

THE ROLE OF CHROMATIN-ASSOCIATED FACTORS IN GENOME TOPOLOGY

Francesca Mugianesi

TESI DOCTORAL UPF / 2022

Thesis supervisors:

Dr. Marc A. Marti-Renom (Structural Genomics)

Dr. Luciano Di Croce (Epigenetic Events in Cancer)

DEPARTMENT OF EXPERIMENTAL AND HEALTH
SCIENCES

GENE REGULATION, STEM CELLS AND CANCER

CENTRE FOR GENOMIC REGULATION

CENTRO NACIONAL DE ANÁLISIS GENÓMICO



*“Nature uses only the longest threads to weave her patterns, so that each small
piece of her fabric reveals the organization of the entire tapestry.”*
Richard P. Feynman

In appreciation

Grazie alla mia famiglia, che mi ha insegnato la dedizione allo studio
fin da bambina, dandomi preziose opportunità,

to my Principal Investigators who believed in me and guided me,

to my colleagues who have constantly been enriching my learning
process,

a las amigas y los amigos que me han acompañado en las risas y las
lagrimas,

als companys i companyes de moltíssimes aventures que em fan viure
plenament,

et enfin merci à mon précieux compagnon de voyage.

Abstract

Gene expression, epigenetic states and topological conformation are three fundamental aspects of genome organization that are tightly regulated in space and time. Epigenetic states, protein occupancy and chromatin modifications are mapped on linear chromatin and constitute a mono-dimensional perspective of chromatin functional states. Importantly, they are linked to the topological conformation of the genome for proper spatiotemporal regulation of gene expression. However, the characterization of the relationship between the genome-wide occupancy of chromatin-associated factors, chromatin states and genome three-dimensional (3D) structure is still elusive. For this purpose, in this thesis, I investigate the role of histone H1 in genome 3D conformation and gene expression and present a novel computational method to integrate chromatin interactions and factor occupancy data with the goal of characterizing chromatin states in 3D.

Resumen

La expresión génica, los estados epigenéticos y la conformación topológica son tres aspectos fundamentales de la organización del genoma, los cuales están estrechamente regulados en el espacio y tiempo. Los estados epigenéticos, la ocupación de proteínas y las modificaciones de la cromatina se estudian de forma lineal y constituyen una perspectiva mono-dimensional de los estados funcionales del genoma. Sin embargo, estos aspectos del genoma están relacionados con la su conformación topológica para permitir la correcta regulación espaciotemporal de la expresión génica. Desafortunadamente, la caracterización de la relación entre la ocupación en el genoma de factores asociados a la cromatina, los estados de la cromatina y la estructura 3D del genoma es todavía difícil de estudiar. En esta tesis, he investigado la función de la histona H1 en la conformación 3D del genoma y en la expresión génica, y presento un nuevo método computacional para integrar datos de interacciones de la cromatina con datos de ocupación de factores, con el objetivo de caracterizar los estados de la cromatina en 3D.

Preface

The vast majority of hereditary information necessary for the development and function of eukaryotic organisms is stored within the cell nucleus. All this information is encoded in large polymer of DNA of about 2 m, which must be organized and compacted at multiple levels to be accommodated in the confined space of the nucleus. Genome organization solves such challenging topological problem, while at the same time it provides the substrate for the correct execution of gene expression programs at the right time, and in the right tissue and cell type. The characterization of the mechanisms underlying how chromatin is organized within the nucleus and how this three-dimensional (3D) architecture is linked to gene regulation, cell fate decisions, and evolution are major questions in cell biology. Topological organization in the 3D space occurs through a hierarchy of structures with increasing complexity, from nucleosomes and chromatin fibers, to chromatin loops, domains, compartments and, finally, chromosome territories. Recent technological developments in quantitative biology, genomics and cell and molecular biology approaches are helping gaining insights into the precise nature of genome topology and its regulatory functions in gene expression and genome maintenance, in development and disease (Bonev & Cavalli, 2016).

This thesis is composed of multiple chapters. In the introduction, we review genome organization within the nucleus and its relationship with genome function across different genomic scales. The introduction encompasses main experimental and computational approaches for the analysis and representation of chromatin 3D organization. Following,

the core of the thesis is articulated in chapter 1 and 2 and presents the results obtained in two projects of the candidate. In chapter 1, we investigate the relationship between histone H1, genome architecture and gene expression. In this study, the candidate has specifically contributed by performing the analysis and 3D modeling of chromatin conformation data. The rest of the experiments, performed by our collaborators at the Jordan Lab (IBMB-CSIC), are also included in the chapter for proper understanding of the results. In chapter 2, we present a novel computational method to characterize chromatin states in 3D by integrating chromatin interactions and protein occupancy data, and we study the evolution of 3D chromatin states during stem cell differentiation. The entirety of the Chapter 2 constitutes the main body of work of the candidate. The thesis is ended with a conclusion chapter highlighting the main contributions to the field of 3D genomics by the candidate. Finally, annexes 1, 2, and 3 contain three published articles, where the candidate specifically contributed by carrying out the computational analyses related to genome 3D conformation.

Objectives

The global goal of this thesis is the exploration of the role chromatin-associated factors in genome 3D organization within the cell nucleus. This main goal has been addressed by three different projects or objectives, which aim at:

1. Studying the consequences of histone H1 variants depletion in human breast cancer cells, to gain insights in the role of histone H1 variants in gene expression, chromatin state and genome 3D conformation (Chapter 1).
2. Developing of a novel and generalized computational tool that integrates chromatin interactions and factor occupancy data with genome structural data, to reveal the contribution of chromatin-associated factors to genome topology (Chapter 2, first half).
3. Applying our new approach to mouse embryonic stem cells (ESCs) and neural progenitor cells (NPCs) to identify and to study genome 3D chromatin state changes during stem cell differentiation (Chapter 2, second half).

Table of contents

Abstract.....	vii
Preface.....	ix
Objectives	xi
INTRODUCTION.....	1
1. The DNA macromolecule: structure and organization	1
2. The Genetic Code	3
3. Chromatin and epigenetics.....	5
3.1 Nucleosomes.....	6
3.2 Chromatin fiber	7
3.3 Modulation of chromatin compaction	8
3.4 Chromatin states.....	13
4. Genome 3D Organization	15
4.1 Nuclear positioning.....	17
4.2 Compartments	19
4.3 TADs.....	22
4.4 Chromatin looping	25
4.5 Biomolecular condensates.....	30
5. Experimental approaches for the analysis of genome 3D organization	32
5.1 Super-resolution microscopy	33
5.2 Sequencing-based methods.....	35
5.2.1 Proximity ligation-based methods: chromosome conformation capture (3C).....	36

5.2.2	Ligation-independent techniques	40
5.3	Heterogeneity and dynamics in chromosome conformation 42	
6.	Computational strategies for the analysis and representation of chromatin organization.....	44
6.1	Analysis of Hi-C data.....	44
6.2	3D modeling approaches	46
7.	The relationship between genome function and structure.....	50
CHAPTER 1: <i>Coordinated changes in gene expression, H1 variant distribution and genome 3D conformation in response to H1 depletion.....</i>		55
CHAPTER 2: <i>CHROMATIC reveals chromatin-associated factors contributing to genome topology.....</i>		109
CONCLUSIONS		173
ANNEX 1		177
ANNEX 2		195
ANNEX 3		221
REFERENCES		251

List of figures

Figure 1. From the DNA macromolecule to the chromatin fiber.	3
Figure 2. Snapshot of the state-of-the-art knowledge about the architecture of the eukaryotic genome.	9
Figure 3. Different types of transcription regulatory loops.....	29
Figure 4. 3C methods and fundamental principles of mammalian chromosome organization.....	37

INTRODUCTION

1. The DNA macromolecule: structure and organization

Despite the incredible diversity characterizing life on Earth, the coding instructions of all living organisms are written in the same language of nucleic acids. In the middle of the XX century, biologists recognized that, whatever its nature, the genetic material must (1) store large amounts of instructions, for all the attributes and functions of an organism, (2) replicate faithfully, to be transmitted to descendant cells with great accuracy, and (3) encode a phenotype, translating into the amino acid sequence of proteins.

The discovery of the double-stranded structure of DNA (Franklin & Gosling, 1953; Watson & Crick, 1953; Wilkins, Stokes, & Wilson, 1953), with its specific base pairing, provided an elegant model that helped to explain how the DNA could store and transmit genetic information. It was found that DNA consists of two complementary and antiparallel strings composed of a large number of repeating units, called nucleotides, joined together by phosphodiester linkages. Each nucleotide contains a pentose deoxyribose sugar, a phosphate group, and a nitrogenous base. The phosphate group and the pentose sugar are the same for all nucleotides and constitute the sugar-phosphate backbone of the DNA molecule. Differently, there are two basic types of nitrogenous bases: purines, that are adenine (A) and guanine (G), and pyrimidines, which are cytosine (C) and thymine (T). Since bases are the variable part of the molecule, they encode for genetic instructions. Also,

they are complementary in pairs: A pairs with T, and C with G. Their pairing by hydrogen bonds allows for the stabilization of the two polynucleotide chains, which have complementary sequences and thus encode for the same biological information. Notably, the complementarity of polynucleotide strands provided an elegant molecular explanation for the ability of DNA to replicate faithfully into two identical copies, and to translate instructions into a phenotype by ultimately specifying the amino acid sequence of proteins. The two strands of nucleotides wound around each other, with the sugars and phosphates in the exterior and the bases in the interior. DNA can adopt a number of different configurations, depending on the conditions in which the molecule is placed and on its base sequence. The B-DNA structure (Fig. 1a) is the most stable configuration for a random sequence of nucleotides under physiological conditions, and most evidence suggests that it is the predominant structure in cells. B-DNA is an alpha helix, with a diameter of around 2 nm, and approximately 10 base pairs (bp) per 360-degree rotation of the spiral. Base pairs are 0.34 nanometer (nm) apart from one another, so one complete rotation of the helix encompasses 3.4 nm. The spiraling of the polynucleotide strands generates major and minor grooves, which are important for the binding of proteins that regulate the expression of genes. B-DNA structure confers advantages both for information accessibility and for DNA packaging (Travers & Muskhelishvili, 2015).

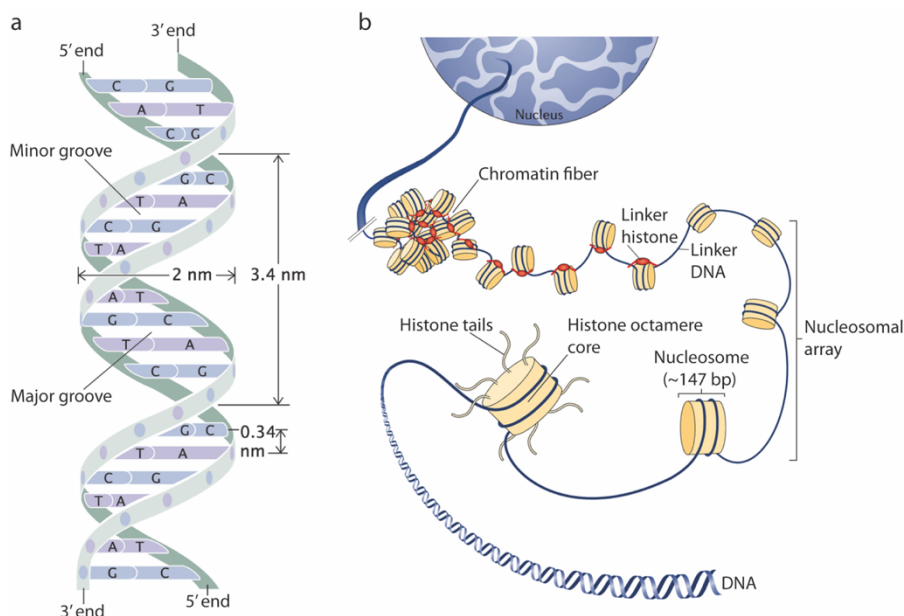


Figure 1. From the DNA macromolecule to the chromatin fiber.

(a) Diagrammatic representation of B-DNA structure. (b) Chromatin compaction within the interphase nucleus occurs through a hierarchy of histone-dependent interactions. The nucleosome is formed by ~147 bp of DNA wrapped around a histone octamer core, composed of two copies of H2A, H2B, H3, and H4. Histone tails are subject to hundreds of different histone post-translational modifications (PTMs) that influence chromatin compactions. Nucleosomal arrays undergo short-range interactions with neighboring nucleosomes to form compacted chromatin fibers. Figure adapted from (Fyodorov, Zhou, Skoultschi, & Bai, 2018; Pierce, 2012).

2. The Genetic Code

Since the revelation of DNA structure, much research focused on how genetic information is encoded, copied and translated. The determination of the human genome reference was a milestone in modern biology. The considerable challenge that derived was to identify and annotate its functional DNA elements. Intriguingly, nearly 99% of the ~3.3 billion nucleotides constituting the human genome does not

code for proteins (Lander et al., 2001). Furthermore, studies of comparative genomics and genome-wide association studies (GWAS) revealed that non-coding elements correspond to the majority of mammalian-conserved and recently adapted regions, and to most of trait-associated loci (Kellis et al., 2014). These findings indicate that non-coding DNA harbors a rich array of functionally significant elements. To better delineate them, the Encyclopedia of DNA Elements (ENCODE) project aims to systematically map cell and tissue repertoires of RNA transcription, chromatin modification and structure, DNA methylation, transcription factors occupancy and RNA-binding proteins, in human and mouse genomes. These data enabled to assign biochemical functions to discrete, linearly ordered sequence features covering around 80% of the genome. Such elements specify either molecular products, like protein-coding genes and non-coding RNAs, or biochemical activities with mechanistic roles in gene regulation, like enhancers and promoters (Consortium et al., 2020). Non-coding RNAs are transcribed RNA molecules that are not translated into proteins, and modulate complex molecular and cellular processes (P. Zhang, Wu, Chen, & Chen, 2019). Enhancers are 10-100 bp regions, target for transcription factors binding, that modulate transcription of target genes in a cell type-specific manner and independently of the enhancer's relative distance. Promoters are ~100 bp protein binding regions upstream of transcription start sites (TSSs) of genes, associated to transcription initiation of the proximal gene (Zabidi & Stark, 2016). Therefore, the past years have witnessed enormous progress in our knowledge about transcriptional regulation, and genome topological rearrangements emerged as an important player

in the communication between different DNA functional elements for correct function of the genetic machinery.

3. Chromatin and epigenetics

Despite having the same linear sequence map, the genome of multicellular organisms must produce different phenotypes in specialized cell types. To do so, information on genome function and gene regulation is also encoded in the way the DNA molecule is condensed in the cell nucleus. To reach this condensed state, genomic DNA in eukaryotic cells is folded up with proteins and RNAs to form chromatin. Chromatin structure is dynamic and exerts profound control over gene expression and other fundamental cellular processes. Indeed, it must ensure to be made accessible for readout by the complex machineries involved in gene transcription, DNA repair and DNA replication.

Maintenance of cell identity during somatic cell division and modulation of cell-type specific gene expression patterns is achieved thanks to the transmission of epigenetic information. Epigenetics can be defined as the study of molecules and mechanisms that can perpetuate alternative gene activity states in the context of the same DNA sequence, encompassing molecular signals peripheral to the DNA such as DNA methylation or histone post-translational modifications (PTMs), as well as gene regulatory signals such as 3D genome organization. Such definition includes both mitotic inheritance of these signals and inheritance across generations via direct replicative mechanisms or indirect reconstruction of the signal in subsequent generations

(Bantignies, Grimaud, Lavrov, Gabut, & Cavalli, 2003; Fitz-James & Cavalli, 2022; Margueron & Reinberg, 2010). The convergence of genetic, biochemical, and cell biological observations have revealed that chromatin epigenetic and architectural states dynamically control genome function at multiple levels of chromatin organization, in normal development (Cheutin & Cavalli, 2019; Ogiyama, Schuettengruber, Papadopoulos, Chang, & Cavalli, 2018) and disease (Loubiere, Martinez, & Cavalli, 2019; Sati et al., 2020).

3.1 Nucleosomes

The building block of chromatin is the nucleosome, which is formed by ~145-147 bp of DNA wrapped around a histone octamer core, composed of two copies of H2A, H2B, H3, and H4 (Luger, Mader, Richmond, Sargent, & Richmond, 1997). Nucleosomes are connected by short segments of linker DNA into nucleosomal arrays, which constitute the primary structure of chromatin (Fig. 1b). Linker histones, such as H1 and its isoforms, bind linker DNA at the base of the nucleosome, near the DNA entry and exit, and are involved in chromatin compaction (Happel & Doenecke, 2009; Willcockson et al., 2021). The amino-termini of core histones are flexible histone tails that extend away from nucleosomal DNA. They interact with neighboring nucleosomes or nuclear factors, and are the site of most PTMs.

An important feature of chromatin is its accessibility, corresponding to the degree at which nuclear macromolecules are able to physically contact chromatin. This parameter is determined by the occupancy and topological organization of nucleosomes and other chromatin-binding

factors that interfere with access to DNA. Since nucleosome and linker histone occupancy and positioning, protein composition of nucleosomes, and nucleosome chemical stability are dynamically variable across the genome, they generate a continuum of DNA accessibility levels that range from closed chromatin to permissive and open chromatin. Pioneer factors are able to bind closed and condensed chromatin, initiate remodeling and increase local accessibility. Accessible chromatin allows transcription factors (TFs) to bind internucleosomal DNA and initiate sequence-specific accessibility remodeling to establish an open chromatin conformation, that in turn allows for the binding of RNA polymerases or other chromatin-binding factors. The landscape of accessibility dynamically changes in response to external stimuli and developmental cues, so it represents a critical determinant of chromatin organization and function (Klemm, Shipony, & Greenleaf, 2019).

3.2 Chromatin fiber

Nucleosomal arrays undergo short-range interactions with neighboring nucleosomes to form compacted chromatin fibers, where DNA is packaged and coiled into a shorter and thicker fiber. For a long time, on the basis of in vitro electron microscopy, nucleosomes were thought to form arrays (often called the 30 nm chromatin fibers) with either solenoid or zigzag shapes (Finch & Klug, 1976; Schalch, Duda, Sargent, & Richmond, 2005; Tremethick, 2007). However, over the years several orthogonal studies have questioned the biological relevance of the 30 nm chromatin fiber (Fussner et al., 2012; T. S. Hsieh et al., 2020; Luger, Dechassa, & Tremethick, 2012; Sanborn et al., 2015; Woodcock, 2005).

Nowadays, the most widely accepted idea is that *in vivo* chromatin is not a stable and periodic structure, but a flexible, heterogeneously organized and unevenly condensed granular chain that is packed together at different concentration densities, with a diameter that ranges from 5 to 24 nm (Cai et al., 2018; Eltsov, Maclellan, Maeshima, Frangakis, & Dubochet, 2008; Ou et al., 2017). Furthermore, stochastic optical reconstruction microscopy (STORM) has shown that at nanoscale level nucleosomes assemble in discrete heterogenous groups of varying sizes, called nucleosome clutches, in the interphase nuclei of mammalian cells (Fig. 2a) (Ricci, Manzo, Garcia-Parajo, Lakadamyali, & Cosma, 2015). Interestingly, large dense clutches are associated to compact heterochromatin and include more linker histone H1, while nucleosome-depleted regions correspond to active chromatin regions. These evidences support the idea that modulation of compactness and accessibility occurs also at the level of the chromatin fiber, is cell type-specific and correlates with chromatin activity states.

3.3 Modulation of chromatin compaction

Chemical modifications to DNA and histone proteins form a complex regulatory network that has profound implications for regulation of key nuclear processes. Changes in nucleosome structure, stability and dynamics affect the compaction of nucleosomal arrays into higher-order structures, which influences how molecular complexes such as

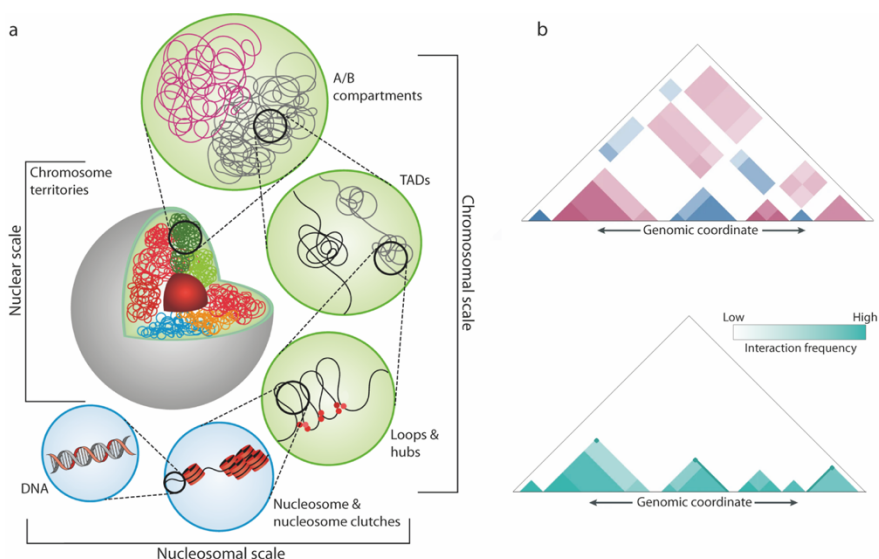


Figure 2. Snapshot of the state-of-the-art knowledge about the architecture of the eukaryotic genome.

(a) Hierarchical chromatin structure. (b) Schematic Hi-C maps representing compartments as a plaid-pattern on top, and TADs as triangles along the diagonal below. Loops at TAD borders appear as enriched punctuate signal at the upper corner of some TADs. Stripes are consistent with the idea that contacts reflect a captured moment of a dynamic process. Figure adapted from (Dogan & Liu, 2018; van Steensel & Furlong, 2019).

the transcriptional and repair machineries interact with DNA and chromatin (Bartke et al., 2010; Luger et al., 2012). Epigenetic modifications refer to the complete repertoire across the genome of these potentially heritable changes (Bernstein, Meissner, & Lander, 2007).

In higher eukaryotes, cytosine DNA methylation at CpG dinucleotides is associated with gene silencing (Chodavarapu et al., 2010). Although the precise relationship between DNA methylation and nucleosome positioning remains poorly understood, methylated DNA presents decreased flexibility and, thus, is less accessible for the transcriptional machinery (Segal & Widom, 2009). Furthermore, the amino acid side chains of core histones that compose the nucleosome are subject to

hundreds of different PTMs, including acetylation, methylation, phosphorylation, and ubiquitination. Such modifications chemically alter histones, they can be added and removed by enzymes, and thus dynamically modulate DNA accessibility. It has become evident that the enzymes responsible histone modifications function in a coordinated pattern to control gene expression, supervising cell fate decisions and differentiation (Margueron, Trojer, & Reinberg, 2005). Indeed, histone modifications may recruit other enzymes and proteins, which in turn recruit nucleosome remodeling complexes and, depending on the specific case, activating or repressing complexes. Histone acetylation contrasts nucleosome array compaction, resulting in an increase in chromatin accessibility and enhanced RNA transcription (Bowman & Poirier, 2015). It acts by reducing the positive charge of histones, thus decreasing the strength of interaction between the negatively charged DNA phosphate backbone and the positively charged histone residues mainly located on histone tails. Moreover, acetylated lysine residues are recognized by protein domains of nucleosome remodeling complexes that favor chromatin accessibility and transcription. Specifically, acetylation of histone 3 lysine 27 (H3K27) is generally associated to both active enhancers and promoters (Z. Wang et al., 2008). Histone methylation, instead, can have different effects depending on which residue is modified. Methylation of histone H3 lysine 4 (H3K4) and H3 lysine 36 (H3K36) is associated with transcribed DNA. Specifically, H3K4me3 marks transcription start sites (TSSs) and promoters of active genes, it stimulates recruitment of the transcriptional and spliceosomal machinery, and is antagonistic to DNA methylation. In contrast, methylation of H3 lysine 9 (H3K9), H3 lysine 27 (H3K27), and H4 lysine 20 (H4K20) generally correlate with repression.

Methylated H3K9 and H3K27 are bound by HP1 and Polycomb, respectively, which mediate chromatin compaction. In pluripotent embryonic stem cells, H3K27me3 and H3K4me3 mark the so-called bivalent promoters of developmental genes, which result repressed in absence of differentiation signals, while poised for timely activation. During differentiation, lineage-specific gene repression and activation are associated with the corresponding loss of H3K4me3 and H3K27me3, respectively (Voigt, Tee, & Reinberg, 2013).

In addition to DNA and histone chemical modifications, a variety of factors influences nucleosome positioning, and thus chromatin compaction, such as DNA sequence preferences, ATP-dependent nucleosome remodeling complexes, transcription factors (TFs) binding, architectural chromatin proteins, Polycomb group proteins (PcG), and histone composition (Segal & Widom, 2009). ATP-dependent chromatin remodellers, such as SWI/SNF or ISWI, are large molecular machines that use the energy of ATP hydrolysis to move, destabilize, eject, or restructure nucleosomes along the DNA (Clapier & Cairns, 2009; Gangaraju & Bartholomew, 2007). TFs can influence nucleosome positioning by competing with them for access to DNA, depending on their relative affinities to the underlying DNA and on their concentrations (Segal & Widom, 2009). Architectural chromatin proteins (ACPs) are abundant nuclear proteins that interact with nucleosomes, influence the three-dimensional arrangement of nucleosomal arrays and orchestrate higher-order chromatin organization through the establishment of interactions between regulatory elements across multiple spatial scales (Gomez-Diaz & Corces, 2014). Their role in 3D genome organization will be discussed in chapter XXX. The best characterized architectural protein in

vertebrates is CCCTC-binding factor (CTCF), which is located at 55,000-65,000 sites in the genome of mammalian cells, mainly located in intergenic regions, introns and exons, and near promoters (T. H. Kim et al., 2007). CTCF binding sites might both insulate different genomic regions, and facilitate enhancer-promoter interactions in a cell type specific manner (Gomez-Diaz & Corces, 2014).

PcG proteins are evolutionarily conserved chromatin-modifying factors that are essential for maintaining epigenetic cellular memory of transcriptional repressed state, and dynamically regulating cellular identity and cell differentiation through epigenetic repression of key developmental regulatory genes. They have been involved in a plethora of cellular processes and have been discovered to orchestrate chromatin architecture at multiple levels (Di Croce & Helin, 2013; Schuettengruber, Bourbon, Di Croce, & Cavalli, 2017). At the scale of the linear genome, PcG proteins modify histones and local chromatin compaction. Even though multiple pathways and mechanisms contribute to recruit PcG proteins in *Drosophila* and vertebrates (Aranda, Mas, & Di Croce, 2015), it is known that in vertebrates hypomethylated CpG islands (CGIs) and long non-coding RNAs (lncRNAs) play a critical role in PcG recruitment, which catalyzes H3K27 trimethylation. Once recruited to their targets, PcG proteins employ diverse mechanisms to regulate their target genes. However, within the most described scenario, H3K27me3 can directly block the deposition of the antagonistically activating acetylation mark on H3K27 (H3K27ac) and can interfere with the recruitment of RNA polymerase II (RNA Pol II) to target promoters.

One more factor influencing chromatin compaction is the incorporation of histone variants. As a demonstration of their

functional relevance, most histone variants are highly conserved between different species (Luger et al., 2012). H1 linker histone variants are the most abundant chromatin-binding proteins. Mammals express 11 different linker histone proteins and they have been reported to be essential for mammalian development. Indeed, whereas deletion of one or two H1 genes does not cause overt phenotypes, simultaneous inactivation of H1-2, H1-3 and H1-4 leads to embryonic lethality (Fan et al., 2005). Several studies have demonstrated that H1 variants are non-randomly distributed in the genome and interact with different protein partners, supporting the idea of functional specificity (Cao et al., 2013; Izzo et al., 2013; Millan-Arino et al., 2014). Moreover, by promoting genomic compaction, their association with chromatin determines nucleosome spacing and controls the balance of repressive and active chromatin domains (Willcockson et al., 2021). Chapter 1 of this thesis is dedicated to the study of the role of histone H1 variants in chromatin compaction and regulation.

Detailed mechanistic insights about how regulatory proteins influence chromatin remodeling and gene regulation are fundamental to characterize of how cells reshape their gene regulatory networks to selectively respond to external signals.

3.4 Chromatin states

Given the central role of chromatin in regulatory signals and control of DNA accessibility, chromatin profiling provides a systematic means of detecting *cis*-regulatory elements. Indeed, specific histone modifications correlate with regulatory binding, transcriptional initiation and elongation, enhancer activity and repression (Barski et al., 2007; Birney

et al., 2007; Guenther, Levine, Boyer, Jaenisch, & Young, 2007; Heintzman et al., 2007; Mikkelsen et al., 2007). In order to segment the genome into biologically meaningful units, unbiased computational approaches like multivariate hidden Markov model (HMM) (Ernst & Kellis, 2012, 2017) have been developed to identify chromatin states, defined as specific combinations of multiple epigenomic datasets. Chromatin states may correspond to known classes of genomic elements, such as enhancers, promoters, transcribed and repressed regions, or may help discover novel classes of elements (Day, Hemmaphardh, Thurman, Stamatoyannopoulos, & Noble, 2007; Ernst & Kellis, 2010; Ernst et al., 2011; Filion et al., 2010; Hon, Wang, & Ren, 2009; mod et al., 2010). Chromatin state annotation has a unique advantage of data reduction, since a large number of datasets involving partially redundant RNA-seq and ChIP-seq data is reduced into a single simple data set, whereby each locus of the genome is annotated with one of several states. Notably, chromatin states build more or less favorable chromatin environments for gene expression, but do not fully determine gene activity.

Mostly depending on the parameters used in the computational analysis, various studies report somewhat different classifications of chromatin types. However, the general consensus is that there are a few types of repressive chromatin, which are Polycomb-bound euchromatin, heterochromatin and a chromatin state that has no strong enrichment for any of the specific analyzed factors or marks (Cavalli & Misteli, 2013). In contrast, it has been more challenging to rigorously classify active or open chromatin states. Typically, at least four types of open chromatin can be distinguished, encompassing enhancers, promoters, transcribed regions and regions bound by chromatin insulator proteins

(Bernstein et al., 2012). Chromatin state annotations for different cell types and tissues are included in the ENCODE project (Siggens & Ekwall, 2014); they provide an important resource for epigenetic and medical genetic studies and represent a useful framework to track regulatory pattern changes across cell types (Ernst et al., 2011).

However, the study of combinatorial patterns of multiple proteins and marks offer a mono-dimensional perspective on chromatin states by considering chromatin as a linear entity, even though, as described in the next sections, chromatin displays a highly organized 3D structure with an important role in gene expression control. To fill this gap, chapter 2 of this thesis presents a novel computational method to characterize combinations of multiple chromatin-associated factors that take place thanks to the 3D folding of the genome and that may contribute to proper gene regulation. At present, there is no other computational tool to identify chromatin states in 3D, and such advancement extends the advantages offered by 1D chromatin segmentation by classifying major types of chromatin interaction that are linked to a specific biological function.

4. Genome 3D Organization

A growing body of work has shown that the genome is a highly organized hierarchical 3D structure, yet involving dynamic conformational changes, that is intimately connected with essential biological functions such as transcription, replication, DNA repair and chromosome translocation (Bickmore & van Steensel, 2013; Gonzalez-Sandoval et al., 2015; Gross, Chowdhary, Anandhakumar, & Kainth,

2015; Pombo & Dillon, 2015; Sexton & Cavalli, 2015; Therizols et al., 2014). These insights have mainly arisen from application of high-resolution microscopy approaches and molecular biology techniques, two complementary classes of techniques that will be discussed in Chapter 5.

Over the multiple scales of loops, hubs, topologically associating domains (TADs), compartments, and nuclear positioning of chromosomes, genome topological organization can be seen as an emergent property of a self-organizing system (Rajapakse & Groudine, 2011), built up from progressive stabilization of homotypic interactions between genes and regulatory elements. Since association of the majority of DNA-bound factors with their cognate sites is transient (Phair & Misteli, 2000), such model of self-organized spatial clustering of related genetic loci may be important for their efficient regulation: a chance encounter between two loci bound by common regulatory factors increases the factors' local concentration, so that when a factor dissociates it is more likely to be re-trapped by the cluster of binding sites within its locale than to diffuse away to another location (Kang et al., 2011; Rajapakse et al., 2009). This model is consistent with the maintenance of active chromatin hubs with expressed genes (Palstra et al., 2003; Schoenfelder et al., 2010), the formation of Polycomb repressive domains (Lanzuolo, Roure, Dekker, Bantignies, & Orlando, 2007), and heterochromatic clustering (Taddei et al., 2009). Furthermore, in this view, beyond preventing aberrant communication between genetic loci, TADs may allow for co-regulated genes to be more efficiently bound by their regulators for prompt transcriptional response, by increasing the local concentrations of diffusible regulatory factors around their sites of activity.

Collectively, mounting evidence demonstrates that chromatin topology can be regulated and exploited by a variety of molecules such as transcription factors, architectural proteins and non-coding RNAs, in order to coordinate underlying gene activity at multiple scales within the nucleus (Sexton & Cavalli, 2015). However, the detailed mechanistic relationship between chromosome folding and genomic functions is still a matter of considerable debate.

4.1 Nuclear positioning

At the scale of the whole nucleus, nuclear positioning of genetic material is not random, is related to gene expression levels, and undergoes changes during physiological processes such as differentiation, development, aging, and in pathological conditions (Cavalli & Misteli, 2013). Depending on their transcriptional activity, genes tend to occupy preferred positions in the 3D nuclear space, relative to other regions in the genome, or to nuclear structures such as the nuclear lamina, domains of heterochromatin or nuclear bodies (Finlan et al., 2008; Lanctot, Cheutin, Cremer, Cavalli, & Cremer, 2007; Misteli, 2007; Peric-Hupkes et al., 2010; Rajapakse & Groudine, 2011).

