothed with a gaussian filter, and the peak and inflection point location of the obtained curve are calculated. Furthermore, we calculate the profile contrast as the ratio between the intensity at the peak and the inflection point. (b) Distribution of single-nucleus profile peak position, relative to the nuclear lamina, for different digestion times. (c) Distribution of single-nucleus profile inflection point position, relative to the nuclear lamina, for different digestion times. (d) Distribution of single-nucleus profile contrast for different digestion times. (e) Single-radius analysis schema. For each nucleus, 200 radii are randomly drawn from the nuclear volume centroid to the fitted nuclear surface. Radii are drawn in a random manner that makes them homogeneously distributed in the 3D space. Then, 100 points are homogeneously distributed along each radius, and the intensity at each point is interpolated from the YFISH signal channel. The intensity at the points is used to build a radial profile for each radius, where peak, inflection point, and contrast are calculated as described in a. (f) Distribution of the standard deviation of single-radius profile peak relative distance from the nuclear lamina. One data point corresponds to one nucleus. (g) Same as f, but for the inflection point. (h) Same as f, but for the profile contrast. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

low, consistent with homogeneous profiles at different radial directions (Figure 4.1.2f-h).

4.1.2 GPSeq - sequencing setup and centrality estimation

After confirming that the GPSeq protocol can indeed produce the expected concentric digestion waves, we proceeded with ligating an NGS-compatible adapter to the restricted samples, instead of the YFISH adapter. The GPSeq sequencing adapter comprises five parts: (1) the sticky ends necessary for ligation to the restricted sites, (2) a condition-specific barcode, (3) a unique molecular identifier (UMI) used for amplification product removal, (4) Illumina RA5 adaptor, and (5) T7 promoter to support *in vitro* transcription (IVT). After adapter ligation, the genomic DNA is extracted, fragmented by sonication, and then amplified by IVT. Finally, the libraries are prepared and loaded at the same concentration in the sequencer (see section 5.1.4.3 for more details, Figure 4.1.3).

Typically, for each GPSeq experiment, all conditions are performed in duplicates, in parallel. One set of samples is reserved for YFISH, while the other to produce sequencing libraries. The ability to image the samples allows avoiding sequencing of experiments where the proper radial profile is not apparent in the corresponding YFISH samples and is especially useful while setting up the procedure.

We performed an additional replicate experiment of the HindIII HAP1 experiment, with the same conditions: 10 min, 15 min, 30 min, 1 h, and 2 h. Moreover, we performed two replicate experiments on the HAP1 cell line using the 4-base cutter MboI enzyme instead. These two additional experiments required new YFISH and sequencing adapters with the sticky hands of the MboI enzyme. Moreover, YFISH revealed that saturation is reached at shorter times for the MboI enzyme, and the conditions used for these replicates 1 min, 5 min, 10 min, and 30 min, instead.

After sequencing, we developed a pre-processing bash-based pipeline (containing some Python and R script), namely `gpseq-seq-gg`, that takes as input the sequencer output (i.e., fastq files) and calculates read counts per restriction site. Briefly, the pre-processing included: quality control report generation, selection of reads with matching prefix and prefix trimming, mapping to hg19 reference genome, removal of low-quality alignments, selection of reads mapped near known restriction sites, removal of reads with low-quality UMIs, and final UMI- and location-based de-duplication (see section 5.2.5 for more details). The pipeline's output is a bed file per digestion time of a GPSeq experiment, with each file containing the number of de-duplicated reads mapped to each restriction site.

We then devised five potential centrality estimation scores*, divided into two categories: respectively probability- and variability-based. Probability-based estimates were formulated on the assumption that internal genomic regions should have a higher probability of being restricted at longer digestion times. On the other hand, the rationale behind variability-based

*Here, we provide only a brief summary of the GPSeq score calculation and on the different devised and tested estimates. For more details, we direct the reader to the Supplementary Note I of Paper II[2].

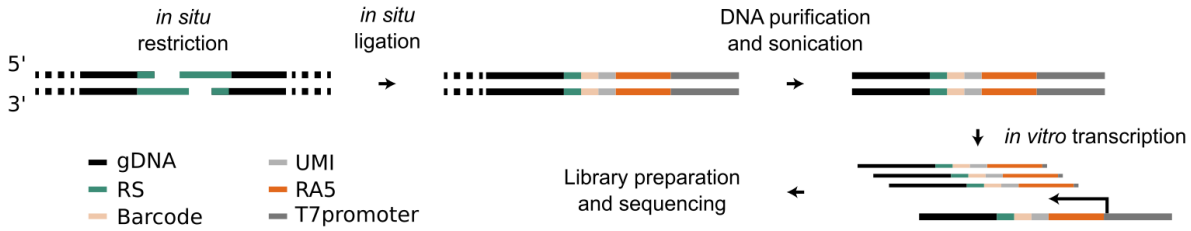

Figure 4.1.3: GPSeq sequencing workflow. After ligation, genomic DNA is purified and sonication. Then, *in vitro* transcription is performed, followed by library preparation and sequencing. gDNA: genomic DNA. RS: restriction site. UMI: Unique Molecular Identifier. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

estimates was that peripheral regions should be fully digested at any digestion time, resulting in a stable and low variability in the number of reads across their restriction sites, while more internal regions would show a higher read count variability. Specifically, we devised two probability-based estimates, a restriction-probability metric (P_S) and a cumulative restriction-probability metric (P_{Sc}), and three variability-based estimates, based on variance (V), coefficient of variation (C_v) and dispersion (C_d), respectively.

Additionally, we devised three different ways to combine the GPSeq conditions when calculating the estimates. Briefly, the "two-point" approach would utilize only the shortest and longest digestion conditions, the "fixed" approach would compare each condition with the shortest digestion, and the "adjacent" approach would compare each condition with the immediately shorter digestion time.

To calculate the different centrality estimate candidates, we developed a Python package, called `gpseqc`, that would take the bed files generated by the pre-processing pipeline as input, and generate tables with the estimated scores. Specifically, `gpseqc` supports centrality estimate calculation in a chromosome-wide manner, genome-wide with customizable binning size and step, or for a set of specific regions. Furthermore, several advanced features are included, like input restriction site outlier removal, masking, and restriction site domain selection (see section 5.2.6 for more details).

4.2 The GPSeq score and its validation

Next, we evaluated how well the different score candidates estimate centrality, select the best score, and then validate the results. We decided to perform score evaluation and selection based on its correlation with the centrality measured by FISH. Ideally, the best score would outperform all other candidates and be highly reproducible across replicates, at both low and high resolution. Then, we proceeded with further validation by comparing the selected score with results obtained by two different sequencing-based techniques: DamID-seq of Lamin B and Hi-C.

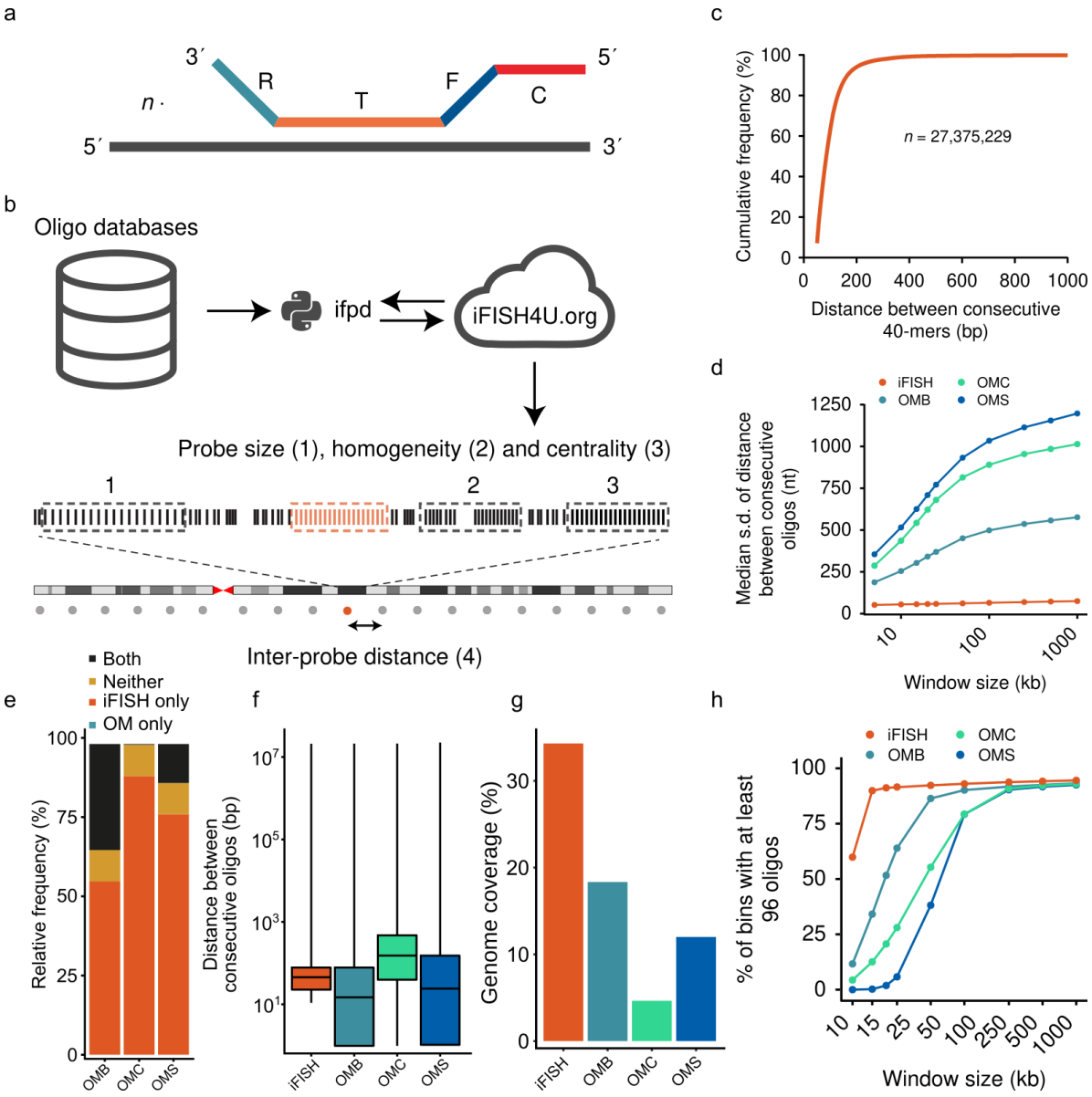

Figure 4.2.1: iFISH design and database comparison. (a) Schema of iFISH probe. n: number of oligonucleotides per probe. R: reverse primer sequence. F: forward primer sequence. C: color sequence. (b) iFISH probe design workflow. The ifpd package reads from the oligonucleotide database and serves to a web-based GUI upon query. The GUI then provides the user with the resulting probe design. (c) Cumulative frequency of distance between consecutive oligonucleotides of the novel database designed with `oligo-picker`. (d) Median distance between consecutive oligonucleotides, calculated across genomic windows of different sizes, for the iFISH and `OligoMiner` (OM) -based databases. (e) Relative frequency of 15 kb genomic windows with at least 96 oligonucleotides only in the `OligoMiner`-based database, only in the iFISH database, in neither, or both. (f) Distribution of linear distance between consecutive oligonucleotides in the four considered databases. (g) Overall genomic coverage of the four considered databases. (h) Percentage of bins with at least 96 oligonucleotides, at different window sizes, for the four considered databases. OMB: `OligoMiner` 'Balance'. OMC: `OligoMiner` 'Coverage'. OMS: `OligoMiner` 'Stringent'. Adapted from Gelali E. & Girelli G., et al., Nature Communications, 2019 (Paper I[1]).

4.2.1 FISH-based evaluation and score selection

We planned to run `gpseqc` to calculate each score candidate for score evaluation at different resolutions, in bins centered around several DNA FISH labeled regions distributed across different chromosomes. Since no tool was available to design FISH probes taking into account features like probe size or homogeneity of oligonucleotide distribution, we focused on developing a DNA FISH resource for probe design and amplification.

4.2.1.1 iFISH: homologous and orthogonal sequence design

We based our oligonucleotide-based iFISH probes on the uniFISH concept we had recently published[135]. Each iFISH probe is composed of several oligonucleotides, each comprising four sequences: a target (T) sequence complementary to the target genomic locus, two PCR-compatible flaps at the 3' and 5' end of the target sequence (F: forward, and R: reverse), and an additional flap (C: color) which acts as an acceptor for fluorescently-labeled oligonucleotides (Figure 4.2.1a). The F, R, and C sequences are orthogonal to the genome of interest to avoid any hybridization that would lead to non-specific signals.

First, we searched the literature for public databases of target sequences. The only publicly available oligonucleotide databases compatible with DNA FISH were designed via `OligoMiner`[139], which provided a relatively low and sparse genome coverage. We implemented a perl-based pipeline, named `oligo-picker`, to design a novel oligonucleotide database to solve this issue. Briefly, the pipeline includes the following steps: identification of unique k-mers[†] from a reference genome by using `JELLYFISH`[157], retrieval of k-mer genomic coordinates with `VMATCH`[158], discarding k-mers with homopolymers, and filtering of k-mers based on GC-content, secondary structure stability (estimated with `OligoArrayAux`[138]), and melting temperature. The pipeline would then generate an SQL database containing the sequence and details of the oligonucleotides passing all the filters (see section 5.2.1 for more details).

We compared the database of 40-mers designed with `oligo-picker` to previously published `OligoMiner` databases. The iFISH database outperformed the `OligoMiner` databases on all the features we evaluated. Specifically, it showed a higher genomic coverage (Figure 4.2.1g), a lower and more consistent distance between consecutive oligos (Figure 4.2.1c-d,f), and a higher fraction of genomic regions (at different resolutions) with at least 96 FISH-suitable oligos (Figure 4.2.1h). Moreover, a comparison of the 15 kb windows containing at least 96 oligonucleotides revealed that the `OligoMiner` databases did not provide any window that was not similarly covered by iFISH. Vice-versa, the iFISH databases provided for at least 50% more windows than any other database (Figure 4.2.1e).

To design the orthogonal 20-mer flap sequences, we developed `ood-fish`, a pipeline using scripts implemented in various languages (R, bash, C) that initially aligns candidate se-

[†]With "k-mer" we indicate an oligonucleotide consisting of k nucleotides.

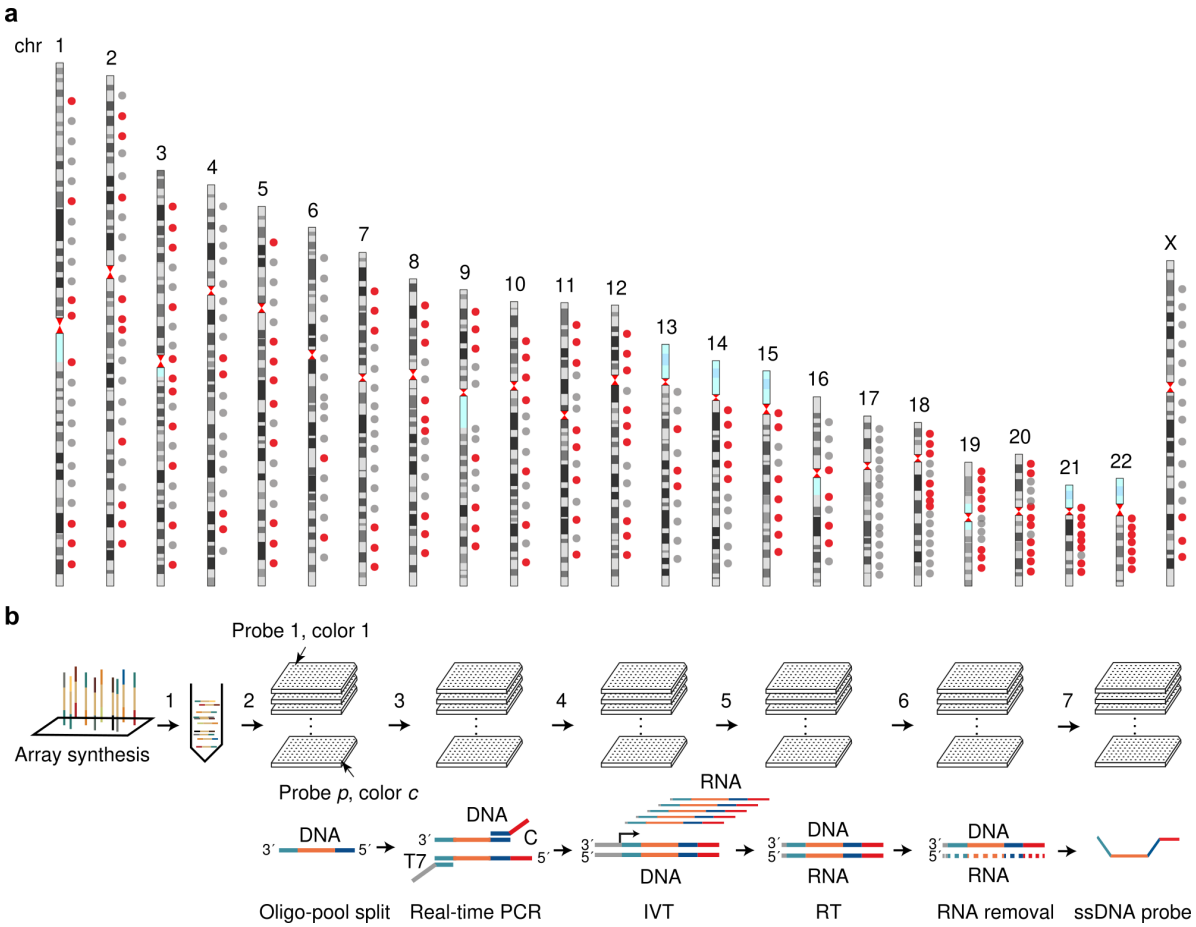

Figure 4.2.2: iFISH probes and amplification. (a) Ideogram with the location of 330 designed iFISH human chromosome spotting probes (circles). Probes labeled in red were individually amplified and tested. (b) Schematic workflow of iFISH probe amplification starting from an oligopool. PCR: Polymerase Chain Reaction. IVT: *In Vitro* Transcription. RT: RetroTranscription. ssDNA: single-stranded DNA. Adapted from Gelali E. & Girelli G., et al., Nature Communications, 2019 (Paper I[1]).

quences to a reference genome using BLAT[159], and then applies a homology filter. Afterward, `ood-fish` calculates the remaining k-mers self-dimerization delta free energy and discards those that can result in stable self-dimers. Finally, the pipeline calculates hetero-dimerization delta free energy of all remaining sequence pairs and identifies the largest set of k-mers that do not form stable hetero-dimers by applying an implementation of the Parallel Maximum Clique algorithm[160]. We applied `ood-fish` to the 20-mers extracted from a previously published list of 240,000 25-mers orthogonal to mammalian genomes[161] and used the generated sequences as F, P, and C flaps (see section 5.2.2 for more details).

4.2.1.2 iFISH probe design

Next, we developed a Python package that would read the database generated by `oligo-picker` and allow us to select oligonucleotides for probe design, either through a web-interface or the command line (Figure 4.2.1b). Following this design, we developed "iFISH-probe-designer," or `ifpd`, a Python package that allows designing single or multiple homo-

geneously spread (i.e., spotting) probes in a genomic region of interest. Moreover, `ifpd` can serve a web based graphical user interface based on `bottle` python web framework.

Briefly, `ifpd` performs single probe design by (1) reading all oligonucleotides in the region of interest from a database of non-overlapping sequences, (2) dividing the oligos into groups of N consecutive oligos, also known as probe candidates, (3) calculating three features for each probe candidate, specifically size, homogeneity, and centrality, (4) selecting candidates in a range around the optimal value of one of the three features, and (5) generating a list of remaining probe candidates ranked by a second feature. We defined probe size as the linear distance between the first and the last genomic loci covered by a candidate's oligos, homogeneity as the reciprocal of the standard deviation of the distance between consecutive oligonucleotides, and centrality as the relative position of the probe midpoint between the boundaries of the region of interest and its midpoint (Figure 4.2.1b). Specifically, `ifpd` will consider optimal the minimal probe size, the maximal probe size, and the maximal homogeneity. The user can select which of the three features is used for the initial probe candidate selection (step 3) and for the candidate ranking (step 4).

When designing N probes, `ifpd` subdivides the region of interest into $N + 1$ windows, and runs the single-probe design workflow on each window. The resulting probes are then assembled into a probe set candidate. Then, the windows are shifted by a fraction of the window size, and the process is iterated until completion. Finally, the generated probe set candidates are ranked based on homogeneity of probe size and inter-probe distance.

With `ifpd`, we designed a total of 330 probes for human chromosome spotting, with an inter-probe distance of approximately 10 Mb for chromosomes 1 to 16 and X, and approximately 5 Mb for chromosomes 17 to 22 (Figure 4.2.2a). Each probe consisted of 96 target sequences with probe-specific forward (F) and reverse (R) 20-mer flaps.

4.2.1.3 Probe amplification and hybridization

We continued by establishing a large-scale parallel iFISH probe amplification protocol, based on previously published workflows for Oligopaints and MERFISH probe amplification[126,140,162]. Briefly, the desired probe oligos are exponentially amplified from an oligopool using flap-specific (F-R) PCR, which also incorporates a T7 promoter and C-flap acceptor sequences to the PCR products. Then, the PCR products are linearly amplified via *in vitro* transcription, followed by reverse transcription and RNA digestion, to produce single-stranded DNA iFISH probes (Figure 4.2.2b, see section 5.1.3.1 for more details).