Fluorescence DNA and DNA in situ hybridization (FISH) have revealed that in the nuclear space interphase chromosomes occupy distinct chromosome territories (CTs) (Cremer & Cremer, 2010), which constitute a basic feature of nuclear architecture. Gene-rich and transcriptionally more active chromosomes tend to be located in the euchromatic interior of the nucleus, whereas gene-poor and less active chromosomes are closer to the predominantly heterochromatic periphery (Lanctot et al., 2007). The observation of CTs was later

validated by genome-wide Hi-C data, which showed that interactions between loci on the same chromosome are much more frequent than contacts in trans between different chromosomes (Lieberman-Aiden et al., 2009).

A variety of orthogonal techniques have uncovered a plethora of long-range interactions between genes that share regulation by a common factor, such as Polycomb-mediated repression (Bantignies et al., 2011; Denholtz et al., 2013), activation by tissue-specific TFs (Papantonis et al., 2012; Schoenfelder et al., 2010), pluripotency-linked TFs (Apostolou et al., 2013; de Wit et al., 2013; Denholtz et al., 2013; Z. Wei et al., 2013), or multiple super-enhancers (Beagrie et al., 2017). Such associations occur specifically in cell types where the regulation is mediated, even when genes occupy different chromosomes. The existence of functional clusters of genes at nuclear foci enriched in their regulatory factors and coalescing around different nuclear bodies such as nuclear speckles may facilitate their coordinate expression, and has emerged as prominent regulatory feature of nuclear architecture (Bantignies et al., 2011; Papantonis et al., 2012; Quinodoz et al., 2021; Quinodoz et al., 2018; Schoenfelder et al., 2010; Vangala et al., 2020).

Due to contrasting evidences about the deterministic link between the spatial position of an individual locus and its activity (Kubben et al., 2012; Peric-Hupkes et al., 2010; Reddy, Zullo, Bertolino, & Singh, 2008; Shachar, Voss, Pegoraro, Sciascia, & Misteli, 2015; Therizols et al., 2014), it is known that nuclear positioning is correlated with and underlies gene expression, but the extent of such relationship is still not fully resolved.

4.2 Compartments

At the genomic scale, the eukaryotic genome is partitioned into chromatin compartments, which are spatially segregated genomic regions, located either on the same or on different chromosomes, with distinct biochemical and functional properties.

Prominent nuclear compartments are heterochromatin and euchromatin, which were originally defined based on differences in apparent chromatin compaction, as visible by microscopy. Generally, transcriptionally inactive or repressed genomic regions are heterochromatic, whereas transcribed regions are euchromatic. Heterochromatin tends to be marked by H3K27me3 mark, or by H3K9me3 and H3K9me2 (Bernstein et al., 2012; Filion et al., 2010). In metazoan cells, heterochromatin marked by H3K9me2 and H3K9me3 is typically concentrated at the nuclear lamina and, to a lesser extent, around nucleoli. Euchromatin regions are densely populated by active genes and enhancer elements, and are typically marked by a multitude of histone modifications, such as methylation of H3K4 and acetylation of various histone lysine residues. Euchromatin is generally located in the nuclear interior, although it can also interact with nuclear pores (van Steensel & Furlong, 2019). Partitioning of euchromatin and heterochromatin has also been reflected in chromosomal contact maps generated by chromosome conformation capture technologies, such as Hi-C. Hi-C contact maps exhibit a chromosome-wide plaid pattern of extensive long-range intrachromosomal and interchromosomal contacts (Fig. 2), which can be >10 Mb apart, corresponding to two major classes of self-associating compartment with little inter-mixing (Lieberman-Aiden et al., 2009). They were termed compartment A and

compartment B, and are enriched in active or inactive chromatin marks, respectively. Lamina-associated domains (LADs) and heterochromatin overlap with compartment B, while euchromatic inter-LAD regions overlap with compartment A.

Subsequently, higher resolution Hi-C and other 3C-based techniques suggested that these two major compartments can be further partitioned into six different subcompartments, with two subcompartments for A compartment and four subcompartments for B compartment (S. S. Rao et al., 2014; Wijchers et al., 2016).

The partitioning between compartments is dynamic and genomic loci can switch between compartments in a cell-type specific manner (Dixon et al., 2015; Lieberman-Aiden et al., 2009). Accordingly, during cell differentiation, hundreds of genes are repositioned from peripheral heterochromatin to the internal euchromatin and vice versa (Shachar & Misteli, 2017), corresponding in most case to their activation and repression, respectively.

In mammalian cells, knockdown of architectural protein cohesin, which is a key factor for chromatin looping (see below), results in strengthening of existing compartmentalization and reduction of TADs (Haarhuis et al., 2017; S. S. P. Rao et al., 2017; Schwarzer et al., 2017). This and similar results suggest that chromatin looping and compartmentalization are distinct and competing mechanisms contributing to chromatin folding (van Steensel & Furlong, 2019). Currently, the most accepted albeit speculative scenario describes local mechanisms such as looping and gene activity as the basis of TAD formation, whereas compartments may be formed by attraction and/or repulsion between individual TADs with similar epigenetic marks. This model is supported by super-resolution microscopy, which showed that

spatial interactions between neighboring TADs with different epigenetic states are remarkably different; for instance, Polycomb-repressed domains are particularly condensed and exclude neighboring domains to a large extent (Boettiger et al., 2016).

Although their mechanistic nature is still unclear, compartments appear to emerge from the superposition of highly stochastic and mutually exclusive interactions between different types of chromatin regions (S. Wang et al., 2016), possibly mediated by mechanisms involving liquid–liquid phase separation (LLPS) of chromatin-associated proteins (Falk et al., 2019; Nuebler, Fudenberg, Imakaev, Abdennur, & Mirny, 2018; Strom et al., 2017; L. Wang et al., 2019).

A variety of proteins contribute to the self-association of heterochromatin in compartment B, including heterochromatin protein HP1 mediating long-range interactions between H3K9me2 and H3K9me3-marked loci (Strom et al., 2017). In mammalian cells, H3K27me3-marked Polycomb domains form intrachromosomal and interchromosomal contacts that can be part of either the A compartment or the B compartment, depending on the cell type (van Steensel & Furlong, 2019).

The role of euchromatin proteins in mediating the self-association of the euchromatic loci is much less established. Despite direct evidence is still lacking, transcription factors, cofactors and the transcription machinery, whose nuclear foci have been known for decades (Jackson, Hassan, Errington, & Cook, 1993), may collectively be responsible for the organization and function of the euchromatin compartment through the formation of condensates (Boehning et al., 2018; Boija et al., 2018; Hnisz, Shrinivas, Young, Chakraborty, & Sharp, 2017; Sabari et al., 2018).

Further genome-wide experiments in mutants deficient in chromatin modifiers and proteins are required to determine the role of different factors and epigenetic marks in genome architecture.

4.3 TADs

Genome-wide 3C technologies have shown that at the sub-megabase scale, chromosomes of many metazoan genomes fold into distinct modules, called topologically associating domains (TADs), that can be considered as functional units of the genome (Dixon et al., 2012; Hou, Li, Qin, & Corces, 2012; Nora et al., 2012; Sexton et al., 2012). They are typically 100kb-1Mb in length, and in Hi-C maps appear as contiguous squares or triangles along the diagonal (Fig. 2). Genomic interactions are extensive within domains but are depleted on crossing the boundary between neighboring TADs.

TADs display dynamics and cell-to-cell variability that cannot be captured by Hi-C data, since this one reflects the population-average folded state of the chromosome in fixed cells. However, domains identified on Hi-C maps show a surprising developmental and evolutionary robustness, suggesting that TADs may be chromosome building blocks required for appropriate genome function. Indeed, most domains correlate well with many linear markers of chromatin activity, such as histone modifications and replication timing (Dixon et al., 2012; Le Dily et al., 2014; Sexton et al., 2012), and coordinated gene expression (Le Dily et al., 2014; Nora et al., 2012). Moreover, TADs may avoid inappropriate enhancer–promoter (E-P) interactions and insulate promoters from the action of enhancers located in neighboring TADs (Sexton & Cavalli, 2015; Shen et al., 2012), by constraining the

effective search space of enhancers and promoters to find each other (Symmons et al., 2016; Symmons et al., 2014). Loss of a TAD boundary could thereby lead to the misexpression of genes in a neighbouring TAD, as observed at some loci (Flavahan et al., 2016; Lupianez et al., 2015).

Furthermore, technological advances have revealed smaller and finer-scale structures, hierarchically nested within TADs, that exhibit high developmental dynamics and may even encompass a single gene unit. Multiple studies called them with different names, such as sub-TADs, mini-domains, microTADs, chromatin nanodomains (CNDs), or 3D nanocompartments (T. S. Hsieh et al., 2020; Krietenstein et al., 2020; Phillips-Cremins et al., 2013; S. S. Rao et al., 2014; Rowley et al., 2017; Szabo et al., 2020; Szabo et al., 2018). Due to the nested structure of TADs, their exact definition is ambiguous, and strongly depends on the resolution of the performed experiment and to some extent on the employed detection method (Soler-Vila, Cusco, Farabella, Di Stefano, & Marti-Renom, 2020).

In general, TADs are characterized by sharp boundaries that correspond to binding sites for CTCF, other chromatin insulator-binding proteins and transcription factors, as well as to active transcriptional start sites (Bonev et al., 2017; Dixon et al., 2012; Krietenstein et al., 2020; Sexton et al., 2012). In mammals, strong chromatin loops, that will be discussed below, are observed at the borders of $\sim 40\%$ of the domains (Fig. 2), suggesting a strong relationship between chromatin loop formation and the demarcation of domain boundaries. The role of boundary factors such as CTCF and loops could thus be to strengthen the stability of the boundaries

between domains of different chromatin types or to sharpen their localization.

Domains of the same type have the tendency to establish strong inter-TAD interactions, whether they are active, Polycomb domains, or HP1-heterochromatic domains (Csink & Henikoff, 1996; Sexton et al., 2012). Polymer physics-based modeling showed that the simple assumption of homotypic interactions between domains of these chromatin types are sufficient to generate polymer structures that mimic those represented in Hi-C contact maps (Jost, Carrivain, Cavalli, & Vaillant, 2014). This result suggests that homotypic interactions between domains may contribute to TADs establishment.

Overall, in higher eukaryotic cells a diversity of mechanisms underlies the existence of physical chromosomal domains, including transcriptional levels, epigenetic compositions, architectural proteins, and chromatin modifying factors like PcG proteins. However, the detailed cause–consequence relationship between these factors is still poorly understood, and it is still unknown whether TADs are dynamically built by transcriptional silencing or activation machineries and chromatin-modifying complexes, or TADs themselves set the stage for cooperative binding of specific chromatin factors to determine gene expression. TADs organization may be explained by the propensity of chromatin to establish preferential transient contacts in the form of loops, that are increasingly likely for smaller distances along the same chromosome, with the specificity added by different chromatin factors that contribute to the separation between types of loops, such as those involving active and repressive chromatin.

Future research, including new genome-engineering tools such as CRISPR/Cas9 and live imaging of chromatin interactions in single cells

following their dynamics over the cell cycle, should tease out the details of TADs formation and functions in different nuclear environments.

4.4 Chromatin looping

Higher-order chromosome organization levels are thought to arise from multiple, dynamic and cell type-specific chromatin interactions, that occur at the kilobase-to-megabase scale between regulatory elements and are crucial for proper gene expression and cell identity (Fig. 2a) (Kieffer-Kwon et al., 2013; G. Li et al., 2012; Palstra et al., 2003; S. S. Rao et al., 2014; Sanyal, Lajoie, Jain, & Dekker, 2012).

The pervasive tendency of chromatin to engage in contacts with other chromatin fibers is reflected in the fact that at the TAD level the predominant structural features are point-like focal interactions or stripe-like structures of hundreds of kb (Fig. 2b), that often connect sequences bound by CTCF and cohesin (de Wit et al., 2015; Fudenberg, Abdennur, Imakaev, Goloborodko, & Mirny, 2017; Guo et al., 2015; S. S. Rao et al., 2014; Vian et al., 2018). Stripes are consistent with the idea that contacts within TADs in individual cells reflect not static loops, but a captured moment of a dynamic process (Giorgetti et al., 2014; Hansen, Cattoglio, Darzacq, & Tjian, 2018).

Chromatin loops appear as a ubiquitous means for enhancer-promoter (E-P) or promoter-promoter (P-P) communication (Fig. 3a) (Tolhuis, Palstra, Splinter, Grosveld, & de Laat, 2002). Although the mechanistic details of enhancers' stimulation of transcription are not yet clarified, distal enhancers carry a large regulatory potential and are bridged with their target gene promoters for the induction of transcription. One well known example is the locus control region (LCR) of the β -globin

cluster, which in erythroid cells, where the β -globin gene is active, forms an active chromatin hub with its target genes (Palstra et al., 2003).

The leading mechanism governing loop formation is thought to involve loop-extruding complexes like cohesin, and border elements such as CTCF architectural protein (Fig. 3b) (Fudenberg et al., 2016; Nichols & Corces, 2015; Sanborn et al., 2015). Interestingly, loop extrusion and compartmentalization appear as two separated principles of chromosome folding (Schwarzer et al., 2017).

Cohesin is large ring-shaped protein complex, that is important for genome stability in dividing cells, and is involved in sister chromatid cohesion and DNA repair (Nasmyth & Haering, 2009). CTCF is a DNA-binding protein that recognizes a specific sequence motif and, among architectural proteins, has probably received the most attention (Ong & Corces, 2014). It is conserved in most bilaterians, is ubiquitously expressed, and is essential for embryonic development (Soshnikova, Montavon, Leleu, Galjart, & Duboule, 2010). Originally, it was characterized as an insulator protein, capable of restricting E-P interactions and establishing discrete functional chromatin domains (Narendra et al., 2015).

In the mechanistic model of loop extrusion, cohesin loads on DNA and bidirectionally extrudes loops until it is blocked in each end of the loop by CTCF proteins binding in convergent orientation (Davidson et al., 2019; Fudenberg et al., 2016; Ganji et al., 2018; Golfier, Quail, Kimura, & Brugues, 2020; Hansen, 2020; Y. Kim, Shi, Zhang, Finkelstein, & Yu, 2019). This model can explain the nesting of domains and loops as the assembly of possible states within a population. Also, it predicts the observed enrichment of CTCF at TAD boundaries, and the consequences of CTCF motif deletion or inversion for loop and domain

formation. Furthermore, it is consistent with changes in 3D chromosome architecture observed in cohesin-depleted or CTCF-depleted cells, where the decrease in TAD insulation most likely results from loss of preferential contacts within TADs, and increased randomness in interactions (Bintu et al., 2018; Nora et al., 2017).

Thus, CTCF-mediated loops are believed to play a fundamental role in maintenance of TAD structure (Giorgetti et al., 2014), and appear to be linked to multiple nuclear processes, such as transcriptional regulation and DNA repair (Oudelaar & Higgs, 2021).

Besides CTCF-loops, nucleosome-resolution interaction maps spotlight dots and ~10-15 kb stripes linking accessible co-expressed loci, such as P-P or E-P sites, that are driven by the transcription machinery and are independent from CTCF and cohesin (T. S. Hsieh et al., 2020). Furthermore, TAD borders often coincide with active promoters but not CTCF sites (Bonev et al., 2017; Dixon et al., 2012; Ramirez et al., 2018; Ulianov et al., 2016). These data suggest a dynamic, reciprocal interplay between genome organization and active transcription (van Steensel & Furlong, 2019).

In mammals and flies, Polycomb complexes have been proven to have an important role in the formation of long-range contacts involving repressed gene promoters in early development (Fig. 3c) (Bantignies et al., 2011; Bonev et al., 2017; Denholtz et al., 2013; Rowley et al., 2017; Schoenfelder et al., 2015; Vieux-Rochas, Fabre, Leleu, Duboule, & Noordermeer, 2015), even at the nucleosomal level (T. S. Hsieh et al., 2020). Despite PcG proteins have been historically described as transcriptional repressors, it has also been found that during *Drosophila* development PcG subunits might support transcriptional activation, by forming 3D loops that involve active promoters and enhancers and

fine-tune their expression (Loubiere, Papadopoulos, Szabo, Martinez, & Cavalli, 2020). Also, other transcriptional regulators have been involved in repressive loop interactions, and in general the molecular principles underlying repressive looping, including Polycomb looping, remain elusive.

Additional types of long-range chromatin contacts with direct functional have been described. In the so-called intragenic loops (Fig. 3d), the 5' end of transcribed genes joins the transcription termination site (TTS). This may allow efficient recycling of the RNA polymerase II (Pol II) and may help establish a short-term memory of the transcriptionally active state for the gene (Mas et al., 2018; Tan-Wong, Wijayatilake, & Proudfoot, 2009; Tan-Wong et al., 2012).

Interestingly, in addition to proteins, long non-coding RNAs (lncRNAs) may participate in the formation of loops, even though it is unclear to what extent. LncRNAs have been shown to mediate the colocalization of several genomic regions located on different chromosomes (Hacisuleyman et al., 2014), and to exploit 3D chromatin organization in order to spread across the X chromosome during X chromosome inactivation (Engreitz et al., 2013; Simon et al., 2013). Future work is required to dissect the precise role of lncRNAs in establishing and maintaining 3D chromatin structure, which has been recently hinted by computational models (Farabella, Di Stefano, Soler-Vila, Marti-Marimon, & Marti-Renom, 2021).

Overall, nested structures are the prevalent folding feature within TADs and, together with constraints provided by the nuclear lamina and sub-nuclear compartments such as speckles and nucleoli, several factors

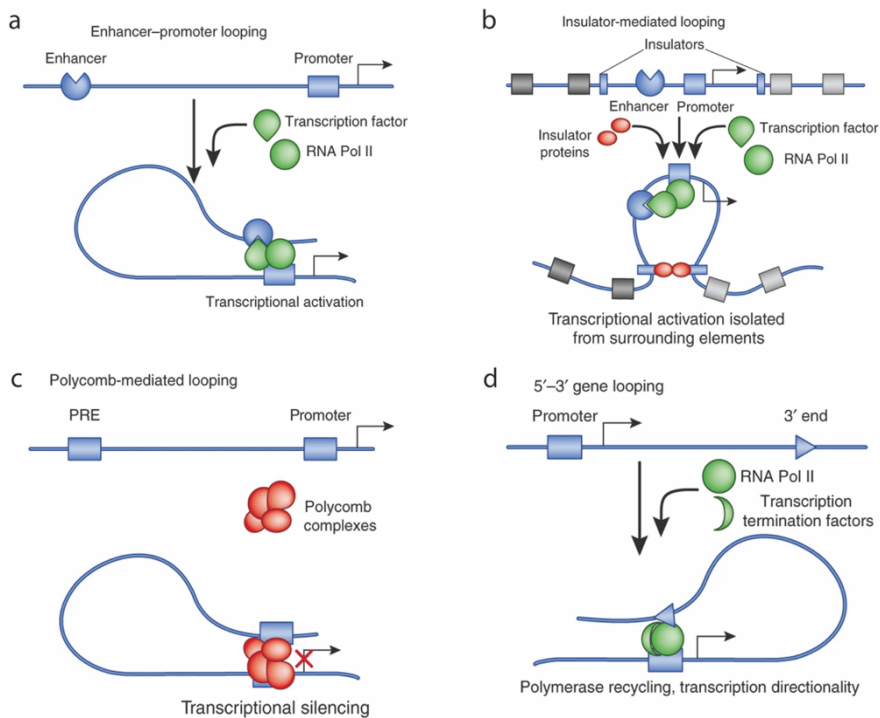


Figure 3. Different types of transcription regulatory loops.

(a) Enhancer-promoter loops leading to transcriptional activation. (b) Insulator-mediated loops may segregate genes and their regulatory elements from surrounding genome landscape, favoring proper gene expression. (c) Loops between Polycomb-bound regions (PREs) and promoters prevent RNA Pol II recruitment and mediate transcriptional silencing. (d) Intragenic loops joining the 5' and 3' end of genes may allow recycling of RNA Pol II and facilitate maintenance of transcriptional directionality. Figure from (Cavalli & Misteli, 2013).

interact together to shape the complex pattern of chromatin interactions of mammalian chromosomes, including the transcriptional machinery, architectural proteins, lncRNAs, TFs and chromatin-remodeling complexes like PcG. However, because of contrasting evidences among different studies, how these interactions are established and regulated is still a matter of considerable debate. It would be important to clarify the contribution of the chromatin

environment and transcription to loop formation, and to investigate whether chromatin loops can be formed also through other processes.

4.5 Biomolecular condensates

Phase separation is emerging as key principle in the spatiotemporal organization of living cells. Increasing evidence indicates that cells also organize membrane-less biomolecular condensates of protein, RNA, and other biomolecules, that are thought to form through the physical process of liquid-liquid phase separation (LLPS) and might operate as versatile biochemical 3D interaction hubs inside the cell (Banani, Lee, Hyman, & Rosen, 2017; Y. Shin & Brangwynne, 2017).

Super-enhancers (SEs) have been proposed to be phase-separated condensates formed by clusters of enhancers, remarkably occupied by interacting master TFs that may cooperatively assemble the transcriptional apparatus to drive robust expression of genes, with prominent roles in cell identity (Hnisz et al., 2017; Sabari et al., 2018). Similarly, transcription factor condensates have been suggested to regulate transcriptional initiation and amplify transcriptional burst frequency and size of expressed genes, being enriched in regulatory elements such as enhancers or silencers, and facilitating the interaction with gene promoters in a cell-specific manner (Beagrie et al., 2017; Javierre et al., 2016; Y. Shin et al., 2018; Stevens et al., 2017).

Interestingly, macromolecular condensates may be dynamically assembled in response to a tunable external stimulus, as in the case of droplets of nuclear receptors TFs, protein kinases and enhancers, in breast cancer hormone responsive cells upon steroid hormone stimulus (Zaurin et al., 2021).

Chromatin interactions driven by ATP-dependent loop extrusion have emerged as key organizing principles of the genome (Di Pierro, Zhang, Aiden, Wolynes, & Onuchic, 2016; Mirny, Imakaev, & Abdennur, 2019). Also, spatial compartmentalization is a hallmark of eukaryotic genomes emerging on various length scales, but its intrinsic physical properties have remained unclear. In conventional nuclei, euchromatin is localized in the nuclear interior and heterochromatin at the nuclear periphery. Inverted nuclei of rods in nocturnal mammals, instead, present the opposite distribution and provided an opportunity to elucidate the mechanisms that underlie compartmentalization. Experiments and modelling suggest that phase separation of the active and inactive genome in inverted and conventional nuclei is achieved thanks to attractions between heterochromatic regions, and chromatin interactions with the lamina are essential to build the conventional architecture from these segregated phases (Falk et al., 2019).

Therefore, in a phase separation-based model for genome organization and regulation, the intrinsic property of chromatin to phase separate within the nucleoplasm may enable establishment and maintenance of distinct chromatin compartments (Gibson et al., 2019), tuned through engagement of cellular factors such as linker histone binding, histone acetylation, interactions with histone tail readers, and spacing of nucleosomes. Functional chromatin states, corresponding to promoters, enhancers, insulators, PcG regions, etc. (Bernstein et al., 2012; Filion et al., 2010) may adopt different phase-separated states with specific structural and dynamic properties that are important for their unique functions in cells and for the formation and segregation of chromatin domains.

Theoretically, phase separation is associated to the heterogeneous mixing of two components, either by spinodal decomposition, or nucleation. In living cells, it has been proposed that intrinsically disordered regions (IDR) are the main driving mechanism promoting LLPS (Boijja et al., 2018). However, a quantitative understanding of the biophysical parameters controlling transcription factor condensation in the living cell nucleus is largely missing.

5. Experimental approaches for the analysis of genome 3D organization

Deciphering the rules of genome folding in the cell nucleus is essential to understand its functions. Insights about genome 3D organization have mostly arisen thanks to major technological breakthroughs in two orthogonal classes of techniques, high-resolution microscopy approaches and sequencing-based methods. Due to their respective strengths and limitations and to the high complexity of genome organization, chromatin architecture is best studied using a combination of approaches, neither of which is comprehensive on its own. Mathematical modelling can complement biological investigation, rationalizing and predicting important aspects of chromatin behavior. Microscopy-based methods provide important information about the relative and radial positioning of genomic regions, as well as the variability of spatial DNA organization within cell populations, but these methods often suffer from limited throughput, coverage and genomic resolution. By contrast, sequencing-based approaches are

genome-wide, but their results may represent a superimposition of individual genome conformations rather than one stable structure.

The simultaneous advances in technological and scientific approaches is leading us to an integrated understanding of the function of the genome and its associated components in development, physiology and disease. The development of novel single cells technologies is essential to capture structural features in rare cell populations, as well as structural changes in dynamic processes, and is helping deepen the characterization of cell type-specific gene regulation. Further improvements of current live imaging may allow tracking of the dynamics of chromatin domains and interactions in live cells in order to investigate conformational changes upon various stimuli and in relation to gene expression. The combination of these tools with functional studies, particularly those made possible by the advent of genome-engineering technologies such as CRISPR–Cas9 (Wright, Nunez, & Doudna, 2016), promises to lead to major advances in the near future. These complex multi-dimensional data generated with different modalities require advanced computational strategies for integration and extensive quantitative analyses.

5.1 Super-resolution microscopy

Remarkable improvements in microscopy techniques are expanding our understanding of the fine-scale structure of the chromatin fiber to a degree that was unthinkable a decade ago. Their incompatibility with sequence determination has been circumvented by a second complementary class of genomic methods, here collectively referred to

as sequencing-based techniques, that will be discussed in the next section.

Historically, fluorescent in situ hybridization (FISH) has mostly been used to study the position and organization of chromosomes, domains and specific loci within the nucleus. Despite its key advantage of direct visualization of the spatial position and arrangement of genomic loci in the nucleus, it has been traditionally limited in throughput and resolution.

Recent advancements in super-resolution microscopy and imaging techniques have enabled direct visualization of the fine-scale structures of the genome of single cells at sub-diffraction resolution and at unprecedented throughput.

Some examples are stochastic optical reconstruction microscopy (STORM) (Rust, Bates, & Zhuang, 2006), photo-activated localization microscopy (PALM) (Betzig et al., 2006), and oligonucleotide arrays such as Oligopaint (Beliveau et al., 2015; Beliveau et al., 2012). Also, HIPMap identifies novel factors affecting the radial positioning of different types of genomic locus, with high-throughput (Shachar et al., 2015). Super-resolution chromosome tracing approaches employing highly multiplexed FISH probes perform distance measurements between thousands of loci in single cells, at unprecedented scales (S. Wang et al., 2016). Oligopaint design in conjunction with STORM (OligoSTORM) (Beliveau et al., 2017; Nir et al., 2018) remarkably improved the resolution, at the same time allowing the analysis of regions at the megabase scale. Furthermore, sequential Oligopaints method in conjunction with super-resolution allows to sequentially label continuous genomic coordinates of the genome at the level of single gene, loops, TADs, compartments (Bintu et al., 2018; Nir et al., 2018).

OligoFISSEQ has remarkably improved high throughput imaging and tracing of genomic loci in thousands of cells (Nguyen et al., 2020). Other flavours of Oligopaint-based methods, such as Hi-M (Cardozo Gizzi et al., 2019), jointly detect of the positioning and transcriptional activity of loci.

In addition to improvements in spatial resolution, live imaging in combination with genome engineering using CRISPR–Cas9 systems facilitates and improves the study of 4D chromatin contact dynamics (changes in 3D chromatin structure over time) of individual loci. Chimeric array of gRNA-oligo (CARGO) and CRISPR–Cas-mediated Live FISH are two examples (Gu et al., 2018; H. Wang et al., 2019).

The integration of single-cell information of spatial positioning of genomic loci with functional genomic and epigenomic features, such as gene activity, or epigenetic states, will enable the tracking of chromatin and transcription dynamics in live cells during cell differentiation (McCord, Kaplan, & Giorgetti, 2020), opening venues for application ranging from basic science to diagnostics.

5.2 Sequencing-based methods

In contrast to microscopy, sequencing-based methods represent an orthogonal approach to investigate large-scale chromatin organization, providing rich sequence context but uncertain spatial context (Belmont, 2014).

5.2.1 Proximity ligation-based methods: chromosome conformation capture (3C)

A major breakthrough in chromatin biology was the establishment of proximity ligation-based methods (Cullen, Kladde, & Seyfred, 1993) and, in particular, chromosome conformation capture (3C) technologies (Dekker, Rippe, Dekker, & Kleckner, 2002), which marked the beginning of the era of high-throughput next-generation sequencing-based methods for the investigation of chromosome conformation. 3C-based methods provide quantitative, high-resolution, genome-wide measurements of physical proximity events within and across chromosomes, generally called chromatin contacts or interactions.

The first step of most 3C-based methods involves the formaldehyde crosslinking of cells, usually followed by *in situ* chromatin fragmentation by digestion with restriction enzymes such as HindIII or DpnII. Then, proximity-based ligation of adjacent DNA ends is followed by determination of pairwise interactions (Fig. 4a). After reverse crosslinking, different approaches can be used (de Wit & de Laat, 2012; Denker & de Laat, 2016). The classical 3C method interrogates a single pair of interacting loci (one-versus-one). In the circular chromosome conformation capture (4C) protocol, genome-wide interactions involving one locus of interest are detected (one-versus-many) (van de Werken et al., 2012). In the carbon copy chromosome conformation capture (5C) approach, chromatin interactions between two sets of loci are captured (many-versus-many) (Dostie & Dekker, 2007). In Capture-C methodology, biotin-labelled probes complementary to specific restriction fragment ends interrogate hundreds of pairs of loci of interest

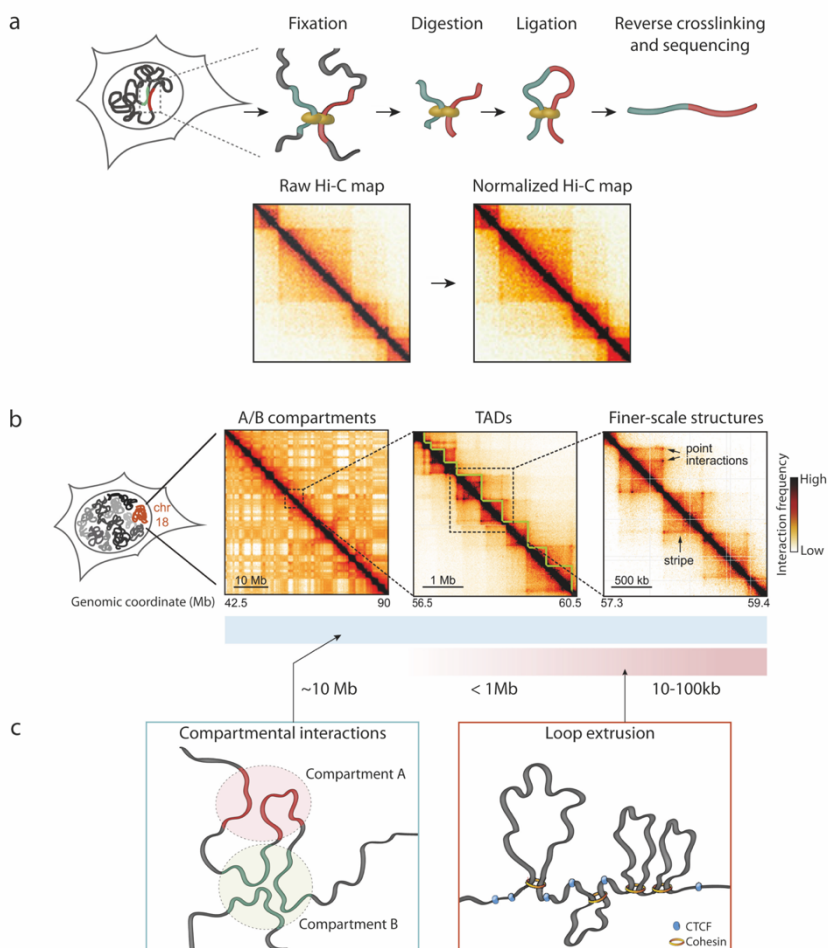


Figure 4. 3C methods and fundamental principles of mammalian chromosome organization.

(a) Scheme of the core steps in 3C protocols. Chromatin is crosslinked in cell nuclei and digested with a restriction enzyme (or endonuclease in the case of Micro-C), followed by ligation and decrosslinking. This results in the formation of hybrid DNA molecules that can be identified by high-throughput sequencing. In the case of Hi-C, the resulting list of genome-wide pairwise contacts can be represented by contact maps. Maps need to be corrected for biases and artifacts. (b) Hi-C contact maps illustrating the folding of mammalian chromosomes into A/B compartments (left), TADs (middle), and finer-scale structures (right). (c) Chromosome folding may be mainly driven by compartmentalization and loop extrusion. Figure adapted from (McCord et al., 2020).

(Hughes et al., 2014; Mifsud et al., 2015). In the Hi-C method, the coupling of 3C to high-throughput sequencing generates genome-wide catalogs of pairwise chromatin interactions (all-versus-all) within populations of billions of nuclei (Lieberman-Aiden et al., 2009). Progressively, these techniques have been tweaked by combination with chromatin immunoprecipitation to allow for enrichment of specific contacts associated to proteins of interest, including chromatin interaction analysis with paired-end tag (ChIA-PET) (Fullwood, Wei, Liu, & Ruan, 2009), HiChIP (Mumbach et al., 2016) and proximity ligation-assisted ChIP followed by sequencing (PLAC-seq) (Fang et al., 2016).

Despite their immense contribution to the field of genome structure, 3C-based approaches currently have limitations. First, resolution is strictly linked to sequencing depth and to the distribution of restriction sites. Techniques that are based on different fragmentation methods, such as DNase-Hi-C with DNase I (X. Deng et al., 2015; Ma et al., 2014; Ramani et al., 2016) and Micro-C with MNase (T. H. Hsieh et al., 2015), have successfully improved Hi-C resolution. Furthermore, it is still unclear what structural features at the cell population level represent at a single cell level (see Section 5.3). The typical maps obtained by 3C-based approaches represent a superimposition of different conformation states present in a population of cells, and the interpretation of the relationship between the number of detected ligation products with actual contact probabilities between genomic sequences has important implications for the biological significance of chromatin contacts, for the determination of an appropriate polymer model from experimental data (Fudenberg & Mirny, 2012), and for how data are normalized (Imakaev et al., 2012). Indeed, 3C-based data lack

internal normalization criteria, limiting the ability to compare contact frequencies different conditions and cell types. Another challenge is due to the fact that 3C-based methods rely on formaldehyde crosslinking and ligation, two molecular processes that represent potential sources of bias (Belmont, 2014; Gavrilov, Razin, & Cavalli, 2015; Williamson et al., 2014). For this reason, products captured by 3C-based approaches do not always reflect spatial proximity, and crosslinking might capture contacts between sequences located hundreds of nanometers apart, a distance range that is about one order of magnitude larger than the typical distance of contacts mediated by direct molecular interactions of the chromatin fiber through protein complexes. Moreover, techniques that rely on standard formaldehyde crosslinking inherently bias fragmentation towards open chromatin regions, and are potentially limited in capturing interactions of proteins with short residence time. Cap-C approach aims at circumventing such problem through dendrimer crosslinking, to achieve uniform fragmentation (You et al., 2021).

Reassuringly, general features of large-scale chromatin organization are generally recapitulated by 3C-based methods, microscopy and ligation-independent methods (see Section 5.2.2) (McCord et al., 2020). Despite significant discrepancies between DNA FISH and Hi-C have occasionally been reported (Williamson et al., 2014), these comparisons also show that Hi-C data are directly proportional to the fraction of cells in the population where a certain contact occurs at the moment of crosslinking, importantly for the development of mechanistic physical models of chromosome folding. Last but not least, standard 3C-based methods are unable to reveal whether multiple regions are interacting simultaneously (cooperativity) or mutually exclusively (exclusion), while

it is known that biologically relevant chromosome interactions may occur between pairs of loci as well as within hubs of cooperative contacts (Schoenfelder et al., 2015; Strom et al., 2017; Sutherland & Bickmore, 2009). Modified 3C versions have been developed to detect multi-contact configurations (Allahyar et al., 2018; Ay et al., 2015; Olivares-Chauvet et al., 2016; Oudelaar et al., 2018; Zheng et al., 2019) and, overall, have revealed common cooperative interactions between multiple loci, as well as multi-contact configurations occurring in a small subset of cells, which would be missed in population-averaged pairwise contact maps. Furthermore, complementary approaches, such as ligation-free genomic methods and super-resolution chromosome tracing, have revealed extensive evidence for cooperative multiway contacts, including highly transcribed regions that form transcription factories and super-enhancers (McCord et al., 2020). This type of data, combined with perturbative studies, may clarify the role of phase separation in the collective spatial partitioning of chromosome regions.

5.2.2 Ligation-independent techniques

The invention of ligation-independent techniques allowed to investigate chromosome conformation at the same time probing the nuclear position of chromatin contacts and multiway contacts, complementing intrinsic limitations and potential source of bias inherent of 3C-based methods (Kempfer & Pombo, 2020). Indeed, in 3C-like methods, genomic fragments ligation prior to sequencing is only partially efficient, and short paired-end sequencing, which does not provide information about multipartite in vivo chromatin interactions.

Among ligation-independent methods there are tyramide signal amplification (TSA), DNA adenine methyltransferase identification (DamID), split-pool recognition of interactions by tag extension (SPRITE), and genome architecture mapping (GAM) (Beagrie et al., 2017; Y. Chen et al., 2018; Guelen et al., 2008; Quinodoz et al., 2018; van Steensel & Henikoff, 2000; L. Zhang et al., 2020).

In SPRITE, crosslinked nuclei are isolated and fragmented, individual crosslinked pieces of chromatin are uniquely barcoded. After high-throughput sequencing, reads carrying the same combination of barcodes represent genomic sites that are a part of the same crosslinked cluster. In GAM, fixed cells are embedded in sucrose, frozen and cryo-sectioned, and DNA is extracted and sequenced from each section. Loci that are closer to each other in the nuclear space are co-sequenced more frequently than distant loci. As sections are taken from multiple nuclei sliced at random orientations, the co-segregation of all possible pairs of loci among a large collection of nuclear section profiles is used to generate a matrices of inferred locus proximities. Such maps are similar to Hi-C ones, even though GAM requires fewer cells — a few hundred nuclei produce maps that approximate those obtained from large populations of cells in Hi-C. Like SPRITE, GAM can identify multiple interactions, thereby enabling the direct study of multivalent enhancer–promoter interactions and of higher-order chromatin structures.

The combination of genome structure analysis with additional omics modalities is likely to offer critical information for revealing nuclear function. For instance, SPRITE has been further adapted into RD-SPRITE (Quinodoz et al., 2021), to enable mapping the interactions of RNA relative to other DNA and RNA, thereby allowing to determine

the relationship of some RNAs with the nuclear landmarks and compartments.

5.3 Heterogeneity and dynamics in chromosome conformation

Sequencing-based assays generally represent snapshots of chromosome conformation at a given time point, averaged over an entire cell population. Their relationship with the actual conformations of the chromatin fiber in single cells and their evolution in time is still unclear. In order to study chromatin conformation stochasticity and inter-cell variability while not compromising high throughput, an increasing number of chromatin analysis techniques are being developed into single-cell applications.

The first of these single-cell adaptations was single-cell Hi-C (scHi-C) (Nagano et al., 2013). Also, ligation-free tools have been migrated towards single cell assays, such as scSPRITE (Arrastia et al., 2021). A major observation from single cell chromatin conformation experiments and computational analyses is the existence of extensive inter-cell conformational variability at all genomic length scales (Finn et al., 2019). At the sub-TAD level, pairwise contacts and CTCF loops occur as stochastic events (Flyamer et al., 2017; Nagano et al., 2013; Ramani et al., 2017; Stevens et al., 2017; Tan, Xing, Chang, Li, & Xie, 2018), with only a subset of the contacts identified by population-average assays being present within an individual cell. Patterns of TADs and compartments in single cells are highly variable as well, and the ones observed in population Hi-C maps emerge when superimposing many single-cell conformations, reflecting preferential interactions in a highly stochastic ensemble of structures (Bintu et al., 2018; Boettiger et al.,

2016; Cardozo Gizzi et al., 2019; Cattoni et al., 2017; Giorgetti et al., 2014; Mateo et al., 2019; Nagano et al., 2013; Nora et al., 2012; Szabo et al., 2020; Szabo et al., 2018).

Thus, the probabilistic nature of higher-order spatial genome organization is a critically important feature, and average interaction maps generated using population-based methods appear as an ensemble of many different genome landscapes pertaining to multiple subpopulations of cells.

Since stochastic interactions between regulatory elements are likely to result in the stochastic transfer of regulatory information, pervasive cell-to-cell structural variability might have important implications for transcriptional regulation. However, single-cell genomics and fixed-cell imaging still generate static snapshots of 3D genome structures in single cells. Therefore, there are still many open questions about the degree of stochasticity and dynamicity of the exchange of regulatory information. For instance, very little is known about the timescale over which enhancer-promoter contacts assemble and disassemble, and how they relate to transcription and other nuclear processes. To address this issue, live-cell imaging (Brandao, Gabriele, & Hansen, 2021) and advances in genomic engineering are opening exciting possibilities to characterize the dynamics of chromatin looping and its link to the dynamic exchange of regulatory information and transcription.

6. Computational strategies for the analysis and representation of chromatin organization

The continuous evolution of experimental methods dedicated to the study of genome 3D organization is accompanied by rapid advancements of specialized algorithms to grasp the full biological significance of the experimental data. Since in the last decades massive amount of Hi-C data has been produced and greatly improved our characterization of nuclear structure.

6.1 Analysis of Hi-C data

Within Hi-C data, a series of factors introduce biases and limits in the resolution for the call of contact regions or domain boundaries. The achievable spatial resolution of Hi-C is affected by sequencing depth, library complexity and the DNA-cutting frequency of the enzyme used for chromatin fragmentation. This typically results in sparse Hi-C matrices, with many null entries, where the genuine absence of contacts and the absence of contacts due to low sequencing depth are undistinguishable. Moreover, uneven restriction fragment sizes and mappability levels across the genome make Hi-C matrices very heterogeneous at different genomic locations, while the decay of interaction frequencies with increase in genomic distance differentially affects Hi-C signal across different distances. Next, I briefly outline the major approaches affecting each of the key steps of Hi-C analysis (Fig. 4a, 4b):

Normalization. For the filtering and normalization of Hi-C data, different tools based on alternative strategies cope with typical Hi-C data biases.

Some of the most common methods are Yaffe and Tanay's one (Yaffe & Tanay, 2011), ICE (Imakaev et al., 2012), HICNorm (Hu et al., 2012), and OneD (Vidal et al., 2018).

Compartment analysis. In the first Hi-C study (Lieberman-Aiden et al., 2009), compartments have been identified by conversion of Hi-C matrices into correlation matrices, followed by principal component analysis to distinguish A and B compartment types. Later studies produced Hi-C maps based on much deeper sequencing, and additionally applied clustering steps such as Gaussian hidden Markov modelling for improved specification of epigenetic compartment signatures, leading to more detailed stratification of the A compartment into two sub-compartments and of the B compartment into three sub-compartments (S. S. Rao et al., 2014).

TAD detection. As regards TAD calling, although it is routinely done, there are numerous TAD callers that are based on different principles. Initial computational approaches, such as the insulation score and the directionality index, could not identify nested TADs (Dixon et al., 2012; H. Shin et al., 2016). Subsequently, other computational approaches were developed to inform on TAD hierarchy, such as Matryoshka (Malik & Patro, 2018), by further development of the linear score approach, ICFinder (Haddad, Vaillant, & Jost, 2017) and TADpole (Soler-Vila et al., 2020), by clustering of contacts' map data, or 3DnetMod (Norton et al., 2018), by graph theory-based algorithms.

Loop analysis. Thanks to the increase in Hi-C maps resolution, it became possible to detect specific chromatin contacts and loops, corresponding to statistically significant enrichment in contact frequency compared with a general background model. One of the first loop-dedicated algorithms, HiCCUPS (S. S. Rao et al., 2014), identifies a chromatin

loop as the most enriched bin compared with its immediate neighborhood. The tool Fit-Hi-C instead assigns statistical confidence to contacts by using random polymer modelling, while accounting for known Hi-C biases such as genomic distance, to find significant interactions (Ay, Bailey, & Noble, 2014). HiCPlus is a machine learning approach based on deep convolutional neural network that enhances Hi-C maps with low-sequence depth, to overcome the resolution limit of Hi-C maps for better loops and TAD borders detection (Y. Zhang et al., 2018).

6.2 3D modeling approaches

Computational modeling provides an important avenue for interpretation of data generated by experimental techniques that probe chromatin conformation and inference of the underlying chromatin 3D structure. The configuration in space of the genome serves as a quantitative framework to integrate information from different types of experimental datasets, often allows to test hypothesis regarding underlying molecular mechanisms and to generate predictions that can be experimentally tested. Importantly, models translate 3C information into a context of distances/space helping discern simultaneous from exclusive contacts, and representing heterogeneity between cells and dynamics across time.

Most modelling approaches subdivide the genome in chunks, by either specific underlying features or a defined genomic length. Each chunk is then represented by connected points or spheres, or alternatively as elements composing a polymer (Oluwadare, Highsmith, & Cheng, 2019). A set of parameters or physics rules constrains such particles to

define how they interact with the rest of particles and how the model folds.

Chromatin modelling approaches can be divided into two main categories: *ab initio* models aim at understanding the processes of genome folding and identifying components shaping the genome, while data-driven models are focused on the more refined analysis of the represented chromatin (Bendandi, Dante, Zia, Diaspro, & Rocchia, 2020; Lin, Bonora, Yardimci, & Noble, 2019; Marti-Renom & Mirny, 2011).

On the one hand, *Ab initio* models use as input statistical features and physics principles to simulate the behavior of chromatin in the 3D space, by applying a conjunction of known and hypothesized properties to the chromatin fiber. Usually, varying levels of packing conformation of the chromatin fiber and the behavior of the bead-spring polymer models are assumed, with defined toughness, elasticity and behavior (Finch & Klug, 1976; Rosa & Everaers, 2008). Polymer folding has been typically modelled as an equilibrium globule (Mirny, 2011), or as a fractal globule (Grosberg, Nechaev, & Shakhnovich, 1988). The last modality is consistent with the first genome-wide chromosome interaction maps, where it was observed that different chromatin regions poorly intermingle, probably allowing for rapid access to active regions by the transcriptional machinery (Lieberman-Aiden et al., 2009). Methods following these approaches have contributed to prove that loop extrusion processes could be sufficient to drive chromatin compaction (Goloborodko, Marko, & Mirny, 2016) and form chromosomal domains (Fudenberg et al., 2016), and that epigenetic features such as chromatin states contribute to the formation of TADs and compartments (Di Pierro, Cheng, Lieberman Aiden, Wolynes, &

Onuchic, 2017; Falk et al., 2019; Jost et al., 2014; Jost & Vaillant, 2018). A current limitation of such models is that typically they do not perform equally well at the different scales of loops, TADs, compartments, chromosome territories, partly due to the considerable computational time that they require.

On the other hand, data-driven models are focused on the treatment and transformation of experimental data into restraints, to reliably reconstruct its 3D organization. Restraints may be inferred from interaction data, such as Hi-C experiments, additional experimental observations such as nuclear dimensions or chromatin-lamina interactions, and physics properties of the chromatin like the bending rigidity of the fiber (Serra et al., 2015). The resolution of the experiment and the computational workload are the main limiting factors. Therefore, when analyzing long chromatin fibers such as the whole human genome, resolution is normally lowered at about a megabase. When instead models are focused on specific selected regions of interest, they typically reach a resolution of few kilobases, closer to the limit defined by the experiment itself (Serra et al., 2015). Scoring functions infer how well the 3D distances between the output model represent the input interaction data, and, finally, the conformations that best satisfy the imposed restraints are retained. Modelling methods are further divided into two classes. Consensus-based modeling approaches analytically provide a single consensus structure that best explains the input interaction data, with reduced computational time. Ensemble-based modeling methods, instead, determine a set of 3D conformations that try to account for the variability of population 3C-based datasets (Lin et al., 2019). Among ensemble-based methods, TADbit (Serra et al., 2017) is well-suited for chromatin 3D modeling from Hi-C data,

serving of the Integrative Modeling Platform (IMP) (Russel et al., 2012) for the application of spatial restraints. First, the input interaction data is normalized and transformed via \log_{10} and Z-score. Then, chromatin is represented as a chain of particles, with a diameter defined by the resolution of the data. A combination of parameters is used to transform the Z-scores of non-consecutive particles into different types of restraints and assign a range of allowed distances to each pair of particles, while consecutive particles are spatially restrained by their occupancy. Finally, the restraints are applied starting from randomly distributed particles, by a series of Monte Carlo rounds combined with standard simulated annealing. The output of this process is an ensemble of models that best fit the input restraints, while minimizing the defined scoring function for the different parameter combinations. Subsequently, the comparison of obtained ensembles with the input interaction matrix allows to optimize the parameters. By using only Hi-C data as input, TADbit was able to generate models at the kilobase scale representing distinct 3D features associated to previously defined epigenetic states (Filion et al., 2010; Serra et al., 2017). Additionally, provided Hi-C data from time course experiments are available, this type of modelling can interpolate the restraints through the various time points to deliver information about chromatin folding dynamics (Di Stefano et al., 2020).

Overall, modelling strategies have mostly focused their attention on technologies like Hi-C, which have been widely used in the last decade. However, other chromatin interaction technologies, like 4C, Promoter Capture Hi-C (PCHi-C) or HiChIP, are being increasingly employed, and, since they produce sparser interaction datasets compared to Hi-C,

the use of appropriate methods is required (Mendieta-Esteban, Di Stefano, Castillo, Farabella, & Marti-Renom, 2021).

Improvements in algorithms and computation power are crucial complementary tools to experimental methods, and will hopefully soon allow to model the dynamics of whole-genome folding at high spatial and temporal resolution.

7. The relationship between genome function and structure

Abundant experimental evidence suggests that chromatin structural dynamics contributes to the specification of distinct gene expression programs and biological functions (Galupa & Heard, 2017; Spitz, 2016). Perturbative studies coupling existing methods, notably 3C-based, with recently developed techniques such as CRISPR–Cas9 technology, give unprecedented opportunities to manipulate genome architecture and explore the mechanistic connections between chromosome structure and nuclear biology. However, because of contradicting evidences, the mechanisms regulating dynamic chromatin changes and the causality between genome topology and transcription are under intense investigation.

Positioning patterns of genes and chromosomes differ between cell types, and undergo changes during physiological processes such as differentiation, development, aging, and in pathological conditions. The genomic landscape in embryonic stem (ES) cells is abundantly associated to active marks (Meshorer et al., 2006; Mikkelsen et al., 2007), and its structure is maintained in a globally open, readily accessible

configuration, allowing for maximum plasticity (Fussner et al., 2011). Upon ES cells differentiation, many of ES cell-specific chromatin hallmarks rapidly disappear.

Despite primary domain architecture seems to be mainly preserved in different cell types and across species (Dixon et al., 2012; S. S. Rao et al., 2014; Sexton et al., 2012), during lineage specification in early stages of human development intra-TAD interactions in some domains are strongly altered, often correlate with relocation of the TAD from one compartment to another, and with changes in chromatin accessibility and transcription status (Dixon et al., 2015). On the same line, in response to the transient stimuli of hormone treatment in breast cancer cells, substantial changes in transcription are accompanied by only few dynamic TAD boundary regions, but TADs respond to the hormone treatment as a unit. Responsive TADs change epigenetic signature, switch between the A and B compartments and undergo changes in their level of compaction, suggesting that the transcription status might be coordinated within a TAD (Le Dily et al., 2014).

It has also been described that mature B cell formation and activation involves a strong relationship between nuclear architecture, TFs, and the epigenetic machinery (Azagra, Marina-Zarate, Ramiro, Javierre, & Parra, 2020; Stadhouders et al., 2018), including the formation of DNA loops between distant regulatory regions mediated by CTCF, and potentially also by lncRNAs (Bunting et al., 2016; Kieffer-Kwon et al., 2017; Ramachandrareddy et al., 2010).

In human primary hematopoietic cell types and embryonic stem cell-derived cardiomyocytes, the promoters interactome has been demonstrated to be crucial for enhancers to contact their target genes in a cell-type specific manner, and for non-coding genome-wide

association study (GWAS) variants to be linked with putative target genes, shedding light on the genomic regulatory mechanisms underlying common diseases (Choy et al., 2018; Javierre et al., 2016).

Several elegant genetic perturbation studies have together favored a model in which TADs ensure proper spatiotemporal regulation of gene expression, by creating insulated neighborhoods that demarcate the enhancer search space for target gene promoters in the appropriate developmental time window (Beagan & Phillips-Cremins, 2020; Norton & Phillips-Cremins, 2017; Symmons et al., 2014). *Hox* gene cluster domains constitute a representative example of this principle. They are among the best-studied Polycomb domains, which in general are formed by clusters of Polycomb-bound sites with preferential interactions (Bantignies et al., 2011; Lanzuolo et al., 2007; Schuettengruber et al., 2014; Sexton et al., 2012). In mouse embryonic stem cells, where *Hox* genes are transcriptionally inactive, they associate into a single Polycomb domain that is well separated from flanking active regions (Vieux-Rochas et al., 2015). Upon *Hox* gene activation during differentiation, active genes progressively segregate into an active TAD, and the transition in spatial configuration coincides with the change of chromatin marks from a repressed to an active state (Noordermeer et al., 2014; Noordermeer et al., 2011). Architectural protein CTCF seems to be a key protein in insulating active and repressed *Hox* clusters into spatially disjoint domains (Narendra et al., 2015).

Although the relationship between TAD boundaries, insulation and disease is not entirely clear, structural variations perturbing TAD boundaries, CTCF binding, and insulation can lead to aberrant gene expression, developmental defects and disease (Akdemir et al., 2020;

Andrey & Mundlos, 2017; Bruneau & Nora, 2018; Despang et al., 2019; Downen et al., 2014; Flavahan et al., 2016; Franke et al., 2016; Hnisz et al., 2016; Kraft et al., 2019; Laugsch et al., 2019; X. S. Liu et al., 2018; Lupianez et al., 2015; Lupianez, Spielmann, & Mundlos, 2016; Narendra et al., 2015; Valton & Dekker, 2016; van Bemmelen et al., 2019; Weischenfeldt et al., 2017).

Chromatin looping has been demonstrated to play a critical role in activation or repression of gene expression, depending on the specific cases. In a landmark study, forcing a loop between the β -globin promoter and its locus control region (LCR) in absence of the TF GATA1, which is normally required for β -globin expression, was sufficient to recruit RNAPII and upregulate the expression of the β -globin gene (W. Deng et al., 2012). In *D. melanogaster*, the prevention of loop formation showed that Polycomb-dependent genomic loops can contribute to gene silencing during development (Ogiyama et al., 2018). Interestingly, acute depletion of CTCF or of cohesin complex subunits results in the disruption of most of loop domains across the genome, while compartmentalization is unaffected or strengthened (Nora et al., 2017; S. S. P. Rao et al., 2017; Schwarzer et al., 2017), and changes in gene expression are unexpectedly modest (Beagan & Phillips-Cremins, 2020). Extensive genome-wide deletions, duplications and inversions in *Drosophila* impact chromatin-domain placement, but generate only minor alterations in gene expression (Ghavi-Helm et al., 2019). These results indicate that possibly not all genes might be regulated through long-range spatial contacts (Beagan & Phillips-Cremins, 2020). Overall, the data available to date indicate a dynamic, reciprocal interplay between transcription and fine-scale genome organization. Loops and domains can modulate function, albeit to a modest degree in some

cases, and genome transcription can also influence looping structures. In contrast, transcription has only moderate effects on domain organization and is not sufficient to create new domain boundaries (van Steensel & Furlong, 2019). As regards A and B compartments, since yet there is no way to prevent their formation without perturbing the nuclear processes by which they form, i.e. self-association and/or phase separation of similarly modified chromatin, experiments to test their functional role are still missing.

Globally, the emerging picture points to a self-organizing function-structure-function model of genome organization. Genome topology might be a modulatory, rather than deterministic, regulator of genome function, consistently with the observed stochastic nature of gene expression. In this model, genome activity would primarily be dictated by DNA sequence, and drive genome topology and epigenetic patterns. Resulting topological and epigenetic features would in turn reinforce genome function, superimposing additional layers of regulation, maintaining the ground state generated by the genetic information and acting as a buffer to potentially detrimental environmental influences, such as cellular stress or aberrant signaling. Possibly, epigenetic and structural mechanisms may alter the functional state of a certain genomic region, such as by placing an active gene into a heterochromatic, repressed environment. Consequently, the newly induced functional state would strengthen the epigenetic and structural features of the genomic region.

CHAPTER 1

Coordinated changes in gene expression, H1 variant distribution and genome 3D conformation in response to H1 depletion

Candidate's contribution: Study of all aspects of the work related to the analysis of Hi-C maps of the 3D genome.

Núria Serna-Pujol[†], Mónica Salinas-Pena[†], Francesca Mugianesi[†], François Le Dily, Marc A. Martí-Renom, Albert Jordan. *Coordinated changes in gene expression, H1 variant distribution and genome 3D conformation in response to H1 depletion*. Nucleic Acids Research, 2022;, gkac226, <https://doi.org/10.1093/nar/gkac226>.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Coordinated changes in gene expression, H1 variant distribution and genome 3D conformation in response to H1 depletion

Núria Serna-Pujol^{1,#}, Mónica Salinas-Pena^{1,#}, Francesca Mugianesi^{2,#}, François Le Dily³, Marc A. Marti-Renom^{2,3,4,5}, Albert Jordan^{1,*}

¹Molecular Biology Institute of Barcelona (IBMB-CSIC), Barcelona, 08028, Spain

²CNAG-CRG, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona, 08028, Spain

³Centre for Genomic Regulation, The Barcelona Institute for Science and Technology, Carrer del Doctor Aiguader 88, Barcelona, 08003, Spain

⁴Pompeu Fabra University, Doctor Aiguader 88, Barcelona, 08003, Spain

⁵ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

[#]These authors contributed equally to this work.

^{*}To whom correspondence should be addressed. Tel: +34 93 402 0487; Fax: +34 93 403 4979; Email: Albert.jordan@ibmb.csic.es

ABSTRACT

Up to seven members of the histone H1 family may contribute to chromatin compaction and its regulation in human somatic cells. In

breast cancer cells, knock-down of multiple H1 variants deregulates many genes, promotes the appearance of genome-wide accessibility sites and triggers an interferon response via activation of heterochromatic repeats. However, how these changes in the expression profile relate to the re-distribution of H1 variants as well as to genome conformational changes have not been yet studied. Here, we combined ChIP-seq of five endogenous H1 variants with Chromosome Conformation Capture analysis in wild-type and H1.2/H1.4 knock-down T47D cells. The results indicate that H1 variants coexist in the genome in two large groups depending on the local GC content and that their distribution is robust with respect to H1 depletion. Despite the small changes in H1 variants distribution, knock-down of H1 translated into more isolated but de-compacted chromatin structures at the scale of topologically associating domains (TADs). Such changes in TAD structure correlated with a coordinated gene expression response of their resident genes. This is the first report describing simultaneous profiling of five endogenous H1 variants and giving functional evidence of genome topology alterations upon H1 depletion in human cancer cells.

INTRODUCTION

DNA is packaged within the nucleus to efficiently regulate nuclear processes. Chromatin packing involves several hierarchical levels of organization that have been mostly described by chromosome conformation capture techniques, among others. First, at megabases scale, the genome can be segregated into the so-called A and B compartments. The A compartment represents active, accessible chromatin with a tendency to occupy a more central position in the

nucleus. The B compartment corresponds to heterochromatin and gene deserts enriched at the nuclear periphery (1). Second, topological associating domains (TADs), which are submegabase structures, interact more frequently within themselves than with the rest of the genome (2–4). TADs are conserved across species and cell types and show a coordinated transcriptional status (5,6). Third, these domains are formed by assemblies of chromatin loops with physical properties that, ultimately, depend on the histone composition and modifications of its resident nucleosomes. In particular, histone H1, which has classically been regarded as a simple condenser, is now known to contribute to the higher-order organization of the genome (7–9). Histone H1 family is evolutionary diverse and human somatic cells may contain up to seven H1 variants (H1.1 to H1.5, H1.0 and H1X). H1.1–H1.5 variants are expressed in a replication-dependent manner while H1.0 and H1X are replication-independent. H1.2 to H1.5 and H1X are ubiquitously expressed, while H1.1 is restricted to certain tissues and H1.0 accumulates in terminally differentiated cells (8,10,11).

Several studies support the idea that H1 variants are not redundant and that functional specificity may exist with H1 variants non-randomly distributed in the genome and interacting with different protein partners (12–18). For example, in breast cancer cells, knock-down (KD) of each individual H1 variant deregulates different subsets of genes (17,19). In mouse embryonic stem cells (ESCs), H1c and H1d (orthologs of the human H1.2 and H1.3, respectively) are depleted from high GC/gene-rich regions and are enriched at major satellites (14). In IMR90 cells, H1.2–H1.5, in contrast to H1.1, are depleted from CpG-dense and regulatory regions (15), with H1.5 binding correlating with depletion of RNA polymerase II (RNAPol II) and repression of target genes in

differentiated cells (13). In skin fibroblasts, H1.0 distribution correlates with GC content and is abundant at gene-rich chromosomes (18). In T47D breast cancer cells, all H1 variants are depleted at promoters of active genes (16) and tagged-H1s are enriched at high GC regions with endogenous H1.2 and H1X resulting in opposite profiles. That is, while H1.2 is found in low GC regions and lamina-associated domains (LADs), H1X strongly correlates with GC content and is associated to RNAPol II binding sites (16,17). Moreover, H1.2 and H1X have an opposite distribution among Giemsa bands (G bands), being H1.2 and H1X associated with low and high GC bands, respectively (20). Finally, a strong correlation has been observed between high H1.2/H1X ratio and the so-called genome B compartment, low GC bands and compact, late-replicating chromatin (20). Although no functional Hi-C experiments have been performed in H1-depleted human cells, the direct involvement of linker histones in chromatin structure has been proved in mouse ESCs. Hi-C experiments were performed in wild-type and H1-triple knockout (TKO) ESCs. In H1 TKO, an increase in inter-TAD interactions correlated with changes in active histone marks, increased number of DNA hypersensitivity sites and decreased DNA methylation (21). These results point to an essential role of histone H1 in modulating local chromatin organization and chromatin 3D organization.

To study the consequences H1 depletion in human cells, we have previously generated a derivative T47D cell line containing a short-hairpin-RNA that affects the expression of several H1 genes as well as the protein levels of mainly H1.2 and H1.4 (22). In such cell line, the H1 total levels are reduced to $\approx 70\%$, which results in heterochromatic repeats including satellites and endogenous retroviruses overexpression

that triggers a strong interferon response. Using this system, here we aim at studying the effects of H1 variant depletion in chromatin organization and nuclear homeostasis. To address this question, we have performed ChIP-seq in T47D breast cancer cells, and Hi-C experiments under basal conditions and after combined depletion of H1.2 and H1.4 (H1 KD). Profiling of endogenous H1 variants revealed that H1.2, H1.5 and H1.0 were abundant at low GC regions while H1.4 and H1X preferentially co-localized at high GC regions. Profiling of H1s within chromatin states showed that all H1 variants were enriched at heterochromatin and low-activity chromatin, but H1X was more abundant at promoters compared to other H1 variants. After H1 KD, chromatin accessibility increased genome-wide, especially at the A compartment where H3K9me3 abundance was reduced. Similarly, the B compartment, where H1.2 was enriched at basal conditions, also showed a more open state. Interestingly, these changes occurred with only slight H1 variant redistributions across the genome. For example, H1.4 profile switched towards the H1.2 group and H1X decreased at heterochromatin and increased in almost all other chromatin states. Our Hi-C results also indicate that upon H1 KD, parts of the genome suffered changes in compartmentalization with no specific direction and TADs increased their internal interactions, which resulted in an increased TAD border strength. In particular, those regions of the genome with high H1.2 overlap resulted in increased local interactions upon H1 KD. Such structural changes were parallel to coordinated gene expression changes within TADs with up-regulated genes enriched in TADs with low basal gene expression and high H1.2 content. Finally, the three-dimensional (3D) modeling of TADs with coordinated gene response indicate that they suffered a general decompaction upon H1

KD. This is the first report describing simultaneous profiling of five endogenous H1 variants within a cell line and giving functional evidence of genome topology alterations upon H1 KD in human cancer cells.