We then proceeded to individually amplify 153 out of the 330 (46%) spotting probes, and tested them on HAP1 cells using different fluorophores (Figure 4.2.3a, see section 5.1.3 for the entire protocol). Specifically, our analysis revealed a highly homogeneous number of FISH dots per nucleus, peaking at 1-2 dots/nucleus, and a probe-independent signal-to-noise ratio which was instead correlated with the fluorophore being used (Paper I, Figure 2).

After validating our novel 330 probes repository, we moved to showcase the power and

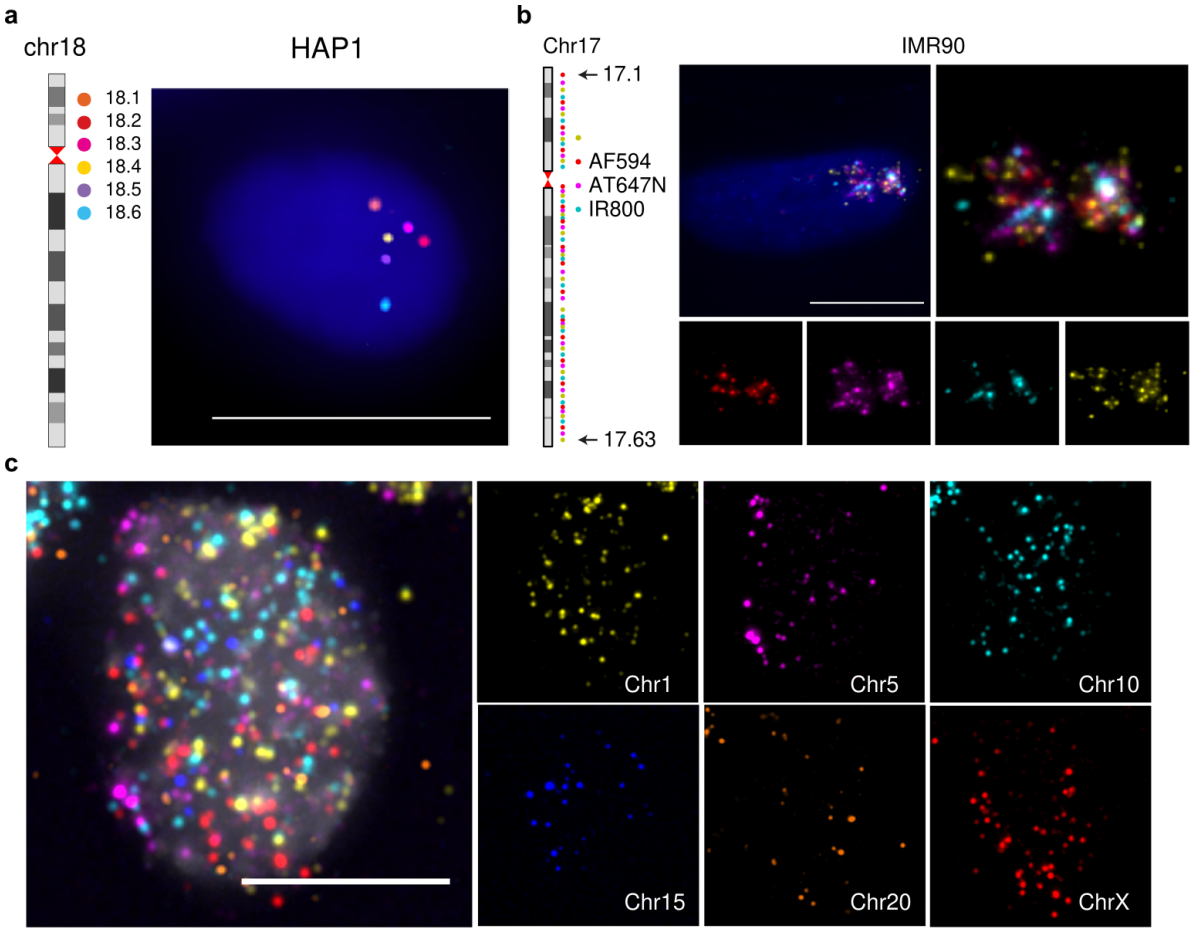

Figure 4.2.3: iFISH applications. (a) Individual iFISH probe validation in HAP1 cells. In this example, the first six chromosome 18 probes (18.1-18.6) are hybridized with different fluorophores. Left: six probes' locations on the linear genome are shown on an ideogram. Right: the mid-section of a HAP1 nucleus after hybridization is reported. (b) The spotting of chromosome 17 in human IMR90 fibroblasts using a set of 63 iFISH probes with an average distance of ~1.25 Mb. Top-left: mid-section composite showing DNA staining channel (blue) and FISH dots. Top-right: composite inset of the labeled chromosome showing the FISH signals. Bottom: inset showing the FISH signal, separated by channel. (c) Multiple chromosome spotting on human ESCs. Left: composite panel showing the DNA staining channel (gray) and FISH dots. Right: each panel shows the FISH signals from a separate channel, each corresponding to a different chromosome. Scale bars: 10 μm. Adapted from Gelali E. & Girelli G., et al., Nature Communications, 2019 (Paper I[1]).

versatility of iFISH in different applications. For example, we visualized chromosome 17 in human IMR90 fibroblast cells with probes designed to have alternating colors along the linear genome (Figure 4.2.3b). Specifically, we designed an additional 46 probes and hybridized a full set of 63 probes, with a mean inter-probe distance of ~1.25 Mb. Convex hull-based chromosome volume estimation using subsets of the hybridized probes revealed that using probes separated by more than 2 Mb results in a clear volume underestimation, highlighting the importance of FISH sampling frequency when investigating volumes of chromatin regions (Paper I, Figure 3c). Furthermore, we performed simultaneous spotting of chromosomes 1, 5, 10, 15, 20, and X in human embryonic stem cells, revealing an apparent lack of chromosome territories in a subset of such undifferentiated cells (Figure 4.2.3c).

4.2.1.4 GPSeq score evaluation by iFISH

To evaluate the formulated centrality score candidates, we selected 68 iFISH probes, distributed across 11 chromosomes (Figure 4.2.4a), and measured the distance from the nuclear lamina of the FISH signal generated by each probe. We then used `gpseqc` to calculate, at different resolutions, the centrality values estimated by the different scores for windows centered on the probes midpoints. As input to the pipeline, we used the pre–processed sequencing data obtained from the first HindIII HAP1 experiment (Exp.1), and calculated the Pearson Correlation Coefficient (PCC) between the scores and the FISH-based measurement at different

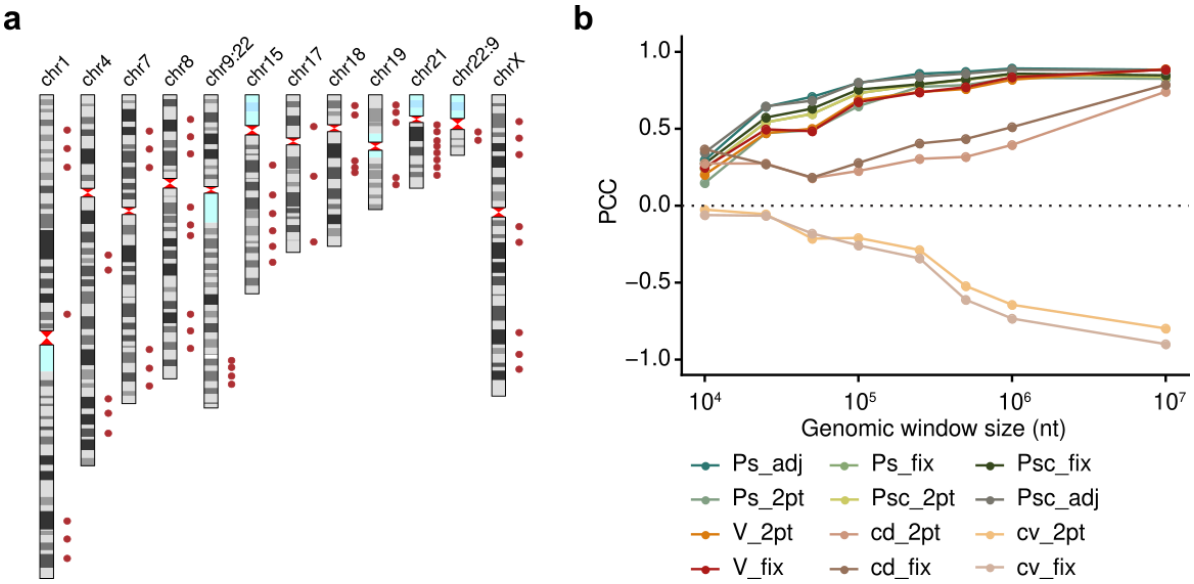

Figure 4.2.4: GPSeq score evaluation by iFISH. (a) The ideograms show the location of 68 iFISH probes selected for GPSeq score evaluation and validation. (b) Pearson Correlation Coefficient (PCC) calculated between the median distance from nuclear lamina measured by iFISH, and the corresponding GPSeq-based centrality estimate, calculated on differently-sized windows centered on the iFISH probes midpoints. Ps: probability-based score. Psc: cumulative probability-based score. V: variance-based score. cd: coefficient of dispersion-based score. cv: coefficient of variation-based score. 2pt: two-point condition combination. fix: fixed condition combination. adj: adjacent condition combination. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

resolutions. Ideally, some score formulations would be suitable for centrality estimation, with one score clearly outperforming the others.

All formulations, except those based on the coefficient of dispersion and coefficient of variation, provided a robust centrality estimation, showing a PCC higher than 0.5 for any resolution between 25 kb and 10 Mb. As expected, the correlation between centrality estimates and FISH-based measurements deteriorated with higher resolutions.

From this analysis, we selected the P_S "adjacent" score formulation as centrality estimate (C, for Centrality). Thus, the centrality of a genomic window (w) is calculated as the sum of ratio between its restriction probability (P_S) in a condition (D_j), and the probability in the immediately shorter condition (D_{j-1}, Figure 4.2.5a). In other words:

$$C(w) = \sum_{j=2}^{n} \frac{P_S(w, D_j)}{P_S(w, D_{j-1})}$$

Where we define the restriction probability of a genomic window (w) in a condition (D_j) as its number of reads ($N_R(w, D_j)$) normalized by the condition's library size ($N_R(D_j)$) and by the number of restriction sites in that window ($N_S(w, D_j)$):

$$P_S(w, D_j) = \frac{N_R(w, D_j)}{N_R(D_j) \times N_S(w, D_j)}$$

When `gpseqc` calculates C, it also provides a rescaled centrality (C_r). This is achieved by identifying outliers (W_o) as follows:

$$W_o = \{w | \, (Q_1(C) - C(w) > 1.5 \times \text{IQR}(C)) \\ \vee \, (C(w) - Q_3(C) > 1.5 \times \text{IQR}(C))\}$$

Where $Q_1(C)$ and $Q_3(C)$ indicate the first and third centrality quartiles, respectively, and IQR(C) indicates the centrality inter-quartile range (IQR). Then, centrality is first rescaled to an intermediate C_s value in such a way that all non-outlier bins span a range from 0 to 1, in other words:

$$(0 \leq C_s(w) \leq 1 \forall w \notin W_o) \wedge (C_s(w) < 0 \vee C_s(w) > 1 \forall w \in W_o)$$

Then, the rescaled centrality C_r is calculated as $C_r(w) = 2^{C_s(w)}$. In other words, the rescaled centrality C_r, which we referred to as "GPSeq score", will satisfy the following:

$$C_r \in (0, \infty]$$

$$(1 \leq C_r(w) \leq 2 \forall w \notin W_o) \wedge (0 < C_s(w) < 1 \vee C_r(w) > 1 \forall w \in W_o)$$

This means that C_r can only assume positive values, with non-outlier bins in the range from 1 to 2, extremes included. The rescaled centrality is useful for comparing centrality across samples and to avoid negative values. When negative values are not an issue, one can plot $C_s = \log_2(C_r) = \log_2(\text{GPSeq score})$. The GPSeq score can thus be interpreted as a centrality estimate where low values are indicative of peripheral regions, while high score values are characteristic of internal genomic regions.

GPSeq score calculated at a resolution of 1 Mb from experiment 1 showed high agreement with centrality measured by FISH (PCC: 0.909 and SCC: 0.920; Figure 4.2.5b). A comparable

correlation with FISH is achieved at an even higher resolution (100 kb) when averaging the GPSeq score across the four HAP1 experiments (PCC: 0.913 and SCC: 0.910; Figure 4.2.5b).

As expected, GPSeq experiments performed using a 4-base cutter (Exp.3-4) showed overall higher correlations with FISH even at low resolutions (e.g., 10 kb) when compared to 6-base cutter experiments (Exp.1-2). Moreover, a high depth 4-base cutter experiment (Exp.4) shows higher agreement with FISH than a replicate sequenced at lower depth (Exp.3, Figure 4.2.5d)[‡]. Moreover, GPSeq appears to be highly reproducible, with correlations higher than 0.8 at 100 kb and 1 Mb resolution for HindIII experiments (Figure 4.2.5e-f).

4.2.2 Validation with sequencing techniques

To further validate GPSeq, we decided to compare the GPSeq score with two orthogonal sequencing-based techniques: DamID-seq of Lamin B, and Hi-C.

Specifically, one would expect a correspondence between genomic regions marked by high DamID-seq signal/control values and regions with low GPSeq scores (i.e., peripheral). Indeed, DamID revealed a higher interaction with Lamin B for genomic regions with a low GPSeq score (peripheral) when compared to more internal regions (Figure 4.2.5g). In other words, LADs appeared to be more peripheral than inter-LADs (iLADs). Additionally, constitutive LADs were found to be more peripheral than facultative ones, while constitutive iLADs were more internal than facultative ones (Figure 4.2.5h)[99].

Interestingly, DamID-seq signal/control values did not show a steep drop from the nuclear periphery towards its interior but rather a gradual monotonically decreasing linear pattern. This can be explained by the bulk (as opposed to single-cell) nature of the DamID data we utilized, as the outcome of the assay is rather a probability, or frequency, of chromatin-lamina interaction among the cells of the sampled population. Moreover, the observed linearity was present only in the first half of the GPSeq score range, hinting at a limited ability of DamID-seq to measure radiality. Furthermore, when calculated in a chromosome-wise manner, we revealed a negative correlation (typically between -0.4 and -0.8) between GPSeq score and DamID-seq signal/control values (for more details, see Paper II, Supplementary Figure 3).

In terms of Hi-C, one could expect a decreasing Hi-C-captured contacts count between genomic regions with increasingly different GPSeq score; the rationale being that the regions with similar radial localization might be in proximity or at the opposite sides of the nucleus, while regions with different radial localization should not be in proximity. Indeed, Hi-C contact count decreases with increasing delta GPSeq score. Importantly, this pattern holds true not only for bins that are in closer linear proximity (adjacent regions) but also when considering all intra-chromosomal contacts or when moving to inter-chromosomal contacts, although being less pronounced (Figure 4.2.5i).

[‡]See Paper II Supplementary Note II for a discussion on how sequencing depth, enzyme choice, and the number of digestion time points can affect the GPSeq score.

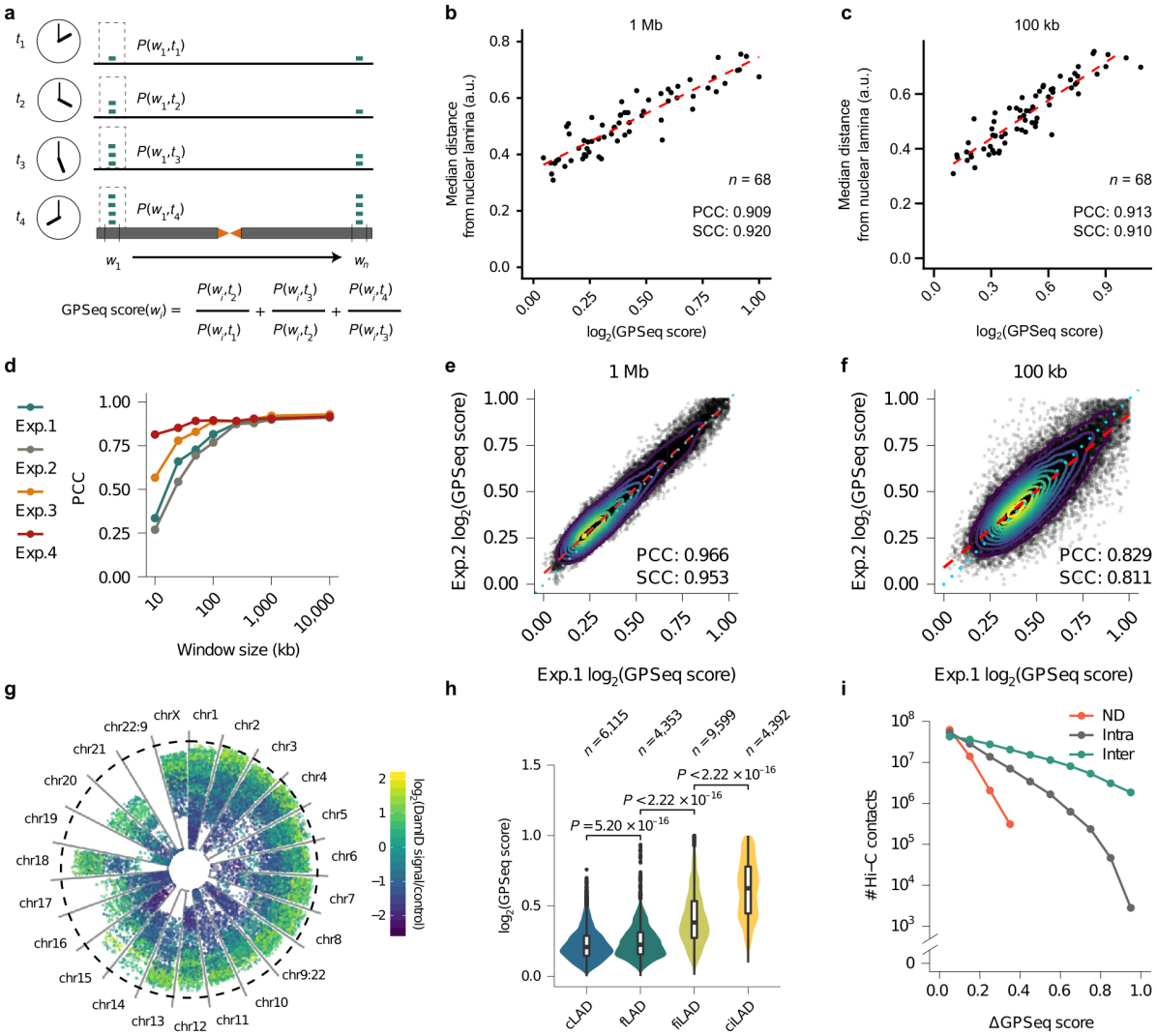

Figure 4.2.5: GPSeq reproducibility and validation. (a) Schematic representation of GPSeq score calculation. (b) Correlation between GPSeq score, calculated on 1 Mb windows, and median distance from the nuclear lamina, as measured by FISH. n: number of FISH probes analyzed. (c) Same as b but with GPSeq score averaged across all HAP1 experiments and calculated on windows of 100 kb. (d) Pearson Correlation Coefficient between the GPSeq score calculated at different resolutions (window size), for different GPSeq experiments, and the median distance from the nuclear lamina measured by FISH. Exp.1-2: HAP1 HindIII GPSeq replicates. Exp.3-4: HAP1 MboI GPSeq replicates. (e) Correlation between genome-wide GPSeq score calculated from HAP1 HindIII replicates, at 1 Mb resolution. (f) Same as e, but at 100 kb resolution. (g) "pizza-plot" depicting the radial distribution of DamID-seq Lamin B signal/control. Each slice of the plot represents a chromosome. The outer dashed dark grey circle represents the nuclear lamina, while the internal one represents the nuclear center. Each data point is a 1 Mb genomic bin placed at a radial position based on their GPSeq score. Each data point is colored based on their DamID-seq Lamin B signal/control value. (h) Distribution of GPSeq score for constitutive (c) and facultative (f) lamina-associated domains (LADs) and inter-LADs (iLADs). (i) Comparison between the number of Hi-C contacts and GPSeq score distance, between 1 Mb genomic region pairs. ND: non-diagonal contacts (i.e., between adjacent regions). Intra: intra-chromosomal contacts. Inter: inter-chromosomal contacts. PCC: Pearson Correlation Coefficient. SCC: Spearman Correlation Coefficient. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

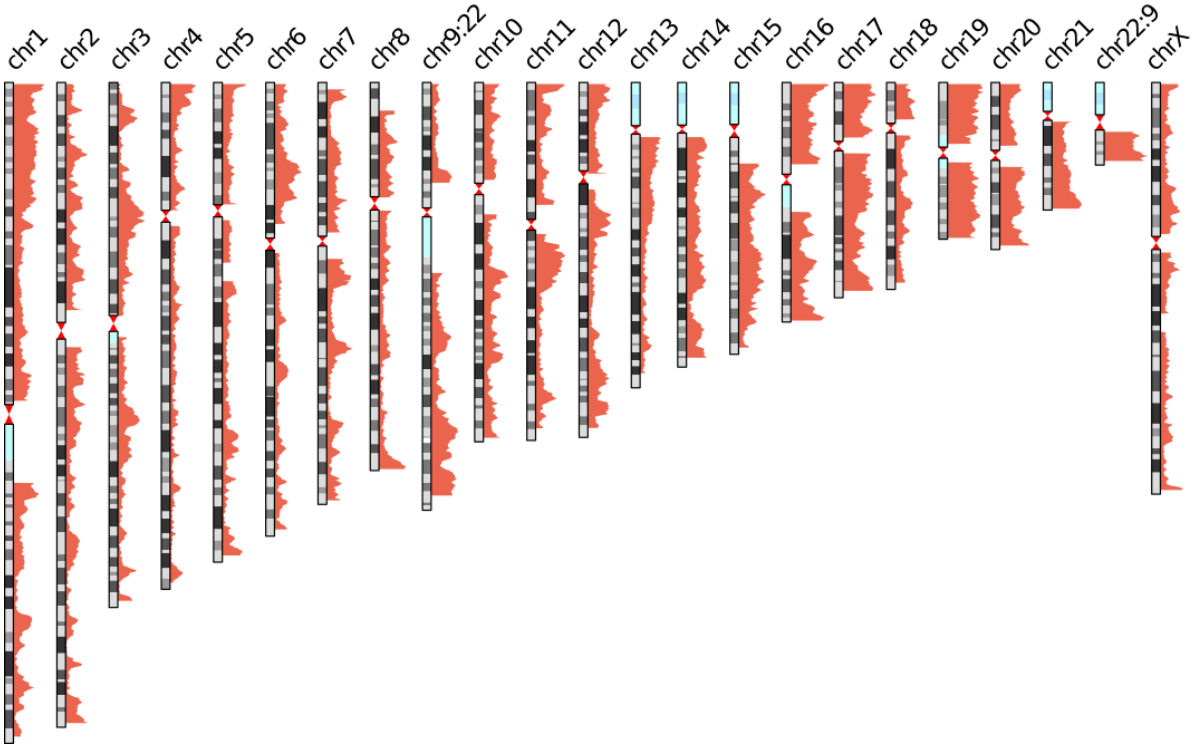

Figure 4.2.6: GPSeq score along ideograms. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

4.3 GPSeq reveals aspects of the radial genome architecture

Observing the GPSeq score distribution along each human chromosome reveals a non-homogeneous scenario rich in peaks (internal regions) and valleys (peripheral regions, Figure 4.2.6). The profiles also hint at the expected more internal (higher GPSeq score) localization of small chromosomes (with chromosome 18 being an expected outlier due to its gene-depleted nature). Indeed, chromosome size and GPSeq score are anti-correlated at chromosome-wide scale (PCC: -0.725, Paper II, Supplementary Figure 5c), though this anti-correlation dramatically weakens already at 1 Mb scale (PCC: -0.376, Figure 4.3.1c).

Gene density and gene expression showed a strong correlation (PCC > 0.69) with GPSeq score at chromosome-wide scale (Paper II, Supplementary Figure 5e-f). At 1 Mb scale, instead, gene density retained a moderate correlation (PCC: 0.645, Figure 4.3.1a), while the correlation with gene expression appeared to be relatively weak, which can be attributed to its bimodal distribution (PCC: 0.420, Figure 4.3.1b). Interestingly, GC-content appeared to be strongly correlated at both chromosome-wide (PCC: 0.840, Figure 4.3.1d) and 1 Mb resolution (PCC: 0.813, Figure 4.3.1e), and surprisingly even at 100 kb resolution (PCC: 0.666, Figure 4.3.1f).

As expected, no single genomic or epigenomic feature appeared to be sufficiently predictive of radial positioning. To overcome this, we built multi-variable linear regression models, combining multiple features to achieve a reliable prediction of the GPSeq score. Specifically, at the chromosome-wide resolution, the best centrality predictions were achieved with a model taking into account chromosome size and GC-content (R2=93.9%, Figure 4.3.1g), with gene

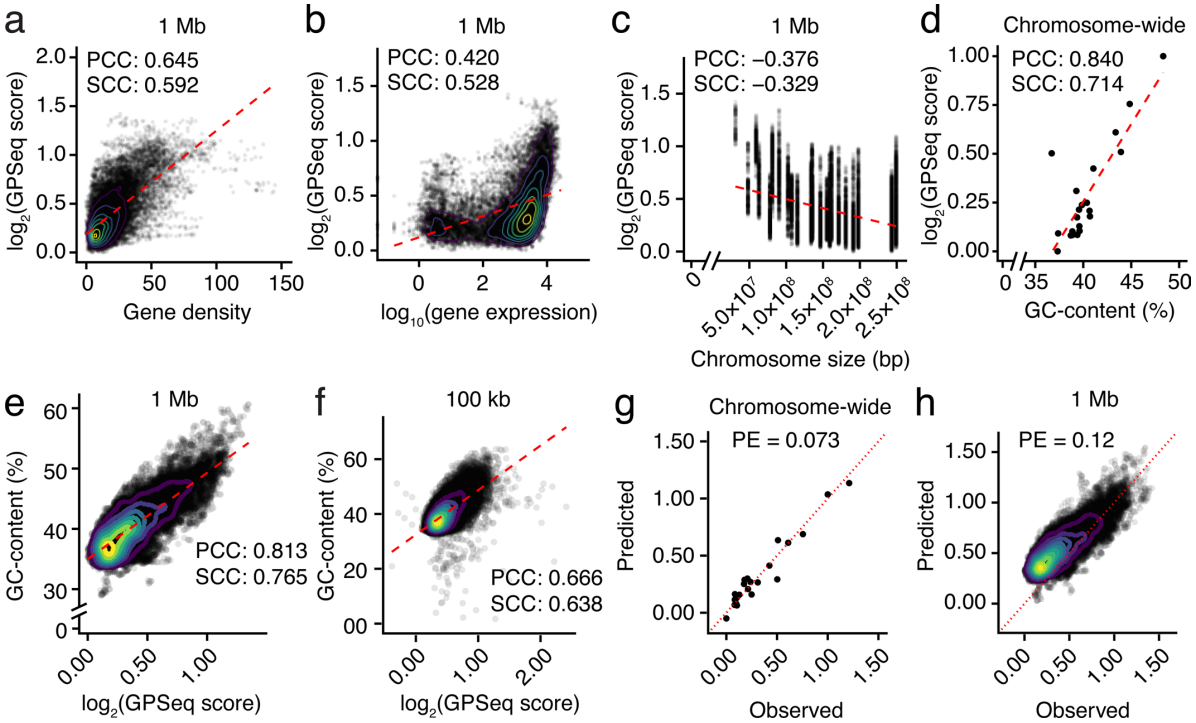

Figure 4.3.1: Predictors of the radial genome architecture. (a) Correlation between GPSeq score and gene density in 1 Mb bins. (b) Same as a, but between GPSeq score and gene expression. (c) Same as b, but between GPSeq score and chromosome size. (d) Correlation between GPSeq score and GC-content of chromosome-wide bins. (e) Same as d, but for 1 Mb bins. (f) Same as d, but for 100 kb bins. (g) Comparison of observed GPSeq score and predictions from our multi-variable model for chromosome-wide resolution. (h) Same as g but for 1 Mb resolution. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

density and expression providing no further improvement. At 1 Mb resolution, instead, the best model combined GC-content, gene density, expression, and chromosome size (R2=74.1%, Figure 4.3.1h).

To evaluate these models' ability to predict the GPSeq score, we performed two additional replicate GPSeq experiments on human GM06990 diploid lymphoblastoid cells with HindIII. The novel experiments (Exp.5-6) showed a high correlation with the HAP1 averaged GPSeq score (PCC: 0.88), and our multi-variable model built on HAP1 data appeared to reliably predict the GM06990 GPSeq score (prediction error = 0.1). These results suggest that cell-type-independent genomic features (e.g., GC-content) might be responsible for establishing a centrality baseline, which is then influenced by cell-type-specific features (e.g., gene expression).

4.3.1 Radial epigenomic patterns

We then moved on to study possible radial patterns of various epigenomic features, initially focusing on A/B genome compartmentalization. In this case, the GPSeq score confirmed the more peripheral distribution of B compartment regions, when compared to A compartment ones, at a genome-wide level (Paper II, Supplementary Figure 6). At the single-chromosome

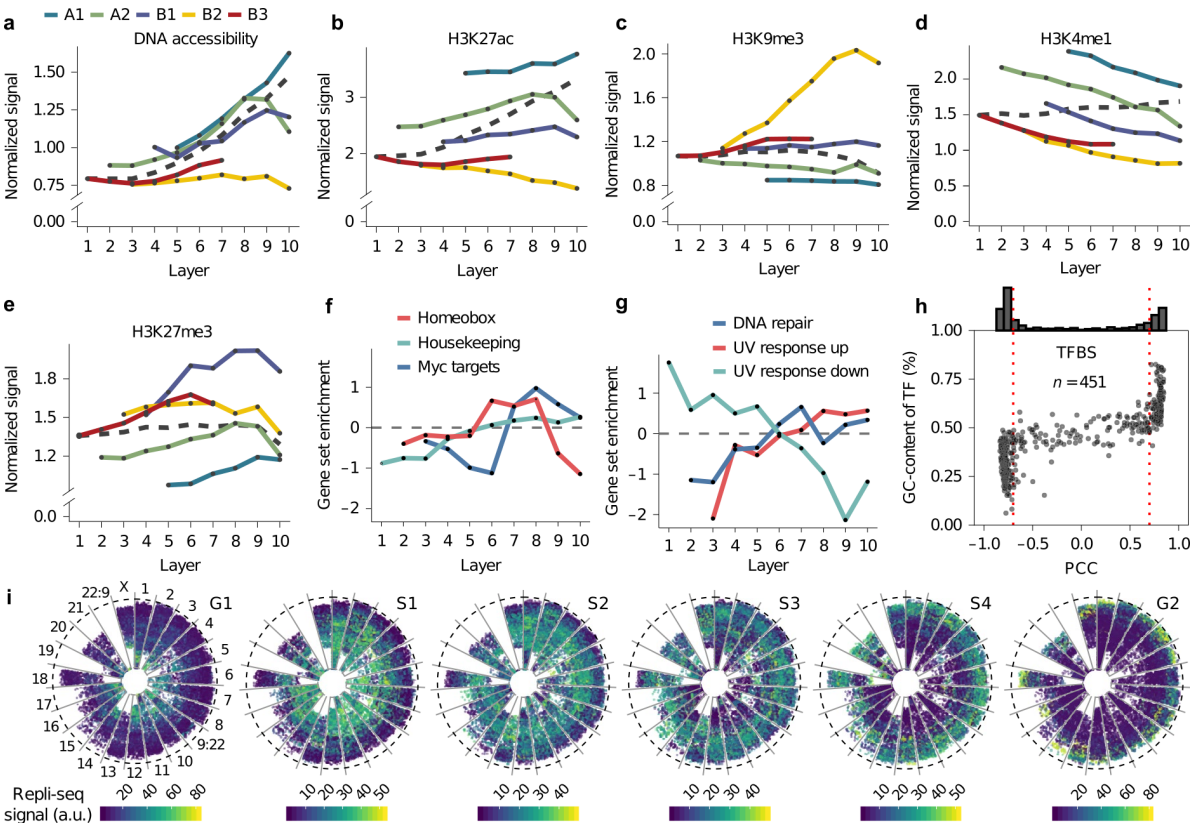

Figure 4.3.2: Radial epigenetic patterns. (a-e) Radial distribution of DNA accessibility and histone marks in 100 kb genomic regions assigned to A/B subcompartments. Dashed line: radial distribution without subcompartment stratification. (f-g) Radial distribution of selected gene groups. (h) Correlation between the log2(GPSeq score) and the TFBS motif count, in 1 Mb bins, as a function of the TF motif's GC-content. n: number of TFBSs. Dashed vertical red lines indicated PCC of ± 0.7. (i) "pizza-plot" showing repli-seq signal of six cell cycle subphases (G1, S1-S4, and G2). All source data for this figure are from HAP1 cells, except for Repli-seq data from K562 cells. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

level, A/B compartment polarization was present in most chromosomes, even when centrally located. Indeed, the difference in radial position of the two compartments appeared to be marked in internal chromosomes, like chromosomes 17 and 19, while they showed similar radial positioning for more peripherally located chromosomes, like chromosomes 10 and 18 (Paper II, Supplementary Figure 6).

However, given that such polarity varied from one chromosome to another and did not reveal a clear pattern, we wondered whether we could obtain a more precise picture by focusing our attention on subcompartments, which divide chromatin into various sub-types characterized by different expression and histone marks. Indeed, when we looked at radiality patterns of the 5 previously-characterized A1-2 and B1-3 subcompartments, their radial profiles on individual chromosomes appeared much more consistent, with the A1 sub-compartment being the most central on all the chromosomes and the B3 one being the most peripheral. We then performed a thorough analysis of the radial patterns of various epigenomic and genetic features, either genome-wide or by stratifying them by different sub-compartments. In general,

active chromatin marks and features (i.e., DNA accessibility, H3K27ac, H3Kme3, genome-bound RNAPII, gene density, and gene expression) showed a genome-wide increase toward the nuclear interior (Figure 4.3.2a-b, Paper II Extended Data Figure 4a-d). At the same time, the heterochromatin marker H3K9me3 showed an overall homogeneous distribution along the nuclear radius, with a mild decrease towards the interior. The radial patterns became more complex when the same analysis was performed on different sub-compartments. For example, DNA accessibility increased towards the center for bins belonging to A1, A2, and B1 chromatin, but remained flat for the B2. Of note, DNA accessibility seemed higher for internally located polycomb-repressed bins (B1) when compared to bins belonging to peripherally located and transcriptionally active A1 chromatin regions. In the case of the H3K9me3, even though the genome-wide trend showed a mild decrease towards the center, this mark appeared to be increasing rather sharply towards the interior of the nuclus in the B2 subcompartment (Figure 4.3.2c). Interestingly, the active and poised enhancer mark H3K4me1 showed a genome-wide increase towards the nuclear interior while showing a substantial decrease in the same direction for all subcompartments (Figure 4.3.2d). The Polycomb marker H3K27me3, thus associated with repressed chromatin, showed an increase in intermediate radial layers, which, as expected, followed the radial distribution of the Polycomb-target homeobox genes (Figure 4.3.2e-f).

We then explored whether genes involved in specific pathways showed peculiar radial arrangement. While most pathways showed a distribution in line with the overall gene distribution, a few cases stood out, for example, genes up- or down-regulated upon UV exposure showed drastically opposite arrangements, with the first ones being enriched at the nuclear interior and vice-versa. On the same line, we explored the radial distribution of predicted transcription factor binding site (TFBS) motifs and found that transcription factors (TFs) can be divided into three main categories. TFs whose binding motifs were enriched at the nuclear periphery (i.e., showing a strong anti-correlation with GPSeq score) were characterized by a low GC-content (i.e., lower than 0.5), while TFs whose binding motifs were enriched at the nuclear interior were characterized by a high GC-content (i.e., higher than 0.5). Interestingly, a minority of transcription factors binding motifs, characterized by an intermediate GC-content (i.e., typically between 0.4 and 0.6) show weak or no correlation with GPSeq score indicating a homogeneous radial distribution.

Additionally, GPSeq highlighted a gradual radial genome replication pattern, moving from the nuclear interior towards its periphery, as DNA replication proceeds (i.e., from late G1 to early G2; Figure 4.3.2i). Interestingly, subcompartments B2 and B3 seem to replicate late even when internally located (Paper II Extended Data Figure 5b), and subcompartments A2 and B1 appeared to be the main drivers of the aforementioned overall gradual radial replication pattern (Paper II Extended Data Figure 5c).

4.4 3D genome modeling

We then moved to evaluate whether GPSeq could represent a valuable contribution to the efforts to generate 3D genome reconstructions from Hi-C data. We reasoned that, due to the Hi-C method's intrinsic tendency to under-detect inter-chromosomal contacts, an additional radial constraint could improve the orientation and localization of modeled chromosome territories. To this end, we initially applied `chromflock`, a 3D genome structure generation tool developed in our lab and based on PGS[163,164], to a publicly available HAP1 Hi-C dataset[68] and generated 10,000 structures, at 1 Mb resolution, with and without GPSeq score integration (Figure 4.4.1a).

Structures generated at 1 Mb without GPSeq score integration showed a strong correlation with Hi-C input (PCC: 0.608, Paper II Supplementary Figure 8a), but only a weak correlation with GPSeq (PCC: 0.404, Paper II Supplementary Figure 9g) or with radial position measured by iFISH (PCC: 0.421, Paper II Supplementary Figure 9h). On the other hand, using the GPSeq score as a radial constraint during structure generation showed improvements in the correlation with the Hi-C input (PCC: 0.676, Figure 4.4.1b-c), with GPSeq (PCC: 0.971, Paper II Extended Data Figure 6d) and with radial position measured by iFISH (PCC: 0.947, Figure 4.4.1d). Moreover, GPSeq-informed structures showed a more polarized A/B compartment arrangement, with B compartments more peripheral than A compartments (Paper II Extended Data Figure 8a-b).

To better investigate our initial hypothesis, where the use of GPSeq data should improve chromosome localization and orientation, we first generated an additional set of 10,000 structures, at 1 Mb resolution, with and without GPSeq score integration, but with complete removal of any inter-chromosomal contacts from the input Hi-C data. As expected, structure generation using only intra-chromosomal Hi-C contacts showed a lack of apparent structure from the averaged structure distance matrix (Paper II Supplementary Figure 9a), and a considerable drop in correlation with GPSeq (PCC: -0.244, Paper II Supplementary Figure 9e) and with iFISH measurements (PCC: -0.131, Paper II Supplementary Figure 9f). Strikingly, the combination of GPSeq and intra-chromosomal Hi-C contacts seemed to rescue the correlation and drastically improved the generated structures' quality. Specifically, these structures showed an increased correlation with Hi-C input (PCC: 0.702, Paper II Extended Data Figure 7c), with GPSeq (PCC: 0.957, Paper II Extended Data Figure 7e), and with iFISH measurements (PCC: 0.942, Paper II Extended Data Figure 7f).

Moreover, we generated 1,000 structures at 100 kb resolution, with and without GPSeq score integration. Not only did these GPSeq-informed structures show a strong correlation with FISH measurements (PCC: 0.914, Figure 4.4.1f), they additionally showed the expected arrangement of subcompartments, with B3 regions at the nuclear periphery and A1 regions in the nuclear interior (Figure 4.4.1g, Paper II Extended Data Figure 8c-d). Furthermore, while both sets of structures generated with and without GPSeq showed somewhat comparable levels

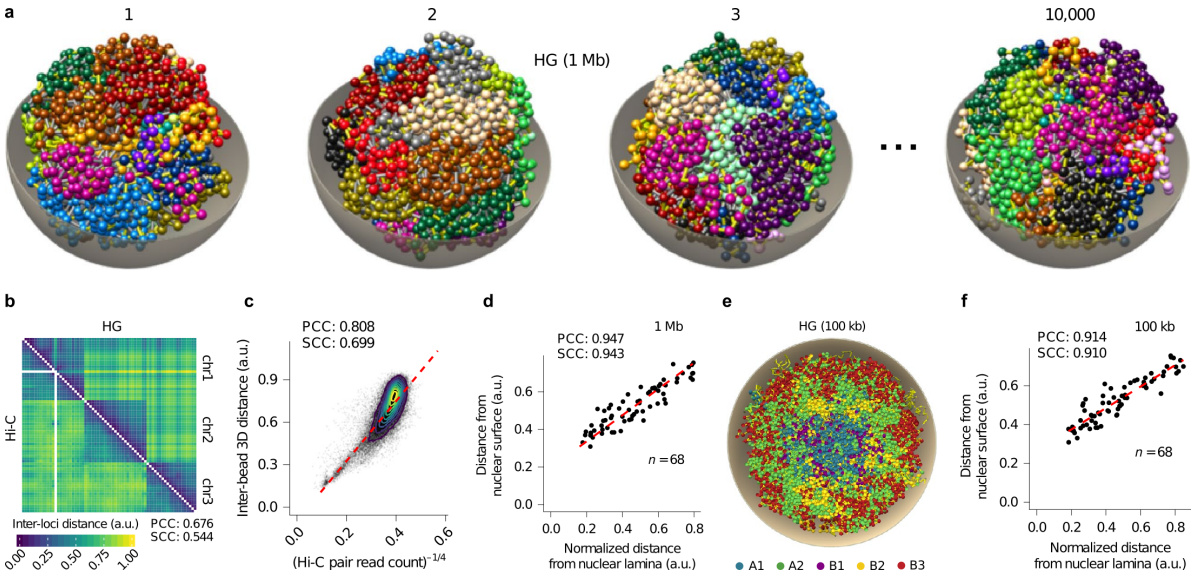

Figure 4.4.1: 3D genome modeling with `chromflock`. (a) Visual representation of four out of the 10,000 structures generated by `chromflock`, at 1 Mb resolution, with GPSeq integration. Each bead corresponds to a 1 Mb genomic region. Beads are colored by chromosome. Sticks connect consecutive beads on the linear genome. (b) Hi-C/structures distance matrix at 10 Mb resolution. The bottom triangle shows the input Hi-C contact frequency converted to distance by raising it to the power of -0.25. The top triangle shows the average distance between two genomic regions across all structures. Correlation coefficients are reported between the input Hi-C and structures-based distance matrices at 1 Mb resolution. (c) Correlation between the input Hi-C and structures-based distance matrices at 10 Mb resolution. (d) Correlation between the average distance between a genomic region and the nuclear surface averaged across all 1 Mb HG structures and FISH-based measurement. (e) Midsection visualization of a 100 kb structure generated by `chromflock`, with beads colored by subcompartment. (f) Same as d, but with 100 kb structures. HG: structures generated by combining Hi-C and GPSeq. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

of A1/B3 subcompartment polarization in each chromosome territory, the radial orientation of this polarization appeared to be correctly preserved only when the additional centrality constraint from GPSeq was in place (Paper II Extended Data Figure 9).

Altogether, these results confirmed that an additional radial constraint, in this case represented by GPSeq, can indeed improve the quality of 3D genome structure generation with the potential to provide novel biological insights.