RESULTS

A stable genome distribution of H1 variants correlates with GC content and chromatin state

It has been previously described that the content of histone H1 variants varies between cell types and along differentiation (8,40). Moreover, its genomic distribution is non-homogeneous and with specific patterns depending on the variants (13–18,20). Therefore, we hypothesize that altering the H1 variants composition in a particular cell type may affect the genomic distribution of the different variants. To test this, we performed ChIP-seq experiments in T47D cells harboring an inducible multiH1 shRNA expression vector which, upon Doxycycline treatment, efficiently depletes H1.2 and H1.4 proteins (H1 KD) (22). After testing the efficacy of H1 KD by Western blot (Figure 1A), we performed ChIP with antibodies against endogenous H1.2, H1.4, H1.5, H1.0 and H1X. The amount of DNA immunoprecipitated with H1.2 and H1.4 antibodies decreased >65% in treated cells compared to untreated, confirming the antibody specificity and the effect of the H1 knock-down. ChIPed DNA was qPCR-amplified with oligonucleotides for TSS and distal promoter regions of *CDK2* (active) and *NANOG* (inactive) genes (Figure 1B), which confirmed that all ChIPs efficiently worked compared to unspecific IgG. The active gene presented the characteristic H1 valley at the Transcription Starting Site (TSS) compared to the distal region, but not the inactive gene (16). Upon H1 KD, the signal of H1.2 and H1.4 significantly decreased, while

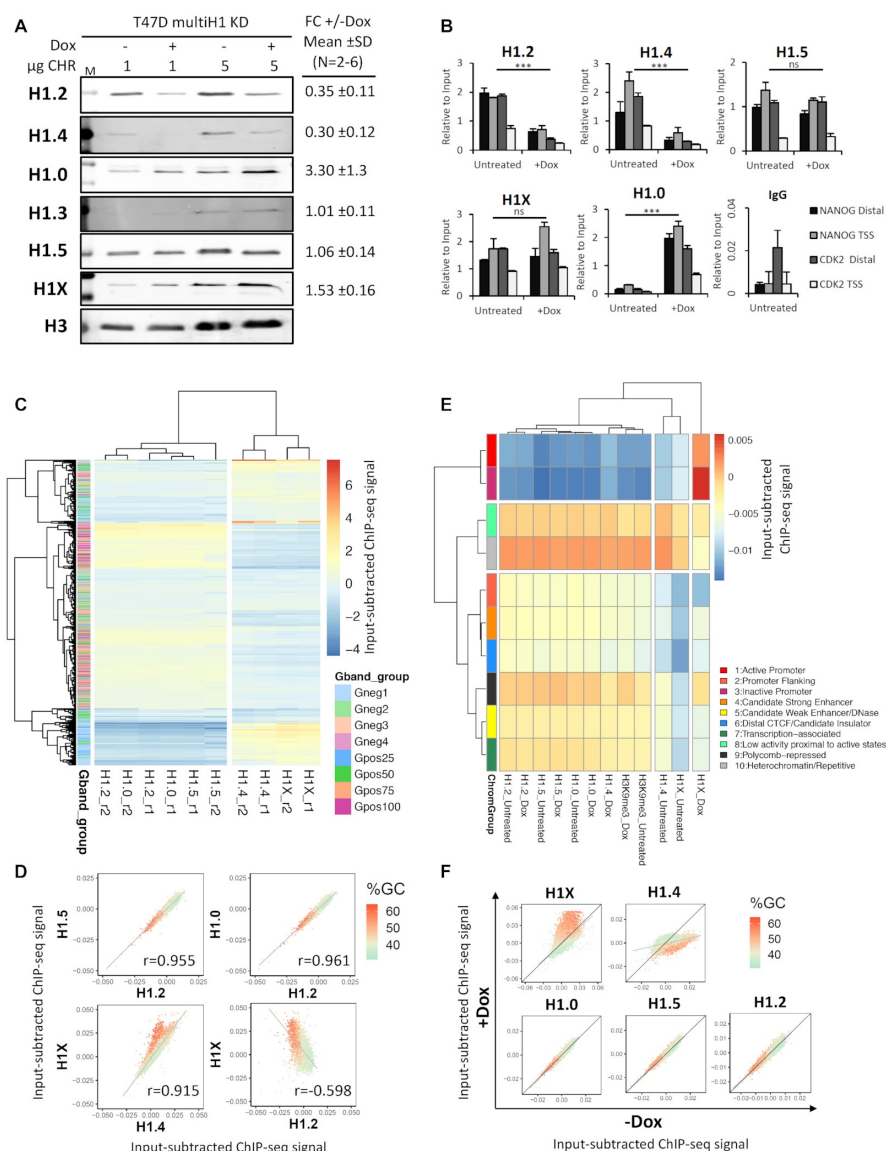


Figure 1. Genomic distribution of histone H1 variants upon H1 knock-down in breast cancer cells. **(A)** Immunoblot analysis of H1 depletion in H1 KD cells. Chromatin extracts (1 or 5μg of protein) from T47D multiH1 KD cells cultured in the presence or not of Doxycycline for 6 days were run in SDS/PAGE and immunoblotted with the indicated antibodies against H1 variants or histone H3 as loading control. ImageJ Immunoblot quantification of multiple experiments is indicated as mean (ratio + Dox/untreated) and SD. Number of biological replicates used for quantification were: $n = 6$ (H1.2, H1.4 and H1.0), $n = 4$ (H1X), $n = 2$ (H1.3, H1.5). **(B)** ChIP-qPCR of H1 variants in multiH1 KD cells. Chromatin from untreated or Dox-treated H1 KD cells was used for ChIP with antibodies against H1

variants and unrelated IgG as a control. Resulting DNA was amplified by qPCR with oligos for distal promoter (3kb upstream TSS) and TSS regions of genes *CDK2* and *NANOG*. ChIP amplification is shown relative to input DNA amplification. A representative experiment quantified in triplicate is shown. Statistical differences between Untreated (-Dox) and +Dox immunoprecipitated DNA for each H1 variant are supported by paired-*t*-test. (***) $P < 0.001$; (ns/non-significant) $P > 0.05$. **(C)** Heat map and cluster analysis of the input-subtracted ChIP-seq abundance of H1 variants within Gpos and Gneg bands from untreated T47D cells. The color of heatmap grids represents the relative input-corrected coverage of the H1 variant indicated at the X-axis within each G band, while the Y-axis shows to which group the band belongs. Two main clusters of H1 variant distribution are formed, one with abundant H1.X and H1.4 at high GC bands, the other with H1.2, H1.5 and H1.0 enriched at low GC bands. Two replicates are shown (r1, r2). **(D)** Scatter plots of the indicated H1 variant pairs input-subtracted ChIP-seq abundance within 100-kb bins of the human genome. The GC content at each bin is color-coded. Pearson's correlation coefficient is shown (P -value < 0.001). **(E)** Heat map and cluster analysis of the input-subtracted ChIP-seq abundance of H1 variants from WT or H1 KD T47D cells (-/+Dox) within 10 chromatin states (ChromHMM segmentation). The profile of H3K9me3 is included. For each heatmap grid, the color represents the input-corrected coverage of the H1 variant identified by the X-axis within each region, while the Y-axis shows the ChromHMM group the region belongs to. **(F)** Scatter plots of H1 variants input-subtracted ChIP-seq abundance within 100-kb bins of the human genome in multiH1 KD cells treated or not with Doxycycline. The GC content at each bin is color-coded. (C, E) Heatmaps were performed by using the R package 'pheatmap'. The 'euclidean' distance measure and the 'complete' cluster method were used in clustering rows and columns.

H1.0 signal increased, in agreement with the Western blot results (Figure 1A-B). The effect of H1 KD as well as the specificity of H1 antibodies were further confirmed RT-qPCR, western blot, ChIP-qPCR and mass-spectrometry (Supplementary Figure S1 and Supplementary Table S2). Cell cycle analysis in H1 KD cells is also shown (Supplementary Figure S1).

We have previously shown that H1.2 strongly correlates with B compartment, late replicating, inaccessible chromatin and low GC bands (20). To further extend this analysis, we measured the ChIP-seq abundance of each H1 within G bands and compared its distribution upon H1 KD. A browser snapshot of the distribution of the H1 variants

in the genome is shown in Supplementary Figure 2. Unsupervised clustering of H1 variant distributions in G bands clearly show the existence of two major clusters of H1 variants within G bands (Figure 1C and Supplementary Figure S3A). In untreated cells, H1.2 was enriched towards low GC content regions (that is, Gpos100/Gneg4, repressed bands), and H1X was enriched at high GC (that is, Gpos25/Gneg1, active bands). Additionally, H1.4 was also enriched at high GC bands, whereas H1.5 and H1.0 were enriched towards low GC bands. These results confirm and expand previous findings on the distribution of H1 variants in the genome (16,20). Interestingly, correlation analysis of the distribution of H1 variants genome-wide using bins of 100-kb confirmed the existence of these two groups of variants (i.e. H1.2, H1.5 and H1.0 in low GC regions as well as H1.4 and H1X in high GC regions) with H1.2 and H1X selected as prototypes of the two groups and opposed distribution within the genome (Figure 1D and Supplementary Figure S3B, C). Next, we assessed whether the clustering distribution of H1 variants would also correlate with genomic chromatin states. To do so, we used as a proxy 10-chromatin states (colors) maps generated elsewhere from several genomic datasets of HeLa-S3 cells (41). Most H1 variants were particularly abundant within *heterochromatin/ repetitive* and *low-activity* chromatin states, but also at *polycomb-repressed*, *transcription-associated* and *weak-enhancer* (Figure 1E and Supplementary Figure S3D). In concordance, such variants were underrepresented at *active* and *inactive promoter* states. This trend was broken by the H1X variant, which was enriched at *promoters*, compared to other variants, confirming that this variant is the most specific of all with respect its

genome-wide distribution. H1.4 was the variant that overlapped the most with H3K9me3 profile within chromatin states.

Changes of H1 variants distribution upon H1 KD were further analyzed within 100-kb bins throughout the genome. Upon H1 KD, H1.0 distribution was unaltered, while H1.2 and H1.5 were slightly increased specially at high GC bins. H1X occupancy increased at high GC bins and decreased at low GC bins, whereas H1.4 decreased at high GC bins (Figure 1F and Supplementary Figures S2, 3A). Similarly, upon H1 KD, H1.2, H1.5 and H1.0 profiles within chromatin states were not altered and H1X profile decreased at *heterochromatin* and increased in almost all other chromatin states, particularly at *Polycomb-repressed* regions and *promoters*, and among them the highest increase occurred at *inactive promoters* (Figure 1E and Supplementary Figure S3D). Finally, H1.4 profile switched towards the H1.2 group. It has to be considered that H1.2 and H1.4 profiles refer to the relative ChIP-seq signal remaining after efficient KD (ca. $\approx 65\%$ of the H1.2 or H1.4 genomic abundance was disappeared).

The average profiles of all H1 variants around transcription start sites (TSS) or termination sites (TTS) and around coding genes was calculated using CEAS software and is shown in Supplementary Figure S4A. All H1 variants showed depletion around TSS of active genes and no changes upon H1 KD, except for H1X, which was enriched around TSS of genes, especially upon H1 KD. Annotation of genomic regions enriched for the different variants showed that H1.2, H1.5 and H1.0 were enriched at intergenic regions both in the absence or presence of Doxycycline, whereas H1X and H1.4 were enriched at promoters and introns, compared to the other variants, in wild-type conditions, but distribution was altered upon H1 KD (Supplementary Figure S4B).

H1X was further enriched at promoters, exons and UTRs upon H1 KD, whereas the remaining H1.4 was decreased from introns and increased at intergenic regions.

Furthermore, differential genomic distribution of H1 variants in T47D cells established by ChIP-seq here is compatible with immunofluorescence imaging of H1 variants within the nuclei (Supplementary Figure S5). H1.2 showed enrichment towards the nuclear periphery and co-localized with lamin A, features of heterochromatin; H1.5 presented a similar pattern. Instead, H1X and H1.4 showed a punctuated pattern inside the nuclei, without lamin A overlapping. Notably, H1X was highly enriched at the nucleolus, as previously reported (17,42). H1.0 was also distributed overall the nucleus but no general overlapping with H1.4 was found, confirming that they occupied different genomic regions. Upon H1 KD, abundances of H1.2 and H1.4 were highly reduced, whereas H1X and H1.0 were increased. However, H1 variants redistribution within chromatin states upon H1 KD was difficult to evaluate with this technique.

In summary, ChIP-seq data in T47D cells demonstrated that H1 variants are differentially distributed through the genome in two profiles: H1.2, H1.5 and H1.0 enriched towards low GC regions and H1X and H1.4 more abundant at high GC regions. Still, all H1 variants are abundant within heterochromatin or inactive regions of the genome. Upon H1.2 and H1.4 depletion, H1.2, H1.0 and H1.5 did not significantly change their genomic distribution, whereas H1X increased at high GC regions, where H1.4 was selectively depleted. H1.0, whose expression and protein levels increased, was homogeneously incorporated throughout the genome.

Changes on genome architecture upon depletion of multiple histone H1 variants

Chromosome conformation capture techniques such as Hi-C allows to detect local and distal contacts within the genome and to establish the position of borders flanking the so-called topologically associating domains (TADs). Hi-C experiments also allow to establish a division of the genome into two compartments, active (A) and inactive (B). To address the consequences of histone H1 depletion on genome architecture, we prepared nuclear DNA from untreated and 6-days Doxycycline-treated multiH1 shRNA cells, in two independent experiments with a total of 3 replicates, and performed the Hi-C protocol (Supplementary Figure S6). After assessing the similarity between Hi-C replicates using HiCRep score (Materials and Methods and Figure 2A), replicates within samples were merged and analysed as a single experiment for WT and H1 KD. Analysis of the average Hi-C interactions as a function of genomic distance indicates that upon H1 depletion there was a decrease in short and medium-range interactions (<30 Mb), and an increase in long-range contacts (>30 Mb) (Figure 2B). To further characterize where those average changes occurred, we segmented the genome first into compartments and then into TADs for WT and H1 KD samples (Supplementary Figure S2).

The segmentation of the genome into the A and B compartments remained largely unchanged upon H1 KD (~80% of the 100-kb bins did not change compartment, Figure 2C). However, significant differences in compartmentalization were observed. For example, about 280 Mb of the genome decompacted (B to A direction) after H1 KD with 1/3 of the bins moving from the B compartment to an A compartment. Conversely, about 294 Mb of the genome compacted (A

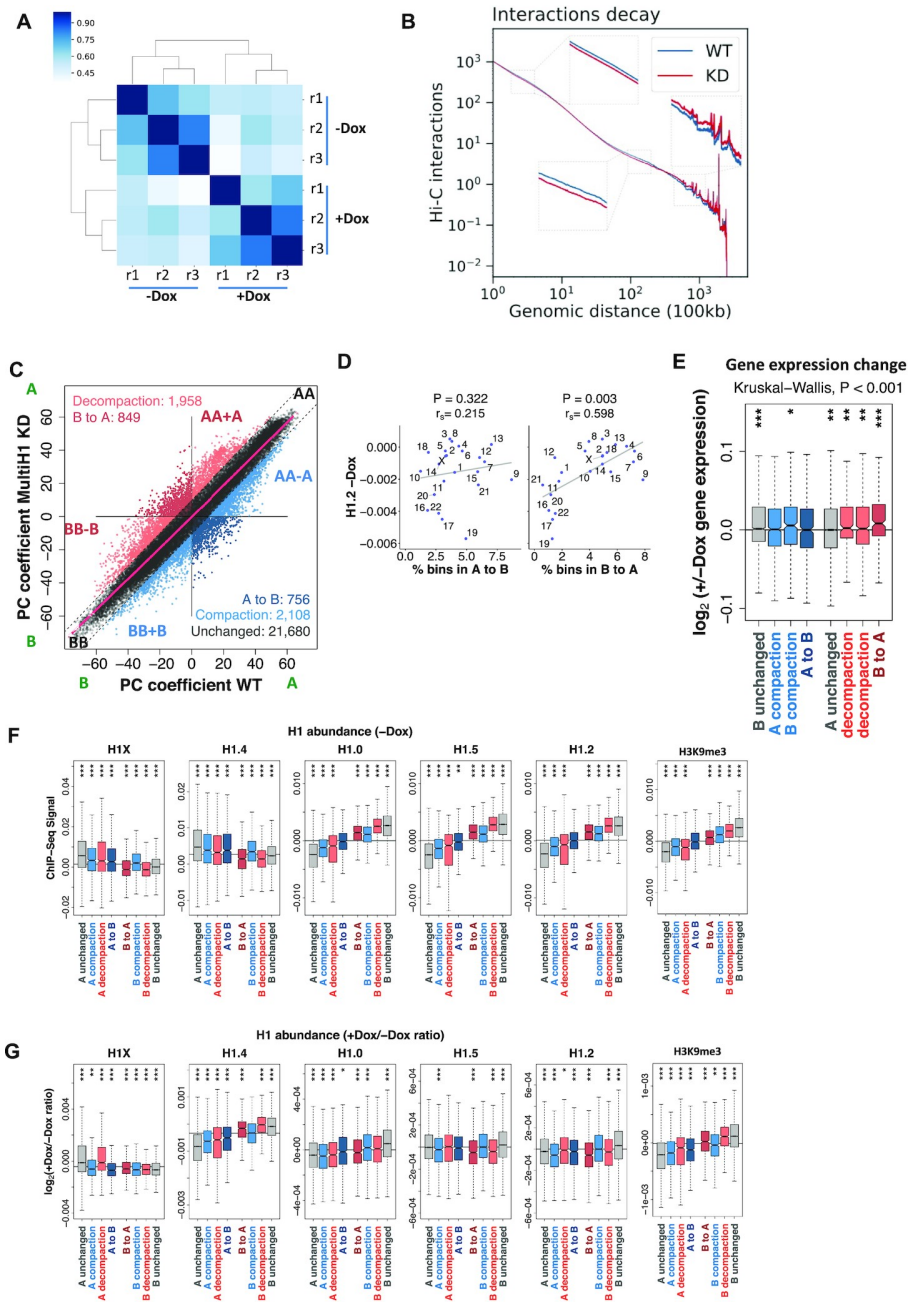


Figure 2. A/B compartments redistribution upon H1 KD. **(A)** Hierarchical clustering of Hi-C replicates from WT (-Dox) and multiH1 KD (+Dox) cells based on the Hi-C reproducibility score between paired experiments. **(B)** Plot comparing the distribution of Hi-C interactions versus genomic distance across the genome for a maximum distance of 500 Mb for WT and H1 KD cells. **(C)** Scatter plot of principal

component (PC) coefficients for 100-kb genomic segments (bins) from WT (–Dox) and H1 KD (+Dox) cells. PC coefficients were used to define A (positive PC) and B (negative PC) compartments, as well as compartment shifting (A-to-B and B-to-A), compaction (blue bins AA-A and BB + B) and decompaction (red bins AA + A and BB-B) upon H1 KD. Unchanged segments AA and BB are black-colored. A polynomial regression line was used to model the relationship between the dependent and the independent variables. **(D)** A/B compartments redistribution within chromosomes. Scatter plot between the percentage of bins that changed from A to B or vice versa upon H1 KD, and the average H1.2 ChIP-seq signal in untreated cells within TADs, for each chromosome. Spearman's correlation coefficient is shown as well as *P*-value. **(E)** Gene expression changes upon H1 KD within bins changing compartment or compaction rate. Normalized RNA-seq reads of coding and non-coding genes before and after Dox-induced H1 KD within 100-kb bins of the eight categories obtained in (C) were used to calculate the +/–Dox fold-change (expressed as log2). (F, G) Box plot showing H1 variants input-subtracted ChIP-seq signal within bins of each category in WT cells (–Dox) **(F)**, or the ratio of change (log2) in H1 KD (+Dox) compared to untreated cells (–Dox) **(G)**. (***) *P* < 0.001; (**) *P* < 0.01; (*) *P* < 0.05. Kruskal–Wallis test determined that there were statistically significant differences between the groups (*P* < 0.001). One-sample Wilcoxon signed-rank test was used to compare each group of bins against the median gene expression changes (E), H1 variants input-subtracted ChIP-seq signal (F), or ratio of change (G).

to B direction) with about 1/4 completely changing compartment category (Figure 2C). Interestingly, these changes in compartmentalization were not homogenous across the genome, being B-to-A shifts upon H1 KD more frequent within chromosomes with high H1.2 content. Notably, the expected anti-correlation for bins moving from the A compartment to the B compartment was not observed, despite chromosomes rich in A compartment were poor in H1.2 (Figure 2D and Supplementary Figure S7). To assess if changes in compartment were related to gene activity, we also explored whether gene expression was altered within bins changing compaction upon H1 KD using RNA-seq data previously acquired in the same cell systems (22). Significant overall gene up-regulation was observed within bins being decompacted (B-to-A and A or B decompaction), but the opposite was not observed for bins being compacted (Figure 2E). We

next wondered whether the changes in compartmentalization and expression were dependent on the basal distribution of H1 variants in the genome as well as their re-distribution upon H1 KD. As expected, we found that H1X and H1.4 were enriched in the A compartment and H1.0, H1.5 and H1.2 were enriched in the B compartment (Figure 2F). Interestingly, such a trend was pronounced for all the bins in the genome which compartmentalization did not change upon H1 KD indicating that the basal state of different H1 variants could determine how compartments respond to H1 depletion. However, was the observed trend upon H1 depletion also accompanied by a change of H1 variant distribution? Interestingly, H1X decreased upon H1 KD in B compartment bins (regardless of their change in compartmentalization) as well as in A-compartment bins that compacted or even moved to the B compartment (Figure 2G), which could indicate that decrease of H1X is associated to B compartmentalization. Similarly, H1.2 decreased in all A compartment bins as well as B compartments that decompacted or even moved to the A compartment, which again indicates that H1.2 decrease is associated to A compartmentalization. To note that, despite H1.4 clear depletion after H1 KD, its changes associated to compartmentalization did not correlate with the observed changes in H1X (Figure 2G). In fact, H1.4 decreased in all A compartment bins and increased in all B compartment bins after H1 depletion, which could indicate a redistribution that could play a significant role in compartmentalization.

Topologically Associating Domains or TADs comprise the next scale of the so-called higher-order organization of chromatin after compartmentalization (43). Similar to the compartmentalization changes, the large majority of TAD borders (i.e. 71.0%) remained

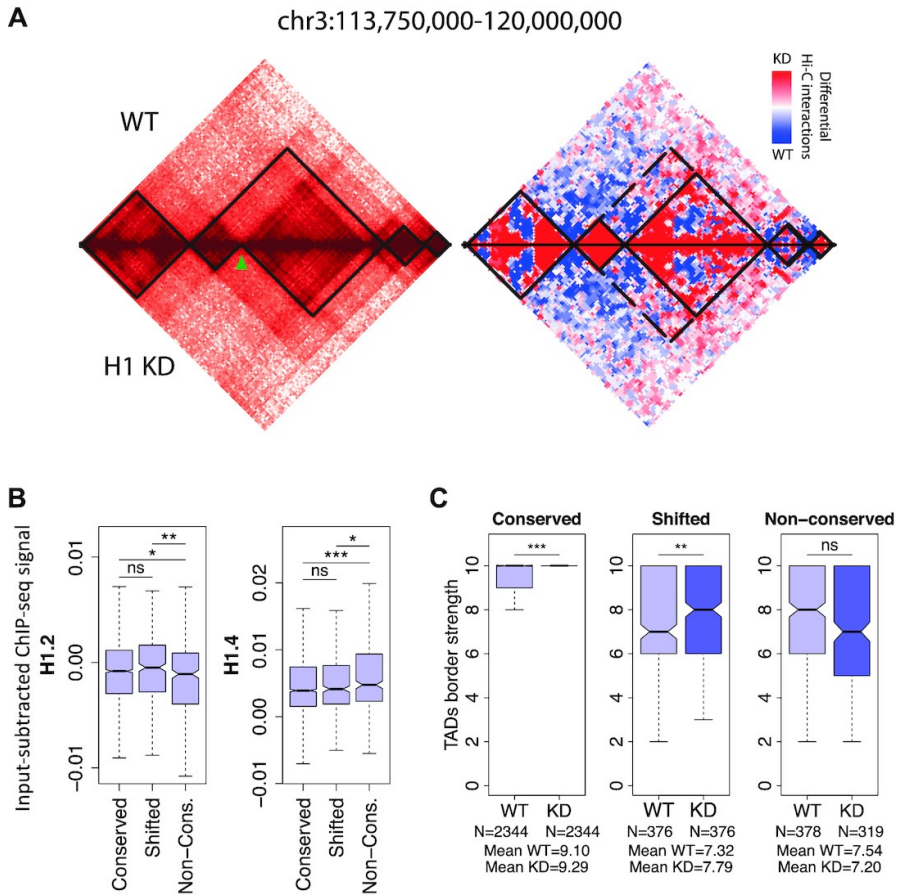


Figure 3. TAD boundaries changes upon H1 KD. (A) Hi-C interaction maps of 6.25 Mb region in chromosome 3 at 50-kb resolution. *Left panel* is a heat map of Hi-C maps normalized by reads coverage in Log2 scale with TADs overlaid by black lines. Top triangle of the map corresponds to Hi-C in WT and lower triangle to H1 KD. Green arrow points to the de-novo detected TAD border in H1 KD. *Right panel*, differential Hi-C map showing the enrichment of internal interaction in the two separated TADs around the new detected border. (B) Box plot showing the H1.2 and H1.4 input-subtracted ChIP-seq signal in WT cells within TADs containing the TAD borders divided in conserved, shifted <100 kb and non-conserved according to their behavior upon H1 KD. (C) TAD border dynamics. Box plot of normalized border strength distribution for TAD borders in WT and H1 KD cells, divided in conserved, shifted <100 kb and non-conserved borders. (***) $P < 0.001$; (**) $P < 0.01$; (*) $P < 0.05$ (Mann–Whitney test).

unchanged upon H1 KD (Supplementary Figure S2), 12.4% shifted by only one 100-kb bin, and 16.5% were not conserved (that is, shifted by

>1 bin, newly formed or disappeared; Figure 3A as example of a de novo detected border after H1 KD). To determine whether those changes could be linked to the basal distribution of H1 variants prior H1 KD, we interrogated the TAD enrichment of H1.2 or H1.4, which we identified above having a role in A/B compartmentalization. The results indicate that H1.2 was significantly depleted at non-conserved compared to conserved TAD borders and H1.4 was higher at TADs with non-conserved borders (Figure 3B). Interestingly, the differences in border position were also associated to changes in border strength. Upon H1 KD there was an increase in border strength for conserved and shifted TAD borders but not for the non-conserved borders, which slightly decreased its border strengths but with no statistically significant differences (Figure 3C). The results suggest, thus, that ‘soft’ borders were prone to be altered upon histone H1 depletion, both in its position as well as in its strength.

The observed increased border strength was associated to an increase in intra-TAD (i.e. local interactions) both within A and B compartments and a decrease of inter-TAD interactions (i.e. non-local interactions) within the A and between A and B compartments (Figure 4A). The increase of local interactions (intra-TAD) with a decrease of non-local interactions (inter-TAD) was also observed with the *D-score*, which measures the differential local interactions per each of the 100-kb bins in a Hi-C matrices. Specifically, the *D-score* is the average of differential interactions between WT and H1 KD of each bin with any other bin within a window of 2 Mb. Thus, it measures if a bin is surrounded by mainly a region in the genome of increased ($D\text{-score} > 0$) or decreased ($D\text{-score} < 0$) interactions (Figure 4B). Next, we compared the basal distribution of H1 variants with the *D-score* and found that H1.2 signal was a strong predictor of the *D-score* across

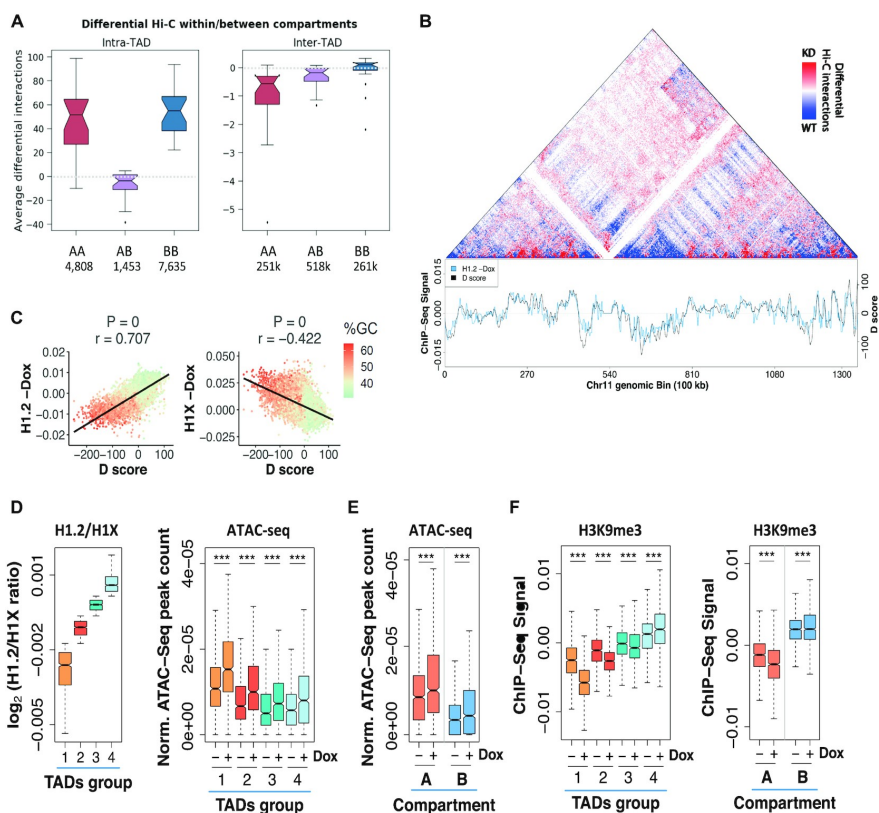


Figure 4. Dynamics of Hi-C genomic interactions and chromatin changes upon H1 KD. **(A)** Box plot showing average number of differential intra- or inter-TAD interactions per chromosome upon H1 KD in different compartments, at 100 kb resolution. The average number per chromosome of differential interactions for each category is indicated in the ticks of X axes. **(B)** *Top*, Differential Hi-C map. Increased (red colored) and decreased (blue colored) interactions in contact matrices of chromosome 11 (0–135 Mb) of H1 KD compared to WT cells, at 100 kb bins resolution. *Bottom*, D score. Profiles of differential interaction D score and input-subtracted H1.2 ChIP-seq abundance from WT cells along chromosome 11, calculated within 100 kb bins. **(C)** Scatter plots between differential interaction D score and H1.2 or H1X abundance from WT cells, genome-wide. Spearman correlation coefficient is shown as well as P-value. The GC content at each bin is color-coded. **(D, E)** Box plots showing the relative number of ATAC-seq peaks (normalized by length) within TADs classified according to H1.2/H1X ratio (Groups 1–4) **(D)** or within A/B compartments **(E)**, at WT and H1 KD (–/+Dox) cells. The ChIP-seq H1.2/H1X signal ratio within TADs in the four groups reported is shown for reference in **(D)**. **(F)** Box plots showing the H3K9me3 input-subtracted ChIP-seq signal within TADs classified according to H1.2/H1X ratio (left) or within A/B compartments (right), at WT and H1 KD (–/+Dox) cells. (***) $P < 0.001$; Wilcoxon

signed-rank test. A compartment, $N = 1032$; B compartment, $N = 1014$; TADs, $N = 756$ TADs per group.

the genome (corr.coef. = 0.707 and Figure 4B, C). Conversely, there was an inverse correlation between the *D-score* and the basal abundance of H1X variant (corr. coeff. = -0.422). In other words, those regions of the genome with high H1.2 overlap are likely to result in increased local interactions once H1 is depleted while regions with high H1X are likely to decrease interactions.

Next, to identify if there was a correlation between the observed changes in H1 variants upon H1 KD within the spatial genome and the underlying chromatin state, we further classified TADs by their content in H1.2 and H1X variants (that is, we generate four discrete groups of TADs from lowest to highest H1.2/H1X ratio; Figure 4D and Supplementary Figure S2). Upon H1 depletion, accessibility measured by ATAC-seq was significantly increased at all TAD categories, but its increase was more pronounced at low H1.2/H1X TADs (Figure 4D). Accordingly, accessibility was also increased at the A and B compartments but most notably in A compartment (Figure 4E and Supplementary Figure S2). As expected, the opposite trend was observed in analyzing the distribution of the repressive mark H3K9me3 upon H1 KD. Indeed, H3K9me3 ChIP-seq signal decreased more in the A compartment compared to the B compartment and in low H1.2/H1X ratio TADs (Figure 4F), which indicates again that chromatin decompaction upon H1 depletion occurs more prominently in already open regions of the genome.

Altogether our findings indicate that the genome structure is not generally but specifically altered upon depletion of H1 variants. First, local and non-local interactions genome-wide were differentially altered

with short and mid-range interactions decreasing and long-range increasing. Second, these changes in interactions correlated with changes in A and B compartments associated to changes of gene expression. Third, intra-TAD interactions increased, mostly within A or B compartments, which resulted in a clear increase of TAD border strength. Fourth, these genome interaction changes were more prominent depending on the basal H1 variant occupancy being the distribution of H1.2 and H1X most informative of the observed changes. Fifth, and final, depletion of H1 variants resulted in an overall increase of accessibility of chromatin, which also depended on the basal occupancy of H1.2 and H1X.