4.5 Radial arrangement of mutations and DNA double-strand breaks

Finally, we aimed at addressing the long-standing hypothesis that heterochromatin could shield euchromatin from DNA damage, also known as the "body-guard hypothesis"[165,166]. In fact, a number of studies have generated data in agreement with this hypothesis by showing that single-nucleotide polymorphisms (SNPs) and cancer-associated single-nucleotide variants (SNVs) are more frequent in heterochromatic regions[167–170]. However, none of these

studies addressed the question of where would such heterochromatic regions be located.

To this end, we focused on four publicly available lists of cancer-associated SNVs, and on SNPs annotated from the 1000 Genomes Project[171]. Using GPSeq to visualize their radial distribution, we revealed that cancer-associated SNVs tend to be enriched at the nuclear periphery (Figure 4.5.1a). At the same time, SNPs show a more homogeneous distribution with a mild increase towards the nuclear interior (Figure 4.5.1b). This is consistent with the previous studies that associated cancer-related SNVs with the H3K9me3 heterochromatin marker[168]. Interestingly, while the increase of SNPs frequency in the nuclear interior seemed indicative of a connection with active chromatin, stratification by subcompartment revealed that the SNP burden was higher in B1 and B2 subcompartments (Figure 4.5.1b). The B1 and B2 subcompartment bins carrying SNPs happened to be enriched in the center of the nucleus. This analysis showed that even though both SNVs and SNPs were enriched in heterochromatic regions, only the SNVs seemed to associate with the peripheral heterochromatin, in agreement with the bodyguard hypothesis. The same does not seem to hold true for the SNPs, which, despite being enriched in heterochromatin, are part of B1 and B2 subcompartment regions enriched

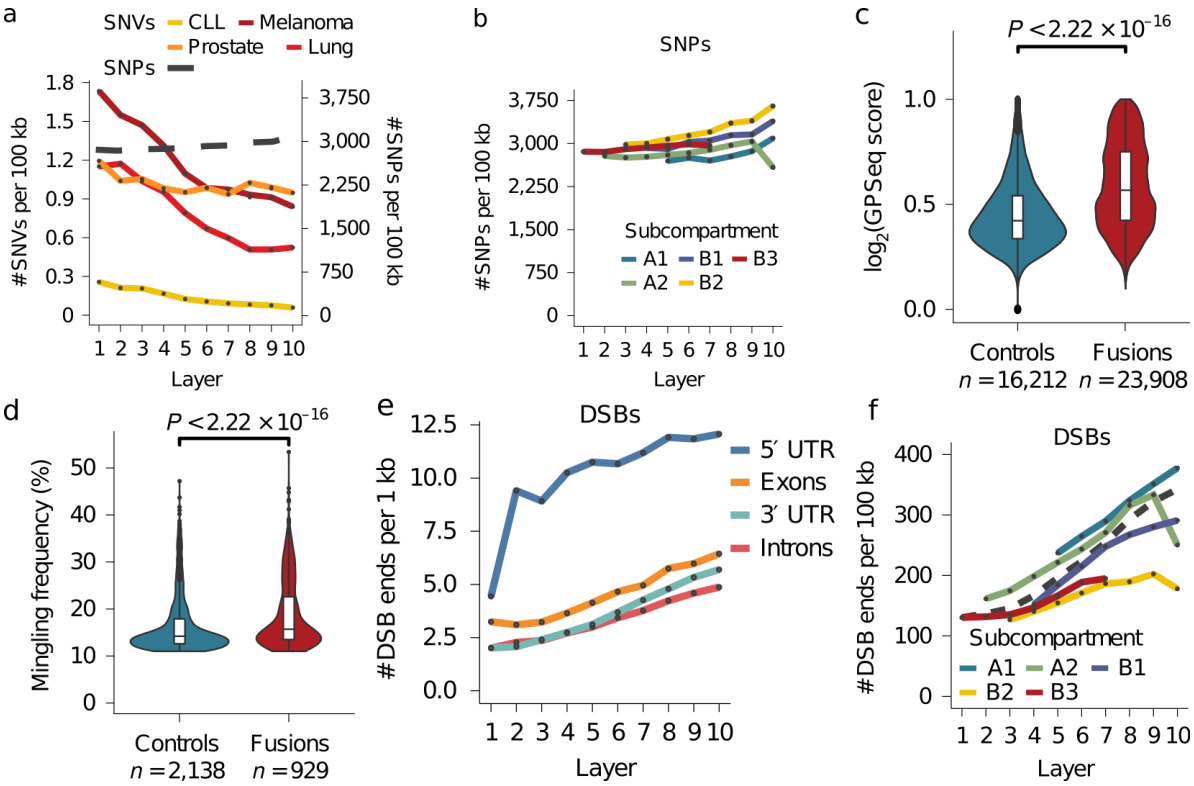

Figure 4.5.1: Radial arrangement of mutations and DNA double-strand breaks. (a) Radial distribution of germline SNPs and cancer-related SNVs, at 100 kb resolution. (b) Germline SNPs radial distribution, at 100 kb resolution, stratified by subcompartment. (c) GPSeq score distribution for genomic regions involved or not in gene fusions. (d) Mingling frequency of genomic regions involved or not in gene fusions from 100 kb structures. (e) Radial distribution of DSBs, at 100 kb resolution, stratified by protein-coding gene part. (f) Same as e, but stratified by subcompartment. Dashed black line: genome-wide unstratified distribution. Adapted from Girelli G., Custodio J. & Kallas T., et al., Nature Biotechnology, 2020 (Paper II[2]).

in the nuclear center. This observation suggests different underlying mutational processes for germline SNPs and cancer SNVs.

We then moved to investigate the radial distribution of genomic loci involved in gene fusions. GPSeq revealed that gene fusion-related genomic regions tend to be more internally located than the rest of the genome (Figure 4.5.1c), consistent with a speculated link between active transcription and gene fusion[172]. Moreover, `chromflock`-generated structures at 100 kb, with GPSeq integration, revealed a higher frequency of mingling for fusion-related regions. Mingling frequency indicates the fraction of structures in which a genomic region enters enters another chromosome's territory (Figure 4.5.1d).

These observations lead us to investigate the radial distribution of double-strand breaks (DSBs), which have been previously implicated in the formation of cancer-related gene fusions[173]. We achieved this by combining GPSeq with a map of endogenous DSBs previously generated with the BLISS method[174]. Indeed, DSBs appeared to be more frequent in gene fusion-related regions (Paper II Extended Data Figure 10d), and DSBs frequency appeared to increase from the nuclear periphery towards its interior, in both genic and non-genic regions (Figure 4.5.1e-f). Additionally, radial analysis of immunofluorescence images labeling the DSBs marker - phosphorylated histone H2A.X - consistently showed the same pattern of signal increasing towards the nuclear interior (Paper II Extended Data Figure 10f).

Altogether, these observations underscore how GPSeq radial maps can unveil novel aspects of genome architecture in different areas of study like evolution or cancer biology.

Chapter 5

Materials and Methods

This chapter briefly summarizes the primary materials and methods, both experimental and computational, used in this thesis's studies. For more details or methods pertaining to specific analyses not presented here, please refer to the main text and supplementary materials of Paper I and II.

5.1 Experimental Materials and Methods

5.1.1 Cell culture

5.1.1.1 For iFISH

A549 lung carcinoma cells, HME human mammary epithelial cells, and IMR90 fetal lung fibroblasts were purchased from ATCC (cat. no. CCL-186, PCS-600–010, and CCL-185, respectively). HAP1 chronic myeloid leukemia cells were purchased from Horizon Discovery (cat. no. C859). Human embryonic stem cells (HS975 40) were derived and used following the donor's written consent and approval from Regional Ethical Review Board in Stockholm (2011/745–31/3).

Cells were cultured following the manufacturer's instructions in 6-well chambered coverslips (custom-made by Grace Bio-Labs) or in coverslips immobilized onto a silicon gasket (CultureWell Multislip Cell Culture System MSI-12, Thermo Fisher Scientific, cat. no. C24760). Specifically: A549 cells were cultured in Ham's F-12K (Kaighn's) Medium (Thermo Fisher Scientific, cat. no. 21127022) supplemented with 10% FBS; HAP1 cells in Iscove's Modified Dulbecco's Medium (Sigma-Aldrich, cat. no. I2911) supplemented with 10% FBS; HME cells were cultured in Medium 171 (Thermo Fisher Scientific, cat. no. M171500) supplemented with Mammary Epithelial Growth Supplement (Thermo Fisher Scientific, cat. no. S0155); IMR90 cells were cultured in Minimum Essential Medium (MEM, Merck, cat. no. M4655) supplemented with 10% FBS, 1% non-essential amino acids (Thermo Fisher Scientific, cat. no. 11140035) and 1% L-glutamine (Thermo Fisher Scientific, cat. no. 25030081); Primed hESCs were cultured in NutriStem hPSC XF Medium containing bFGF and TGFβ

(Biological industries, cat. no. 05–100–1 A) on coverslips pre-coated with 10 µg/ml Human recombinant laminin-521 (BioLamina, cat. no. LN521–03).

5.1.1.2 For YFISH and GPSeq

HAP1 and GM06990 cells were obtained from Horizon Discovery (cat. no. C859) and Coriell Cell Repository (cat. no. GM06990), respecctively.

HAP1 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, Merck, cat. no. 51471C) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher Scientific, cat. no. F2442). GM06990 cells were cultured in Roswell Park Memorial Institute Medium 1640 supplemented with 2 mM L-glutamine (RPMI, Sigma, cat. no. R8758) and 15% fetal bovine serum (Thermo Fisher Scientific, cat. no. F2442).

5.1.2 Sample preparation

5.1.2.1 For iFISH

The samples were processed once the cells reached 80% confluence in each well (60% in case of hESCs) following an adapted version of the protocol for 3D-FISH[135]. Briefly, cells were fixed for 10 min at room temperature (in 1x PBS (Thermo Fisher Scientific, cat. no. AM9625)/4% formaldehyde (EMS, cat. no. 15710)) followed by quenching for 5 min at room temperature (in 1x PBS/125 mM glycine). Then, the samples were washed and permeabilized (in 1x PBS/0.5% Triton X-100 for 20 min at room temperature). After overnight incubation (in 1x PBS/20% glycerol at room temperature), the samples were subjected to five cycles of freeze-and-thaw in liquid nitrogen (30 s in liquid nitrogen, thawing in ambient air, 2–3 min in 1x PBS/20% glycerol at room temperature), and then washed. Next, the samples were incubated in 0.1 N HCl for 5 min and washed before being incubated overnight at room temperature (in 2x SSC/50% formamide/50 mM sodium phosphate). Finally, the samples were kept for one week at +4 °C in 2x SSC/50% formamide/50 mM sodium phosphate. hESCs were additionally incubated in 2x SSC/50% formamide/50 mM sodium phosphate/ 0.1% Tween20 for 24 h at +4 °C in order to reduce auto-fluorescence in the AlexaFluor 594 channel. Lastly, the buffer was exchanged to 2x SSC at +4 °C, and the samples were stored in it for up to 1 month.

5.1.2.2 For YFISH and GPSeq

HAP1 cells were seeded directly onto 22x22 mm coverslips placed in 6-well plates and grown until ~70% confluency was reached in each well. Instead, GM06990 were first centrifuged for 5 min at 300 g, resuspended in 1x PBS, dispensed onto 22x22 mm coverslips pre-coated with Poly-L-Lysine (Sigma, cat. no. P8920-100 ml), placed inside a 6-well plate, and incubated for 10 min at room temperature (RT). Briefly, cells were fixed for 10 min at room temperature (in 1x PBS (Thermo Fisher Scientific, cat. no. AM9625)/4% formaldehyde (EMS, cat. no.

15710)) followed by quenching for 5 min at room temperature (in 1x PBS/125 mM glycine). Then, the samples were washed and permeabilized (in 1x PBS/0.5% Triton X-100 for 20 min at room temperature). After overnight incubation (in 1x PBS/20% glycerol at room temperature), the samples were subjected to four cycles of freeze-and-thaw in liquid nitrogen (30 s in liquid nitrogen, thawing in ambient air, 2–3 min in 1x PBS/20% glycerol at room temperature), and then washed. Next, the samples were incubated in 0.1 N HCl for 5 min and washed. Finally, the samples were rinsed in 2x SSC buffer (Thermo Fisher Scientific, cat. no. AM9763) and stored in 2x SCC/0.05% NaN_3 up to one month at 4 °C.

5.1.3 iFISH protocol

A step-by-step protocol is available on Research Square [175].

5.1.3.1 Probe amplification

iFISH probe oligonucleotides were initially purchased as 12 K oligo-pools from CustomArray Inc. Each oligo-pool was diluted and dispensed equally in volumes in 96-well plates, with each well corresponding to a specific probe. The oligos in each well were then amplified by real-time PCR with the SYBR Select Master Mix (Thermo Fisher Scientific, cat. no. 4472913). PCR primers were designed to anneal to the F and R adapters (specific for each probe in the pool) and incorporated the C adapter and T7 promoter sequence on the 5' side of the F and R adapters, respectively. All primers were purchased from Integrated DNA Technologies (IDT) as standard desalted oligos. The PCR products in each well were separately purified with Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63881), and their DNA concentration was quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. no. Q32854). Next, each PCR product was amplified to single-stranded RNA using the HiScribe T7 Quick high yield RNA synthesis kit (NEB, cat. no. E2040S). Each reaction was carried out at 37 °C for 12–16 h, in a final volume of 30 µL containing 1 µg of purified PCR product, 6.67 mM of dNTPs, 2 Units of RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific, cat. no. 10777019) and 2 µL of T7 RNA polymerase mix. The amplified RNA was purified with Agencourt RNAClean XP beads (Beckman Coulter, cat. no. A63987), and its concentration was measured with the Qubit RNA BR assay (Thermo Fisher Scientific, cat. no. Q10210). The purified RNA was converted into cDNA by reverse transcription (RT) using the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific, cat. no. EP0751) and a primer with the C adapter sequence (RT primer). Each reaction was carried out at 50 °C for 1 h, in a volume of 20 µL containing 15 µg of purified RNA, 1.5 mM of dNTPs, 20 µM of the corresponding primer, 1x reverse transcription buffer, 10 Units of Maxima H reverse transcriptase, and 2 Units of RNaseOUT. Reaction enzymes were inactivated via incubation at 85 °C for 5 min. Template RNA was removed by adding 20 µL of 0.5 M EDTA and 20 µL of 1 M NaOH directly into the RT reaction and then incubating at 95 ° C for 15 min. Afterward, the single-

stranded DNA was immediately purified using Oligo binding buffer (Zymo Research, cat. no. D4060–1–40) and Zymo-Spin IC columns (Zymo Research, cat. no. C1004). Each probe was eluted in 40 µL nuclease-free water and its concentration was measured with the Qubit ssDNA Assay Kit (Thermo Fisher Scientific, cat. no. Q10212). The probes' lengths were matched to the expected by running them on Novex TBE-Urea Gels, 15% (Thermo Fisher Scientific, cat. no. EC6885BOX). Probes were then stored at −20 °C before DNA FISH experiments.

5.1.3.2 Single-locus probe FISH

The samples were covered with pre-hybridization buffer (PHB: 2x SSC/5x Denhardt's solution/50 mM sodium phosphate buffer/1 mM EDTA/100 ng/µL ssDNA/50% formamide, pH 7.5–8.0) and incubated for 1 h at 37 °C in a humidity chamber. The PHB was then replaced with 10 µL of hybridization mix (HM-1: 1 pmol of each probe added 1:9 v/v ratio to 2.2x SSC/5.5x Denhardt's solution/55 mM sodium phosphate buffer/1.1 mM EDTA/111 ng/µL ssDNA/55% formamide/11% dextran sulfate, pH 7.5–8.0) and the coverslip sealed on top of a microscopy slide with fixogum (MP Biomedical, cat. no. 11FIXO0125). Once the fixogum solidified, DNA was denatured by placing the samples on a heating block, pre-heated at 75 °C, for 3 min, followed by a 15-18 h incubation at 37 °C. Next, the samples were washed, and the second hybridization was performed by incubating the samples in a humidity chamber with 100 µL of the second hybridization mix (HM-2: containing fluorescently-labeled oligonucleotides, each at a final 20 nM concentration, in 2xSSC/25% formamide/10% Dextran sulfate/1 mg/mL E.coli tRNA/0.02% bovine serum albumin/10 mM Vanadyl-ribonucleoside complex) at 30 °C for 3 h. Afterward, the samples were washed and stained for DNA (30 min at 30 °C in 1.23 ng/mL Hoechst 33342 in 2x SSC/25% formamide) before mounting and imaging.

5.1.3.3 Chromosome-spotting probe FISH

The samples were first subjected to two consecutive pre-hybridization steps: (1) incubation for 1 h at room temperature in a drop of Image-iT FX Signal Enhancer (Thermo Fisher Scientific, cat. no. I36933), followed by (2) removal of signal enhancer and incubation in PHB for 1 h at 37 °C in a humidity chamber. Then, the protocol follows the same steps as for the single-locus probes (section 5.1.3.2), with HM-1 containing each probe at a final concentration of 6 nM.

5.1.4 GPSeq protocol

A step-by-step GPSeq protocol is available on Research Square[176].

5.1.4.1 Digestion and ligation

Briefly, *in situ* restriction was performed using either 10 µl of HindIII-HF (NEB, cat. no. R3104S) or 8 µl of MboI (NEB, cat. no. R0147M) in 400 µl at 37 °C for different durations,

ranging from 1 min to 30 min, in the case of MboI, and 6 h in the case of HindIII. The restriction reaction was stopped by placing the samples in ice-cold 1x PBS/50 mM EDTA/0.01% Triton X-100 and washing them multiple times on ice. Afterward, the samples were dephosphorylated by incubating them in 400 µl of 1x calf intestinal alkaline phosphatase buffer containing 6 µl of calf intestinal alkaline phosphatase (Promega, cat. no. M1821) for 2 h at 37 °C. Next, ligation was performed with YFISH adapters at a final concentration of 0.2 µM in 300 µl of 1x T4 DNA ligase buffer containing 36 µl of T4 DNA ligase (Thermo Fisher Scientific, cat. no. EL0014), and by incubating the samples for 18 h at 16 °C. The next day, unligated adapters were washed by incubating the samples in 10 mM Tris-HCl/1 M NaCl/0.5% Triton X-100, pH 8, five times for 1 h each at 37 °C while shaking.

5.1.4.2 YFISH and imaging

After digestion and ligation (section 5.1.4.1), the hybridization mix was prepared by diluting the labeled oligonucleotide to 200 nM in a hybridization buffer containing 2x SSC/25% formamide/10% dextran sulfate/1 mg ml^{-1} *E. coli* tRNA/0.02% bovine serum albumin. The coverslips were placed onto a piece of Parafilm, with cells facing a 300-µl droplet of hybridization mix, and incubated in a humidity chamber for 18 h at 30 °C. The next day, the samples were washed in washing buffer containing 2x SSC/25% formamide for 1 h at 30 °C. Finally, the samples were incubated in 2x SSC/25% formamide/0.1 ng $µl^{-1}$ Hoechst 33342 (Thermo Fisher Scientific, cat. no. H3570) for 30 min at 30 °C, rinsed twice in 2 × SSC and mounted in ProLong Gold Antifade Mountant (Thermo Fisher Scientific, cat. no. P36930) before imaging. All samples were imaged using either wide-field epifluorescence microscopy or STED microscopy (see Paper II Supplementary Methods for more details).

5.1.4.3 Library preparation and sequencing

After digestion and ligation (section 5.1.4.1), cells were scraped off the coverslips and digested in 110 µl of 10 mM Tris-HCl/100 mM NaCl/50 mM EDTA/1% SDS, pH 8, containing 10 µl of Proteinase K (NEB, cat. no. P8107S), for 18 h at 56 °C. The next day, the reaction enzymes were inactivated by increasing the temperature to 96 °C for 10 min. Genomic DNA (gDNA) was purified using phenol-chloroform extraction and precipitated gDNA using glycogen (Sigma, cat. no. 10901393001) and sodium acetate, pH 5.5 (Life Technologies, cat. no. AM9740) in ice-cold ethanol (VWR, cat. no. 20816.367) for 18 h at –80 °C. The DNA pellets were then resuspended in 100 µl of TE buffer and sonicated in a Bioruptor Plus machine with the following settings: 30 s ON, 90 s OFF, high mode, 16 cycles. Afterward, gDNA was concentrated to a final volume of 8 µl in nuclease-free water, using AMPure XP (Beckman Coulter, cat. no. A63881). *In vitro* transcription (IVT) was performed, separately on each sample, with the MEGAscript T7 Transcription Kit (Thermo Fisher Scientific, cat. no. AM1334-5), using the same amount of gDNA (between 50 and 300 ng) for each sample in

a final volume of 20 µl and incubating the samples for 14 h at 37 °C. After IVT, 1 µl of DNAse I (Thermo Fisher Scientific, cat. no. AM2222) was added to the samples and incubated for 15 min at 37 °C. RNA was then purified with Agencourt RNAClean XP beads (Beckman Coulter, cat. no. A63987). Lastly, sequencing libraries were prepared using the TruSeq Small RNA Library Preparation Kit (Illumina, cat. no. RS-200-0012), following the manufacturer's instructions (modifications to the manufacturer's protocol are described in the step-by-step protocol). All the libraries were sequenced with the NextSeq 500 system (Illumina) using the NextSeq 500/550 High Output v2 kit (75 cycles) (Illumina, cat. no. 20024906).

5.1.5 Image acquisition and pre-processing

5.1.5.1 For iFISH

All samples were imaged using a 100x 1.45 NA objective mounted on a custom-built Eclipse Ti-E inverted microscope system (Nikon) controlled by the NIS Elements software (Nikon) and equipped with an iXON Ultra 888 ECCD camera (Andor Technology). Multiple image stacks were acquired for each sample, each consisting of 49–60 focal planes spaced 0.3 µm apart. Raw images were converted from ND2 (Nikon) format to uncompressed TIFF format using the `nd2_to_tiff` script from our custom-developed `pygpseq` Python3 package available on GitHub: https://github.com/ggirelli/pygpseq/releases/tag/v3.3.5. Out-of-focus image stacks were removed by using the `tiff_findoof` script of the `pygpseq` package. The script identifies and discards stacks in which the peak of the gradient magnitude of the stack intensity over Z does not fall in a range of 50% of the stack around the mid slice. To correct chromatic aberrations and shifts between channels, TetraSpeck Microspheres (0.1 µm, fluorescent blue/green/orange and dark red, Thermo Fisher Scientific, cat. no. T7279) were imaged before or after each imaging session. The DNA stain channel was used as the reference channel, and the location of the beads was determined by fitting a 2D Gaussian profile in (x,y) and a 1D Gaussian in z by optimizing a maximum-likelihood functional using the Nelder–Mead method[177]. Signals corresponding to the same bead were identified by clustering the strongest dots detected in the DNA channel with the strongest dots in each other channel. After determining the shift between the channels, outliers were detected and discarded. Finally, a second-order 2D polynomial deformation was fitted to the (x,y) plane using the remaining point pairs. Along the z directions, only a shift correction was applied[178]. After chromatic aberration correction, automated 3D segmentation of cell nuclei stained with Hoechst 33342 was performed. To this end, the DNA staining channel was first deconvolved using the Huygens Professional v17.04 Software (Scientific Volume Imaging), with the following parameters: CMLE algorithm, null background, and signal-to-noise ratio (SNR) of 7, in 50 iterations. A theoretical point spread function was estimated using the same software, considering both the microscope setup and the optical configuration used to acquire the images. After deconvolution, 3D segmentation of the nuclei in each field of view was performed using the `tiff_auto3dseg`

script of the `pygpseq` package. The script combines with a logical AND operation two binary masks, generated with the `threshold_otsu` and `threshold_local` methods from the `scikit-image.filters` package[179]. Then, it discards objects touching the XY contour of the image, fills any holes in the masks, and performs a dilate-fill-erode operation. To identify putative G1-phase cells, a previously published approach[180] that selects nuclei based on the integral of DAPI intensity over the nuclear volume was adapted. Briefly, a sum of Gaussians is first fitted to the distribution of both total DNA staining intensity over the nuclear volume and nuclear area in the z-projection of the segmented images. In cases when the fitting fails, a single Gaussian is fitted instead. Lastly, all nuclei falling in a range of ±3x standard deviations around the mean of the major fitted Gaussian in both distributions were labeled as G1.

5.1.5.2 For YFISH

All YFISH, 3D DNA FISH, and IF samples were imaged a 100x 1.45 NA objective mounted on the custom-built Eclipse Ti-E inverted microscope system (Nikon) described above for iFISH (section 5.1.5.1). For YFISH and 3D DNA FISH, multiple image stacks were acquired per sample, each consisting of 40–110 focal planes spaced 0.2 or 0.3 µm apart. All images were corrected for chromatic aberrations as described above for iFISH (section 5.1.5.1).

Super-resolution 3D-STED imaging of YFISH samples was performed on a Leica SP8 3X STED system equipped with lasers for the depletion of fluorophores emitting in the blue/green (592 nm, MPB Communications Inc.), orange (660 nm, Laser Quantum), and red/far-red (775 nm, OneFive GmbH). Specifically, YFISH signal was imaged by exciting the ATTO647N fluorophore with a tunable pulsed white-light fiber laser (Leica Microsystems), and an excitation wavelength of 640 nm together with 775 nm stimulated emission depletion. Similarly, DNA staining with Hoechst 33342 was imaged with a 405 nm excitation laser delivered by a diode-laser. All the samples were imaged with a chromatically optimized oil-immersion objective (HC PL APO 100X/1.40 OIL STED WHITE, Leica Microsystems). The fluorescence signals were passed through a 0.9–1.0 Airy unit pinhole and detected using sensitive photodetectors (Leica Hybrid Detectors). The emitted light was filtered by appropriate dichroic mirrors and selecting an appropriate wavelength window (Hoechst 33342: 420–480 nm; ATTO647N: 650–730 nm) in the Acousto-Optical Beam Splitter (AOBS, Leica Microsystems). Excess STED laser light was blocked with a sharp notch filter in front of the ATTO647N detector. Dual-color axial stacks were acquired sequentially, frame-by-frame, at a scan speed of 400 lines per sec. The super-resolution STED pixel size was tuned based on the depletion power applied laterally and axially (80% laterally and 20% axially). Before analysis, all stacks were deconvolved with the Huygens Software (Scientific Volume Imaging).

5.2 Computational Materials and Methods

5.2.1 Homologous sequence design

All 40-mers of the human reference genome (Grch37/hg19 GCA_000001405.1) were extracted using `JELLYFISH` v2.2.6[157] and a custom-made pipeline in Perl, which we dubbed `oligo-picker`[181]. The extracted 40-mers were retained only when they did not contain a homopolymer stretch of 7 or more bases, and they had a GC-content within the 35–80% interval. Then, the average melting temperature of the remaining 40-mers was calculated. Subsequently, sequences with a homology of 70% or higher to more than one genomic location were discarded using `VMATCH` v2.2.4[158]. Afterward, the delta free energy (ΔG) of the most stable secondary structure at 65 °C was calculated using `OligoArrayAux` v3.8[138], and 40-mers with a negative ΔG were discarded. Only the 40-mers with a melting temperature in a range of 20 °C around the previously calculated average temperature were retained for further analysis. Lastly, overlapping 40-mers were discarded by starting from the first one and iterating through. All retained 40-mers were stored in a sqlite3 database for easy access.

An additional database was generated by flagging 40-mers based on their off-target count. Briefly, to achieve this, `bowtie` v1.2.2[182] was used to align the 40-mers to the human genome reference, allowing up to 6 mismatches with the following command: `bowtie -f -seedlen 10 -seedmms 3 -maqerr 200 -k 1000 -sam -mm hg19.genome.fa unique_oligonucleotides_2016_run40mer_h70.fasta`. The `-seedmms` , `-seedlen` and `-maqerr` parameters correspond to the seed length, the number of maximum mismatches allowed in the seed sequence, and the threshold for the total number of mismatches in the final alignment, respectively. `sambamba` v0.6.7[183] was used to sort and index the alignment file and a custom Python script to post-process and to collect statistics for each entry in our initial dataset. Briefly, each oligonucleotide sequence was flagged if it had more than 10 off-targets with up to 5 mismatches, based on the "`NM:i < N >`" flag.

5.2.2 Orthogonal sequence design

All 20-mers from a previously published list of 240,000 25-mers orthogonal to the human genome[161] were extracted using a custom-made pipeline (`ood-fish` v0.0.2, https://github.com/ggirelli/ood-fish/releases/tag/v0.0.2). The 20-mers were aligned to the human reference genome (Grch37/hg19 GCA_000001405.1) using `BLAT` v36x1[159] with the following parameters: `-tileSize=6 -stepSize=1 -minMatch=1 -oneOff=1 -minScore=0 -minIdentity=0 -maxGap=0 -repMatch=131071 -noHead`. All the 20-mers with at least one alignment with maximum homology (defined as the fraction of single-base matches over the sequence length, i.e., 20 nt) equal to or higher than 80% of the oligonucleotide length (i.e., 16 nt) were discarded. The self-dimerization free energy (ΔG_S) of all the 20-mers was then calculated, and only sequences with self-dimerization free energy

$\Delta G_S > -5$ kcal/mol were retained. Then, the lowest hetero-dimerization free energy (ΔG_H) for every pair of 20-mers and their reverse complement sequence was calculated. The largest set of 20-mers with $\Delta G_S \geq -9$ kcal/mol was identified using the Parallel Maximum Clique library[160]. Both ΔG_S and ΔG_H values were calculated using the nearest-neighbor method for the longest stretch of matching nucleotides, assuming an oligonucleotide concentration equal to 0.25 M and a sodium concentration equal to 50 mM.

5.2.3 iFISH data analysis

First, out-of-focus image stacks were discarded, DNA staining channels deconvolved, nuclei automatically segmentated in 3D, FISH signals detected in 3D, and chromatic aberrations corrected, as described above for iFISH (section 5.1.5.1).

Then, the output of DOTTER was passed to the `gpseq_fromfish` script of the `pygpseq` package to measure the radial position of FISH dots. Briefly, for each FISH dot, the script computes its distance from the nuclear edge (i.e., the contour of the segmented DNA stain) as well as from the nuclear center. The distances are then normalized in a FISH dot-wise manner by dividing them by the sum of the corresponding distances from the nuclear edge and center. In this case, the nuclear center was defined as the set of voxels in the top percentile of the distribution of distances from the nuclear edge. This analysis was implemented as a `snakemake` workflow[184], also available on GitHub: https://github.com/ggirelli/iFISH-singleLocus-analysis/releases/tag/v1.0.

5.2.4 YFISH image analysis

First, out-of-focus image stacks were discarded, DNA staining channels deconvolved, and nuclei automatically segmentated in 3D, as described above for iFISH (section 5.1.5.1).

The `gpseq_anim` script of the `pygpseq` package was then run on the YFISH images. The script first estimates the background level of both DNA and YFISH channels as the median background intensity and subtracts these values from the original image (with negative intensity voxels set to zero). The script estimates the volume, shape, surface, flattened size (in Z-projection), the sum of intensity, average intensity, a shape descriptor, and the center of mass for each segmented nucleus. Specifically, the script was run with the following parameters: `--compressed -rn -a 200 130 130 --an-type 3d - uy --nuclear-sel -t 10`. In particular, the `-u` parameter allows extracting all the nuclear bounding boxes for further analysis. For each voxel, the pipeline produced all channel images, nuclear masks, and EDT-based matrices with absolute distance from the nuclear lamina and center (defined as the 1% most distant voxels from the lamina).

The `pygpseq-scripts` suite of Python3 and R scripts[185] was used to analyze the extracted nuclear bounding boxes. First, the `extract_nuclear_features.py` script was run to obtain all extracted nuclei characteristics, which were fed to the `select_nuclei.R` script, with pa-

rameter `-k 2`, to select only nuclei from cells in the G1 phase of the cell cycle. This script performs a nuclear selection as described above for iFISH (section 5.1.5.1). Only nuclei selected in such a manner were retained for further analysis. Afterward, the `extract_nuclear_vx.py` script was used to obtain the following measurements for every single voxel of the selected nuclei: intensity from each channel, absolute distance from the nuclear lamina, absolute distance from nuclear center, normalized distance from the nuclear lamina (defined as the absolute distance from the nuclear lamina divided by the sum of the absolute distance from lamina and center).

Next, the voxel data to three scripts that build radial profiles from the nuclear periphery inwards for: (1) each condition (e.g., time of digestion) by pooling all the nuclear voxels (`extract_condition_profiles.py`); (2) each nucleus (`extract_nuclear_profiles.py`); and (3) single straight-line trajectories, going from the nuclear center of mass (CoM) to the nuclear surface (`extract_nuclear_radii.py`). Specifically, the `extract_nuclear_radii.py` script was run with default parameters to draw 200 straight trajectories and sample 100 points homogeneously in each of them, for 500 randomly picked nuclei for each condition. The 200 trajectories all departed from the nuclear CoM, were homogeneously spread in space using a spherical Fibonacci lattice, and terminated at their point of intersection with the nuclear surface, determined by a triangular meshing algorithm over the nuclear 3D mask. In all three cases, radial profiles were built by first dividing the nuclear radius into 200 bins and then by assigning the intensity values of the voxels (or sampled points) of interest to the corresponding bins, based on the normalized distance from the nuclear lamina. Then, the median intensity of each bin was calculated and reported alongside the midpoint of each bin. Finally, the `extract_profile_descriptors.R` script was run to retrieve information on the peak, inflection point, and contrast of the profiles. Specifically, the script fits a 5 th-degree polynomial curve to the profile data and uses the `uniroot.all` function from the `rootSolve` R package (v1.7) to identify profile's peak and inflection point, and the relative intensities. The profile contrast is then calculated as the ratio of the intensity at the peak over the intensity at the inflection point.

5.2.5 GPSeq pre-processing

Raw sequencing data were demultiplexed based on the RA5 indexes, either using the BaseSpace Sequence Hub cloud service of Illumina, or manually with `bcl2fastq` (v2.18). The generated fastq files were quality-checked with `fastqc`(v0.11.4). Reads containing the full prefix (UMI_barcode_restriction site) were selected using `scan_for_matches`. Afterward, the reads were trimmed to remove the prefix (including the restriction site), and the remaining part was aligned against the human reference genome (Grch37/hg19 GCA_000001405.1) using `bwa-mem`. Primary alignments were retained only with a mapping quality equal to or higher than 30, while the following were discarded: unmapped reads, chimeric reads, and reads mapped to chrY (not present in HAP1 cells). The reads were selected for those whose 5'

end is less than 20 bp away from the position of a HindIII or MboI recognition site (RS) in the reference human genome, depending on which enzyme was used. The reads are not strictly required to align precisely to RS's position because the T7 polymerase and reverse transcriptase used during the GPSeq library preparation are prone to occasionally skip some bases, leading to the resulting reads aligning slightly downstream of the RS. Afterward, the UMI sequence of each aligned read was recovered, and the reads were filtered based on the quality of the UMI sequence using an approach similar to the one used by the `fastq_quality_filter` tool (http://hannonlab.cshl.edu/fastx_toolkit/index.html). UMI sequences mapped to the same restriction site were de-duplicated, and a BED file containing the genomic coordinate and number of de-duplicated UMIs associated with each restriction site was generated. All the above steps were performed using gpseq-seq-gg (v2.0.2), a bash/Python/R custom-designed pipeline available on GitHub (https://github.com/ggirelli/gpseq-seq- gg). A newer version of the pipeline (v2.0.3) with improvements to the UMI de-duplication script's efficiency is available at DOI: 10.5281/zenodo.3264757. This version is optimized to deal with datasets generated with 4-base cutters such as MboI used in this study. Finally, the BED files generated by `gpseq-seq-gg` were corrected for the presence of the t(9;22) translocation using an ad hoc Python script available at http://github.com/ggirelli/bed-fix-chrom- rearrangement (v0.0.1, DOI: 10.5281/zenodo.3365906), using the following parameters: `-1 chr9:133681295 -2 chr22:23632359`.

5.2.6 Genome centrality estimation

After pre-processing (see section 5.2.5), centrality scores were estimated with the `gpseqc_estimate` script of the `gpseqc` Python3 package[186]. This analysis was also implemented as a `snakemake`[184] workflow[187]. Briefly, the script first discarded restriction sites (AAGCTT in Exp.1 and 2 with HindIII and GATC in Exp. 3 and 4 with MboI) associated with an abnormally high number of de-duplicated UMIs for a given digestion time (i.e., condition) by identifying outliers with a chi-squared method and a significance of 0.01. The reference genome was then binned using either 1-Mb overlapping windows sliding in 100 kb steps (1 Mb resolution) or non-overlapping 100-kb windows (100 kb resolution). For each condition, all the restriction sites that had been cut were considered to calculate a digestion probability, based on which the GPSeq score in turn was calculated. The script generated a BED-like file containing the GPSeq score per window and then masked it based on a manually curated mask of repetitive and low-complexity regions (Paper II Supplementary Table 7). For comparison purposes, the calculated GPSeq score as described before (see section 4.2.1.4) and in Paper II Supplementary Note 1.

Part III

Final remarks

Chapter 6

Discussion

In this thesis's scope, we established two novel methods for investigating the mammalian genome architecture and beyond.

The first method, namely **iFISH** (**Paper I**[1]), represents a powerful resource available to the research community enabling easy design and large-scale amplification of DNA FISH probes. Specifically, iFISH presents a framework for a straightforward design of both single probes or multiple homogeneously distanced (spotting) probes. We implemented this framework in the form of the `ifpd` Python package, which provides access to the probe design tools via a locally hosted web graphical user interface, complementing the two pipelines that we developed to design of homologous (`oligo-picker`) and orthogonal (`ood-fish`) sequences. We applied `oligo-picker` to establish a novel database of human genome homologous sequences, which showed a higher genome coverage, essential for the design of probes targeting small genomic regions when benchmarked against state-of-the-art `OligoMiner`-based databases. Additionally, we designed, individually tested, and released a repository of 330 DNA FISH probes, covering all human chromosomes in a uniform manner (every 10 Mb for chromosomes 1 to 16 X; every 5 Mb for chromosomes 17 to 22). Furthermore, we showcased the flexibility and power of iFISH in several applications, revealing a lack of chromosome territories in a subset of human embryonic stem cells and the necessity for high labeling frequency when estimating chromatin volume from spotting probes.

The second method, namely **Genomic loci Positioning by sequencing** (**GPSeq**, **Paper II**[2]), is a novel molecular biology assay allowing for genome-wide measurement of chromatin radial localization. GPSeq is based on a simple yet elegant concept, in which the application of differential enzymatic restriction times to different samples generates growing crown-shaped genomic digestion waves from the nuclear periphery towards its interior. We provide an initial proof-of-concept via our novel YFISH approach, which enables the visualization of digested sites via fluorescence microscopy. To provide a quantitative assessment of the radial digestion waves, we developed `pygpseq`, a Python package containing full-stack image analysis tools, from conversion to TIFF format from microscope proprietary formats (e.g., ND2, CZI) to the construction of radial intensity signal profiles.

We then established a sequencing-ready library preparation protocol applied to samples subjected to different durations of restriction time. After sequencing, the reads generated from each condition are pre-processed via a `bash`-based pipeline (`gpseq-seq-gg`) that generates UMI-based de-duplicated read counts per restriction site, per condition. We initially designed multiple potential mathematical formulations that combined these read counts from the different digestion times into potential centrality estimates in an unbiased manner. We evaluated these formulations against the distance from the nuclear lamina measured for 68 genomic loci, across 11 chromosomes, by exploiting our iFISH method. Most of our formulations were strongly correlated with the iFISH-based measurements, based on which we selected the top-performing formulation, which we then called "GPSeq score." We performed additional validation of the GPSeq score by comparing it with two orthogonal sequencing-based techniques, Hi-C and DamID-seq, which yielded the expected results and intriguingly showed a different radial distribution of constitutive and facultative lamina-associated domains (LADs) and inter-LADs.

We believe that GPSeq represents a clear step forward compared to the other techniques allowing for some degree of measurement of radial chromatin localization. For example, while the TSA-seq[154] method can be used to measure the distance of genomic loci from Lamin B1, it has been reported to measure up to a maximum distance of 1.5 μm. Similarly, DamID-seq[98,99] performed on Lamin B1 can measure chromatin radial localization along a portion of the nuclear radius but is limited to loci containing the GATC sequence and requires transgenic samples. On the same line, GAM[149], single-cell Hi-C[32], and Dip-C[33] allow for the estimation of chromatin radial localization. However, GAM represents a relatively complex protocol, based on cryosectioning and laser dissection, limiting the number of cells that can be easily processed, and leaves questions on the achievable radial resolution of the technique, as this is dependent on the number of sections per nucleus. Similarly, the single-cell Hi-C and Dip-C, with the latter providing higher-resolution results, are expensive and laborious methods that can be applied only to a limited number of single cells.

On the other hand, GPSeq is a relatively inexpensive and straightforward method that allows measuring chromatin radial localization along the full nuclear radius. Furthermore, GPSeq potentially allows for the combination of multiple restriction enzymes, thus overcoming the limitation due to restriction enzyme motif distribution, and does not require any complicated sample preparation, being applicable also to primary cell lines.

We applied GPSeq to haploid HAP1 human cells and confirmed previously known low-resolution genome architecture patterns, like preferential internal localization of small chromosomes and active chromatin. Using the A/B compartment annotations provided by Hi-C, GPSeq confirmed the previously described radial polarization where B-compartments are more peripheral than A-compartments[40,41,188]. Moreover, GPSeq allowed to reveal the radial distribution of subcompartments, showing a more refined organization than the simple B-peripheral/A-internal dichotomy, with B1 being typically more internally located than A2

subcompartment regions.

We combined GPSeq and publicly available RRBS-Seq and ChIP-seq data to investigate the radial arrangement of chromatin marks and DNA methylation. As expected, repressed and inactive chromatin marks are enriched at the nuclear periphery, while active chromatin marks are enriched at the nuclear interior. Stratification of radial chromatin mark radial distribution revealed interesting and novel patterns, with the H3K9me3 heterochromatin mark decreasing genome-wide towards the nuclear interior but sharply increasing for B2-subcompartment regions. Similarly, the H3K4me1 active and poised enhancer mark increased genome-wide towards the nuclear interior but showed a monotonic decrease in the same direction for all subcompartments when taken separately. Moreover, we combined GPSeq and Repli-Seq to reveal an A2/B1 subcompartment-driven gradient in replication timing, moving from the nuclear interior towards the periphery.

The genome-wide centrality maps drawn thanks to GPSeq allowed us to build accurate and simple multi-variable linear regression models to predict radial positioning at different resolutions. Specifically, we identified GC-content and chromosome size as the main predictors of chromosome radial localization. At 1 Mb resolution, instead, the best model combined GC-content, gene density, gene expression, and chromosome size. Applying the HAP1-based models to human GM06990 lymphoblasts showed low prediction error compared to GPSeq performed on that cell line. Altogether, this suggested that cell-type-independent genomic features like GC-content and chromosome size might be drawing a radial localization blueprint, further shaped by cell-type-dependent features like gene expression.