Gene expression is coordinately altered within TADs upon H1 KD

As previously observed, H1 variant depletion resulted in deregulation of hundreds of genes with about one third of the up-regulated genes associated to transcriptional response to interferon (22). In our experiments, a total of 1089 and 1254 genes were up-regulated and down-regulated, respectively ($FC \geq 1.4$, adjusted P -value ≤ 0.05 , Figure 5A). Interestingly, groups of regulated genes were more often than expected co-localized within the same TAD. The 2,343 deregulated genes were distributed across 1,292 TADs with an enrichment of TADs with either only up or down regulated genes (Supplementary Figure S8A). For example, there was 531 TADs with at least one up-regulated genes and no down-regulated genes (here called 'Up'). Similar numbers were observed for down-regulated TADs with 520 with at least one down-regulated gene and no up-regulated genes (here called 'Dw'). Finally, a total of 241 TADs contained at least 2 genes deregulated with mix directions (here called 'UpDw'). UpDw

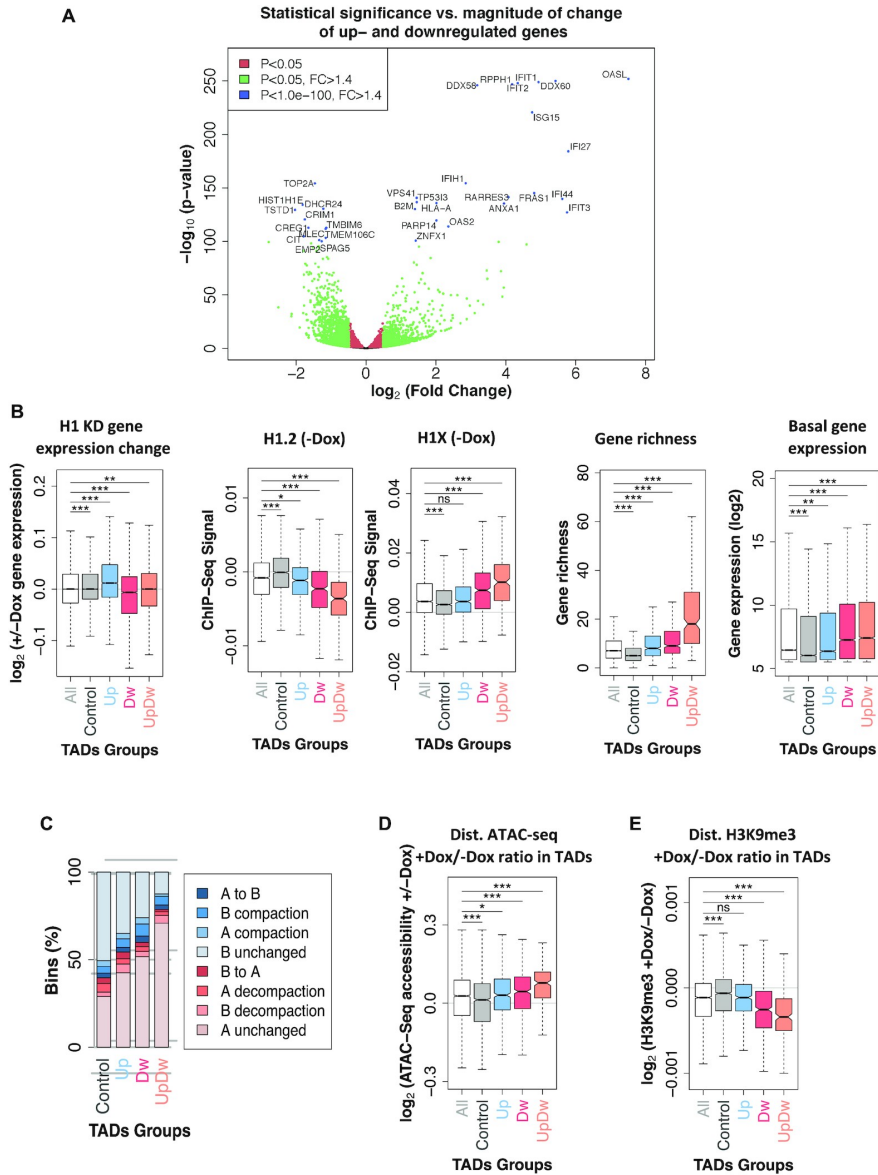


Figure 5. Gene expression is coordinately altered within TADs upon H1 KD. (A) *Top*, Histogram of the frequencies of TADs for the observed (gray) or randomized (purple) position of genes, for TADs containing an increasing proportion of genes per TAD with positive FC. Observed and expected values were compared using Pearson's chi-square test. Gene locations were randomized 10 000 times, constraining by chromosome, not allowing overlapping, and only considering TADs with ≥ 4 genes. *Bottom*—Ratio of observed versus expected frequencies of TADs with distinct proportions of genes with positive or negative H1 KD-induced FC; $FC > 1$ or $FC < -1$. (B) TADs with ≥ 4 genes where at least 90% of genes are down- (left) or up-regulated

(right) with $FC \leq -1$ or $FC > 1$, respectively (total $N = 115$). Log2 of gene expression FC is shown. TADs are ordered from low to high abundance of genes per TAD. Dashed lanes indicate $FC = -1.4$ or $FC = 1.4$. Red dots represent ISGs. Example genes shown in (C) are located within TADs marked with an arrow. **(C)** Examples of TADs with biased coordinated response to H1 KD. Fold change +Dox/-Dox (log2) is shown for all coding and non-coding genes present within a representative TAD containing 90–100% of genes with negative (left) or positive (right) FC, respectively. Genes are ordered according to their position within the genome. Red asterisk represents ISGs. **(D)** Box plot showing the H1.2 ChIP-seq signal in untreated cells within TADs in the 10 groups described in (A). **(E)** Box plot showing the ATAC-seq accessibility gain upon H1 KD (+/-Dox) within TADs in the 10 groups described in (A). Kruskal–Wallis test determined that there were statistically significant differences between the groups in (D) and (E). Comparison between each group of TADs and the median ChIP-seq H1.2/H1X log2 ratio (D) or the ATAC-seq accessibility changes (E) was performed using the one-sample Wilcoxon signed-rank test (***) $P < 0.001$; (**) $P < 0.01$. **(F)** Bar plots showing the frequency of overlap between all the TAD groups described in (A) and genome segments within A/B compartment categories described in Figure 2C that changed compaction upon H1 KD. The observed and expected count of bins of the different groups of TADs were significantly different ($P < 0.001$, Pearson's chi-squared test).

TADs corresponded to higher gene density and lower H1.2 content compared to either TAD-Up or TAD-Dw. Finally, TADs without deregulated genes (here called Control) had the highest H1.2 content as well as the lowest gene richness (Figure 5B). Accordingly, H1X was significantly enriched within UpDw TADs and depleted from Control TADs contrary to the observed trend for H1.2 variant. Most TADs containing significantly deregulated genes upon H1 KD were located before KD within the A compartment (Figure 5C), while Control TADs were enriched at the B compartment. Chromatin remodeling also followed the expected trends for the TAD groups classified by their change in expression of the resident genes. For example, upon H1 KD, ATAC-seq accessibility increased globally in all TADs, especially in UpDw type (Figure 5D). Conversely, H3K9me3 abundance significantly decreased in Dw and UpDw TADs (Figure 5E). The same

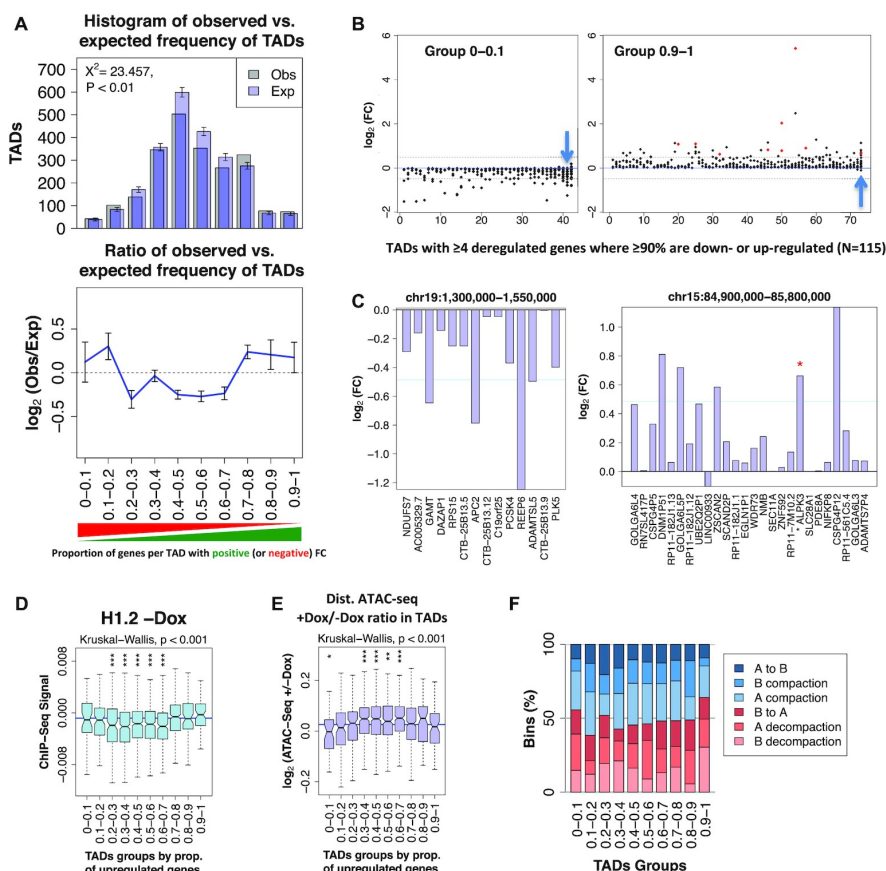


Figure 6. Gene expression is coordinately altered within TADs upon H1 KD. **(A)** *Top*, Histogram of the frequencies of TADs for the observed (gray) or randomized (purple) position of genes, for TADs containing an increasing proportion of genes per TAD with positive FC. Observed and expected values were compared using Pearson's chi-square test. Gene locations were randomized 10 000 times, constraining by chromosome, not allowing overlapping, and only considering TADs with ≥ 4 genes. *Bottom*– Ratio of observed versus expected frequencies of TADs with distinct proportions of genes with positive or negative H1 KD-induced FC; $\text{FC} > 1$ or $\text{FC} < -1$. **(B)** TADs with ≥ 4 genes where at least 90% of genes are down- (left) or up-regulated (right) with $\text{FC} < -1$ or $\text{FC} > 1$, respectively (total $N = 115$). \log_2 of gene expression FC is shown. TADs are ordered from low to high abundance of genes per TAD. Dashed lanes indicate $\text{FC} = -1.4$ or $\text{FC} = 1.4$. Red dots represent ISGs. Example genes shown in **(C)** are located within TADs marked with an arrow. **(C)** Examples of TADs with biased coordinated response to H1 KD. Fold change +Dox/–Dox (\log_2) is shown for all coding and non-coding genes present within a representative TAD containing 90–100% of genes with negative (left) or positive (right) FC, respectively. Genes are ordered according to their position within the genome. Red asterisk represents ISGs. **(D)** Box plot showing the H1.2 ChIP-seq signal in untreated cells

within TADs in the 10 groups described in (A). **(E)** Box plot showing the ATAC-seq accessibility gain upon H1 KD (+/-Dox) within TADs in the 10 groups described in (A). Kruskal–Wallis test determined that there were statistically significant differences between the groups in (D) and (E). Comparison between each group of TADs and the median ChIP-seq H1.2/H1X log2 ratio (D) or the ATAC-seq accessibility changes (E) was performed using the one-sample Wilcoxon signed-rank test (***) $P < 0.001$; (**) $P < 0.01$. **(F)** Bar plots showing the frequency of overlap between all the TAD groups described in (A) and genome segments within A/B compartment categories described in Figure 2C that changed compaction upon H1 KD. The observed and expected count of bins of the different groups of TADs were significantly different ($P < 0.001$, Pearson's chi-squared test).

correlations were obtained using TADs containing genes deregulated upon H1 KD considering a $FC \geq 2$ (Supplementary Figure S8B-F).

As previously described in T47D cell lines (6), we observed an intra-TAD coordinated response of gene expression. Indeed, we found an enrichment of gene-rich TADs (that is, with at least four genes) where most of its genes changed expression in the same direction ($FC > \pm 1$). Specifically, TADs with over 70% of their genes up-regulated or at least 80% down-regulated were observed in proportions beyond random expectation (Figure 6A). These correspond to TADs where all or most of the genes changed expression in the same direction upon H1 KD, including Interferon stimulated genes (ISGs) such as ISG20, CMPK2, DDX60 or GBP3 (Figure 6B, C and Supplementary Figure S9A). This could result from two hypothetical scenarios: (i) upon H1 depletion, the whole TADs were (architecturally) affected and most resident genes became up- or down-regulated coordinately; (ii) upon H1 depletion, some gene within a TAD became deregulated and, consequently, neighbor genes within the same TAD changed expression in the same direction. To discern between these two scenarios, we characterized the groups of TADs with most coordinated changes of expression upon depletion of H1. Generally, these were poor in gene density, low in GC

content, low in basal expression (except group 0–0.1), and high in H1.2 (Figure 6D and Supplementary Figure S9B–F). Moreover, the selected TADs were poor in H1X and H1.4 (Supplementary Figure S9G). Interestingly, these TADs suffered less prominent changes in H1 variant distribution or ATAC-seq coverage than non-coordinated response TADs (Figure 6E and Supplementary Figure S9H). Despite this, coordinated TADs were enriched in regions of the genome that suffered decompaction as measured by the Hi-C compartmentalization analysis (Figure 6F).

Altogether, the results support that upon H1 depletion the majority of the genome does not alter its expression. However, genes located in regions of high H1.2 content harbored more genes whose expression was coordinated within entire TADs. Therefore, our results indicate that upon H1 depletion, the entire TADs were architecturally altered and most resident genes were coordinately deregulated.

3D modeling of TADs with coordinated transcriptional response

To further characterize architecturally changes within TADs with coordinated transcriptional response to H1 KD, we next generated 3D models of genomic regions harboring TADs that contained at least 90% of genes down or up-regulated (group 0–0.1, ‘*d*’, $N = 42$; group 0.9–1, ‘*u*’, $N = 73$; Figure 6B), both in WT and H1 KD conditions. As a control, we also modeled TADs with the most extreme H1.4 decrease upon H1 KD (group ‘*b1*’, $N = 100$), TADs with a bidirectional transcriptional response (‘*bi*’, $N = 174$, picked from groups 0.4–0.6; Supplementary Figure S9C), TADs with minimum gene expression changes (‘*mi*’, $N = 100$), and TADs with no annotated genes (‘*mi*’, $N = 12$). Models were built based on our Hi-C data at 10 kb

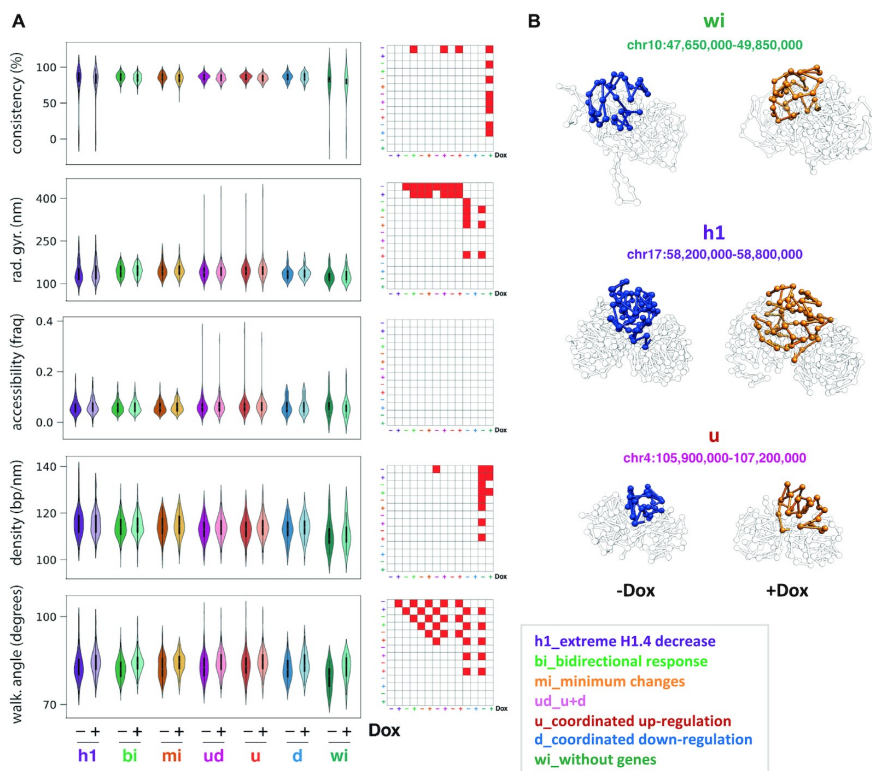


Figure 7. Structural properties of TADs. **(A)** Violin plots of structural properties measured on the 3D models computed for seven classes of TADs, both in WT and H1 KD conditions (-/+Dox): TADs with the most extreme H1.4 decrease upon H1 KD (*h1*, $N = 100$), TADs presenting a bidirectional transcriptional response to H1 KD (*bi*, $N = 174$), TADs presenting the minimum gene expression changes (*mi*, $N = 100$), TADs presenting a coordinated transcriptional response to H1 KD (*u*, only up-regulated genes, $N = 73$; *d*, only down-regulated genes, $N = 42$; *ud*, TADs *u* and *d* together, $N = 115$), TADs without genes (*wi*, $N = 12$). For each TAD 1000 models have been generated and clustered, and plotted measures are relative to the main cluster of models. Reported measures are consistency, radius of gyration, accessibility, density and walking angle. Matrices next to violin plots indicate classes of TADs that are significantly different for each measure. Statistical significance of the difference between distributions was computed with Kolmogorov-Smirnov test (P -value < 0.01). See Materials and Methods for details. **(B)** 3D models of the indicated TADs within chr10, chr17 and chr4, from the *wi*, *h1* and *u* groups, respectively, in WT (blue) and KD (orange) conditions. The 3D modelling reflected a tendency to chromatin opening upon H1 KD.

resolution using TADbit as previously described (29). Several structural measures were computed and compared between groups of modeled TADs, such as: consistency, radius of gyration, accessibility, density and walking angle (Figure 7 and Materials and Methods for the definition of the structural measures). Additionally, the analyzed TAD groups were characterized in terms of H1 abundance, gene expression changes, GC content and ATAC-seq accessibility for comparison with the structural data (Supplementary Figure S10). All modeled TAD groups resulted in highly consistent models, this indicates that the input Hi-C data did not contain many contradictory interactions and that fairly structural similar conformations were obtained from the ensemble of models for all cases. Only TADs harboring no genes resulted in 3D models with lower consistency measures indicating that more different conformations could satisfy the input restraints (Figure 7A). Interestingly, TADs with the highest H1.4 decrease upon H1 KD ('*b1*' group) as well as TADs with no genes ('*w*' group) overall resulted in more different structural properties. Specifically, both *b1* and *w* TADs are more compact (lower radius of gyration) compared to the rest of the groups (Figure 7A). However, *b1* results in the densest DNA (bp per nanometer) models compared to the *w*, which are the least dense of all. Other groups have similar density values and between these two extremes. It is important to note that there is an apparent discrepancy in TAD structural features and ATAC-seq data for some TAD groups such as *b1*. These TADs result in models that tend to be dense/compact while highly accessible in the ATAC experiment. Nevertheless, at the level of resolution of the 3D models (that is 10 kb) it is impossible to assess whether the apparent discrepancy is due to the data or the modeling exercise. The reason is that the measures are averaged over entire TADs and the comparison

of both datasets cannot be done directly as ATAC-seq data is ~ 100 base pair resolution and our models are 10 kb resolution. The high gene expression rate and GC content in *b1* TADs group explains the high ATAC accessibility at low resolution. However, at 10 kb resolution the ‘density’ measure says that at the TAD level the DNA fiber is more compacted.

Finally, the models indicate that upon H1 KD and across all types of TAD groups there is a significant increase of the walking angle measure indicating a change of stiffness of the chromatin (Figure 7A). In general, changes in the structural properties of TADs reflected a tendency to chromatin opening upon H1 KD, such as the significant increase of chromatin walking angle and tendency to increase of the radius of gyration. The observed changes are exemplified in three models from the *ni*, *b1* and *ud* groups (Figure 7B). In general, we observed no significant differences in structural changes upon H1 KD in TADs with no genes (*ni*), while the changes were more evident in the *b1* group and also in the *ud* group independently of the direction of the changes in gene expression. Indeed, although without significance, changes in the TADs with a coordinated transcriptional response to H1 KD (*u*, *d*) have the same trends, indicating that TADs that were coordinately up- or down-regulated were similarly structurally altered upon H1 KD. Our 3D models indicate that all TADs are altered in a similar way due to H1 KD, with different consequences in gene expression deregulation that might depend on local features or distinct H1 abundance.

DISCUSSION

In this study, we have analyzed the genomic distribution of five endogenous H1 variants within T47D breast cancer cells by ChIP-seq using specific antibodies. This is almost the whole somatic H1

complement of this cell line with the exception of H1.1, which is not expressed in these cells, and H1.3 that was not profiled due to the lack of ChIP-grade antibodies. This is, to our knowledge, the first time that most of endogenous variants have been profiled in a mammalian cell. Antibodies for H1.2, H1X and H1.0 were used before on ChIP-seq experiments (16–18,20); H1.4 and H1.5 antibodies have been used here for the first time, to our knowledge. Specificity of all H1 antibodies used has been assayed extensively (Supplementary Figure S1).

In previous studies, we mapped endogenous H1.2 and H1X, demonstrating that they have different distributions across the genome (16,17,20). On the one side, H1.2 is enriched within intergenic, low gene expression regions and lamina-associated domains. On the other side, H1X is enriched at gene-rich chromosomes, RNA polymerase II enriched sites, coding regions and hypomethylated CpG islands. The apparent differential distribution of the two H1 variants in active versus inactive chromatin, also correlates with the GC content of the regions where they localize. Indeed, we have observed here that H1.5 and H1.0 colocalize with H1.2, at low GC regions, while H1.4 distribution is similar to H1X with the exception of H1X being highly enriched at high GC regions. Previously, we profiled H1.0 and H1.4 fused to an HA tag at C-termini, stably expressed through a lentiviral vector into T47D cells. Using this technique, both H1.4-HA and H1.0-HA were enriched at high GC regions, indicating that profiling exogenous, tagged H1 proteins may give different results than endogenous proteins (16). In apparent contradiction, H1.0 has been profiled in human skin fibroblasts, being enriched at high GC regions (18) while in mouse, tagged, knocked-in H1c (H1.2), H1d (H1.3) and H1.0 have been profiled in ESCs and found enriched at low GC regions (14).

Altogether, this suggests that H1 variants distribution might be different among cell types, and could be explained by the relative levels of expression of the different variants. Extensive profiling of H1 variants among different cell types with the same methodology should be done to clarify whether the observed distribution of H1 variants is cell type-specific or universal for some of the variants.

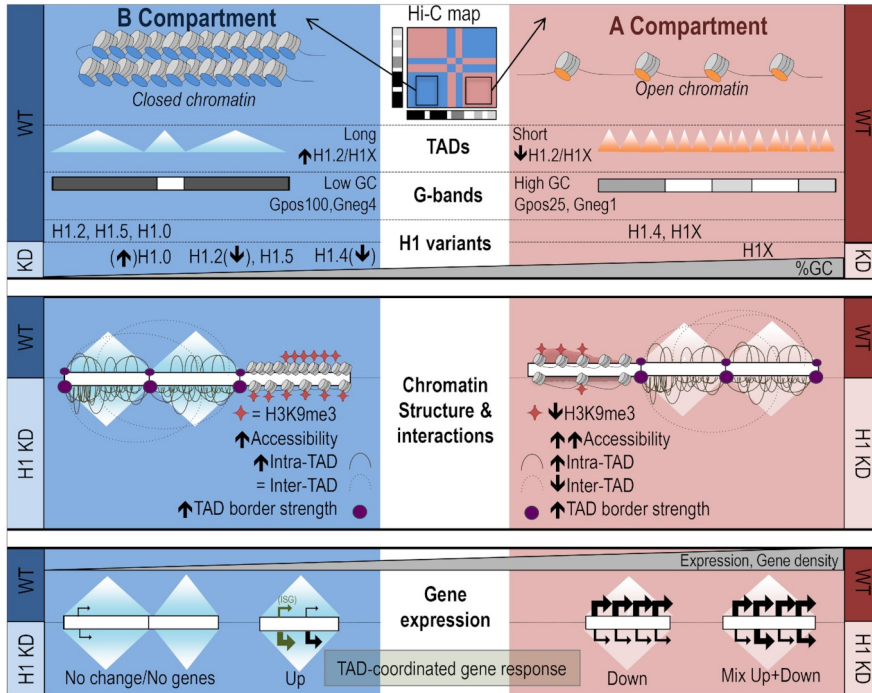
In T47D cells, the H1 content was estimated to be 9% for H1.0, 23% for H1.2, 13% for H1.3, 24% for H1.4 and 31% for H1.5 (19). Our distribution analysis thus indicates that most of H1 variants we profiled are located in low GC regions, which supports its role as heterochromatic protein. However, and as previously described (17), H1X is enriched at high GC regions suggesting its possible role as regulatory H1. We also found that the enrichment of H1.4 at high GC regions is intriguing as it was suggested that, because of its K26 residue which may be methylated and bind HP1, it could be related to heterochromatin (44,45). Still, a fraction of H1.4 is at low GC regions, and even at high GC bands it could have a role in repression at particular sites. In fact, when profiled within chromatin states, H1.4 overlapped H3K9me3 distribution, a *bona fide* heterochromatin marker.

To study whether alteration of the total H1 content and relative abundance of the different variants affected the genomic localization of remaining histones, we performed ChIP-seq in T47D cells knocked-down for H1.2 and H1.4 with an inducible system, previously characterized (22). Interestingly, upon H1 KD, H1.4 preferentially remained at low GC regions, supporting its putative role in heterochromatin, and was displaced from high GC regions. In parallel, H1X redistributed to high GC regions. H1.0 maintained its distribution across the genome despite its expression and protein levels increased to

compensate the H1 overall $\approx 30\%$ decrease. Overall, H1.2 was depleted but did not change much its relative genomic distribution. Profiling within chromatin states showed that H1.4 slightly switched towards the H1.2 group upon H1 KD, and H1X decreased at heterochromatin and increased in almost all other chromatin states. The redistribution of remaining H1.4 upon H1 KD (i.e its preferential depletion from high GC regions), is puzzling. An alternative explanation could be that the H1.4 antibody, upon depletion of its specific epitope, cross-reacted with other variants located at low GC regions (H1.0, H1.2 or H1.5). Our specificity analysis, so far, does not support this hypothesis (Supplementary Figure S1). Whether this H1.4-distribution occurs in other cell types would be interesting to investigate.

Immunofluorescence analysis of H1 location within the nuclei confirmed the expression changes described upon H1 KD, but any redistribution within chromatin states is difficult to pick up with this technique. Still, it was possible to confirm that H1.4 and H1X localization differs from H1.2, H1.5 and H1.0. Further studies at super resolution fluorescence microscopy might help to characterize, in the future, the differential localization of H1 variants and their role in chromatin organization and genomic functions.

In this work, we have shown that H1 KD caused changes in chromatin accessibility and H3K9me3 distribution, shifts in A/B compartments and TAD borders, and changes in the 3D architecture of TADs (Figure 8). Some of these changes were dependent on the compaction or GC content of genomic domains. In fact, we have previously shown that A and B compartments positively correlate with the measured H1.2/H1X ratio (20). Here we have further shown that the A/B compartments present different abundance of H1 variants and respond



gene density. Gene-dense TADs contained both up- and down-regulated genes simultaneously. TADs with only down-regulated genes showed intermediate features.

differently to H1 depletion. Upon H1 KD, ATAC-seq chromatin accessibility increased genome-wide but more markedly at A compartment. Accordingly, the repressive histone mark H3K9me3 decreased majorly from A compartment. Recent reports have shown that H1 depletion in mouse T cells and germinal centre B cells lead to B-to-A compartment shifting (46,47). These could be due to the fact that differentiated cells present a well-constituted heterochromatin rich in histone H1, compared to pluripotent and cancer cells where chromatin may be more plastic, partially because of a lower H1 content (48,49). H1-mediated compartmentalization may be established along differentiation, sequestering the stem cell programs within the B compartment. Deregulation of H1 levels and compartmentalization may occur in cancer and along reprogramming (40,50,51). The observation of A-to-B and B-to-A shifting in our cancer model T47D cells in similar proportions could be due to an overall less compacted chromatin, or to the simultaneous depletion of H1 variants assayed here to occupy distinct genomic compartments. We here show that H1.2 is abundant at the B compartment and its depletion in H1 KD cells resulted in decompaction and B-to-A shifts, accompanied by gene induction and local increase of DNA interactions. Conversely, H1.4 is abundant at the A compartment and its depletion upon H1 KD preferentially accompanied A-to-B shifts or compaction. However, as A decompaction also occurred in regions with H1.4 occupancy, our results could suggest a dual role of this H1 variant, which requires further investigation.

We reported before that multiH1 KD (H1.2 + H1.4) effects were more drastic than the simple addition of H1.2 or H1.4 KD effects, e.g. in the number of genes being deregulated, or in causing the induction of the interferon response due to de-repression of heterochromatic repeats (Supplementary Figure S11) (22). This appeared to be due to the synergistic function of these two variants, more than to the total amount of H1 being depleted, because other H1 variants KD combinations did not produce the observed effects. Using our RNAseq data we have explored whether genes that showed coordinated expression within TADs in multiH1 KD cells (Figure 6C) changed expression in H1.2 or H1.4 individual KD cells. The result was that these changes did not occur, neither in intensity nor sense (Supplementary Figure S12). This is an indirect demonstration that single H1 KDs would not alter TADs in the manner shown here for multiH1 KD. All this would support our hypothesis that effects on accessibility or topology would be seen importantly in multiH1 KD but not on the single H1 KD cells.

Previously, we and others have shown that epigenetic states and H1 distribution are more homogeneous within a TAD, suggesting that TAD borders prevent the spreading of these features (6,20). In our work, TADs hardly changed its size or distribution upon H1 KD, however, a clear increased TAD border strength and intra-TAD contacts was observed. Interestingly, the concomitant inter-TAD contacts reduced more predominantly in A compartment compared to the B compartment. Indeed, several reports have also shown that TAD organization remains largely unchanged when disturbing chromatin homeostasis, including mouse H1-depleted cells or epithelial-to-mesenchymal transition (6,21,28,52). However, our work now highlights novel relevant changes in TAD organization due to depletion

of H1 variants, including an increase in border strength accompanied by an increase of intra-TAD interactions.

Severe H1 depletion causes cell cycle arrest and transcription-replication conflicts (22,53,54). One could speculate that this could be the basis for the observed changes in genome topology. It has been shown that topology changes along the cell cycle in ES cells analyzed at single-cell resolution (55). Upon transition of ES cells from G1 to S phase there is a gradual decrease of TAD insulation and a gradual increase on compartmentalization peaking at G2 phase. If we were comparing T47D cells completely shifting from S to G1, we could speculate that observed changes are due to those described in ES cells, but this is not the case. Normal T47D cells cycle slowly and in basal conditions (–Dox) show a $\approx 50\%$ of cells in G1. Upon H1 KD, G1 increases up to $\approx 60\%$ (Supplementary Figure S1). In addition, we have found some of the topological changes enriched at regions abundant in H1 variants that have been depleted in the H1 KD, concomitant with chromatin opening. For all this, we consider that changes observed are compatible with the depletion of H1 from the genome more than with changes linked to cell cycle shift.

We have shown here and previously that H1 variants selective depletion results in changes in expression of hundreds of genes, including repression of intergenic and intronic RNAs, as well as heterochromatic repeats and ERVs, which leads to the induction of the interferon response (22). Moreover, we have shown that responsive genes are non-randomly located throughout the genome but enriched in a limited number of TADs with their resident genes coordinately changing expression to H1 depletion. We have previously reported that, upon H1 KD in T47D cells, the interferon response is induced with many ISGs

being up-regulated. This is due to the accumulation of RNAs from repeats and ERVs, which stimulate the response at cytoplasm mimicking a viral infection and resulting in the transcription of many genes involved in such response. A part of this direct effect of H1 depletion, our results may also indicate that other genes not directly related to such response may be deregulated due to structural changes, chromatin decompaction, or simply by co-existing within the same TAD with genes directly activated. Indeed, we show that ISGs up-regulated genes co-exist within TADs with other genes that coordinated respond to H1 KD. Despite this observation, we also found that many TADs with a coordinated response do not contain annotated ISGs genes, so we propose that the response may be a consequence of architectural changes upon H1 KD. This result is further supported by the 3D modeling of TADs.

Overall, our results indicate that the non-random genomic distribution of H1 variants, their re-location upon variant depletion, and the subsequent genome structural changes have a read-out in their direct (but also indirect) change of the gene expression program.

DATA AVAILABILITY STATEMENT

Hi-C and ChIP-seq data on T47D breast cancer cells reported here and deposited in NCBI's Gene Expression Omnibus are accessible through GEO Series accession number GSE172618 and GSE156036/GSE166645, respectively. A link to a UCSC genome browser session displaying the uploaded ChIP-seq tracks is provided: https://genome.ucsc.edu/s/NSP/GSE156036_GSE166645.

SUPPLEMENTARY DATA AND FIGURES

Supplementary data and figures are available at NAR online.

AUTHOR CONTRIBUTIONS

F.M. performed the analysis and 3D modeling of Hi-C data. The rest of authors performed the rest of analysis and all experiments.

ACKNOWLEDGEMENTS

We acknowledge Elena Rebollo for technical support with the light microscopy experiments and Carles Bonet for help with testing the variant specificity of H1 antibodies. We acknowledge Generalitat de Catalunya and the European Social Fund for AGAUR-FI predoctoral fellowships [to M.S.-P. and to F.M.].

FUNDING

Spanish Ministry of Science and Innovation [BFU2017-82805-C2-1-P and PID2020-112783GB-C21 to A.J., BFU2017-85926-P and PID2020-115696RB-I00 to M.A.M.-R. (AEI/FEDER, EU)]; European Union's Seventh Framework Programme the ERC [609989 to M.A.M.-R., in part]; European Union's Horizon 2020 research and innovation programme [676556 to M.A.M.-R.]; Generalitat de Catalunya Suport Grups de Recerca [AGAUR 2017-SGR-597 to A.J. and 2017-SGR-468 to M.A.M.-R.]; C.R.G. acknowledges support from ‘Centro de Excelencia Severo Ochoa 2013–2017’; SEV-2012-0208; CERCA Programme/Generalitat de Catalunya as well as support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III; Generalitat de Catalunya through Departament de Salut and

Departament d'Empresa i Coneixement; Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014–2020 Smart Growth Operating Program. Funding for open access charge: Spanish Ministry of Science and Innovation.

CONFLICT OF INTEREST STATEMENT

M.A.M.-R. receives consulting honoraria from Acuity Spatial Genomics, Inc. The rest of the authors declare no conflicts of interests.

MATERIALS AND METHODS

Cell lines, culturing conditions and H1 knock-down

Breast cancer T47D-MTVL derivative cell lines, which carry one stably integrated copy of luciferase reporter gene driven by the MMTV promoter, were grown at 37°C with 5% CO₂ in RPMI 1640 medium, supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin, and 100 µg/ml streptomycin, as described previously (19). HeLa and HCT-116 cell lines were grown at 37°C with 5% CO₂ in DMEM GlutaMax medium, supplemented with 10% FBS and 1% penicillin/streptomycin. The T47D-MTVL multiH1 shRNA cell line (22) was used as a model for H1 depletion. This cell line contains a drug-inducible RNA interference system that leads to the combined depletion of H1.2 and H1.4 variants at protein level although it reduces the expression of several H1 transcripts. Construction, establishment and validation of single-H1 knock-downs have been previously described (19). Specifically, shRNA expression was induced with 6 days

treatment of Doxycycline (Dox), in which cells were passaged on day 3. Dox (Sigma) was added at 2.5 $\mu\text{g}/\text{ml}$.

Immunoblot

Chromatin samples were exposed to SDS-PAGE (14%), transferred to a PVDF membrane, blocked with Odyssey blocking buffer (LI-COR Biosciences) for 1 h, and incubated with primary antibodies overnight at 4°C as well as with secondary antibodies conjugated to fluorescence (IRDye 680 goat anti-rabbit IgG, Li-Cor) for 1 h at room temperature. Bands were visualized in an Odyssey Infrared Imaging System (Li-Cor). Coomassie staining or histone H3 immunoblotting were used as loading controls. ImageJ software was used for immunoblot quantification.

Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation was performed according to the Upstate (Millipore) standard protocol. Briefly, cells were fixed using 1% formaldehyde for 10 min at 37°C, chromatin was extracted and sonicated to generate fragments between 200 and 500 bp. Next, 30 μg of sheared chromatin was immunoprecipitated overnight with the indicated antibody. Immunocomplexes were recovered using 20 μl of protein A magnetic beads, washed and eluted. Cross-linking was reversed at 65°C overnight and immunoprecipitated DNA was recovered using the IPure Kit (Diagenode). Genomic regions of interest were identified by real-time PCR (qPCR) using SYBR Green Master Mix (Invitrogen) and specific oligonucleotides in a Roche 480 Light Cycler machine. Each value was corrected by the corresponding input chromatin sample. Oligonucleotide sequences are detailed in previous studies (17).

ChIP-Seq

Library construction and sequencing: Qualified ChIP and Input samples were subjected to end-repair and then 3' adenylated. Adaptors were ligated to the ends of these 3' adenylated fragments. Fragments were PCR-amplified and PCR products were purified and selected with the Agencourt AMPure XP-Medium kit. The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA (ssCir DNA) were formatted as the final library and then quality-checked. The library was amplified to make DNA nanoball (DNB) which had more than 300 copies of one molecular. The DNBs were loaded into the patterned nanoarray and single end 50 bases reads were generated in the way of sequenced by combinatorial Probe-Anchor Synthesis (cPAS).

ChIP-seq data analysis: Single-end reads were quality-checked via FastQC (v0.11.9) and aligned to the human GRCh37/hg19 reference genome using Bowtie2 (v2.3.5.1) (23) with default options. SAMtools (v1.9) (24) utilities were used to filter out the low-quality reads with the flag 3844. Input, H1 variants, and H3K9me3 genome coverage was calculated and normalized by reads per million with BEDTools (v2.28.0) (25), and regions with zero coverage were also reported in the ChIP-Seq annotation (*genomcov -ibam -bga -scale*). MACS2 (v2.1.2) (26) was used to subtract input coverage from H1 variants and H3K9me3 to generate signal tracks (*bdgcmp -m subtract*).

ChIP-Seq data on histone H1 variants and H3K9me3 epigenetic modification from T47D multiH1 shRNA cells treated or not with Dox has been deposited in NCBI's Gene Expression Omnibus and is accessible through GEO Series accession number GSE156036. ChIP-Seq data on histone H1 variants from WT T47D cells is at GSE166645.

Antibodies

Specific antibodies recognizing human H1 variants used for ChIP/ChIP-seq were: anti-H1.0/H5 clone 3H9 (Millipore, 05–629-I), anti-H1.2 (Abcam, ab4086), anti-H1.4 (Invitrogen, 702876), anti-H1.5 (Invitrogen, 711912) and anti-H1X (Abcam, ab31972). ChIP-seq of H3K9me3 was performed using anti-H3K9me3 (Abcam, ab8898). Other antibodies used were: anti-H1.0 (Abcam, ab11079), anti-H1.3 (Abcam, ab24174), anti-H1.5 (Abcam, ab24175), anti-H3 (Abcam, ab1791) and anti-Lamin A (Abcam, ab8980).

In situ Hi-C

Hi-C libraries were generated from T47D multiH1 shRNA cells treated or not with Dox, as single replica (r1) or duplicate (r2 and r3), as previously described (27,28). In brief, adherent cells were cross-linked with 1% formaldehyde in PBS for 10 min at room temperature and glycine 0.125 M was added for 5 min at room temperature and for 15 min at 4°C to stop the cross-link reaction. Before permeabilization, cells were treated for 5 min with trypsin. Nuclei digestion was performed with 400 units of MboI restriction enzyme. The ends of restriction fragments were labeled using biotinylated nucleotides and ligated with T4 DNA ligase. After reversal of cross-links, DNA was purified and sheared (Diagenode BioruptorPico) to obtain DNA fragments between 300 and 500 bp and ligation junctions were pull-down with streptavidin beads. Hi-C libraries were amplified, controlled for quality and sequenced on an Illumina HiSeq 2500 sequencer (r1) or DNBseq (r2,r3).

Hi-C data pre-processing, normalization and generation of interaction matrices

The analysis of Hi-C data, from FASTQ files mapping to genome segmentation into A/B compartments and TADs, was performed

using *TADbit* (29), which started by performing a quality control on the raw data in FASTQ format. Next, sequencing reads were mapped to the reference genome (GRCh37/hg19) applying an iterative strategy (30) and using the GEM mapper (31). Mapped reads were filtered to remove those resulting from unspecified ligations, errors or experimental artefacts. Specifically, nine different filters were applied using the default parameters in *TADbit*: self-circles, dangling ends, errors, extra dangling-ends, over-represented, too short, too long, duplicated and random breaks (29). Hi-C data were next normalized with OneD correction to remove Hi-C biases and artifacts (32). Filtered read-pairs were binned at the resolutions of 1 Mb, 500, 100 and 10 kb, applying biases from the normalization step and decay correction to generate interaction matrices. Hi-C data on T47D breast cancer cells has been deposited in NCBI's Gene Expression Omnibus and is accessible through accession number GSE172618. A summary of the number of valid reads obtained per replica and filtered artifacts is shown as Supplementary Table S1. Replicates were compared and merged with *TADbit merge* that implements the HiCRep score (33).

Genome segmentation into Topologically Associating Domains (TADs)

TADs were identified at the resolution of 50 kb using *TADbit segment* with default parameters. Briefly, *TADbit* segments the genome into constitutive TADs after analyzing the contact distribution along the genome using a BIC-penalized breakpoint detection algorithm (29). This algorithm leads to a ~99% average genome coverage. To assign a strength value to each TAD border, *TADbit* repeats the dynamic programming segmentation 10 times after the optimum is reached, each time decreasing the by a fix off-set the optimal TAD border detection

path. The strength of a TAD border is then calculated as the number of times it was included in the optimal pathway. If a TAD border is found in all 10 sub-optimal paths, then the score of the border is equal to 10, if it was found only one time, the score is 1. Finally, TADbit also returns a TAD density score as the ratio between the number of interactions within TADs and the number of interactions of the rest of the genome.

Genome segmentation into A/B compartments

A/B compartments were identified at 100kb resolution using HOMER (34). Briefly, HOMER calculates correlation between the contact profiles of each bin against each other, and performs principal component analysis (PCA) on chromosome-wide matrices. Normally, the A compartment is assigned to genomic bins with positive first principal component (PC1), and the B compartment is assigned to genomic bins with negative PC1. However, in some chromosomes and in cell lines with aberrant karyotypes, the PC1 is reversed in the sign, with A compartment corresponding to negative PC1, and B compartment corresponding to positive PC1. Additionally, sometimes the PC1 captures other correlations in the chromosome that do not correspond to the compartments. For these reasons, all PC1 and PC2 for all chromosomes were visually inspected and correctly assigned to decipher the proper segmentation of the genome into the A and B compartments.

3D modelling of TADs based on Hi-C data

TADbit model (29) was used with default parameters to generate 3D models of selected TADs at the resolution of 10 kb. Hi-C interaction maps were transformed into a set of spatial restraints that were then used to build 3D models of the TADs that satisfied as best as possible the imposed restraints, as previously described (35,36). For each TAD,

we generated 1000 models, structurally aligned and clustered them in an unsupervised manner, to generate sets of structurally related models. For every TAD, we used the main cluster to compute consistency, accessibility, density, radius of gyration, and walking angle (29). Consistency quantifies the variability of the position of particles across the considered set of models. Accessibility measures with a fraction from 0 to 1 how much each particle in a model is accessible to an object (*i.e.* a protein complex) with a radius of 100 nm. Density measures a proxy for local DNA compactness as the ratio of DNA base pairs and the distances between two consecutive particles in the models – the higher the density, the more compact the DNA. Walking angle measures the angle between triplets of consecutive particles—the higher the value, the straighter the models—and can be used as a proxy for the stiffness of the chromatin fiber. Finally, radius of gyration measures 3D structure compactness as the root mean square distance of the all particles in a model from its center of mass.

ATAC-Seq data analysis

ATAC-Seq data identified by the accession number GSE100762 was reprocessed as previously described (37) with slight modifications. Paired-end reads were quality-checked via FASTQC (v0.11.9), trimmed, and subsequently aligned to the human GRCh37/hg19 reference genome using Bowtie2 (v2.3.5.1) (23). SAMtools (v1.9) (24) was used to filter out the low-quality reads with the flag 1796, remove reads mapped in the mitochondrial chromosome and discard those with a MAPQ score below 30. The peak calling was performed with MACS2 (v.2.1.2) (26) by specifying the *-BAMPE* mode. Filtered BAM files were also used to compute the ATAC-Seq genome coverage, which was normalized to reads per million (*genomecov -ibam -bga -scale*).

Genomic data retrieval

Genome-wide GC content, G bands coordinates at 850 bands per haploid sequence (bphs) resolution and chromosomes coordinates were obtained from the UCSC human genome database (38,39). G bands were classified as G positive (Gpos25 to Gpos100, according to its intensity upon Giemsa staining), and G negative (unstained), which were further divided into four groups according to their GC content (Gneg1 to Gneg4, from high to low GC content). HeLa-S3 genome segmentation by ChromHMM (ENCODE) was obtained from UCSC human genome database (38,39). RNA-seq and ATAC-seq datasets were download from GEO (accession numbers GSE83277 and GSE100762, respectively) an parsed as previously described (22).

REFERENCES

1. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragozy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326, 289–293.
2. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485, 376–380.
3. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., Van Berkum,N.L., Meisig,J., Sedat,J. et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485, 381–385.
4. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-

dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148, 458–472.

5. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159, 1665–1680.

6. Le Dily,F.L., Baù,D., Pohl,A., Vicent,G.P., Serra,F., Soronellas,D., Castellano,G., Wright,R.H.G., Ballare,C., Filion,G. et al. (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.*, 28, 2151–2162.

7. Bednar,J., Horowitz,R.A., Grigoryev,S.A., Carruthers,L.M., Hansen,J.C., Koster,A.J. and Woodcock,C.L. (1998) Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc. Natl Acad. Sci. U.S.A.*, 95, 14173–14178.

8. Millan-Ari ´ no,L., Izquierdo-Bouldstridge,A. and Jordan,A. (2016) ~ Specificities and genomic distribution of somatic mammalian histone H1 subtypes. *Biochim. Biophys. Acta - Gene Regul. Mech.*, 1859, 510–519.

9. Fyodorov,D.V, Zhou,B.-R., Skoultchi,A.I. and Bai,Y. (2018) Emerging roles of linker histones in regulating chromatin structure and function. *Nat. Rev. Mol. Cell Biol.*, 19, 192–206.

10. Izzo,A., Kamieniarz,K. and Schneider,R. (2008) The histone H1 family: specific members, specific functions? *Biol. Chem.*, 389, 333–343.

11. Happel,N. and Doenecke,D. (2009) Histone H1 and its isoforms: contribution to chromatin structure and function. *Gene*, 431, 1–12.
12. Kalashnikova,A.A., Winkler,D.D., McBryant,S.J., Henderson,R.K., Herman,J.A., DeLuca,J.G., Luger,K., Prenni,J.E. and Hansen,J.C. (2013) Linker histone H1.0 interacts with an extensive network of proteins found in the nucleolus. *Nucleic Acids Res.*, 41, 4026–4035.
13. Li,J.Y., Patterson,M., Mikkola,H.K.A., Lowry,W.E. and Kurdistan,S.K. (2012) Dynamic distribution of linker histone H1.5 in cellular differentiation. *PLoS Genet.*, 8, e1002879.
14. Cao,K., Lailier,N., Zhang,Y., Kumar,A., Uppal,K., Liu,Z., Lee,E.K., Wu,H., Medrzycki,M., Pan,C. et al. (2013) High-resolution mapping of H1 linker histone variants in embryonic stem cells. *PLoS Genet.*, 9, e1003417.
15. Izzo,A., Kamieniarz-Gdula,K., Ram´irez,F., Noureen,N., Kind,J., Manke,T., van Steensel,B. and Schneider,R. (2013) The genomic landscape of the somatic linker histone subtypes H1.1 to H1.5 in human cells. *Cell Rep.*, 3, 2142–2154.
16. Millan-Ari ´ no,L., Islam,A.B.M.M.K., Izquierdo-Bouldstridge,A., ~ Mayor,R., Terme,J.M., Luque,N., Sancho,M., Lopez-Bigas,N. and ´ Jordan,A. (2014) Mapping of six somatic linker histone H1 variants in human breast cancer cells uncovers specific features of H1.2. *Nucleic Acids Res.*, 42, 4474–4493.
17. Mayor,R., Izquierdo-Bouldstridge,A., Millan-Ari ´ no,L., Bustillos,A., ~ Sampaio,C., Luque,N. and Jordan,A. (2015) Genome distribution of replication-independent histone H1 variants shows H1.0 associated with nucleolar domains and H1X associated with RNA polymerase II-enriched regions. *J. Biol. Chem.*, 290, 7474–7491.

18. Torres,C.M., Biran,A., Burney,M.J., Patel,H., Henser-Brownhill,T., Cohen,A.H.S., Li,Y., Ben-Hamo,R., Nye,E., Spencer-Dene,B. et al. (2016) The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity. *Science*, 353, aaf1644.
19. Sancho,M., Diani,E., Beato,M. and Jordan,A. (2008) Depletion of human histone H1 variants uncovers specific roles in gene expression and cell growth. *PLoS Genet.*, 4, e1000227.
20. Serna-Pujol,N., Salinas-Pena,M., Mugianesi,F., Lopez-Anguita,N., Torrent-Llagostera,F., Izquierdo-Bouldstridge,A., Marti-Renom,MA. and Jordan,A. (2021) TADs enriched in histone H1.2 strongly overlap with the B compartment, inaccessible chromatin and AT-rich Giemsa bands. *FEBS J.*, 288, 1989–2013.
21. Geeven,G., Zhu,Y., Kim,B.J., Bartholdy,B.A., Yang,S.M., Macfarlan,T.S., Gifford,W.D., Pfaff,S.L., Verstegen,M.J.A.M., Pinto,H. et al. (2015) Local compartment changes and regulatory landscape alterations in histone H1-depleted cells. *Genome. Biol.*, 16, 289.
22. Izquierdo-Bouldstridge,A., Bustillos,A., Bonet-Costa,C., Aribau-Miralbes,P., García-Gomis,D., Dabad,M., Esteve-Codina,A., Pascual-Reguant,L., Peiro,S., Esteller,M. et al. (2017) Histone H1 depletion triggers an interferon response in cancer cells via activation of heterochromatic repeats. *Nucleic Acids Res.*, 45, 11622–11642.
23. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9, 357–359.
24. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
25. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

26. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoutte,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. et al. (2008) Model-based analysis of Chip-Seq (MACS). *Genome Biol.*, 9, R137.
27. Vara,C., Paytuví-Gallart,A., Cuartero,Y., Le Dily,F., Garcia,F., Salva-Castro,J., Gomez,H.L., Julia,E., Moutinho,C., Aiese Cigliano,R. et al. (2019) Three-Dimensional genomic structure and cohesin occupancy correlate with transcriptional activity during spermatogenesis. *Cell Rep*, 28, 352–367.
28. Pascual-Reguant,L., Blanco,E., Galan,S., Le Dily,F., Cuartero,Y., Serra-Bardenys,G., Di Carlo,V., Iturbide,A., Cebria-Costa,J.P., Nonell,L. et al. (2018) Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. *Nat. Commun.*, 9, 3420.
29. Serra,F., Bàu,D., Goodstadt,M., Castillo,D., Filion,G. and Martí-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.*, 13, e1005665.
30. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, 9, 999–1003.
31. Marco-Sola,S., Sammeth,M., Guigo,R. and Ribeca,P. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9, 1185–1188.
32. Vidal,E., le Dily,F., Quilez,J., Stadhouders,R., Cuartero,Y., Graf,T., Martí-Renom,M.A., Beato,M. and Filion,G.J. (2018) OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.*, 46, e49.

33. Yang,T., Zhang,F., Yardımcı,G.G., Song,F., Hardison,R.C., Noble,W.S., Yue,F. and Li,Q. (2017) HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, 27, 1939–1949.
34. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-Regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38, 576–589.
35. Russel,D., Lasker,K., Webb,B., Velazquez-Muriel,J., Tjio,E., Schneidman-Duhovny,D., Peterson,B. and Sali,A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, 10, e1001244.
36. Baù,D. and Marti-Renom,M.A. (2012) Genome structure determination via 3C-based data integration by the integrative modeling platform. *Methods*, 58, 300–306.
37. Corces,M.R., Trevino,A.E., Hamilton,E.G., Greenside,P.G., Sinnott-Armstrong,N.A., Vesuna,S., Satpathy,A.T., Rubin,A.J., Montine,K.S., Wu,B. et al. (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, 14, 959–962.
38. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W. and David Haussler,W.J.K. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, 32, 493D–496.
39. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. et al. (2021) The UCSC genome browser database: 2021 update. *Nucleic Acids Res.*, 49, D1046–D1057.

40. Terme,J.M., Sese,B., Millán-Ariño,L., Mayor,R., Belmonte ~ Izpis´ua,J.C., Barrero,M.J. and Jordan,A. (2011) Histone H1 variants are differentially expressed and incorporated into chromatin during differentiation and reprogramming to pluripotency. *J. Biol. Chem.*, 286, 35347–35357.
41. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9, 215–216.
42. Stoldt,S., Wenzel,D., Schulze,E., Doenecke,D. and Happel,N. (2007) G1 phase-dependent nucleolar accumulation of human histone H1x. *Biol. Cell*, 99, 541–552.
43. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, 14, 390–403.
44. Ryan,D.P. and Tremethick,D.J. (2018) The interplay between H2A.Z and H3K9 methylation in regulating HP1 binding to linker histone-containing chromatin. *Nucleic Acids Res.*, 46, 9353–9366.
45. Daujat,S., Zeissler,U., Waldmann,T., Happel,N. and Schneider,R. (2005) HP1 binds specifically to Lys26-methylated histone H1.4, whereas simultaneous Ser27 phosphorylation blocks HP1 binding. *J. Biol. Chem.*, 280, 38090–38095.
46. Yusufova,N., Kloetgen,A., Teater,M., Osunsade,A., Camarillo,J.M., Chin,C.R., Doane,A.S., Venters,B.J., Portillo-Ledesma,S., Conway,J. et al. (2021) Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature*, 589, 299–305.
47. Willcockson,M.A., Heaton,S.E., Weiss,C.N., Bartholdy,B.A., Botbol,Y., Mishra,L.N., Sidhwani,D.S., Wilson,T.J., Pinto,H.B., Maron,M.I. et al. (2021) H1 histones control the epigenetic landscape by local chromatin compaction. *Nature*, 589, 293–298.

48. Woodcock,C.L., Skoultchi,A.I. and Fan,Y. (2006) Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res.*, 14, 17–25.
49. Meshorer,E., Yellajoshula,D., George,E., Scambler,P.J., Brown,D.T. and Misteli,T. (2006) Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev. Cell*, 10, 105–116.
50. Pan,C. and Fan,Y. (2016) Role of H1 linker histones in mammalian development and stem cell differentiation. *Biochim. Biophys. Acta Protein Struct. Mol. Enzymol.*, 1859, 496–509.
51. Scaffidi,P. (2016) Histone H1 alterations in cancer. *Biochim. Biophys. Acta - Gene Regul. Mech.*, 1859, 533–539.
52. Chandra,T., Ewels,P.A., Schoenfelder,S., Furlan-Magaril,M., Wingett,S.W., Kirschner,K., Thuret,J.-Y., Andrews,S., Fraser,P. and Reik,W. (2015) Global reorganization of the nuclear landscape in senescent cells. *Cell Rep.*, 10, 471–483.
53. Bayona-Feliu,A., Casas-Lamesa,A., Reina,O., Bernues,J. and Azorín,F. (2017) Linker histone H1 prevents R-loop accumulation and genome instability in heterochromatin. *Nat. Commun.*, 8, 283.
54. Almeida,R., Fernandez-Justel,J.M., Santa-María,C., Cadoret,J.C., Cano-Aroca,L., Lombrana,R., Herranz,G., Agresti,A. and ~ Gomez,M. (2018) Chromatin conformation regulates the ´ coordination between DNA replication and transcription. *Nat. Commun.*, 9, 1590.
55. Nagano,T., Lubling,Y., Varnai,C., Dudley,C., Leung,W., Baran,Y., Mendelson Cohen,N., Wingett,S., Fraser,P. and Tanay,A. (2017) Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547, 61–67.

CHAPTER 2

CHROMATIC reveals chromatin-associated factors contributing to genome topology

Candidate's contribution: Design, development and coding of the CHROMATIC method. Application of CHROMATIC to biologically relevant samples. Analysis and interpretation of the results.

CHROMATIC reveals chromatin-associated factors contributing to genome topology

Francesca Mugianesi^{1,2}, Ivano Mocavini², Enrique Blanco², Luciano Di Croce^{2,3,4,*}, and Marc A. Marti-Renom^{1,2,3,4,*}

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldori i Reixac 4, 08028 Barcelona, Spain.

2. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.

3. Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain.

4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

*To whom correspondence should be addressed: M.A.M-R. martirenom@cnag.crg.eu and L.D.C. Luciano.DiCroce@crgeu

ABSTRACT

Chromatin-associated factors play a fundamental role in chromatin long-range interactions, which in turn are key for proper spatiotemporal regulation of gene expression. However, the identification of factor-associated chromatin interactions and the characterization of their role in transcription are still elusive. Here we introduce CHROMATIC, a

novel computational method that integrates Hi-C and ChIP-seq data to study chromatin three-dimensional (3D) interactions associated with any factor of interest. CHROMATIC is faster and less expensive than performing experiments that probe protein-directed genome architecture, such as HiChIP. Thanks to the deconvolution of the Hi-C data into factor-specific interactions, our strategy allows discerning the role of each studied factor in genome 3D structure in a cell-type-specific manner. Furthermore, the classification of 3D colocalization patterns of factors using CHROMATIC identifies types of functional 3D interactions, that we call ‘3D-types’. 3D-types may reflect already known interactions between different chromatin factors or may help discover new associations between molecules with specific functional roles. By applying our algorithm to mouse embryonic stem cells (ESCs) and neural progenitor cells (NPCs), we analyzed changes in the types of 3D interactions during early stages of neuronal cell differentiation. We found that pluripotency transcription factors (TFs) play a major role in the genome structure of pluripotent stem cells. When differentiating from ESCs to NPCs, cells switch to a less plastic and more specialized configuration. Overall, the CHROMATIC tool unifies factor occupancy and genome topology analyses, to shed light on their link with gene expression.

INTRODUCTION

Gene expression, epigenetic states, and topological conformation are three facets of the genome that tightly operate in space and time (Zheng & Xie, 2019). Unfortunately, the detailed characterization of the link between them is still largely missing.

The 3D architecture of eukaryotic genomes is organized in multiple layers with a relevant role in gene expression control (Bonev et al., 2017; Rowley & Corces, 2018). At the chromosomal scale, the genome is partitioned into regions of preferential long-range interactions, called A and B compartments, which resemble euchromatin and heterochromatin, respectively (Lieberman-Aiden et al., 2009). A compartment is enriched in histone post-translational modifications (PTMs) associated with transcriptional activity, while B compartment in chromatin modifications that are typical of transcriptional repression. At the sub-megabase scale, Topologically Associating Domains (TADs), or domains in general, are self-interacting regions considered functional units of the genome (Dixon et al., 2012; Nora et al., 2012). Compartments and domains emerge as a result of multiple, dynamic, and cell-type-specific interactions between distal regulatory elements, such as gene promoters and enhancers, driven by different classes of proteins that tightly interact with DNA via either specific or unspecific sequence recognition (Cavalli & Misteli, 2013).

Chromatin interactions can be either loops between pairs of DNA loci or hubs of multiple DNA loci that are clustered together (Bonev & Cavalli, 2016; Rao et al., 2014), likely via loop-extrusion (Alipour & Marko, 2012; Fudenberg et al., 2016; Sanborn et al., 2015) and phase separation (Banani, Lee, Hyman, & Rosen, 2017; Shin & Brangwynne, 2017) mechanisms. Indeed, the loop extrusion model proposed that CTCF and cohesin mediate loop interactions, which are important for accurate gene regulation (Alipour & Marko, 2012; Fudenberg et al., 2016; Sanborn et al., 2015). Also, Polycomb group proteins (PcG) repress target genes via their clustering into repressive 3D hubs known as Polycomb bodies (Blackledge et al., 2020; Eagen, Aiden, & Kornberg,

2017; Huseyin & Klose, 2021; Kundu et al., 2017; Ogiyama, Schuettengruber, Papadopoulos, Chang, & Cavalli, 2018). Additionally, key ESC TFs such as NANOG, SOX2, OCT4, KLF4, and ESRRB are found at most pluripotency genes in ESCs (Whyte et al., 2013) and are associated with 3D enhancer rewiring and transcriptional changes during reprogramming (Stadhouders et al., 2018). Overall, protein-associated chromatin interactions are fundamental to ensure proper gene expression (Di Giammartino et al., 2019), but their identification and the characterization of the underlying mechanisms are still lacking. Hi-C is an experimental technique combining DNA proximity ligation (Cullen, Kladde, & Seyfred, 1993; Dekker, Rippe, Dekker, & Kleckner, 2002) with high-throughput sequencing, that is mostly used to probe compartments, domains, loops, and hubs (Lieberman-Aiden et al., 2009). The result of a population-based Hi-C experiment is a list of DNA-DNA contacts between pairs of loci in at least hundreds of thousands of cells, usually represented by a map of contact frequencies. To elucidate the map of chromatin long-range interactions driven by specific proteins, several methods have been developed that combine ChIP-seq with 3C-based (Chromosome Conformation Capture) experiments (Furey, 2012). Currently, HiChIP is the most suitable strategy (Mumbach et al., 2016), in which ChIP is performed on the Hi-C library of proximity-ligated DNA fragments. The comparison of HiChIP of cohesin subunit SMC1a with Hi-C revealed that this method enhances the signal-to-background ratio, enriching the signal at chromatin loops associated with cohesin and depleting it elsewhere (Mumbach et al., 2016).

The study of combinations of multiple proteins and marks in linear chromatin (1D) has been fundamental to annotate chromatin states,

discover regulatory regions, and characterize their cell type-specific patterns (Day, Hemmaplardh, Thurman, Stamatoyannopoulos, & Noble, 2007; Ernst & Kellis, 2010; Ernst et al., 2011; Filion et al., 2010; mod et al., 2010). These combinatorial patterns often capture known classes of genomic elements, such as enhancers, promoters, transcriptionally active and repressed regions, or can help discover novel classes of elements. ChromHMM learns chromatin states from multiple ChIP-seq epigenomic tracks using a multivariate hidden Markov model (HMM) and is the most widely used software for this purpose (Ernst & Kellis, 2012, 2017). However, ChromHMM offers a mono-dimensional perspective on chromatin states by considering chromatin as a linear entity. Thus, it does not take advantage of the insights gained from studying chromatin in its 3D context. A recent study based on machine learning and polymer physics discovered a combinatorial code linking 3D chromatin architecture to 1D chromatin states, that allows to derive models of genome 3D conformations from 1D chromatin states through physics mechanisms, outperforming the 3D modeling based on epigenetic linear segmentation only (Esposito et al., 2022). Hence, 1D chromatin states and genome architecture are intimately linked, but at present there is no computational method to characterize chromatin states directly in 3D, by integrating chromatin interactions and factor occupancy.

To address this limitation, we have developed CHROMATIC, a computational method to characterize chromatin functional states in 3D. CHROMATIC systematically integrates chromatin structure data from Hi-C interaction matrices and genome-wide factor occupancy data from ChIP-seq profiles, to identify chromatin 3D interactions associated with proteins and histone post-translational modifications

(PTMs). Importantly, since different chromatin factors may cooperate for proper gene regulation thanks to their colocalization in the 3D genome, the application of CHROMATIC to a certain set of factors reveals ‘3D-types’, i.e., types of 3D interactions associated with specific combinations of factors.

To demonstrate its applicability, we used CHROMATIC on a comprehensive set of 37 ChIP-seq tracks of chromatin factors in two different cell types, mouse ESCs and NPCs. We characterized four major types of functional 3D interactions for each cell line. Finally, by comparing the results obtained for the two, we identified factors that mostly contribute to genome structure in a cell type-specific manner and analyzed changes in types of 3D interactions during early stages of neuronal cell differentiation.

CHROMATIC is fast and conceptually simple, resulting in the classification of major types of chromatin interactions that are linked to a specific biological function. Thus, CHROMATIC constitutes an inexpensive and reliable alternative to study factor-associated interactions compared to performing experiments such as HiChIP.

METHODS

Experimental datasets

Genome interaction maps for mouse ES cells (mESC) and neural progenitor cells (NPC) were obtained from *in situ* Hi-C experiments previously generated (GEO database accession number GSE96107) (Bonev et al., 2017). ChIP-seq datasets were previously generated by our (Stevens et al., 2017) and other labs and are available in the GEO database with accession codes GSE99530 (Mas et al., 2018), GSE79606 (Beringer et al., 2016), GSE42466 (Morey, Aloia, Cozzuto, Benitah, &

Di Croce, 2013), GSE44288 (Whyte et al., 2013), GSE22557 (Kagey et al., 2010), GSE11431 (Chen et al., 2008), GSE89575 (C. Huang et al., 2017), GSE53542 (Aloia et al., 2014), GSE57186 (McAninch & Thomas, 2014), GSE35496 (Lodato et al., 2013), GSE65462 (Nishi et al., 2015), GSE96107 (Bonev et al., 2017), GSE36203 (Phillips-Cremins et al., 2013), GSE74330 (Kloet et al., 2016). Interaction datasets for SMC1 and OCT4 HiChIP were also downloaded from GEO (GSE80820) (Mumbach et al., 2016). Constitutive Lamin Associating Domains (LADs) dataset was downloaded from GEO (GSE17051) (Peric-Hupkes et al., 2010) and converted from mouse MGSCv37/mm9 to GRCm38/mm10 reference genome using the Lift Genome Annotation tool (Kent et al., 2002). To find active enhancers, the command *intersect -v* from BEDTools toolkit (Quinlan & Hall, 2010) was used to select peaks of H3K27ac not overlapping with peaks of H3K4me3, the command *genomeDistribution* from SeqCode toolkit (Blanco, Gonzalez-Ramirez, & Di Croce, 2021) was used to calculate the distribution of the selected peaks into different genomic features, and the peaks annotated as intergenic and intronic were considered as active enhancers. Analogously, poised enhancers were found as intergenic and intronic peaks H3K27me3. To find active promoters, the command *intersect* from BEDTools toolkit (Quinlan & Hall, 2010) was used to find the overlap between peaks of H3K4me3 and H3K27ac, that was then intersected with the position of the Transcription Start Site (TSS) of RefSeq genes ($TSS \pm 500bp$). For bivalent promoters, the same process was applied, considering instead the overlap between H3K4me3 and H3K27me3 peaks. Super-enhancers were computed using HOMER (Heinz et al., 2010) on H3K27ac ChIP-seq data and

relative to the control, with the command *findPeaks -style super -o auto*. All datasets were parsed as described in the following sections.

ChIP-seq data processing

Single-end reads were aligned to mouse GRCm38/mm10 reference genome using Bowtie2 (Langmead & Salzberg, 2012) with default options. SAMtools utilities (Li et al., 2009) were used to filter out unaligned reads with the flag -F 0x4. The command *buildChIPprofile* from SeqCode toolkit (Blanco et al., 2021) was used to generate BedGraph profiles from BAM files, where the total number of reads of the experiment was used to normalize the height of the resulting profile. To avoid mapping artifacts, the set of genomic regions reported in the ENCODE blacklist was removed from the aligned sequences (Amemiya, Kundaje, & Boyle, 2019). To reduce the influence of outliers, we held out values that are more than five standard deviations higher than the average, since their probability is significantly low ($p=3 \times 10^{-7}$ in case of normal distribution) and possibly correspond to artifacts. We set this selected group of values to the maximum value among the retained ones, corresponding to five standard deviations above the average value. Next, ChIP-seq BedGraph tracks were binned at a resolution of 5 kb for their subsequent integration with the Hi-C data. The resulting tracks were further divided by the corresponding tracks of control (IgG, WCE, GFP) for normalization, and linearly transformed in the range of values between zero and one to be able to compare the output obtained for different ChIP-seq tracks. Finally, the MAC2 software (Y. Zhang et al., 2008) was used for peak calling from BAM files against controls using the command *callpeak --nomodel --extsize 150*, with the *--broad* option for H3K36me3, H3K27me3, RNA Pol II Serine 5P, and RNA Pol II tracks.

Hi-C and HiChIP data processing

Hi-C datasets were processed using TADbit (Serra et al., 2017). Specifically, for both mESC and NPC datasets, paired-end FASTQ files of four Hi-C replicates, previously assessed for reproducibility (Bonev et al., 2017), were merged and mapped to mouse GRCm38/mm10 reference genome applying a fragment-based iterative strategy (Imakaev et al., 2012) using the GEM mapper (Marco-Sola, Sammeth, Guigo, & Ribeca, 2012). Mapped reads were filtered using TADbit with default parameters, which removed self-circles, dangling ends, duplicated and random breaks among other minor artifactual reads (Serra et al., 2017). After mapping and filtering, the resulting Hi-C matrices contained a total of 1,537,751,681 valid pairs for mESC (**Supplementary Table 2**) and 3,974,901,849 for NPC (**Supplementary Table 3**). The resulting raw Hi-C interaction matrices were next normalized using OneD (Vidal et al., 2018) at the resolution of 5kb, which removed experimental Hi-C biases. Similarly, the available merged FASTQ files from four replicates of SMC1 and OCT4 HiChIP (Mumbach et al., 2016) were processed, mapped, filtered, and normalized with TADbit with default parameters, which resulted in a total of 219,998,058 valid pairs for SMC1 and 252,920,123 valid pairs for OCT4 (**Supplementary Table 4**).