We then confirmed our hypothesis that GPSeq might improve 3D genome reconstruction by providing an additional centrality-based constraint to the construction algorithm. Specifically, we implemented `chromflock`: an algorithm based on PGS [163,164] that allows for the deconvolution of a Hi-C contact matrix into an ensemble of haploid putative single-cell-like 3D genome structures, further allowing for seamless GPSeq integration. Indeed, chromflock was able to reproduce previously published results, and the additional GPSeq constraint greatly improved the radial organization of the 3D reconstructions. Interestingly, combining GPSeq radial maps and intra-chromosomal Hi-C contacts was sufficient for the construction algorithm to recover the Hi-C-detected inter-chromosomal contacts. Moreover, we showed that compartment and sub-compartment polarization and radial orientation was following previously described patterns only when the GPSeq constraint was in place. These results support the view that 3D genome reconstruction from Hi-C can benefit from additional constraints deriving from orthogonal techniques, to yield more reliable 3D structures compatible with the investigation of novel biological aspects.

Finally, we focused on the radial distribution of mutations, gene fusions, and double-strand breaks. Specifically, genome-wide GPSeq-based radial maps allowed us to approach, for the first time, the so-called "bodyguard hypothesis" [165,166], where the peripheral heterochromatin shields internal euchromatin from exogenous mutagens. In fact, GPSeq revealed that cancer-

related SNVs tend to be enriched in the nuclear periphery, in direct contrast with germline SNPs, which are enriched at the nuclear interior. Moreover, internally located germline SNPs seemed to be more frequently located in B1/B2 subcompartments and not at active chromatin A compartment regions. These results hint at a different underlying mechanism from which germline SNPs and cancer SNVs might arise. Moreover, we showed that gene fusion-related regions tend to be located more internally and mingle more frequently than other genomic regions, in line with an increased double-strand break frequency and phosphorylated H2A.X histone IF signal in internal regions. Altogether, these results underline the relevance of genome-wide radial maps in the context of genome architecture-related studies.

Chapter 7

Conclusion and Future Perspectives

In short, in this thesis, we presented two novel methods, iFISH and GPSeq, to investigate the mammalian genome architecture. We envision some future improvements for both methods.

In iFISH, the `ifpd` package can be improved to provide more advanced design options, and faster read/write operations. Moreover, we envision using a database containing overlapping oligonucleotides, which might help in the design of probes in problematic areas by performing a better selection of non-overlapping oligonucleotides during probe design. We are currently developing these and other features in a new version of the package, namely `ifpd2`[189], available on GitHub.

For GPSeq, we intend to extend the method by applying it to diploid cells and less rounded nuclear shapes. Specifically, GPSeq application on diploid chimeric samples would directly allow for allele-calling, revealing whether different alleles show different radial localizations. In terms of nuclear shapes, GPSeq could be initially applied to cells with characteristically elongated nuclei (e.g., human IMR90 fibroblast cells) and then to more complex shapes (e.g., white blood cells poly-lobated nuclei). Additionally, GPSeq could easily be adapted to be applied on various samples, like tissues or other clinical samples, or even on condensed mitotic chromosomes, to unveil aspects of radial chromatin organization after compaction into chromatids. Potentially, the simple enzyme diffusion concept at the basis of GPSeq could be adapted to investigate the radial organization of other nuclear components, like RNAs or proteins, the latter via a combination with mass spectrometry.

Many improvements could be applied to the image analysis, library pre-processing, and GPSeq score calculation from the technical perspective. Specifically, the image analysis dedicated `pygpseq` package should be updated to utilize more standardized formats (e.g., OME-TIFF) and a more robust object-oriented programming architecture for a more flexible and easily usable resource. We are currently implementing a new version of the package (called `radiantkit`) available on GitHub (https://github.com/ggirelli/radiantkit). Recently, we also implemented a `data.table`-based R script, called `GPSeq-RadiCal`[190], which replaces the `gpseqc`[186] package while providing the same functionalities in a faster and more memory-efficient manner.

Furthermore, we envision establishing novel and more reproducible GPSeq pre-processing and iFISH-based probe design pipelines by relying on cutting-edge containerization and workflow management tools like Docker[191] and Nextflow[192].

In conclusion, we believe that, while leaving space for further improvements, the current versions of both iFISH and GPSeq methods represent crucial tools for the community to study the forces that shape the genome architecture thanks to their ease of use, flexibility, and extensive applicability.

Understanding one hundred percent of everything is impossible. That's why we spend all our lives trying to understand the thinking of others. That's what makes life so interesting.

Ryoji Kaji

This thesis was written in LyX (v2.3.4.2)
on a GK68XS Epomaker Keyboard,
customized with Kailh Box Jade thick click switches.

Acknowledgments

This thesis would have never seen the light of day if not for the support from a great number of great people. As the Ph.D. journey is a long one (from 2015/2016 to 2021!), I will do my best to mention all the friends I shared good times with, which kept me going through the hard ones. If we spent time together and you made me laugh, know that you managed to achieve an almost impossible feat! And if I do not mention you here, it is simply because it would take another book to write you all down. Thank you all from the bottom of my heart!

First and foremost, I want to thank my main supervisor **Magda Bienko**. I do not know how to express how grateful I am for the opportunity to join BiCro lab, and your support and help over the years. Thanks to that, I improved as a scientist and a person and reached this huge achievement. Thinking back, now, at how everything started, it feels a bit like destiny played a part in you meeting Chiara at a summer school after reading my application.

Then, I would like to thank my co-supervisor **Marianne Farnebo**. I appreciated our meetings, and I want you to know that knowing that you were always available in case of need was extremely helpful. I am grateful for that.

I would like to thank all examination board members, **Anita Göndör, Sten Linnarson, Ilaria Testa**, and my thesis opponent **Juan Manual Vaquerizas** for the time and energy spent reading the thesis and discussing it with me on the day of the defense. I also thank the defense chairperson **Simon Elsässer** for his support and help in hosting the defense.

Of course, I have to thank everyone from **BiCroLab** that accompanied me on this adventure over the years. First and foremost, my second group leader, **Nicola Crosetto**. Thank you for your support, expertise, and always spot-on feedback and insights.

Then, I want to thank the people who shared the adventure of publishing in the competitive research environment, my co-first authors **Quim**, **TK**, and **Eleni**. We really shared the joys and pains of going through manuscript writing and rounds of revision. In particular, I want to thank **TK** for his never-ending sarcasm and for sharing great times together in front of pizzas, dumplings, and much more. **RezaMi**, thank you for your constant support and positive attitude. Even during tough times, you always gave me the energy to hold on and push forward! **Silvano**, you never said no when I was looking for some help with statistics or some math. Thank you. **Erik**, for being so supportive and such a person to learn from, for sharing music, books, and chunks of sourdough :) **Xiaolu** for being my Chinese-language lǎoshī, and never complaining when I was clumsily trying to put a couple of Chinese words together. **Michi**,

for all the good times spent together, board games, beers, and chats! You are almost here; hang in there! ;) **BB**, my mentor and stroopwafel provider. Thank you for your support and for always being willing to listen to my rants. And thank you for trying to teach me how to correctly say [ay]! **Maud**, my favorite French person! And **Su**, it's funny to see how small the world is and that we ended up working together after meeting briefly during iGEM years before. Thank you both for your constant support, positive energy, and all the good times spent together! We need to organize another trip once the world is back on its feet! :D **Masa-kun**, arigatō gozaimasu for your support and for sharing good times and the iFISH journey together. I will never forget celebrating Hanami together in Kungsträdgården! **Roberto**, for the board game nights at SciLifePub, oh how I miss them! **Ana, Emma, Fede, Luuk** (or is it **Look**?), **Hin, Eleonora, Solrun, Ning, Katta, Xiaoze**, thank you for being such amazing colleagues. Working with you is always a pleasure! I also want to thank all the visiting students that I had the pleasure to meet and spend some time together over the years, to mention a few: **Olivia, Pontus, Julie, Theresa, Diana, Carla, Jesko, Merula**, and **Thu**!

Now that I am on a roll, I want to thank the people that we shared an open-office/laboratory with: **Banu, Dörte, Angelo, Anna-Maria, Kyle**, and everyone else from the **Elsässer lab**. I also want to thank everyone I had the pleasure of working with or interact with at SciLifeLab. Especially **Ann-Sofie, Valeria**, and **Dimitris** for their positive energy, **Jaime** for the lovely chats about books, the **SciLifePub crew** for all the good times organized together, and the **Café Delta ladies** for the fantastic chats and for doing your best in providing my favorite ice cream :)

Then I want to thank all the people I had the pleasure to meet and to call friends here, in the lands north of the Wall.

First and foremost, I must thank the creek of "***disagio e purè***": **Pede**, it is truly thanks to you that I got here; **Andrea and Kiki**, amazing friends, always supportive, and together with me catalysts and dispellers of "disagio"; **Beatroce and Denisio**, I miss you both so much, especially the big laughs we had together.

Then, I cannot forget to thank the best family ever: the ***Larsberg Familia***! Thank you **Luisa, Lorenzo a.k.a. DjAmmooore, Leo** and **Mirco**. We had such great times together, from cooking to chess matches, from parties to sketching together. It's so awesome to still keep in touch and update each other on the latest things happening. Miss you all!

The ***caxxos***, my big crew of friends, I would not have survived to see this Ph.D. thesis without you all. I thank you from the bottom of my heart for your constant support, even during the more challenging times. **Anna**, it is funny to think about the first time we met at KTH to play a camel-something board game with the other Erasmus students. How much our friendship has grown. **Marco**, the biggest Dragonball fan I ever met, thanks for the time spent together on the Dolomites, the chess matches, and the long chats on anime & co. **Nuria** for the chess matches (have I ever won?) and introducing me to great TV series :D **Maria**, thanks for the great times and for "*precipitevolissimevolmente*"!! **Joep, Jemina, Esther, Alice & Ale,**

Harpa, Shane, Alek, and all the people that were part of the crew but left for new adventures. I am so bad at writing this down, as I cannot really express how truly and deeply grateful I am for having met you and having the luck to live so many adventures together. I think we could fill many books with everything we have done, and I hope many more for future adventures!

To the whole ***DSA 2020 team***, a huge huge thank you. Although we mainly met digitally due to the circumstances, you were truly a fantastic team to work with. I really enjoyed writing the minutes of our discussion, even though often I would joke about it. Thank you, **Sebastian, Leonie, Valeria, Yujiao, Abishek, Jan, Yildiz, Amineh**, and everyone else involved in our activities!

Thank you, **Luca & Sara**, for your support, help, friendship, and simply for being such amazing people! I can't believe I was so lucky to meet you and to be able to call you my friends.

Thank you to **Ilias, Stefania**, **Alba**, and the rest of the pub quiz crew. You have been such a great group of people to spend some time and take a break with in the past few months! I would like to thank the whole ***kickboxing crew***, especially **Luis** and **Nancy**. The SSSIF lectures were the perfect way to get rid of stress, exercise, and laugh together. **Xinge & Chen**, for your support and friendship. I still hold dear that fantastic hotpot we had together; I hope we can do it again sometime soon! To **Ida, Prem, Adeline**, and **Alex**, thank you for your friendship, all the adventures, and all the good memories. I would not have survived the first half of the Ph.D. endeavor without your support! Thank you **Federica, Jerry, Denis**, and **Stefano**. You were one of the first groups of friends I met here in Sweden; thank you for all the good times. I will never forget us dancing at the sound of "andiamo a comandare" with you in Helsinki!

I want to conclude Stockholm's friends' part by thanking **Mirco** especially. It is difficult to find a part of my life in Stockholm or an adventure that I did not have the pleasure to share with you. I just want to thank you for your friendship, support, occasional small fights, and for everything you have done over the years. It is not easy to find a great person like you and to be able to call them friends. I really feel fortunate to be able to do that with you.

Okay, now it's time to switch gear.

Un grazie dal profondo del cuore ai miei confratelli e consorelle di ***B.I.R.***: **Elisa**, la mia "cuginetta", **Paco** e **Pingu**, i miei fratelli "from another mother and father", **Quiri** e **Chiara**. Chi dalla nascita, chi dalle superiori, chi dall'Università, mi avete accompagnato fino a questo momento. Senza il vostro supporto e la vostra amicizia non sarei arrivato fino a questo punto. **Eliå**, **Ilaria** & **Moin**, sono così felice di avervi conosciuti e di aver avuto la possibilità di ospitarvi qua a Stoccolma. Non vedo l'ora di un'altra avventura insieme. Grazie **Andrea** (**Leo**) per un'amicizia cominciata in quel di Trento e continuata anche in queste fredde lande del Nord.

Cere, **Gine**, **Toso**, **Isy**, e **Mara**, non sapete quanto mi mancate; vorrei esservi fisicamente vicino più spesso. Vi considero più di amici, siete ormai di famiglia. Senza di voi molte

cose sarebbero andate diversamente e non sarei arrivato a questo traguardo. Grazie per il supporto, l'aiuto, i consigli, e l'ascolto. **Matteo**, questa tesi è dedicata *in primi*s a te. Quasi non riesco a ricordare la vita senza la tua amicizia. **Danijela**, dopo una pausa fin troppo lunga abbiamo riallacciato l'amicizia. Non so esprimere quanto felice sia di questo. Non vedo l'ora ti poterci rivedere di persona. **Manuel**, sei un grandissimo amico e vorrei poter essere in zona più spesso. **Carla**, la mia sorellona, amica, insegnante. Grazie per la tua amicizia e costante supporto.

Anna e Nicolò. Cosa potrei scrivere? Non sarebbe mai all'altezza degli amici che ho trovato in voi. Questa tesi è dedicata *in primis* anche a voi. Non so come facciate ancora a sopportarmi, ma vi ringrazio dal profondo del cuore per tutto quello che avete fatto per me. E per avermi fatto conoscere **Silvia**, **Claudio**, **Matteo**, **Max & Ste**, e **Elisa & Christian**!

Inoltre, vorrei ringraziare nuovamente **Francesco** e **Anna** per l'aiuto nella revisione del riassunto a fini divulgativi incluso all'inizio della tesi.

Lore, **PP**, e **Lavi**: non avrei mai osato immaginare che un tirocinio di due mesi mi avrebbe fatto conoscere degli amici splendidi come voi. Spero riusciremo a organizzare una rimpatriata il prima possibile! **Fracca** e **Lia**, uno dall'altra parte del mondo e l'altra nel centro del continente, grazie per il vostro supporto e la vostra amicizia. **Famiglie Antonini**, **Menini**, **Ceradini**, e **Guarnati**. Praticamente delle seconde famiglie. Grazie per il supporto durante questi anni.

A **Nella** e **Nora**, due persone semplicemente fantastiche, venute purtroppo a mancare nei mesi antecedenti alla difesa di questa tesi. Mi avete conosciuto fin dall'infanzia e siete state come parte della famiglia, una zia e una nonna adottive. Non vi dimenticherò mai.

Annalisa, **Francesco**, e **Ilaria**, la miglior famiglia che si potesse chiedere. Nonostante i bisticci, i pasticci, i dispetti. Sí, potremmo fare una sitcom basata su di noi. Vi voglio bene e vi ringrazio per l'infinito amore e supporto che ho sempre sentito da parte vostra e nei vostri confronti. **Nonno Guido e Luisa**, **Zia Eleonora e Claudio**, **Marco**, **Michele**, **Zio Tano e Barbara**, **Zio Alfonso e Francesca**, **Zii Angelo e Bruno**, tutti i cugini e parenti. Grazie per essere parte della mia famiglia e per il vostro supporto.

Finally, I would like to thank everyone who helped me or supported me during the years that led to making this Ph.D. thesis a reality. Thank you from the bottom of my heart.

And, as I concluded my MS thesis, I will also end this thesis.

I survived.

Bibliography

[1] Eleni Gelali, Gabriele Girelli, Masahiro Matsumoto, Erik Wernersson, Joaquin Custodio, Ana Mota, Maud Schweitzer, Katalin Ferenc, Xinge Li, Reza Mirzazadeh, Federico Agostini, John P. J.P. Schell, Fredrik Lanner, Nicola Crosetto, and Magda Bienko. iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. *Nature Communications*, 10(1), dec 2019. ISSN 20411723. doi: 10.1038/s41467-019-09616-w.

[2] Gabriele Girelli, Joaquin Custodio, Tomasz Kallas, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota, Solrun Kolbeinsdottir, Eleni Gelali, Nicola Crosetto, and Magda Bienko. GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nature Biotechnology*, 38(10):1184–1193, 2020. ISSN 15461696. doi: 10.1038/s41587-020-0519-y.

[3] Emil Heitz. *Das heterochromatin der moose*. Bornträger, 1928.

[4] E. Passarge. Emil Heitz and the concept of heterochromatin: Longitudinal chromosome differentiation was recognized fifty years ago. *American Journal of Human Genetics*, 31(2):106–115, 1979. ISSN 00029297.

[5] Frédéric Berger. Emil Heitz, a True Epigenetic Pioneer. *Nature Reviews Molecular Cell Biology*, 20(10): 572, 2019. ISSN 14710080. doi: 10.1038/s41580-019-0170-y.

[6] Horng D. Ou, Sébastien Phan, Thomas J. Deerinck, Andrea Thor, Mark H. Ellisman, and Clodagh C. O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349), 2017. ISSN 0036-8075. doi: 10.1126/science.aag0025.

[7] Shane A. Evans, Jeremy Horrell, and Nicola Neretti. The three-dimensional organization of the genome in cellular senescence and age-associated diseases. *Seminars in Cell and Developmental Biology*, 90: 154–160, 2018. ISSN 10963634. doi: 10.1016/j.semcdb.2018.07.022.

[8] R. D. Goldman, D. K. Shumaker, M. R. Erdos, M. Eriksson, A. E. Goldman, L. B. Gordon, Y. Gruenbaum, S. Khuon, M. Mendez, R. Varga, and F. S. Collins. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. *Proceedings of the National Academy of Sciences*, 101(24):8963–8968, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0402943101.

[9] Abhijit Chakraborty and Ferhat Ay. The role of 3D genome organization in disease: From compartments to single nucleotides. *Seminars in Cell and Developmental Biology*, 90:104–113, 2019. ISSN 10963634. doi: 10.1016/j.semcdb.2018.07.005.

[10] Anton Krumm and Zhijun Duan. Understanding the 3D genome: Emerging impacts on human disease. *Seminars in Cell and Developmental Biology*, 90:62–77, 2019. ISSN 10963634. doi: 10.1016/j.semcdb.2018.07.004.

[11] Omar L. Kantidze, Katerina V. Gurova, Vasily M. Studitsky, and Sergey V. Razin. The 3D Genome as a Target for Anticancer Therapy. *Trends in Molecular Medicine*, 26(2):141–149, feb 2020. ISSN 1471499X. doi: 10.1016/j.molmed.2019.09.011.

[12] Sabine Mai. The three-dimensional cancer nucleus. *Genes Chromosomes and Cancer*, 58(7):462–473, 2019. ISSN 10982264. doi: 10.1002/gcc.22720.

[13] Aishwarya Sivakumar, Jose I. de las Heras, and Eric C. Schirmer. Spatial genome organization: From development to disease. *Frontiers in Cell and Developmental Biology*, 7(MAR):1–12, 2019. ISSN 2296634X. doi: 10.3389/fcell.2019.00018.

[14] Ezgi Süheyla Doğan and Chang Liu. Three-dimensional chromatin packing and positioning of plant genomes. *Nature Plants*, 4(8):521–529, 2018. ISSN 20550278. doi: 10.1038/s41477-018-0199-5.

[15] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. 6th ed. edition, 2014. ISBN 978-0-8153-4453-7.

[16] Elizabeth H. Blackburn and Joseph G. Gall. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *Journal of Molecular Biology*, 120(1):33–53, 1978. ISSN 00222836. doi: 10.1016/0022-2836(78)90294-2.

[17] L Hayflick and PS Moorhead. The Serial Cultivation of Human Diploid Cell Strains. *Experimental Cell Research1*, 25:585–621, 1961.

[18] A. M. Olovnikov. A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *Journal of Theoretical Biology*, 41(1):181–190, 1973. ISSN 10958541. doi: 10.1016/0022-5193(73)90198-7.

[19] Junli Feng, Walter D. Funk, Sy Shi Wang, Scott L. Weinrich, Ariel A. Avilion, Choy Pik Chiu, Robert R. Adams, Edwin Chang, Richard C. Allsopp, Jinghua Yu, Siyuan Le, Michael D. West, Calvin B. Harley, William H. Andrews, Carol W. Greider, and Bryant Villeponteau. The RNA component of human telomerase. *Science*, 269(5228):1236–1241, 1995. ISSN 00368075. doi: 10.1126/science.7544491.

[20] Teresa Szczepińska, Anna Maria Rusek, and Dariusz Plewczynski. Intermingling of chromosome territories. *Genes Chromosomes and Cancer*, 58(7):500–506, 2019. ISSN 10982264. doi: 10.1002/gcc.22736.

[21] Philipp G. Maass, A. Rasim Barutcu, and John L. Rinn. Interchromosomal interactions: A genomic love story of kissing chromosomes. *Journal of Cell Biology*, 218(1):27–38, 2019. ISSN 15408140. doi: 10.1083/jcb.201806052.

[22] T Boveri. Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität. *Archiv für Zellforschung*, 3:181–268, 1909.

[23] Laura Manuelidis. Individual interphase chromosome domains revealed by in situ hybridization. *Human Genetics*, 71(4):288–293, 1985. ISSN 03406717. doi: 10.1007/BF00388453.

[24] Margit Schardin, T. Cremer, H. D. Hager, and M. Lang. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Human Genetics*, 71(4):281–287, 1985. ISSN 03406717. doi: 10.1007/BF00388452.

[25] T. Cremer, P. Lichter, J. Borden, D. C. Ward, and L. Manuelidis. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Human Genetics*, 80(3):235–246, 1988. ISSN 03406717. doi: 10.1007/BF01790091.

[26] D Pinkel, J Landegent, C Collins, J Fuscoet, R Segraves, J Lucas, and J Gray. Fluorescence in situ hybridization with human chromosome-specific libraries: Detection of trisomy 21 and translocations of chromosome 4. *Proc. Nati. Acad. Sci.*, 85:9138–9142, 1988.

[27] Thomas Cremer, Marion Cremer, Steffen Dietzel, Stefan Müller, Irina Solovei, and Stanislav Fakan. Chromosome territories - a functional nuclear landscape. *Current Opinion in Cell Biology*, 18(3):307–316, 2006. ISSN 09550674. doi: 10.1016/j.ceb.2006.04.007.

[28] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):1–22, 2010. ISSN 19430264. doi: 10.1101/cshperspect.a003889.

[29] Carl Rabl. Uber zelltheilung. *Morphol. Jahrb.*, 10:214–330, 1885.

[30] D. E. Comings. Arrangement of chromatin in the nucleus. *Human Genetics*, 53(2):131–143, 1980. ISSN 03406717. doi: 10.1007/BF00273484.

[31] Maxime Pouokam, Brian Cruz, Sean Burgess, Mark R. Segal, Mariel Vazquez, and Javier Arsuaga. The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific Reports*, 9(1):1–10, 2019. ISSN 20452322. doi: 10.1038/s41598-019-42967-4.

[32] Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G.S. S Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, Ernest D. Laue, P Liam, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O Shaughnessy-kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G.S. S Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, and Brian Hendrich. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):1–21, 2017. ISSN 0028-0836. doi: 10.1038/nature21429.

[33] Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X. Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018. ISSN 0036-8075. doi: 10.1126/science.aat5641.

[34] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M a Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid a Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009. ISSN 0036-8075. doi: 10.1126/science.1181369.

[35] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. D Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.11.021.

[36] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015. ISSN 0028-0836. doi: 10.1038/nature14222.

[37] Johanna Gassler, Hugo B Brandão, Maxim Imakaev, Ilya M Flyamer, Sabrina Ladstätter, Wendy A Bickmore, Jan-Michael Peters, Leonid A Mirny, and Kikuë Tachibana. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *The EMBO Journal*, 36(24):3600–3618, 2017. ISSN 0261-4189. doi: 10.15252/embj.201798083.

[38] Gordana Wutz, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, Gregor Jessberger, Matthias Muhar, M Julius Hossain, Nike Walther, Birgit Koch, Moritz Kueblbeck, Jan Ellenberg, Johannes Zuber, Peter Fraser, and Jan-Michael Peters. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, 36(24):3573–3599, 2017. ISSN 0261-4189. doi: 10.15252/embj.201798004.

[39] Judith HI Haarhuis and Benjamin D Rowland. Cohesin: building loops, but not compartments. *The EMBO Journal*, 36(24):3549–3551, 2017. ISSN 0261-4189. doi: 10.15252/embj.201798654.

[40] Siyuan Wang, Jun-Han Su, Brian J. Beliveau, Bogdan Bintu, Jeffrey R. Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf8084.

[41] Teresa R Luperchio, Michael EG Sauria, Xianrong Wong, Marie-Cécile Gaillard, Peter Tsang, Katja Pekrun, Robert A Ach, N Alice Yamada, James Taylor, and Karen Reddy. Chromosome Conformation Paints Reveal The Role Of Lamina Association In Genome Organization And Regulation. *bioRxiv*, page 122226, 2017. doi: 10.1101/122226.

[42] Jean-Philippe Fortin and Kasper D Hansen. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome biology*, 16(1):180, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0741-y.

[43] Xiaobin Zheng and Yixian Zheng. CscoreTool: Fast Hi-C compartment analysis at high resolution. *Bioinformatics*, 34(9):1568–1570, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/btx802.

[44] Kyle Xiong and Jian Ma. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications*, 10(1), dec 2019. ISSN 20411723. doi: 10.1038/s41467-019-12954-4.

[45] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. ISSN 0028-0836. doi: 10.1038/nature11082.

[46] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. Van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012. ISSN 00280836. doi: 10.1038/nature11049.

[47] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4), 2019. ISSN 23752548. doi: 10.1126/sciadv.aaw1668.

[48] Jesse R. Dixon, David U. Gorkin, and Bing Ren. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, 62(5):668–680, 2016. ISSN 10972765. doi: 10.1016/j.molcel.2016.05.018.

[49] M. Jordan Rowley and Victor G. Corces. Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 19:789, 2018. ISSN 1471-0056. doi: 10.1038/s41576-018-0060-8.

[50] Johannes Nuebler, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A. Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29):E6697–E6706, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1717730115.

[51] Kyle P. Eagen. Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences*, 43(6):469–478, 2018. ISSN 13624326. doi: 10.1016/j.tibs.2018.03.006.

[52] Jacques Serizay and Julie Ahringer. Genome organization at different scales: nature, formation and function. *Current Opinion in Cell Biology*, 52:145–153, 2018. ISSN 18790410. doi: 10.1016/j.ceb.2018.03.009.

[53] Eileen E. M. Furlong and Michael Levine. Developmental enhancers and chromosome topology. *Science*, 361(6409):1341–1345, 2018. ISSN 0036-8075. doi: 10.1126/science.aau0320.

[54] Yun Li, Ming Hu, and Yin Shen. Gene regulation in the 3D genome. *Human Molecular Genetics*, 27(R2):R228–R233, 2018. ISSN 0964-6906. doi: 10.1093/hmg/ddy164.

[55] Alexandra Despang, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerkovic, Christina Paliou, Wing-Lee Chan, Bernd Timmermann, Lars Wittler, Martin Vingron, Stefan Mundlos, and Daniel M. Ibrahim. Functional dissection of TADs reveals non-essential and instructive roles in regulating gene expression. *bioRxiv*, page 566562, 2019. doi: 10.1101/566562.

[56] Dusan Racko, Fabrizio Benedetti, Julien Dorier, and Andrzej Stasiak. Are TADs supercoiled? *Nucleic Acids Research*, 47(2):521–532, 2019. ISSN 13624962. doi: 10.1093/nar/gky1091.

[57] Ruifeng Li, Yifang Liu, Tingting Li, and Cheng Li. 3Disease Browser: A Web server for integrating 3D genome and disease-associated chromosome rearrangement data. *Scientific Reports*, 6(October):1–11, 2016. ISSN 15327361. doi: 10.1016/j.surg.2015.02.024.

[58] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161:1012–1025, 2015. ISSN 10974172. doi: 10.1016/j.cell.2015.04.004.

[59] Martin Franke, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, 2016. ISSN 0028-0836. doi: 10.1038/nature19800.

[60] David Umlauf and Raphaël Mourad. The 3D genome: From fundamental principles to disease and cancer. *Seminars in Cell and Developmental Biology*, 2018. ISSN 10963634. doi: 10.1016/j.semcdb.2018.07.002.

[61] Philippe Collas, Tharvesh M. Liyakat Ali, Annaël Brunet, and Thomas Germier. Finding Friends in the Crowd: Three-Dimensional Cliques of Topological Genomic Domains. *Frontiers in Genetics*, 10(June):1–11, 2019. doi: 10.3389/fgene.2019.00602.

[62] Neva C. Durand, Muhammad S. Shamim, Ido Machol, Suhas S P Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, 3(1):95–98, 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.07.002.

[63] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9(1), may 2014. ISSN 17487188. doi: 10.1186/1748-7188-9-14.

[64] Caleb Weinreb and Benjamin J. Raphael. Identification of hierarchical chromatin domains. *Bioinformatics*, 32(11):1601–1609, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btv485.

[65] François Serra, Davide Baù, Mike Goodstadt, David Castillo, Guillaume Filion, and Marc A. Marti-Renom. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Computational Biology*, 13(7):1–17, 2017. ISSN 15537358. doi: 10.1371/journal.pcbi.1005665.

[66] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, 14(7):679–685, 2017. ISSN 15487105. doi: 10.1038/nmeth.4325.

[67] Bas van Steensel and Eileen E.M. Furlong. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology*, 20(6):327–337, 2019. ISSN 14710080. doi: 10.1038/s41580-019-0114-6.

[68] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1518552112.

[69] Kevin Van Bortle, Michael H. Nichols, Li Li, Chin Tong Ong, Naomi Takenaka, Zhaohui S. Qin, and Victor G. Corces. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome biology*, 15(6):R82, 2014. ISSN 1474760X. doi: 10.1186/gb-2014-15-5-r82.

[70] Artem V. Luzhin, Ilya M. Flyamer, Ekaterina E. Khrameeva, Sergey V. Ulianov, Sergey V. Razin, and Alexey A. Gavrilov. Quantitative differences in TAD border strength underly the TAD hierarchy in Drosophila chromosomes. *Journal of Cellular Biochemistry*, 120(3):4494–4503, 2019. ISSN 10974644. doi: 10.1002/jcb.27737.

[71] Quentin Szabo, Daniel Jost, Jia Ming Chang, Diego I. Cattoni, Giorgio L. Papadopoulos, Boyan Bonev, Tom Sexton, Julian Gurgo, Caroline Jacquier, Marcelo Nollmann, Frédéric Bantignies, and Giacomo Cavalli. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, (4):eaar8082, 2018. ISSN 23752548. doi: 10.1126/sciadv.aar8082.

[72] Zhenhua Zhu and Xiangdong Wang. Roles of cohesin in chromosome architecture and gene expression. *Seminars in Cell and Developmental Biology*, 90(August 2018):187–193, 2019. ISSN 10963634. doi: 10.1016/j.semcdb.2018.08.004.

[73] K. Nasmyth. Disseminating the genome: Joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annual Review of Genetics*, 35:673–745, 2001. ISSN 00664197. doi: 10.1146/annurev.genet.35.102401.091334.

[74] Edward J. Banigan and Leonid A. Mirny. Limits of chromosome compaction by loop-extruding motors. *Physical Review X*, 9(3):031007(10), 2019. doi: 10.1103/PhysRevX.9.031007.

[75] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo. Extrusion without a motor: A new take on the loop extrusion model of genome organization. *Nucleus*, 9 (1):95–103, 2018. ISSN 19491042. doi: 10.1080/19491034.2017.1421825.

[76] Robert V. Skibbens. Condensins and cohesins – one of these things is not like the other! *Journal of Cell Science*, 132(3):jcs220491, 2019. ISSN 0021-9533. doi: 10.1242/jcs.220491.

[77] Sarah G. Swygert, Seungsoo Kim, Xiaoying Wu, Tianhong Fu, Tsung Han Hsieh, Oliver J. Rando, Robert N. Eisenman, Jay Shendure, Jeffrey N. McKnight, and Toshio Tsukiyama. Condensin-Dependent Chromatin Compaction Represses Transcription Globally during Quiescence. *Molecular Cell*, 73(3):533–546.e4, 2019. ISSN 10974164. doi: 10.1016/j.molcel.2018.11.020.

[78] Mahipal Ganji, Indra A. Shaltiel, Shveta Bisht, Eugene Kim, Ana Kalichava, Christian H. Haering, and Cees Dekker. Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384):102–105, 2018. ISSN 10959203. doi: 10.1126/science.aar7831.

[79] Diane C. Wang, William Wang, Linlin Zhang, and Xiangdong Wang. A tour of 3D genome with a focus on CTCF. *Seminars in Cell and Developmental Biology*, 90:4–11, 2019. ISSN 10963634. doi: 10.1016/j.semcdb.2018.07.020.

[80] Matthew T. Mawhinney, Runcong Liu, Fang Lu, Jasna Maksimoska, Kevin Damico, Ronen Marmorstein, Paul M. Lieberman, and Brigita Urbanc. CTCF-Induced Circular DNA Complexes Observed by Atomic Force Microscopy. *Journal of Molecular Biology*, 430(6):759–776, 2018. ISSN 10898638. doi: 10.1016/j.jmb.2018.01.012.

[81] Aleksandra Pekowska, Bernd Klaus, Wanqing Xiang, Jacqueline Severino, Nathalie Daigle, Felix A Klein, Malgorzata Oleś, Rafael Casellas, Jan Ellenberg, Lars M Steinmetz, Paul Bertone, and Wolfgang Huber. Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency. *Cell Systems*, 7: 482–495, 2018. doi: 10.1016/j.cels.2018.09.003.

[82] Elphège P. Nora, Anton Goloborodko, Anne Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169(5): 930–944.e22, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.05.004.

[83] Suhas S.P. Rao, Su Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong Rim Kieffer-Kwon, Adrian L. Sanborn, Sarah E. Johnstone, Gavin D. Bascom, Ivan D. Bochkov, Xingfan Huang, Muhammad S. Shamim, Jaeweon Shin, Douglass Turner, Ziyi Ye, Arina D. Omer, James T. Robinson, Tamar Schlick, Bradley E. Bernstein, Rafael Casellas, Eric S. Lander, and Erez Lieberman Aiden. Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2):305–320.e24, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.09.026.

[84] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, Wolfgang Huber, Christian H. Haering, Leonid Mirny, and Francois Spitz. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551 (7678):51–56, 2017. ISSN 14764687. doi: 10.1038/nature24281.

[85] Leonid A. Mirny, Maxim Imakaev, and Nezar Abdennur. Two major mechanisms of chromosome organization. *Current Opinion in Cell Biology*, 58:142–152, jun 2019. ISSN 18790410. doi: 10.1016/j.ceb.2019.05.001.

[86] Yuki Ogiyama, Bernd Schuettengruber, Giorgio L Papadopoulos, Jia-ming Chang, Giacomo Cavalli, Yuki Ogiyama, Bernd Schuettengruber, Giorgio L Papadopoulos, Jia-ming Chang, and Giacomo Cavalli. Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Article Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Molecular Cell*, 71(1):73–88.e5, 2018. ISSN 1097-2765. doi: 10.1016/j.molcel.2018.05.032.

[87] William W. Greenwald, He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Erin N. Smith, and Kelly A. Frazer. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications*, 10(1):1–17, 2019. ISSN 20411723. doi: 10.1038/s41467-019-08940-5.

[88] Jayoung Ryu, Hyunwoong Kim, Dongchan Yang, Andrew J. Lee, and Inkyung Jung. A new class of constitutively active super-enhancers is associated with fast recovery of 3D chromatin loops. *BMC Bioinformatics*, 20(Suppl 3), 2019. ISSN 14712105. doi: 10.1186/s12859-019-2646-3.

[89] Emanuela V Volpi, Edith Chevret, Tania Jones, Racist Vatcheva, Jill Williamson, Stephan Beck, R Duncan Campbell, Michelle Goldsworthy, Stephen H Powis, Jannis Ragoussis, John Trowsdale, and Denise Sheer. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *Journal of Cell Science*, 113:1565–1576, 2000.

[90] Jiazhi Hu, Yu Zhang, Lijuan Zhao, Richard L Frock, Zhou Du, Robin M Meyers, Fei-long Meng, David G Schatz, and Frederick W Alt. Chromosomal Loop Domains Direct the Recombination of Antigen Receptor Genes. *Cell*, 163(4):947–959, 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.10.016.

[91] Irina Solovei, Katharina Thanisch, and Yana Feodorova. How to rule the nucleus: divide et impera, 2016. ISSN 18790410.

[92] Irina Solovei, Moritz Kreysing, Christian Lanctôt, Süleyman Kösem, Leo Peichl, Thomas Cremer, Jochen Guck, and Boris Joffe. Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell*, 137(2):356–368, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.01.052.

[93] Ada L. Olins, Monika Zwerger, Harald Herrmann, Hanswalter Zentgraf, Amos J. Simon, Marc Monestier, and Donald E. Olins. The human granulocyte nucleus: Unusual nuclear envelope and heterochromatin composition. *European Journal of Cell Biology*, 87(5):279–290, 2008. ISSN 01719335. doi: 10.1016/j.ejcb.2008.02.007.

[94] Yina Zhu, Ke Gong, Matthew Denholtz, Vivek Chandra, Mark P. Kamps, Frank Alber, and Cornelis Murre. Comprehensive characterization of neutrophil genome topology. *Genes and Development*, 31(2):141–153, 2017. ISSN 15495477. doi: 10.1101/gad.293910.116.

[95] Sergey V. Ulianov, Semen A. Doronin, Ekaterina E. Khrameeva, Pavel I. Kos, Artem V. Luzhin, Sergei S. Starikov, Aleksandra A. Galitsyna, Valentina V. Nenasheva, Artem A. Ilyin, Ilya M. Flyamer, Elena A. Mikhaleva, Mariya D. Logacheva, Mikhail S. Gelfand, Alexander V. Chertovich, Alexey A. Gavrilov, Sergey V. Razin, and Yuri Y. Shevelyov. Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nature Communications*, 10(1):1–11, 2019. ISSN 20411723. doi: 10.1038/s41467-019-09185-y.

[96] Yuri Y. Shevelyov and Sergey V. Ulianov. The Nuclear Lamina as an Organizer of Chromosome Architecture. *Cells*, 8(2):136, 2019. doi: 10.3390/cells8020136.

[97] Brian Burke and Colin L. Stewart. The nuclear lamins: Flexibility in function. *Nature Reviews Molecular Cell Biology*, 14(1):13–24, 2013. ISSN 14710072. doi: 10.1038/nrm3488.

[98] Jop Kind, Ludo Pagie, Havva Ortabozkoyun, Shelagh Boyle, Sandra S. De Vries, Hans Janssen, Mario Amendola, Leisha D. Nolen, Wendy a. Bickmore, and Bas Van Steensel. Single-cell dynamics of genome-nuclear lamina interactions. *Cell*, 153(1):178–192, 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.02.028.

[99] Jop Kind, Ludo Pagie, Sandra S. de Vries, Leila Nahidiazar, Siddharth S. Dey, Magda Bienko, Ye Zhan, Bryan Lajoie, Carolyn A. De Graaf, Mario Amendola, Geoffrey Fudenberg, Maxim Imakaev, Leonid A. Mirny, Kees Jalink, Job Dekker, Alexander Van Oudenaarden, and Bas van Steensel. Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell*, 163(1):1–14, sep 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.08.040.

[100] Jop Kind and Bas van Steensel. Genome-nuclear lamina interactions and gene regulation. *Current Opinion in Cell Biology*, 22(3):320–325, 2010. ISSN 09550674. doi: 10.1016/j.ceb.2010.04.002.

[101] Wouter Meuleman, Daan Peric-hupkes, Jop Kind, Jean-bernard Beaudry, Ludo Pagie, Manolis Kellis, Marcel Reinders, Lodewyk Wessels, and Bas Van Steensel. Constitutive nuclear lamina – genome interactions are highly conserved and associated with A / T-rich sequence. *Genome research*, pages 270–280, 2013. ISSN 1549-5469. doi: 10.1101/gr.141028.112.

[102] Xiaobin Zheng, Jiabiao Hu, Sibiao Yue, Lidya Kristiani, Miri Kim, Michael Sauria, James Taylor, Youngjo Kim, and Yixian Zheng. Lamins Organize the Global Three-Dimensional Genome from the Nuclear Periphery. *Molecular Cell*, 71(5):802–815.e7, 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.05.017.

[103] Andrey Poleshko, Parisha P. Shah, Mudit Gupta, Apoorva Babu, Michael P. Morley, Lauren J. Manderfield, Jamie L. Ifkovits, Damelys Calderon, Haig Aghajanian, Javier E. Sierra-Pagán, Zheng Sun, Qiaohong Wang, Li Li, Nicole C. Dubois, Edward E. Morrisey, Mitchell A. Lazar, Cheryl L. Smith, Jonathan A. Epstein, and Rajan Jain. Genome-Nuclear Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. *Cell*, 171(3):573–587.e14, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.09.018.

[104] Kelvin See, Yemin Lan, Joshua Rhoades, Rajan Jain, Cheryl L. Smith, and Jonathan A. Epstein. Lineage-specific reorganization of nuclear peripheral heterochromatin and H3K9me2 domains. *Development*, 146 (3):dev174078, 2019. ISSN 0950-1991. doi: 10.1242/dev.174078.

[105] Silke J.A. Lochs, Samy Kefalopoulou, and Jop Kind. Lamina Associated Domains and Gene Regulation in Development and Cancer. *Cells*, 8(3):271, 2019. doi: 10.3390/cells8030271.

[106] Daniele Zink, Margarida D. Amaral, Andreas Englmann, Susanne Lang, Luka A. Clarke, Carsten Rudolph, Felix Alt, Kathrin Luther, Carla Braz, Nicolas Sadoni, Joseph Rosenecker, and Dirk Schindelhauer. Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei. *Journal of Cell Biology*, 166(6):815–825, 2004. ISSN 00219525. doi: 10.1083/jcb.200404107.

[107] Laura Pascual-reguant, Enrique Blanco, Silvia Galan, François Dily, Yasmina Cuartero, Gemma Serrabardenys, Valerio Carlo, Ane Iturbide, Joan Pau Cebrià-costa, Lara Nonell, Antonio García Herreros, Luciano Croce, Marc A Marti-renom, and Sandra Peiró. Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. *Nature Communications*, (2018). ISSN 2041-1723. doi: 10.1038/s41467-018-05912-z.

[108] Frida Forsberg, Annaël Brunet, Tharvesh M.Liyakat Ali, and Philippe Collas. Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. *Nucleus*, 10(1):7–20, 2019. ISSN 19491042. doi: 10.1080/19491034.2019.1570810.

[109] A. F. Naylor, D. Warburton, and F. E. Warburton. Spatial relations of human chromosomes identified by quinacrine fluorescence at metaphase. *Human Genetics*, 18(4):307–313, 1973. ISSN 0340-6717. doi: 10.1007/bf00291127.

[110] Luc Hens, Micheline Kirsch-Volders, Luc Verschaeve, and Charles Susanne. The central localization of the small and early replicating chromosomes in human diploid metaphase figures. *Human Genetics*, 60 (3):249–256, 1982. ISSN 03406717. doi: 10.1007/BF00303012.