Genome segmentation into A/B compartments

A/B compartments were identified at the resolution of 100kb using TADbit (Serra et al., 2017). Briefly, TADbit calculates the correlation between the contact profiles of each bin against each other and performs principal component analysis (PCA) on chromosome-wide matrices. Normally, the A compartment is assigned to genomic bins with positive first principal component (PC1), and the B compartment

is assigned to genomic bins with negative PC1. However, in some chromosomes the PC1 is reversed in the sign, with A compartment corresponding to negative PC1, and B compartment corresponding to positive PC1. Additionally, sometimes the PC1 captures other correlations in the chromosome that do not correspond to compartments. For these reasons, and since GC content correlates with A compartment, PC1 and PC2 were compared to GC content for all chromosomes, visually inspected, and correctly assigned to decipher the proper segmentation of the genome into A and B compartments.

The CHROMATIC pipeline

The CHROMATIC pipeline takes as input a Hi-C normalized interaction matrix and integrates it with a series of ChIP-seq tracks to identify colocalization of groups of marks in the 3D space of the nucleus. Specifically, the pipeline is composed of several steps:

1. *ChIP-seq and Hi-C pre-processing.* For each studied factor, CHROMATIC takes as input the ChIP-seq track c of the factor and an intra-chromosomal Hi-C matrix H obtained from the same cell type and obtained as described above. Before the integration of the two types of data, the input Hi-C matrix is smoothed with the function *medfilt* from *scipy.signal* package with *kernel_size*=5. Also, the input ChIP-seq values are re-scaled as follows to increase the spread of the signal:

$$f(c_i) = \begin{cases} \sqrt[3]{-abs(c_i - 0.5)} + 0.8, & \text{if } c_i < 0.5 \\ 3^{c_i}, & \text{if } c_i \geq 0.5 \end{cases}$$

where c_i is the ChIP-seq signal at bin i . To choose the re-scaling function, we employed a heuristic strategy where several different transformations were applied to our ChIP-seq

to identify the one that best separated low and high ChIP-seq values (**Suppl. Figure 1**).

2. *Hi-C re-weighting*. Intra-chromosomal Hi-C matrices were next re-weighted using the following formula:

$$C_{ij} = H_{ij} \times f(c_i) \times f(c_j)$$

where C_{ij} is the re-weighted Hi-C interaction between bins i and j , H_{ij} is the Hi-C normalized interaction frequency, and $f(c_i)$ and $f(c_j)$ are the transformed ChIP-seq values of bins i and j . To minimize the computational burden for re-weighting large Hi-C matrices, a sliding window of 2,000 bins of 5 kb resolution was used allowing the re-weighting of interactions as far as 10 Mb in sequence using a single computer.

3. *Detection of patches of 3D interaction*. To detect chromatin interactions associated with a given factor, CHROMATIC generated a binary matrix P for each chromosome and for each factor, whose pixels P_{ij} were equal to 1 if there was a ChIP-seq peak in at least one of the bases of the interaction or their adjacent bins and its C_{ij} values were equal to or larger than 0.2. Next, a series of four operations from morphological image processing was applied to the P_{ij} matrix, with *scipy.ndimage* python package: 1) a binary opening using a square 4 by 4 structuring element, 2) a binary closing with a cross-shaped 3 by 3 structuring element, 3) a binary dilation with a square 5 by 5 structuring element, and 4) a binary closing with a square 5 by 5 structuring element. The resulting matrix L , thus, included all

patches of interactions corresponding to significant integration of ChIP-seq and Hi-C signals.

4. *Identification of '3D-types'.* Next, CHROMATIC implements Latent Semantic Analysis (LSA) (Dumais, 2005), a technique in natural language processing. LSA analyzes relationships between a set of documents and their contained terms to automatically identify sets of topics with shared terms. As implemented in CHROMATIC, LSA aims to identify types of 3D interaction ("topics") based on the overlap of 3D interactions associated with different factors ("terms"). In general, LSA generates a *document-term* array describing the terms contained in each document. In our case, it generates an *interaction-factor* array describing the factors participating in each interaction. Next, LSA uses Singular-Value Decomposition (SVD) to find the main combinations of terms that define topics. Thus, in CHROMATIC it finds the main combination of factors in the detected genome interactions, which represent the types of 3D interactions (or '3D-types'). This is performed by two functions: first *TfidfVectorizer* from the *sklearn.feature_extraction.text* package (*stop_words='english', max_df = 1.0, smooth_idf=True*), and then *TruncatedSVD* from *sklearn.decomposition* package (*algorithm='randomized', n_iter=100, random_state=122*). CHROMATIC applied to the ESC, found 18 3D-types, which corresponds to the maximum allowed with 19 factors. In NPC, CHROMATIC found 17 3D-types for a total of 18 factors. As output, CHROMATIC generates two matrices: the *interaction-3Dtype*, which allowed to associate each 3D interaction to one of the identified 3D-types and the *3Dtype-*

factor arrays, which described the composition of each 3D-type in terms of enrichment or depletion of the studied factors.

5. *Overlap of 3D-types with functional genomic features.* 3D-types identified by CHROMATIC were next mapped into genomic loci, which allowed to assess their overlap with functional genomic features for active enhancers (AE), active promoters (AP), super-enhancers (SE), poised enhancers (PE), bivalent promoters (BP), and constitutive LADs (CL). The enrichment of selected 3D-types in each of the functional genomic features was measured by its odds ratio (OR), which quantified the level of association between two events. OR was defined as:

$$OR = \frac{a/c}{b/d}$$

where a is the number of bins of overlap between a functional state and a 3D-type, c is the number of bins of the functional state that do not overlap with the 3D-type, b is the number of bins of 3D-type that are not of the functional type, and d is the number of classified bins that are neither of the functional type or the 3D-type.

6. *Clustering of 3D-types into major types of 3D interactions.* The \log_{10} of the OR of the overlap between 3D-types and functional genomic features (AE, BP, SE, PE, BP, CL) was used as input data for a Principal Component Analysis (PCA). Data standardization was performed by the function *StandardScaler* of the *sklearn.preprocessing* package and fitting of data was done by using *PCA().fit* of the *sklearn.decomposition* module. For each cell type, the minimum number of principal components (PC)

explaining more than 80% of the variance was chosen. Specifically, in ESC the first two PC were kept, capturing 92.2% of the variance, while in NPC the first three PC were kept, explaining 90.9% of the variance. Then, PCA was performed with the chosen number of components and the obtained data was used for K-means clustering to cluster the 3D-types. To determine the number of clusters to compute, the K-means algorithm was run multiple times with a different number of clusters: from 1 to 18 in ESC, where 18 3D-types were identified, and from 1 to 17 in NPC, where 17 3D-types were classified. For each solution, the Within Cluster Sum of Squares (WCSS) was computed. To determine the number of clusters to use, the approach known as the *Elbow* method was used, which consists of looking for a kink or elbow in the plot of the values of WCSS against the number of clusters. The elbow point is identified by the different exponential of the descent on the left and the right of the plot. In both ESC and NPC, the elbow appeared in correspondence with 4 clusters of major types of interactions (**Suppl. Figure 2**).

Functional characterization of major types of 3D interactions

To assign the identified major types of interactions to their biological function, the following analyses were performed:

1. *Overlap with AE, AP, SE, PE, BP, CL*. Each major 3D-type was mapped to 1D genomic loci, whose overlap with AE, AP, SE, PE, BP, and CL was measured by odds ratio as described above.
2. *Proportion of highly-, lowly-expressed and silent genes*. The command *matchpeaksgenes* from SeqCode toolkit (Blanco et al., 2021) was used to match the genomic loci corresponding to each major

3D-types to mouse genes (mm10), within the promoter, 2.5 kb upstream of TSS, and the body of genes. Subsequently, in ESC and NPC, genes were divided into three categories based on whether they were highly-, lowly-expressed or silent. For this purpose, 2 RNA-seq replicates per cell type were analyzed. So, for each gene, the average between the RPKM values of the two replicates was computed, and the gene was considered highly-, lowly-expressed or silent if the RPKM value was respectively $RPKM > 10$, $1 < RPKM \leq 10$, and $RPKM \leq 1$.

3. *Proportion of overlap with A/B compartments.* Genomic loci corresponding to each major 3D-types were intersected with the list of loci assigned to the A compartment or the B compartment as calculated from the Hi-C maps using TADbit (Serra et al., 2017).

Statistical tools for benchmarking

For the comparison between CHROMATIC output and HiChIP data, linear regression was performed by the *polyfit* function from *numpy.polynomial* package. The function *ks_2samp* from *scipy.stats* package was used to perform two-sample Kolmogorov-Smirnov test, that compares the distribution of HiChIP values corresponding to loops and hubs detected by CHROMATIC against the distribution of HiChIP values in the rest of the interaction matrix.

Statistical tools for clustering

Unsupervised hierarchical clustering results were obtained by the *clustermap* function from *seaborn* library (*method=average, metric=euclidean*). To assess how the resulting clustering resembled an ideal one (*i.e.*, when factors cluster according to their functional role), we used the *normalized_mutual_info_score* from *sklearn.metrics* package (NMIS). Thus,

from the obtained hierarchical clustering the ideal number of clusters was computed and then compared with the ideal solution. NMIS value was between 0 (no correlation between clusters) and 1 (perfect correlation).

Cell culture and differentiation

Sox1:GFP E14Tg2a mouse embryonic stem cells (mESC) (Ying, Stavridis, Griffiths, Li, & Smith, 2003) were routinely cultured in Serum/LIF conditions using Glasgow minimum essential medium (Sigma, G5154) supplemented with 20% inactivated fetal bovine serum (Cytiva HyClone SV30160.03), Glutamax (Gibco, 35050-038), Pen/Strep (Gibco, 15140-122), non-essential amino acids (NEAA, Gibco, 11140-050), β -Mercaptoethanol (Gibco, 31350-010), and Leukemia inhibitory factor (LIF, produced and titrated in-house) on cell culture treated plates coated with 0.1% gelatin (Millipore, ES-006-B). Neural precursor cells (NPCs) were obtained as described in ref. (Ying et al., 2003). Briefly, 1.8×10^3 cells cm^{-2} were plated on gelatin-coated plate in Serum/LIF conditions. After 24h the medium was changed to N2B27 differentiation medium, composed of a 1:1 mixture of Neurobasal (Gibco, 21203-049) supplemented with N2 (17502-048) and DMEM-F12 (11320-074) supplemented with B27 (17504-044), to which Glutamax, Pen/Strep, NEAA, and 0.33% BSA fraction V (15260-037) were added. Differentiation medium was changed every other day. Cell differentiation was monitored via cytometre using the Sox1:GFP internal reporter, that marks early neuroectoderm committed cells (Wood & Episkopou, 1999; Ying et al., 2003) and harvested after 6 days of differentiation.

Gene expression analysis

RNA extraction was performed using the RNeasy kit (Qiagen, 74134) according to manufacturer's instructions. For RNA-seq application, RNA samples were processed as follows: samples were quantified using the Nanodrop spectrophotometer (Thermo Fisher Scientific) and libraries were prepared using the TruSeq stranded mRNA Library Prep (Illumina, 20020595) according to the manufacturer's protocol. Libraries were sequenced on a single end for 50+8bp on Illumina's HiSeq2500. A minimum of 40×10^6 reads per sample was generated. Next, raw sequencing data was analyzed as follows: RNA-seq samples were mapped against the mm10 mouse genome assembly using TopHat (Trapnell, Pachter, & Salzberg, 2009) with the option `-g 1` to discard reads that could not be uniquely mapped. DESeq2 (Love, Huber, & Anders, 2014) was run to quantify the expression of every annotated transcript using the RefSeq catalogue of exons (O'Leary et al., 2016) and to identify each set of differentially expressed genes between two conditions. Raw counts and mapped statistics are provided as supplementary material (**Supplementary Table 1**).

RESULTS

Overview of the CHROMATIC algorithm

CHROMATIC characterizes chromatin states in 3D by combining Hi-C data with ChIP-seq tracks of proteins and histone PTMs. First, Hi-C and ChIP-seq data are pre-processed and normalized to remove biases and artifacts (**Methods**). Second, for each factor, CHROMATIC combines the Hi-C map with its ChIP-seq track, generating a new matrix (C) where the coefficient for each pair of bins is given by the

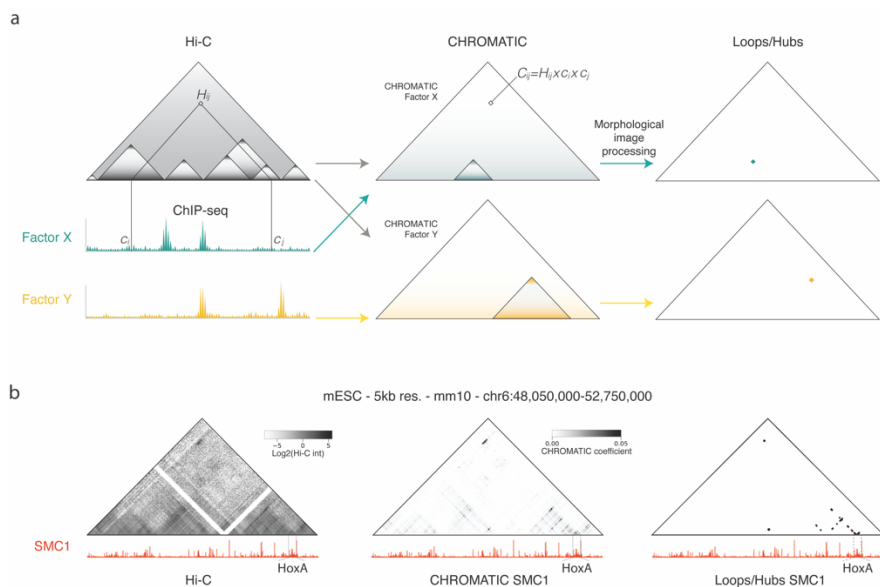


Figure 1. Overview of the CHROMATIC algorithm. **a** Schematic representation of the fundamental steps of CHROMATIC. Left, normalized Hi-C matrix and ChIP-seq tracks of two generic factors X and Y. Center, for each factor the combination of the Hi-C map with the ChIP-seq track generates a new matrix resembling Hi-ChIP maps (CHROMATIC map), where the coefficient C_{ij} is high if bins i and j interact in 3D and have also ChIP-seq enrichment for the factor. Right, a series of morphological image processing operations detects loops and hubs associated with each factor. **b** Specific example for SMC1 in ESCs at the resolution of 5kb, in a region containing *HoxA* gene cluster. Left, normalized Hi-C map and ChIP-seq track of SMC1. Center, CHROMATIC map generated for SMC1. Right, loops and hubs detected for SMC1.

corresponding normalized Hi-C coefficient multiplied by the transformed ChIP-seq values of the two anchoring bins (**Fig. 1a**) (**Methods**). Thus, the value C_{ij} in matrix C is high if loci in bins i and j interact in 3D and exhibit enrichment for the factor signal. Next, to automatically detect loops and hubs associated with the factor, a series of operations from morphological image processing is applied to matrix C . Such operators are applied to identify loops/hubs up to 10 Mb in sequence range and with at least a ChIP-seq peak at one of the two

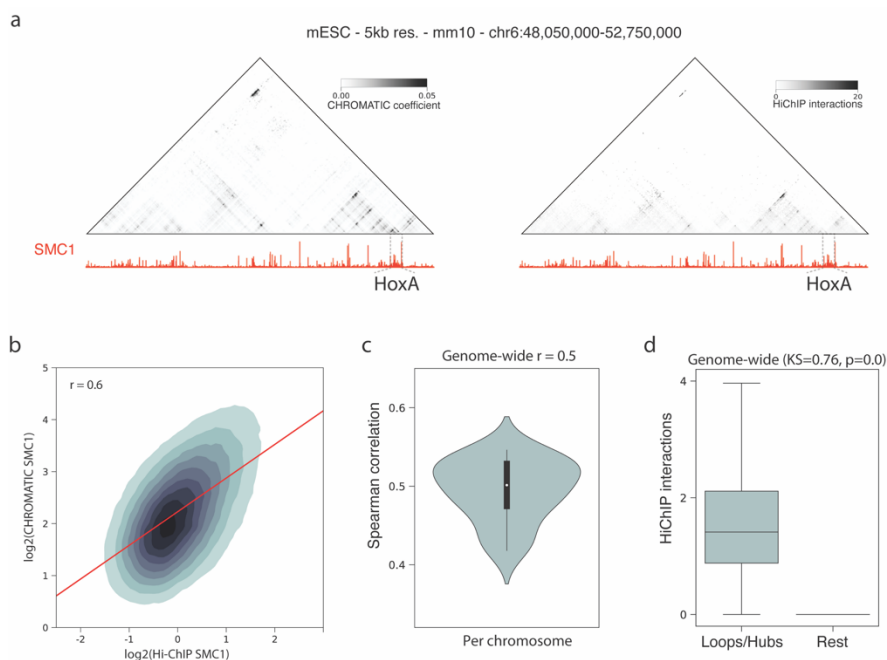


Figure 2. CHROMATIC interactions correlate with HiChIP. **a** Comparison of CHROMATIC SMC1 maps (left) to HiChIP data (right) in a region in chromosome 6 including *HoxA* cluster, at the resolution of 5kb for ESC. SMC1 ChIP-seq is shown underneath each matrix. **b** Correlation between CHROMATIC coefficients and HiChIP values for SMC1 in chromosome 6. Spearman correlation coefficient $r=0.6$ ($p\text{-value} = 0$). **c** Spearman correlation coefficients per chromosome. Genome-wide median $r=0.5$ genome-wide. **d** Boxplots of HiChIP values of SMC1 from detected CHROMATIC SMC1 patches compared HiChIP interactions elsewhere in the matrix (statistically different distributions as for Kolmogorov-Smirnov statistical test= 0.76 , $p\text{-val}=0$).

anchors (**Methods**). As an example, CHROMATIC efficiently detects loops associated with SMC1 on the *HoxA* locus (**Fig. 1b**).

CHROMATIC interactions correlate with HiChIP

To benchmark the CHROMATIC detection of significant interactions mediated by a given factor, we used already published HiChIP datasets for the structural protein SMC1 and the TF OCT4 in mESC (Mumbach et al., 2016) (**Fig. 2a** and **Suppl. Figure 3**). For each chromosome,

CHROMATIC interactions were compared with the corresponding HiChIP maps (**Fig. 2b**), resulting in a median Spearman correlation coefficient of 0.5 genome-wide (**Fig. 2c**). To further assess the accuracy of CHROMATIC detection of significant interactions directed by SMC1, next, we studied the frequency of HiChIP interactions within CHROMATIC detections compared to sites with no detection (**Fig. 2d**). Our analysis clearly indicated that sites with detected CHROMATIC interaction corresponded to pairs of loci that highly interact in HiChIP (Kolmogorov-Smirnov statistical test=0.76, p -val~0). Similar results were obtained for the OCT4 factor (**Suppl. Figure 3**). Overall, CHROMATIC accurately identifies factor-associated chromatin interactions experimentally determined by HiChIP.

CHROMATIC identifies 3D chromatin functional interactions

CHROMATIC can be regarded as a Hi-C matrix deconvolver where the original interaction map is separated into a series of layers associated with each of the different analyzed factors (**Fig. 3a**). This deconvolution exercise allows CHROMATIC to efficiently identify interactions associated with a given factor, which would have been difficult to detect from the original Hi-C map. CHROMATIC was indeed applied genome-wide to 5 kb Hi-C maps and two distinct ChIP-seq datasets of 19 factors and 18 factors in mouse ESC and neural progenitor cells (NPC), respectively (**Fig. 3b** and **Suppl. Figure 4a**). ChIP-seq data included Polycomb group proteins, pluripotency and neuronal TFs, architectural proteins, and chromatin marks related to both activity and repression.

Genome-wide, CHROMATIC detected 49,597 and 46,850 patches of interactions in ESC and NPC, respectively (that is, 5.5% less in NPC

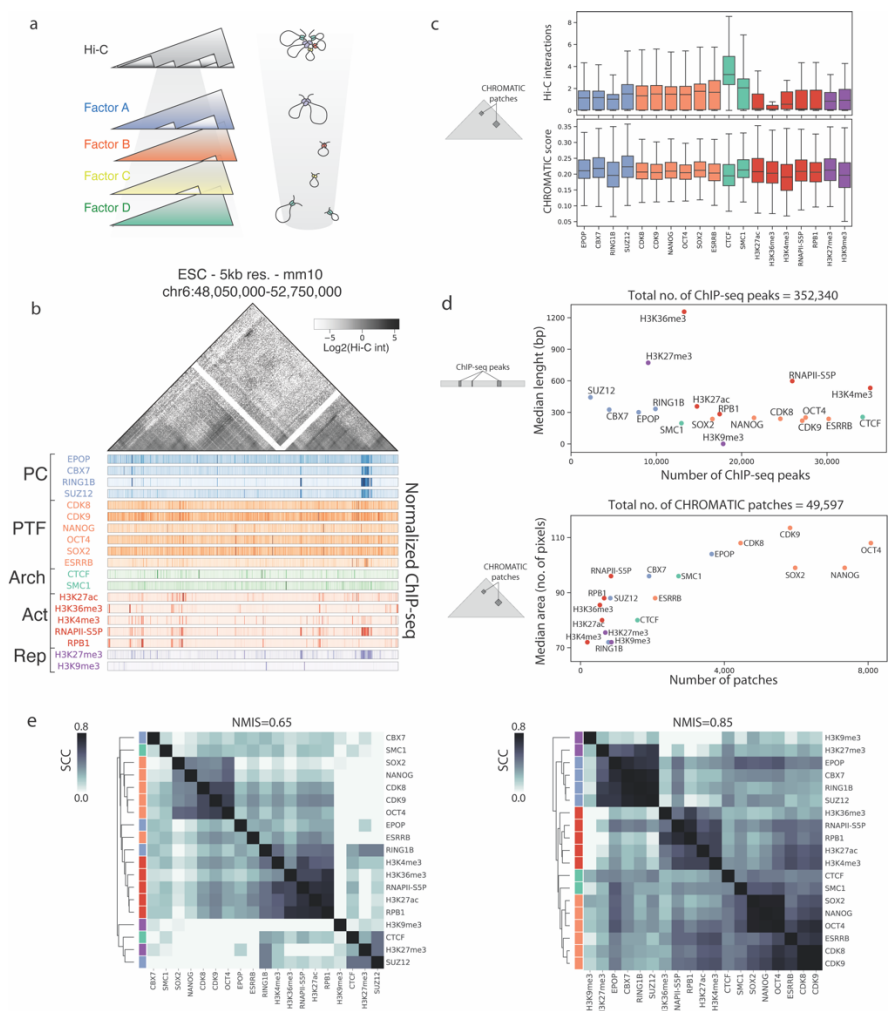


Figure 3. CHROMATIC identifies 3D chromatin functional interactions. **a** In essence, CHROMATIC deconvolves the original Hi-C matrix in layers associated with each of the analyzed factors, allowing to efficiently identify interactions associated with each of them. **b** Example of CHROMATIC applied to Hi-C interaction maps and ChIP-seq profiles at the resolution of 5kb, for 19 factors in ESC. Factors are colored according to their factional role. **c** Top, value distributions of original Hi-C interactions corrected by decay and median filter in ESC, before CHROMATIC processing, in correspondence of the patches detected by CHROMATIC. For each patch, the average of the corresponding Hi-C values is considered. Bottom, CHROMATIC coefficient distributions in ESC, in correspondence of the detected patches. **d** Top, number of ChIP-seq peaks for each factor with respect to their median length (base pairs), in ESC. Bottom, number of patches detected genome-wide by CHROMATIC for each factor with respect to their median area (number of

5kb x 5kb pixels). **e** Unsupervised hierarchical clustering of factors studied in ESC based on their genome-wide pair-wise correlation, of ChIP-seq tracks on the left and of CHROMATIC maps on the right.

than in ESC). The values of original Hi-C, before CHROMATIC processing, in correspondence with the detected patches, reveal that architectural proteins, especially CTCF and SMC1, are associated with the strongest Hi-C interactions (**Fig. 3c** and **Suppl. Fig. 4b**, top panels). This result agrees with the fact that CTCF-driven peaks were the first ones to be systematically discovered in Hi-C maps (Rao et al., 2014). However, the values of the CHROMATIC score are appreciably balanced among different factors (**Fig. 3c** and **Suppl. Fig. 4b**, bottom panels). CHROMATIC can, thus, detect significant interactions also for factors whose interactions appear comparatively weak in the Hi-C map such as H3K36me3 in ESC (**Fig. 3c**).

To explore the relative contribution of the studied factors to the spatial organization of the genome, we next compared the number and length of ChIP-seq peaks to the number and area of CHROMATIC detected patches (**Fig. 3d** and **Suppl. Fig. 4c**). In ESC, histone PTMs related to transcriptional activity, RNA Pol II-Ser5P, and RNA Pol II subunit RPB1 have a high number of mid-sized ChIP-seq peaks. However, they appear in fewer and smaller 3D interactions compared to other factors (**Fig. 3d** bottom). This may indicate that, in ESCs, histone marks related to transcriptional activity, RNA Pol II-Ser5P, and RNA Pol II subunit RPB1 may not play a genome-wide structural role and could be rather considered of a more specific functional role. Instead, pluripotency TFs may play a more relevant role in ESC genome topology than previously reported (*i.e.*, they result in the highest number of CHROMATIC patches of largest size, **Fig 3d** bottom). Interestingly, in NPC, histone

PTM H3K27ac and neuronal TF OLIG2 result in the most abundant CHROMATIC interactions and of larger sizes, indicating that they may have a more prominent structural role (**Suppl. Fig. 4c** bottom). Notably, at the structural level H3K27ac may be more related to organizing the genome structure (more patches and of larger size) in NPC compared to ESC. In NPC, CTCF is also found in a large number of CHROMATIC interactions of mid-size (**Suppl. Fig. 4c** bottom). The apparent stronger correlation between factor type and their role in 3D genome organization, especially in ESC, prompted us to further analyze the correlation of the factors at ChIP-seq tracks (1D) and CHROMATIC (3D) levels (**Fig. 3e** and **Suppl. Fig. 4d**). The results indicate that the dendrogram of unsupervised hierarchical clustering of factors based on CHROMATIC correlations better separates the functional role of factors, especially in ESC. For example, EPOP and SUZ12, two Polycomb (PcG) proteins, are known to co-bind the same set of loci, same for CBX7 and RING1B. Importantly, although this is not reflected in the clustering of ChIP-seq signals, CHROMATIC correctly associates PcG components together.

In summary, CHROMATIC allows the discovery of factor-specific interactions by deconvolving the Hi-C signal into factor-specific signals otherwise hidden by the background levels of the experimental data. The detection of CHROMATIC signal results in the identification of factors that may contribute to genome structure in a cell type-specific manner. Finally, the CHROMATIC interactions detected are more informative of the functions of the studied factors.

CHROMATIC classifies functional types of chromatin 3D interactions

To study the function of the analyzed factors in mediating genome structure, we next aimed to classify all CHROMATIC interactions into a limited number of interaction types similar to what ChromHMM does in 1D signal (Ernst & Kellis, 2012). To do so, we implement a Latent Semantic Analysis (LSA) approach that aims at identifying the cooperativity of signals in the spatial genome (**Fig. 4a** and **Methods**). LSA is conceptually simple and computationally fast. Indeed, for its genome-wide application, the computing time was less than 10 minutes in a single modern workstation for both studied cell types. The LSA output is represented as a heatmap (**Fig. 4b** and **Suppl. Fig. 5a**) defining a set of types of chromatin 3D interactions ('3D-types') based on specific combinations of factors found in those genome interactions. In ESC, 3D-type '1' is the most abundant and is concomitantly enriched in pluripotency factors OCT4, NANOG, SOX2, CDK8, and CDK9, while it is moderately depleted of the rest of factors. 3D-type '2', instead, is enriched in the association of NANOG and SOX2, and is depleted of CDK9, CDK8, and OCT4 (**Fig. 4b**). To functionally characterize the identified 3D-types, all CHROMATIC interactions were mapped into their genomic coordinates and their overlap with functional genomic features was computed. These included active enhancers (AE), active promoters (AP), super-enhancers (SE, *i.e.*, 3D clusters of enhancers), poised enhancers (PE), bivalent promoters (BP), and constitutive LADs (CL) (**Fig. 4c** and **Suppl. Fig. 5b**). Interestingly, in ESC 3D-type '4' is characterized by the presence of SOX2 and the absence of NANOG and OCT4, and has a strong overlap with CLs. 3D-type '5' is enriched in PcG components EPOP and CBX7, together

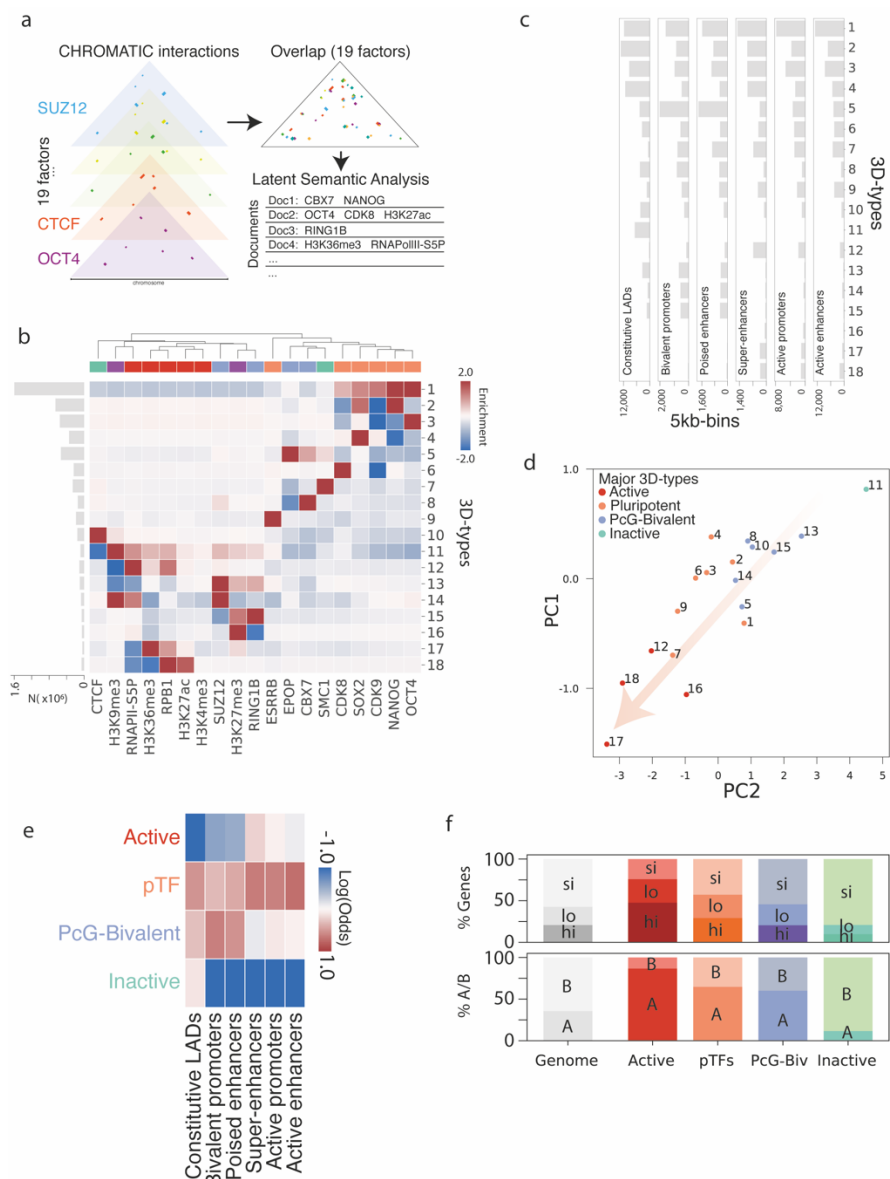


Figure 4. CHROMATIC classifies functional types of chromatin 3D interactions. **a** Patches detected by CHROMATIC for different factors colocalize in 3D in a set of limited combinations detected by latent semantic analysis (LSA). **b** Resulting emissions of LSA in ESC defining sets of types of 3D interactions (3D-types) in terms of enrichment (in red) or depletion (in blue) of factors. Factors are colored according to their functional role as in **Fig. 3b**. Left, bar plot indicates the number of 5kb x 5kb pixels associated with each 3D-type. **c** Overlap in number of 5kb-bins between 1D loci corresponding to each 3D-type and chromatin types, in

ESC. **d** LogOdds of the overlap between 1D loci corresponding to each 3D-type and the functional genomic features in **c** were used as input for principal component analysis. The first two PC were considered, capturing 92.2% of the total variance. Plots depict the values of principal components 1 and 2 (PC1, PC2) for the different 3D-types, which were further classified by K-means unsupervised clustering, in ESC. The arrow shows the direction from inactive to active for the identified 3D-types. **e** LogOdds of the overlap between 1D loci corresponding to each major 3D-type and the functional genomic features in ESC. **f** Percentage of silent (si), lowly-expressed (lo), and highly-expressed (hi) genes, and percentage of A and B compartments for the whole genome (in gray) and for the 4 major 3D-types, in ESC.

with SMC1, while it is depleted in pluripotency TFs and strongly overlaps with BPs and PEs. Conversely, 3D-type ‘7’ is enriched in SMC1 while it is depleted of EPOP, and overlaps more with AEs and APs compared to 3D-type ‘5’. Similar to 3D-type ‘5’, 3D-types ‘8’, ‘13’, ‘14’ and ‘15’ are enriched in Polycomb components and mainly overlap with BPs and PEs. Unexpectedly, 3D-type ‘10’ is enriched in CTCF and, to a lesser extent, in H3K9me3, and overlaps mainly with CLs. 3D-type ‘11’ is strongly enriched in H3K9me3, partially enriched in RNA Pol II subunits, histone marks related to transcriptional activation, and PcG proteins, at the same time that mildly depleted of pluripotency TFs, and sharply overlap with CLs. 3D-type ‘14’, where H3K9me3 and Polycomb Repressive Complex 2 (PRC2) subunit SUZ12 are strongly enriched simultaneously, with depletion of H3K36me3, H3K27ac, and H3K4me3, overlaps with BPs and PEs. 3D-types ‘12’, ‘17’ and ‘18’ are among the least abundant types, are enriched in histone marks related to transcriptional activity (*i.e.*, H3K36me3 or H3K27ac and RNA Pol II subunit RPB1), and mainly overlap with SEs. Surprisingly, 3D-type ‘16’ is enriched in H3K27me3 and weakly in RNAPII-Ser5P, and overlaps with APs and SEs.