[111] C. Wollenberg, M. P. Kiefaber, and K. D. Zang. Quantitative studies on the arrangement of human metaphase chromosomes. IX. Arrangement of chromosomes with and without spindle apparatus. *Human Genetics*, 62(4):310–315, 1982. ISSN 03406717. doi: 10.1007/BF00304545.

[112] Jenny A Croft, Joanna M Bridger, Shelagh Boyle, Paul Perry, Peter Teague, and Wendy A Bickmore. Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *J Cell Biol*, 145(6):1119–1131, 1999. ISSN 0021-9525, 1540-8140. doi: 10.1083/jcb.145.6.1119.

[113] Shelagh Boyle, Susan Gilchrist, Joanna M Bridger, Nicola L Mahy, Juliet A Ellis, and Wendy A Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Technical Report 3, 2001.

[114] Hideyuki Tanabe, Felix A. Habermann, Irina Solovei, Marion Cremer, and Thomas Cremer. Non-random radial arrangements of interphase chromosome territories: Evolutionary considerations and functional implications. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 504(1-2): 37–45, 2002. ISSN 00275107. doi: 10.1016/S0027-5107(02)00077-5.

[115] Jeffrey M. Craig and Wendy A. Bickmore. The distribution of CpG islands in mammalian chromosomes. *Nature Genetics*, 7(3):376–382, 1994. ISSN 15461718. doi: 10.1038/ng0794-376.

[116] Nick Gilbert, Susan Gilchrist, and Wendy A. Bickmore. Chromatin organization in the mammalian nucleus. In *International Review of Cytology*, volume 242, page 283. 2004. ISBN 9780123646460. doi: 10.1016/S0074-7696(04)42007-5.

[117] Nicolas Sadoni, Sabine Langer, Christine Fauth, Giorgio Bernardi, Thomas Cremer, Bryan M. Turner, and Daniele Zink. Nuclear organization of mammalian genomes: Polar chromosome territories build up functionally distinct higher order compartments. *Journal of Cell Biology*, 146(6):1211–1226, 1999. ISSN 00219525. doi: 10.1083/jcb.146.6.1211.

[118] S. Goetze, J. Mateos-Langerak, H. J. Gierman, W. de Leeuw, O. Giromus, M. H. G. Indemans, J. Koster, V. Ondrej, R. Versteeg, and R. van Driel. The Three-Dimensional Structure of Human Interphase Chromosomes Is Related to the Transcriptome Map. *Molecular and Cellular Biology*, 27(12):4475–4487, 2007. ISSN 0270-7306. doi: 10.1128/MCB.00208-07.

[119] Katrin Küpper, Alexandra Kölbl, Dorothee Biener, Sandra Dittrich, Johann von Hase, Tobias Thormeyer, Heike Fiegler, Nigel P. Carter, Michael R. Speicher, Thomas Cremer, and Marion Cremer. Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma*, 116(3):285–306, 2007. ISSN 00095915. doi: 10.1007/s00412-007-0098-4.

[120] Wendy A. Bickmore. The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics*, 2013. ISSN 1527-8204. doi: 10.1146/annurev-genom-091212-153515.

[121] Takumi Takizawa, Prabhakar R. Gudla, Liying Guo, Stephan Lockett, and Tom Misteli. Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker GFAP. *Genes & Development*, 22 (4):489–498, 2008. ISSN 08909369. doi: 10.1101/gad.1634608.

[122] Takumi Takizawa, Karen J. Meaburn, and Tom Misteli. The Meaning of Gene Positioning. *Cell*, 135(1): 9–13, 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.09.026.

[123] Pierre Therizols, Robert S Illingworth, Celine Courilleau, Shelagh Boyle, Andrew J Wood, and Wendy A Bickmore. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, 346(6214):1238–1242, 2014. doi: 10.1126/science.1259587.

[124] Longzhi Tan, Dong Xing, Nicholas Daley, and X. Sunney Xie. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nature Structural & Molecular Biology*, apr 2019. ISSN 1545-9993. doi: 10.1038/s41594-019-0205-2.

[125] Jeffrey R. Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P. Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):201612826, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1612826113.

[126] J. R. Moffitt and X. Zhuang. *RNA Imaging with Multiplexed Error-Robust Fluorescence in Situ Hybridization (MERFISH)*, volume 572. Elsevier Inc., 1 edition, 2016. ISBN 9780128022924. doi: 10.1016/bs.mie.2016.03.020.

[127] Xiaowei Zhuang. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature Methods*, 18(1):18–22, 2021. doi: 10.1038/s41592-020-01037-8.

[128] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10): 987–990, 2019. ISSN 15487105. doi: 10.1038/s41592-019-0548-y.

[129] Ludvig Larsson, Jonas Frisén, and Joakim Lundeberg. Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods*, 18(1):15–18, 2021. ISSN 15487105. doi: 10.1038/s41592-020-01038-7.

[130] Iain Williamson, Soizik Berlivet, Ragnhild Eskeland, Shelagh Boyle, Robert S. Illingworth, Denis Paquette, Joséee Dostie, Wendy A. Bickmore, Soizik Berlivet, Denis Paquette, and Joséee Dostie. Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes and Development*, 28(24):2778–2791, 2014. ISSN 1549-5477. doi: 10.1101/gad.251694.114.

[131] Rieke Kempfer and Ana Pombo. Methods for mapping 3D chromosome architecture, apr 2020. ISSN 14710064.

[132] Kashif Ahmed, Hesam Dehghani, Peter Rugg-Gunn, Eden Fussner, Janet Rossant, and David P. Bazett-Jones. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS ONE*, 5(5), 2010. ISSN 19326203. doi: 10.1371/journal.pone.0010531.

[133] Scott A. Tomlins, Daniel R. Rhodes, Sven Perner, Saravana M. Dhanasekaran, Rohit Mehra, Xiao Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, Charles Lee, James E. Montie, Rajal B. Shah, Kenneth J. Pienta, Mark A. Rubin, and Arul M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005. ISSN 00368075. doi: 10.1126/science.1117679.

[134] Alistair N. Boettiger, Bogdan Bintu, Jeffrey R. Moffitt, Siyuan Wang, Brian J. Beliveau, Geoffrey Fudenberg, Maxim Imakaev, Leonid A. Mirny, Chao Ting Wu, and Xiaowei Zhuang. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586):418–422, 2016. ISSN 14764687. doi: 10.1038/nature16496.

[135] Eleni Gelali, Joaquin Custodio, Gabriele Girelli, Erik Wernersson, Nicola Crosetto, and Magda Bienko. An Application-Directed, Versatile DNA FISH Platform for Research and Diagnostics. In *CpG Islands*, pages 303–333. Humana Press, New York, NY, 2018. doi: 10.1007/978-1-4939-7768-0_17.

[136] J. Roohi, M. Cammer, C. Montagna, and E. Hatchwell. An improved method for generating BAC DNA suitable for FISH. *Cytogenetic and Genome Research*, 121(1):7–9, 2008. ISSN 14248581. doi: 10.1159/000124374.

[137] Brian J. Beliveau, Alistair N. Boettiger, Maier S. Avendaño, Ralf Jungmann, Ruth B. McCole, Eric F. Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y. Fonseka, Jelena Erceg, Mohammed a. Hannan, Hien G. Hoang, David Colognori, Jeannie T. Lee, William M. Shih, Peng Yin, Xiaowei Zhuang, and Chao-ting Wu. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nature Communications*, 6:7147, 2015. ISSN 2041-1723. doi: 10.1038/ncomms8147.

[138] Nicholas R Markham and Michael Zuker. UNAFold: software for nucleic acid folding and hybridization. In *Bioinformatics, Volume II: Structure, Function and Applications*, volume 453, pages 3–31. 2008. ISBN 978-1-60327-428-9. doi: 10.1007/978-1-60327-429-6.

[139] Brian J. Beliveau, Jocelyn Y Kishi, Guy Nir, Hiroshi M Sasaki, Sinem K Saka, Son C Nguyen, Chao-ting Wu, and Peng Yin. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization. *Proceedings of the National Academy of Sciences*, 115(10):E2183–E2192, 2018.

[140] Brian J. Beliveau, Eric F. Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y. Fonseka, Ruth B. McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R. Williams, Jean Marie Rouillard, and Chao Ting Wu. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52):21301–21306, 2012. ISSN 00278424. doi: 10.1073/pnas.1213818110.

[141] Marco Passaro, Martina Martinovic, Valeria Bevilacqua, Elliot A. Hershberg, Grazisa Rossetti, Brian J. Beliveau, Raoul J.P. Bonnal, and Massimiliano Pagani. OligoMinerApp: A web-server application for the design of genome-scale oligonucleotide in situ hybridization probes through the flexible OligoMiner environment. *Nucleic Acids Research*, 48(1):W332–W339, 2020. ISSN 13624962. doi: 10.1093/NAR/GKAA251.

[142] Job Dekker, Marc a Marti-Renom, and Leonid a Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403, 2013. ISSN 1471-0064. doi: 10.1038/nrg3454.

[143] Josef Redolfi, Yinxiu Zhan, Christian Valdes, Mariya Kryzhanovska, Isabel Misteli Guerreiro, Vytautas Iesmantavicius, Guido Tiana, Tim Pollex, Jop Kind, Sebastien Smallwood, Wouter de Laat, and Luca Giorgetti. Modeling of DNA methylation in cis reveals principles of chromatin folding in vivo in the absence of crosslinking and ligation. *bioRxiv*, pages 1–41, 2018. doi: 10.1101/407031.

[144] Job Dekker and Leonid Mirny. The 3D Genome as Moderator of Chromosomal Communication. *Cell*, 164(6):1110–1121, 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.02.007.

[145] James Fraser, Iain Williamson, Wendy A. Bickmore, and Josée Dostie. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and molecular biology reviews*, 79(3): 347–72, 2015. ISSN 1098-5557. doi: 10.1128/MMBR.00006-15.

[146] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, 16(1):259, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0831-x.

[147] Kai Kruse, Clemens B. Hug, and Juan M. Vaquerizas. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biology*, 21(1), dec 2020. ISSN 1474760X. doi: 10.1186/s13059-020-02215-9.

[148] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013. ISSN 00280836. doi: 10.1038/nature12593.

[149] Robert A. Beagrie, Antonio Scialdone, Markus Schueler, Dorothee C. A. Kraemer, Mita Chotalia, Sheila Q. Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R. Branco, James Fraser, Josée Dostie, Laurence Game, Niall Dillon, Paul A. W. Edwards, Mario Nicodemi, and Ana Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017. ISSN 0028-0836. doi: 10.1038/nature21411.

[150] Warren Winick-Ng, Alexander Kukalev, Izabela Harabula, Luna Zea Redondo, Mandy Meijer, Leonid Serebreni, Simona Bianco, Dominik Szabo, Andrea Chiariello, Ibai Irastorza Azcarate, Luca Fiorillo, Francesco Musella, Christoph Thieme, Ehsan Irani, Elena Torlai Triglia, Aleksandra Kolodziejczyk, Andreas Abentung, Galina Apostolova, Eleanor Paul, Vedran Franke, Rieke Kempfer, Altuna Akalin, Sarah Teichmann, Georg Dechant, Mark Ungless, Mario Nicodemi, Gonçalo Castelo-Branco, and Ana Pombo. Cell-type specialization in the brain is encoded by specific long-range chromatin topologies. 2020. doi: 10.1101/2020.04.02.020990.

[151] Izabela Harabula and Ana Pombo. The dynamics of chromatin architecture in brain development and function. *Current Opinion in Genetics and Development*, 67:84–93, 2021. ISSN 18790380. doi: 10.1016/j.gde.2020.12.008.

[152] Sofia A. Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, Alexander A. Shishkin, Prashant Bhat, Yodai Takei, Vickie Trinh, Erik Aznauryan, Pamela Russell, Christine Cheng, Marko Jovanovic, Amy Chow, Long Cai, Patrick McDonel, Manuel Garber, and Mitchell Guttman. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, 174(3):744–757.e24, 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.05.024.

[153] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, Annelies De Klein, Lodewyk Wessels, Wouter De Laat, and Bas Van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197): 948–951, 2008. ISSN 14764687. doi: 10.1038/nature06947.

[154] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K. Brinkman, Stephen A. Adam, Robert Goldman, Bas Van Steensel, Jian Ma, and Andrew S. Belmont. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology*, 217(11):4025–4048, 2018. ISSN 15408140. doi: 10.1083/jcb.201807108.

[155] Liguo Zhang, Yang Zhang, Yu Chen, Omid Gholamalamdari, Yuchuan Wang, Jian Ma, and Andrew S. Belmont. TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome Research*, 31(2):251–264, 2021. ISSN 1088-9051. doi: 10.1101/gr.266239.120.

[156] Joseph Dopie, Michael J. Sweredoski, Annie Moradian, and Andrew S. Belmont. Tyramide signal amplification mass spectrometry (TSA-MS) ratio identifies nuclear speckle proteins. *Journal of Cell Biology*, 219(9), 2020. ISSN 15408140. doi: 10.1083/jcb.201910207.

[157] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr011.

[158] Stefan Kurtz. The Vmatch large scale sequence analysis software-a manual, 2010.

[159] W James Kent. BLAT — The BLAST -Like Alignment Tool. *Genome research*, 12:656–664, 2002. ISSN 1088-9051. doi: 10.1101/gr.229202.

[160] Ryan a. Rossi, David F. Gleich, Assefaw H. Gebremedhin, and Md. Mostofa Ali Patwary. Parallel Maximum Clique Algorithms with Applications to Network Analysis and Storage. *arXiv*, page 11, 2014. ISSN 10957200. doi: 10.1137/14100018X.

[161] Qikai Xu, Michael R Schlabach, Gregory J Hannon, and Stephen J Elledge. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proceedings of the National Academy of Sciences*, 106(7):2289–2294, 2009.

[162] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 1363(2014):1360–1363, 2015. ISSN 1095-9203. doi: 10.1126/science.aaa6090.

[163] Harianto Tjong, Wenyuan Li, Reza Kalhor, Chao Dai, Shengli Hao, Ke Gong, Yonggang Zhou, Haochen Li, Xianghong Jasmine Zhou, Mark A. Le Gros, Carolyn A. Larabell, Lin Chen, and Frank Alber. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1512577113.

[164] Nan Hua, Harianto Tjong, Hanjun Shin, Ke Gong, Xianghong Jasmine Zhou, and Frank Alber. Producing genome structure populations with the dynamic and automated PGS software. *Nature Protocols*, 13(5):915–926, 2018. ISSN 17502799. doi: 10.1038/nprot.2018.008.

[165] T C Hsu. A possible function of constitutive heterochromatin: the bodyguard hypothesis. *Genetics*, 79:137–150, 1975.

[166] Elodie Gazave, Philippe Gautier, Susan Gilchrist, and Wendy A. Bickmore. Does radial nuclear organisation influence DNA damage? *Chromosome Research*, 13(4):377–388, 2005. ISSN 09673849. doi: 10.1007/s10577-005-3254-9.

[167] John A. Stamatoyannopoulos, Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev. Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393–395, 2009. ISSN 10614036. doi: 10.1038/ng.363.

[168] Benjamin Schuster-Böckler and Ben Lehner. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507, aug 2012. ISSN 00280836. doi: 10.1038/nature11273.

[169] Lin Liu, Subhajyoti De, and Franziska Michor. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature Communications*, 4: 1–9, 2013. ISSN 20411723. doi: 10.1038/ncomms2502.

[170] Sandro Morganella, Ludmil B. Alexandrov, Dominik Glodzik, Xueqing Zou, Helen Davies, Johan Staaf, Anieta M. Sieuwerts, Arie B. Brinkman, Sancha Martin, Manasa Ramakrishna, Adam Butler, Hyung Yong Kim, Åke Borg, Christos Sotiriou, P. Andrew Futreal, Peter J. Campbell, Paul N. Span, Steven Van Laere, Sunil R. Lakhani, Jorunn E. Eyfjord, Alastair M. Thompson, Hendrik G. Stunnenberg, Marc J. Van De Vijver, John W.M. Martens, Anne Lise Børresen-Dale, Andrea L. Richardson, Gu Kong, Gilles Thomas, Julian Sale, Cristina Rada, Michael R. Stratton, Ewan Birney, and Serena Nik-Zainal. The topography of mutational processes in breast cancer genomes. *Nature Communications*, 7(May):1–11, 2016. ISSN 20411723. doi: 10.1038/ncomms11383.

[171] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015. ISSN 14764687. doi: 10.1038/nature15393.

[172] Henrike Johanna Gothe, Britta Annika Maria Bouwman, Eduardo Gade Gusmao, Rossana Piccinno, Giuseppe Petrosino, Sergi Sayols, Oliver Drechsel, Vera Minneker, Natasa Josipovic, Athanasia Mizi, Christian Friberg Nielsen, Eva Maria Wagner, Shunichi Takeda, Hiroyuki Sasanuma, Damien Francis Hudson, Thomas Kindler, Laura Baranello, Argyris Papantonis, Nicola Crosetto, and Vassilis Roukos. Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Molecular Cell*, 75(2):267–283.e12, jul 2019. ISSN 10974164. doi: 10.1016/j.molcel.2019.05.015.

[173] Fredrik Mertens, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15(6):371–381, 2015. ISSN 14741768. doi: 10.1038/nrc3947.

[174] Winston X. Yan, Reza Mirzazadeh, Silvano Garnerone, David Scott, Martin W. Schneider, Tomasz Kallas, Joaquin Custodio, Erik Wernersson, Yinqing Li, Linyi Gao, Yana Federova, Bernd Zetsche, Feng Zhang, Magda Bienko, and Nicola Crosetto. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications*, 8(May):1–9, 2017. ISSN 20411723. doi: 10.1038/ncomms15058.

[175] Nicola Crosetto, Magda Bienko, Eleni Gelali, Gabriele Girelli, Masahiro Matsumoto, Erik Wernersson, Joaquin Custodio, Ana Mota, Maud Schweitzer, Katalin Ferenc, and Others. iFISH: a publically available resource enabling versatile DNA FISH to study genome architecture. 2019.

[176] Joaquin Custodio, Tomasz Kallas, Gabriele Girelli, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota, Solrun Kolbeinsdottir, Eleni Gelali, Nicola Crosetto, and Others. GPSeq reveals the radial organization of chromatin in the cell nucleus. 2020.

[177] Anish V Abraham, Sripad Ram, Jerry Chao, E S Ward, and Raimund J Ober. Quantitative study of single molecule location estimation techniques. *Optics Express*, 17(26):23352, 2009. ISSN 1094-4087. doi: 10.1364/OE.17.023352.

[178] M. Kozubek and P. Matula. An efficient algorithm for measurement and correction of chromatic aberrations in fluorescence microscopy. *Journal of Microscopy*, 200(3):206–217, 2000. ISSN 0022-2720. doi: 10.1046/j.1365-2818.2000.00754.x.

[179] Stéfan Van Der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. Scikit-image: Image processing in python. *PeerJ*, 2014(1):1–18, 2014. ISSN 21678359. doi: 10.7717/peerj.453.

[180] V Roukos, G Pegoraro, T C Voss, and T Misteli. Cell cycle staging of individual cells by fluorescence microscopy. *Nature Protocols*, 10(2):334–348, 2015. ISSN 1754-2189. doi: 10.1038/nprot.2015.016.

[181] Gabriele Girelli. ggirelli/oligo-picker: August 2018. feb 2019. doi: 10.5281/ZENODO.2565952. URL `https://zenodo.org/record/2565952`.

[182] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), 2009. ISSN 14747596. doi: 10.1186/gb-2009-10-3-r25.

[183] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btv098.

[184] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, 2021. ISSN 2046-1402. doi: 10.12688/f1000research.29032.1.

[185] Gabriele Girelli. ggirelli/pygpseq-scripts: [0.0.1] - 2019-08-11. aug 2019. doi: 10.5281/ZENODO.3365634. URL `https://zenodo.org/record/3365634`.

[186] Gabriele Girelli. ggirelli/gpseqc: Manuscript version. mar 2020. doi: 10.5281/ZENODO.3727192. URL `https://zenodo.org/record/3727192`.

[187] Gabriele Girelli. ggirelli/gpseqc-snakemake: Manuscript version. mar 2020. doi: 10.5281/ZENODO.3727194. URL `https://zenodo.org/record/3727194`.

[188] TR Luperchio, MEG Sauria, VE Hoskins, X Wong, E DeBoy, M-C Gaillard, P Tsang, K Pekrun, RA Ach, NA Yamada, J Taylor, and KL Reddy. The repressive genome compartment is established early in the cell cycle before forming the lamina associated domains. *bioRxiv*, pages 1–45, 2018. doi: http://dx.doi.org/10.1101/481598.

[189] Gabriele Girelli. ggirelli/ifpd2: [1.0.0-alpha] - 2021-02-26. feb 2021. doi: 10.5281/ZENODO.4563935. URL `https://zenodo.org/record/4563935`.

[190] Gabriele Girelli. ggirelli/GPSeq-RadiCal: [0.0.6]. jul 2020. doi: 10.5281/ZENODO.3952504. URL `https://zenodo.org/record/3952504`.

[191] Dirk Merkel. Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood. *Linux Journal*, 2014(239):2–7, 2014. URL `http://delivery.acm.org.ezproxy.library.wisc.edu/10.1145/2610000/2600241/11600.html?ip=128.104.46.196{&}id=2600241{&}acc=ACTIVESERVICE{&}key=066E7B0AFE2DCD37.066E7B0AFE2DCD37.4D4702B0C3E38B35.4D4702B0C3E38B35{&}{_}{_}acm{_}{_}=1557803890{_}216b4a0168a6b29b8f2e7a74`.

[192] Paolo DI Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. ISSN 15461696. doi: 10.1038/nbt.3820.