In NPC, 3D-type ‘1’ is the most abundant of all and is simultaneously enriched in neuronal TF OLIG2, PRC1 component RING1B, PRC2

component EZH2 and heterochromatic protein CBX3. 3D-type ‘2’ is again enriched in OLIG2 but it is depleted in PcG components, and compared to 3D-type ‘1’ it is more associated with AEs, APs, and SEs. 3D-type ‘3’ is enriched in EZH2 and CBX3, depleted in RING1B and OLIG2, and mainly overlaps with CLs and APs. 3D-type ‘4’ is enriched in CBX3 and depleted in EZH2, it overlaps with CLs similar to 3D-type ‘3’, but also with BPs. 3D-type ‘5’ is enriched in SMC1, neuronal TF NKX6.1, activator ZRF1, and active mark H3K4me3, is depleted of factors enriched in 3D-types ‘1’-‘4’ and strongly overlaps with BPs, APs, and SEs. 3D-type ‘7’ is similar to 3D-type ‘5’, but it is enriched only in H3K4me3 and ZRF1, while being depleted in SMC1 and NKX6.1. 3D-type ‘9’ is enriched in Pol II, CTCF, and H3K27me3, and overlaps with AEs, APs, SEs, and BPs. 3D-type ‘10’ is enriched in CTCF and H3K27me3 but it is depleted in Pol II, and overlaps more with PEs and CLs. Surprisingly, 3D-type ‘11’ is enriched in H3K27me3, H3K9me3, and neuronal TF NKX2.2, and mainly overlaps with APs. 3D-type ‘14’ is enriched in SOX2, PcG proteins SUZ12 and PCGF2 and active mark H3K27ac, and overlaps with PEs and BPs.

Next, the results of the overlap between loci corresponding to each 3D-type and AEs, APs, SEs, PEs, BPs, and CLs were used as input for principal component analysis (PCA) to reduce the dimensionality of the data, which finally was further classified by K-means unsupervised clustering (**Fig. 4d**, **Suppl. Fig. 5c**, and **Methods**). In each cell type, 3D-types were clustered into four groups of interactions with different functional roles according to their enrichment or depletion of functional marks (**Fig. 4e** and **Suppl. Fig. 5d**). The PCA analysis indicates that for both cell types, the detection of interaction types by CHROMATIC allows for functionally classifying the 3D-types from

inactive to active regions of the 3D genome (*i.e.*, in ESC the axis composed of PC1 and PC2, while for NPC PC1 axis can be considered the path from inactive to active interactions). Therefore, the four identified clusters based on the PCA analysis correspond to four functional types: Active, cell-type-specific Transcription Factors, PcG-bivalent and Inactive.

To assess whether the four types of spatial interactions indeed represent ranges of activity in the genome, mouse genes were assigned to their 3D-type/s (**Methods**). Once mapped, genes were classified as silent (“si”, RPKM<1), lowly (“lo”, 1<RPKM<10), and highly expressed (“hi”, RPKM>10) and their proportion in each of the four groups of 3D-type interactions was assessed (**Fig. 4f** and **Suppl. Fig. 5e**). For both cell types, and as expected by their chromatin states (both 1D and 3D), there is a correlation between the expression of the resident genes and the type of 3D interaction they concur. The “Active” 3D-type is enriched in active genes and occurs more often in the A compartment compared with the genome-wide distribution. Conversely, the “Inactive” 3D-type is enriched in silent genes and occurs more often in the B compartment (**Fig. 4f** and **Suppl. Fig. 5e**).

In total, 5,216,011 5Kb x 5Kb patches are classified in ESC, while 6,710,882 are classified in NPC (22.3% less in ESC than in NPC). In ESC the vast majority (73.4%) of 3D interactions are associated with pluripotency TFs, 20.1% is associated with a bivalent state characterized by the presence of PcG proteins, and only 6.5% is specialized in either active or inactive states (**Fig. 5a** left panel). In NPC, TFs are associated with structure (40.2%), but to a lower extent compared to ESC, there are many more 3D interactions that are specialized in either active or

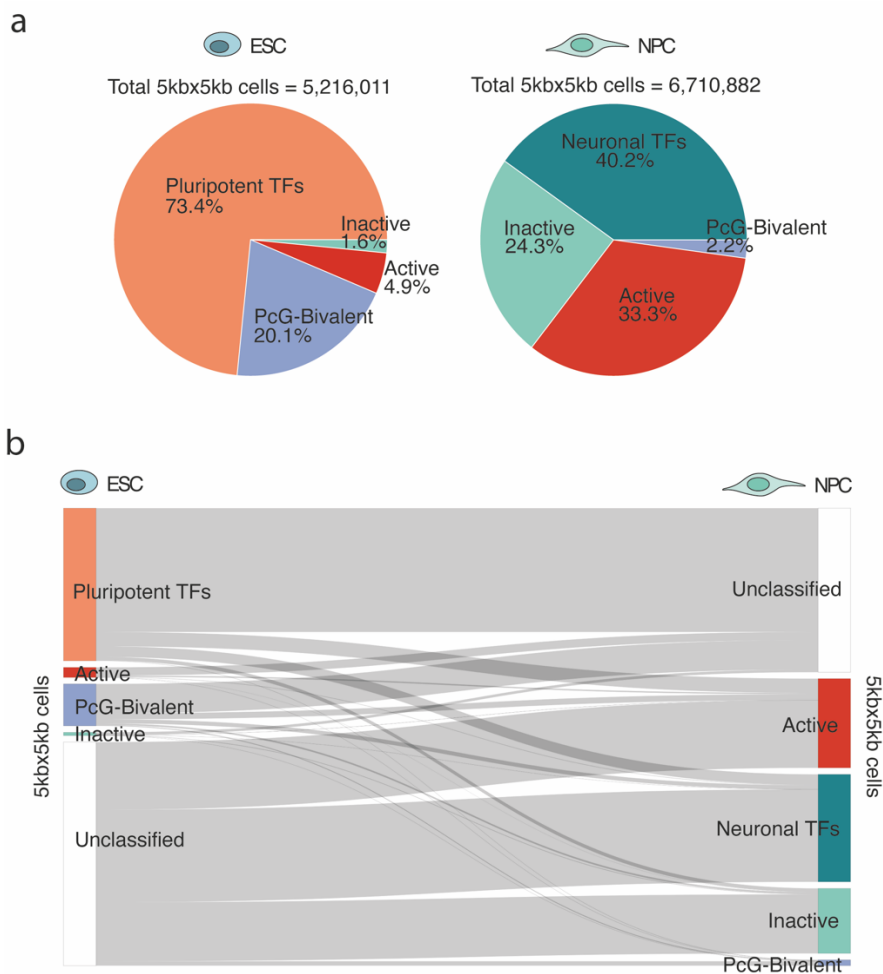


Figure 5. Changes in 3D interaction types during mouse neural development.
a Distribution of the four major types of interactions in ESC (left) and NPC (right). Percentages refer to the number of 5kb x 5kb cells of each type compared to the total number of 5kb x 5kb cells classified in each cell type (reported at top of the pie charts).
b Sankey plot describing the transitions between the different types of 3D interactions, between ESC (left) and NPC (right). “Unclassified” cells indicate 5kb x 5kb cells that were not classified by CHROMATIC.

inactive states (57.6%), and bivalent interactions have a 10-fold decrease (2.2%) (**Fig. 5a** right panel). Most interactions that are classified in ESC are unclassified in NPC, and vice versa (**Fig. 5b**). Considering only interactions that are classified in both cell types, each major 3D-type

identified in ESC mainly transitions into “Active” and “Neuronal TFs” in NPC, except for “Inactive” interactions that mainly remain “Inactive” (**Suppl. Fig. 6**).

Altogether, CHROMATIC classification into 3D interaction types allows for further investigation of structural interactions with functional meaning. Such classes can be regarded as the 3D chromatin state of a cell type in a similar way that ChromHMM classifies linear chromatin states based solely on co-occupancy of ChIP-seq tracks (Ernst & Kellis, 2012, 2017).

Changes in complex functional 3D hubs during mouse neural development

To assess whether specific loci alter their chromatin states in 3D, we studied the changes on CHROMATIC identified interactions between ESC and NPC of two loci of interest for their involvement in the development of neurons (that is, the *Zfp608* and *HoxA* loci). The *Zfp608* locus is a neural-specific region where, during differentiation, a novel domain boundary is formed at the TSS of *Zfp608*, concomitantly with the activation of the gene (Bonev et al., 2017). In ESC, the gene is involved in a few interactions that CHROMATIC classified as PcG-bivalent, while in NPC it participates in a larger number of interactions that were classified as active and associated with neuronal TFs involving H3K27ac (**Fig. 6**). In contrast to the *Zfp608* locus, the structural changes between ESC and NPC of the *HoxA* locus are less dramatic. However, the chromatin binding of factors changes significantly, which is identified by the altered 3D-types of interactions as determined by CHROMATIC. In ESC, the *HoxA* cluster genes are not expressed and are found within a bivalent domain associated with PcG proteins. In

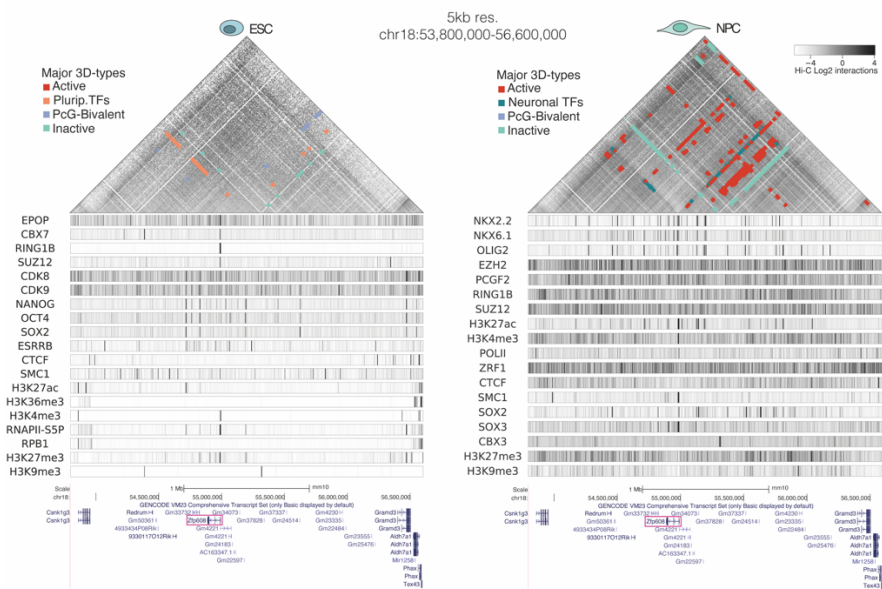


Figure 6. Changes in complex functional 3D hubs during mouse neural development. Interactions classified into major 3D-types, in ESC (left) and NPC (right), in a neural-specific region in chromosome 18. Hi-C maps (top) and ChIP-seq tracks (below) are in grey. Major 3D-types classified by CHROMATIC are in colors. The *Zfp608* gene is highlighted. During differentiation, a novel TAD boundary is formed at the TSS of *Zfp608*, concomitantly with the activation of the gene. In ESC, there are no CHROMATIC interactions involving the gene. In NPC, multiple interactions are classified as active or associated with neuronal TFs, possibly allowing the gene to scan downstream putative enhancers.

NPC, instead, they are still enclosed by a bivalent cage, but in the interior a small active domain appears, in agreement with the activation of a small group of HoxA genes (Noordermeer et al., 2014; Noordermeer et al., 2011) (**Suppl. Fig. 7**).

DISCUSSION

To better capture the relationship between gene expression, epigenetic states, and genome topology, here we presented CHROMATIC, a novel computational method that for the first time integrates Hi-C and ChIP-

seq data in a single map of *in silico* HiChIP. CHROMATIC results in fast, inexpensive, and accurate identification of factor-associated chromatin interactions in the 3D space, which agree with those already determined experimentally by HiChIP. Effectively, CHROMATIC deconvolves the Hi-C signal into factor-specific interactions otherwise hidden by the background levels of the Hi-C experimental data. Moreover, compared to the analysis of ChIP-seq data alone, the detected CHROMATIC interactions will provide more information on the function(s) of these factors on chromatin.

The application of CHROMATIC between two or more different cell types also helps to identify factors that contribute to genome topology in a cell type-specific manner. Thus, we applied it to a total of 37 different factors in ESC and NPC. Among the analyzed factors, pluripotency TFs in ESC and H3K27ac and neuronal TF OLIG2 in NPC are associated with an unexpectedly large fraction of 3D chromatin interactions, suggesting that they may play the most relevant structural role. On the one hand, in line with pieces of evidence from other studies (Kim & Shendure, 2019), TFs might play a crucial role in shaping the genome, especially in pluripotent cells, to properly regulate genes in a cell-type manner. On the other hand, H3K27ac, which decorates AEs and APs, results in a more prominent structural role in NPC compared to ESC, which may be explained by the fact that in ESC it is dispensable for enhancer activity (T. Zhang, Zhang, Dong, Xiong, & Zhu, 2020), in agreement with the remarkable structural role observed for this mark in NPC. Some other factor may intervene to bridge together H3K27ac enhancers in NPC.

Based on the 3D colocalization of the studied factors, we identified different types of functional 3D interactions. In ESC, 3D-type ‘1’ is the

most abundant and is enriched in OCT4, SOX2, NANOG, CDK8, and CDK9. This may represent the cooperative association of pluripotency TFs that is known to be crucial for the efficiency of stem cell transcriptional regulation (Chronis et al., 2017; X. Huang & Wang, 2014; Yeo & Ng, 2013), which could correspond to condensates of pluripotency TFs (Boija et al., 2018; Hnisz, Shrinivas, Young, Chakraborty, & Sharp, 2017). 3D-type ‘2’, instead, is enriched exclusively in SOX2 and NANOG, whose cooperative interaction has been already reported to be central to ESC self-renewal (Gagliardi et al., 2013; Yesudhas D, Anwar MA, & S, 2019). 3D-type ‘4’ is enriched in SOX2, depleted in NANOG and OCT4, and mainly overlaps with CLs. Interestingly, SOX2 has been shown to act also as a transcriptional repressor in neural stem cells (Liu et al., 2014), thus 3D-type ‘4’ interactions could help SOX2 to exert its repressive role. As expected, 3D-types that are enriched in Polycomb components (3D-types ‘5’, ‘8’, ‘13’, ‘14’, and ‘15’) mainly overlap with BPs and PEs. As it happens for most of the analyzed proteins, SMC1 participates in 3D interactions with different functional roles depending on its 3D-colocalizing factors. When it associates in 3D with Polycomb proteins, the involved loci strongly overlap with BPs and PEs (3D-type ‘5’). In absence of Polycomb, SMC1 interactions overlap more with APs and AEs (3D-type ‘7’). Notably, the 3D association of CTCF and SMC1 described in the loop-extrusion model is not particularly enriched in any identified 3D-type. However, this does not exclude that it is a participant in more than one 3D-type of interaction. For example, 3D-types ‘5’ and ‘7’ are enriched in SMC1 and show mild enrichment in CTCF. Thus, 3D-types ‘7’ might correspond also to the well-known CTCF-SMC1 extruded loops. 3D-type ‘10’ instead, which is strongly enriched in CTCF and

H3K9me3 and mainly overlaps with CLs, might capture the repressive role of CTCF which has been previously described (Lutz et al., 2000). 3D-type ‘11’ is particularly enriched in H3K9me3 and sharply overlaps with CLs. It is interesting that, despite being the most inactive 3D-type in ESC (**Fig. 4d**), it is also enriched in RNA Pol II subunits, histone marks related to transcriptional activation, and PcG proteins. This might indicate that in pluripotent stem cells inactive regions are not completely silent, but are instead ready to be activated at the right time during differentiation, reflecting the high plasticity characteristic of ESCs. Differently, in 3D-type ‘14’, where H3K9me3 and PRC2 subunit SUZ12 are simultaneously enriched together with the depletion of H3K36me3, H3K27ac, and H3K4me3, chromatin interactions are associated with a bivalent state. Finally, 3D-types ‘12’, ‘17’ and ‘18’ are enriched in histone marks related to transcriptional activity H3K36me3 or H3K27ac and RNA Pol II subunit RPB1, and mainly overlap with SEs. They are among the least abundant 3D-types, meaning that only a portion of SEs may be exclusively characterized by H3K36me3, H3K27ac, RPB1, while most SEs may also be enriched in pluripotency TFs (3D-types ‘1’ to ‘4’). Surprisingly, 3D-type ‘16’ includes 3D interactions marked by H3K27me3 in the absence of Polycomb and involving active loci. Further investigations are needed to properly interpret such observation.

In NPC, 3D-type ‘1’ is enriched in neuronal TFs OLIG2 and NKX6.1, RING1B, EZH2, and CBX3. This result agrees with the fact that PRC2 component EZH2 colocalizes with OLIG2 in neurogenic astroglia (Hwang et al., 2014). *Olig2* is a direct target of EZH2, and its repression is critical for neuronal differentiation. Thus, regions involved in 3D-type ‘1’ might include genes like *Olig2* that will be shut down for mature

neuron differentiation. Conversely, 3D-type ‘2’ is enriched in OLIG2 and NKX6.1 but it is depleted of RING1B, EZH2, and CBX3 and overlaps more with AEs, APs and SEs compared to 3D-type ‘1’. OLIG2 can function either as a repressor or an activator in oligodendrocyte formation (Wei et al., 2021) and this might be reflected in the different functional roles of 3D-types ‘1’ and ‘2’. 3D-types ‘3’ and ‘4’ are enriched in heterochromatic protein CBX3 and show marked overlap with CLs. However, 3D-type ‘3’ is also enriched in PRC2 subunit EZH2 and slightly in H3K4me3, and overlaps more with APs. This might reflect the fact that, beyond its well-known repressive function, PRC2 binds APs and contacts nascent RNAs (Kaneko, Son, Shen, Reinberg, & Bonasio, 2013). 3D-type ‘5’ and ‘7’ are enriched in ZRF1 and active mark H3K4me3, with 3D-type ‘5’ involving also SMC1 and NKX6.1. Both 3D-types mainly overlap with BPs, APs, and SEs, and might involve loci that are important for the establishment and maintenance of neural progenitor identity (Aloia et al., 2014). 3D-type ‘9’ is enriched in RNA Pol II, CTCF, and mildly in H3K27me3, and overlaps with AEs, APs, SEs, and BPs. It might correspond to regions that in ESC were covered by H3K27me3 and kept in a bivalent state and that began to be expressed in NPC. 3D-type ‘10’ is also enriched in CTCF and H3K27me3 but it is depleted in Pol II, and indeed it overlaps more with PEs and CLs. Such type of 3D interaction is consistent with the observed role of CTCF-based loops in the spreading of repressive H3K27me3 mark at distant micro-domains that repress euchromatic genes (Heurteau et al., 2020). Surprisingly, 3D-type ‘11’ is enriched in H3K27me3, H3K9me3, and neuronal TF NKX2.2, and mainly overlaps with APs. NKX2.2 can function both as a transcriptional repressor and activator, depending on temporal and cellular context (Doyle & Sussel,

2007), thus regions involved in 3D-type ‘11’ interactions might be repressed in ESC and start to be expressed in NPC. Further analyses are needed to characterize this type of interaction. 3D-type ‘14’ is enriched in SOX2, PcG proteins, and H3K27ac and overlaps with PEs and BPs; like 3D-type ‘4’ in ESC, SOX2 might act as a transcriptional repressor in such interactions and, thanks to Polycomb and H3K27ac, contribute to set the involved loci in a bivalent state.

The study of combinatorial patterns of multiple proteins and chromatin marks has been fundamental to annotate chromatin states, discover novel regulatory elements and characterize their cell type-specific patterns (Day et al., 2007; Ernst & Kellis, 2010; Ernst et al., 2011; Filion et al., 2010; mod et al., 2010). Chromatin states have recently been linked to genome 3D conformation by machine learning and polymer physics approaches (Esposito et al., 2022), but they continue to be considered as a 1D entity. CHROMATIC follows principles that are similar to the ones of ChromHMM, but it extends the potential of such combinatorial approaches being the first computational method to offer a 3D perspective on chromatin states. Identified 3D-types may indeed reflect already known interactions between different chromatin factors, or may help discover new associations between molecules with specific functional roles that need to be validated by specific experiments.

To further investigate the functional implications of chromatin interaction types, in each studied cell we grouped chromatin interaction types into four major functional classes: Active, TFs-associated, PcG-bivalent, and Inactive. Such classes can be regarded as the 3D chromatin states of a cell type, similar to how we consider linear chromatin states (Ernst & Kellis, 2012, 2017). Overall, ES cells result in about 50% of all genome interactions as unclassified (that is, with no CHROMATIC

significant interaction type), which is about 20% fewer classified pixels compared to NPC. Hence, the structure of the NPC genome is more restrained by functional interactions compared to the ESC genome. Moreover, most ESC interactions are associated with pluripotency TFs and with a Polycomb-bivalent state, leaving only 6.5% of the genome associated with active or inactive states. In NPC, instead, most of the classified interactions are active or inactive (57.6%), while bivalent interactions have a 10-fold decrease compared to ESC. Overall, this suggests that ESC transitions from a mostly flexible, open, plastic state to a more specialized configuration when differentiating to NPC. Interestingly, most of the interactions that are classified in ESC are unclassified in NPC, and vice versa, pointing to substantial changes in the overall chromatin 3D conformation and factor occupancy between the two cell types. Considering only interactions that are classified in both cell types, each major 3D-type identified in ESC mainly transitions into the Active or Neuronal-TFs state in NPC. However, most of the interactions that are Inactive in ESC remain Inactive in NPC, which suggests that a subgroup of 3D interactions associated with a repressed transcriptional state in NPC was already present in ESC.

Finally, beyond global changes in structure and factor occupancy, we explored changes in complex functional 3D hubs occurring at specific loci during early stages of neural cell differentiation. The *Zfp608* gene is specifically activated in NPC, concomitantly with the appearance of a novel TAD border at its transcription starting site. CHROMATIC identifies that the gene promoter site switches from a configuration where it is involved in a few PcG-bivalent interactions in ESC, to one with a large number of interactions mainly classified as active and associated with neuronal TFs. Eventually, this structural change might

be driven by the active factors and neuronal TFs, and might allow the gene to scan putative enhancers marked by peaks of H3K27ac.

Our classification is limited by 5 kb resolution that we employed for computational feasibility. Furthermore, the integration of data from different experimental assays, such as chromatin accessibility and DNA methylation assays, would provide a more complete picture of 3D chromatin states. Overall, we consider that CHROMATIC will allow researchers to have a better understanding of the link between chromatin states, genome topology, and gene transcription in the studied cell type.

ACKNOWLEDGMENTS

MAMR acknowledges support by the Spanish Ministerio de Ciencia e Innovación (PID2020-115696RB-I00) and the National Human Genome Research Institute of the National Institutes of Health under Award Number RM1HG011016. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. LDC acknowledges support by the Spanish Ministry of Science and Innovation (PID2019-108322GB-I00) and AGAUR SGR 2017. FM acknowledges Generalitat de Catalunya and the European Social Fund for AGAUR-FI predoctoral fellowship. Finally, CRG acknowledges support from ‘Centro de Excelencia Severo Ochoa 2013-2017’, SEV-2012-0208 and the CERCA Programme/ Generalitat de Catalunya as well as support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and the EMBL partnership, the Generalitat de Catalunya through Departament de Salut and Departament d’Empresa i Coneixement, and the Co-financing with funds from the European

Regional Development Fund (ERDF) by the Spanish Ministry of Science and Innovation corresponding to the Programa Operativo FEDER Plurirregional de España (POPE) 2014-2020 and by the Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement of the Generalitat de Catalunya corresponding to the programa Operatiu FEDER Catalunya 2014-2020.

REFERENCES

- Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res*, *40*(22), 11202-11212. doi:10.1093/nar/gks925
- Aloia, L., Di Stefano, B., Sessa, A., Morey, L., Santanach, A., Gutierrez, A., . . . Di Croce, L. (2014). Zrf1 is required to establish and maintain neural progenitor identity. *Genes Dev*, *28*(2), 182-197. doi:10.1101/gad.228510.113
- Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep*, *9*(1), 9354. doi:10.1038/s41598-019-45839-z
- Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*, *18*(5), 285-298. doi:10.1038/nrm.2017.7
- Beringer, M., Pisano, P., Di Carlo, V., Blanco, E., Chammas, P., Vizan, P., . . . Di Croce, L. (2016). EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell*, *64*(4), 645-658. doi:10.1016/j.molcel.2016.10.018
- Blackledge, N. P., Fursova, N. A., Kelley, J. R., Huseyin, M. K., Feldmann, A., & Klose, R. J. (2020). PRC1 Catalytic Activity Is

- Central to Polycomb System Function. *Mol Cell*, 77(4), 857-874 e859. doi:10.1016/j.molcel.2019.12.001
- Blanco, E., Gonzalez-Ramirez, M., & Di Croce, L. (2021). Productive visualization of high-throughput sequencing data using the SeqCode open portable platform. *Sci Rep*, 11(1), 19545. doi:10.1038/s41598-021-98889-7
- Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., . . . Young, R. A. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, 175(7), 1842-1855 e1816. doi:10.1016/j.cell.2018.10.042
- Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nat Rev Genet*, 17(11), 661-678. doi:10.1038/nrg.2016.112
- Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., . . . Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3), 557-572 e524. doi:10.1016/j.cell.2017.09.043
- Cavalli, G., & Misteli, T. (2013). Functional implications of genome topology. *Nat Struct Mol Biol*, 20(3), 290-299. doi:10.1038/nsmb.2474
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., . . . Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6), 1106-1117. doi:10.1016/j.cell.2008.04.043
- Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., . . . Plath, K. (2017). Cooperative Binding of Transcription Factors

- Orchestrates Reprogramming. *Cell*, 168(3), 442-459 e420.
doi:10.1016/j.cell.2016.12.016
- Cullen, K. E., Kladde, M. P., & Seyfred, M. A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. *Science*, 261(5118), 203-206. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8327891
- Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoiyannopoulos, J. A., & Noble, W. S. (2007). Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11), 1424-1426. doi:10.1093/bioinformatics/btm096
- Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558), 1306-1311. doi:10.1126/science.1067799
- 295/5558/1306 [pii]
- Di Giammartino, D. C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., . . . Apostolou, E. (2019). KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nat Cell Biol*, 21(10), 1179-1190. doi:10.1038/s41556-019-0390-6
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376-380. doi:10.1038/nature11082
- nature11082 [pii]
- Doyle, M. J., & Sussel, L. (2007). Nkx2.2 regulates beta-cell function in the mature islet. *Diabetes*, 56(8), 1999-2007. doi:10.2337/db06-1766

- Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188-230. doi:10.1002/aris.1440380105
- Eagen, K. P., Aiden, E. L., & Kornberg, R. D. (2017). Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc Natl Acad Sci U S A*, 114(33), 8764-8769. doi:10.1073/pnas.1701291114
- Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8), 817-825. doi:10.1038/nbt.1662
- Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3), 215-216. doi:10.1038/nmeth.1906
- Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, 12(12), 2478-2492. doi:10.1038/nprot.2017.124
- Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shores, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345), 43-49. doi:10.1038/nature09906
- Esposito, A., Bianco, S., Chiariello, A. M., Abraham, A., Fiorillo, L., Conte, M., . . . Nicodemi, M. (2022). Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states. *Cell Rep*, 38(13), 110601. doi:10.1016/j.celrep.2022.110601
- Filion, G. J., van Bommel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., . . . van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in

- Drosophila cells. *Cell*, 143(2), 212-224. doi:10.1016/j.cell.2010.09.009
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep*, 15(9), 2038-2049. doi:10.1016/j.celrep.2016.04.085
- Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, 13(12), 840-852. doi:10.1038/nrg3306
- Gagliardi, A., Mullin, N. P., Ying Tan, Z., Colby, D., Kousa, A. I., Halbritter, F., . . . Chambers, I. (2013). A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J*, 32(16), 2231-2247. doi:10.1038/emboj.2013.161
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4), 576-589. doi:10.1016/j.molcel.2010.05.004
- Heurteau, A., Perrois, C., Depierre, D., Fosseprez, O., Humbert, J., Schaak, S., & Cuvier, O. (2020). Insulator-based loops mediate the spreading of H3K27me3 over distant micro-domains repressing euchromatin genes. *Genome Biol*, 21(1), 193. doi:10.1186/s13059-020-02106-z
- Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, 169(1), 13-23. doi:10.1016/j.cell.2017.02.007

- Huang, C., Su, T., Xue, Y., Cheng, C., Lay, F. D., McKee, R. A., . . . Carey, M. (2017). Cbx3 maintains lineage specificity during neural differentiation. *Genes Dev*, 31(3), 241-246. doi:10.1101/gad.292169.116
- Huang, X., & Wang, J. (2014). The extended pluripotency protein interactome and its links to reprogramming. *Curr Opin Genet Dev*, 28, 16-24. doi:10.1016/j.gde.2014.08.003
- Huseyin, M. K., & Klose, R. J. (2021). Live-cell single particle tracking of PRC1 reveals a highly dynamic system with low target site occupancy. *Nat Commun*, 12(1), 887. doi:10.1038/s41467-021-21130-6
- Hwang, W. W., Salinas, R. D., Siu, J. J., Kelley, K. W., Delgado, R. N., Paredes, M. F., . . . Lim, D. A. (2014). Distinct and separable roles for EZH2 in neurogenic astroglia. *Elife*, 3, e02439. doi:10.7554/eLife.02439
- Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., . . . Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9(10), 999-1003. doi:10.1038/nmeth.2148
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., . . . Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314), 430-435. doi:10.1038/nature09380
- Kaneko, S., Son, J., Shen, S. S., Reinberg, D., & Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol*, 20(11), 1258-1264. doi:10.1038/nsmb.2700

- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, 12(6), 996-1006. doi:10.1101/gr.229102
- Kim, S., & Shendure, J. (2019). Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol Cell*, 76(2), 306-319. doi:10.1016/j.molcel.2019.08.010
- Kloet, S. L., Makowski, M. M., Baymaz, H. I., van Voorthuijsen, L., Karemaker, I. D., Santanach, A., . . . Vermeulen, M. (2016). The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat Struct Mol Biol*, 23(7), 682-690. doi:10.1038/nsmb.3248
- Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J. T., Sadreyev, R. I., . . . Kingston, R. E. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell*, 65(3), 432-446 e435. doi:10.1016/j.molcel.2017.01.009
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357-359. doi:10.1038/nmeth.1923
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of

- the human genome. *Science*, 326(5950), 289-293. doi:10.1126/science.1181369
- Liu, Y. R., Laghari, Z. A., Novoa, C. A., Hughes, J., Webster, J. R., Goodwin, P. E., . . . Scotting, P. J. (2014). Sox2 acts as a transcriptional repressor in neural stem cells. *BMC Neurosci*, 15, 95. doi:10.1186/1471-2202-15-95
- Lodato, M. A., Ng, C. W., Wamstad, J. A., Cheng, A. W., Thai, K. K., Fraenkel, E., . . . Boyer, L. A. (2013). SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. *PLoS Genet*, 9(2), e1003288. doi:10.1371/journal.pgen.1003288
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12), 550. doi:10.1186/s13059-014-0550-8
- Lutz, M., Burke, L. J., Barreto, G., Goeman, F., Greb, H., Arnold, R., . . . Renkawitz, R. (2000). Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Res*, 28(8), 1707-1713. doi:10.1093/nar/28.8.1707
- Marco-Sola, S., Sammeth, M., Guigo, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, 9(12), 1185-1188. doi:10.1038/nmeth.2221
- Mas, G., Blanco, E., Ballare, C., Sanso, M., Spill, Y. G., Hu, D., . . . Di Croce, L. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet*, 50(10), 1452-1462. doi:10.1038/s41588-018-0218-5
- McAninch, D., & Thomas, P. (2014). Identification of highly conserved putative developmental enhancers bound by SOX3 in neural

- progenitors using ChIP-Seq. *PLoS ONE*, 9(11), e113361. doi:10.1371/journal.pone.0113361
- mod, E. C., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., . . . Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012), 1787-1797. doi:10.1126/science.1198374
- Morey, L., Aloia, L., Cozzuto, L., Benitah, S. A., & Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. *Cell Rep*, 3(1), 60-69. doi:10.1016/j.celrep.2012.11.026
- Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods*, 13(11), 919-922. doi:10.1038/nmeth.3999
- Nishi, Y., Zhang, X., Jeong, J., Peterson, K. A., Vedenko, A., Bulyk, M. L., . . . McMahon, A. P. (2015). A direct fate exclusion mechanism by Sonic hedgehog-regulated transcriptional repressors. *Development*, 142(19), 3286-3293. doi:10.1242/dev.124636
- Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., & Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *Elife*, 3, e02557. doi:10.7554/eLife.02557
- Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., & Duboule, D. (2011). The dynamic architecture of Hox gene clusters. *Science*, 334(6053), 222-225. doi:10.1126/science.1207194

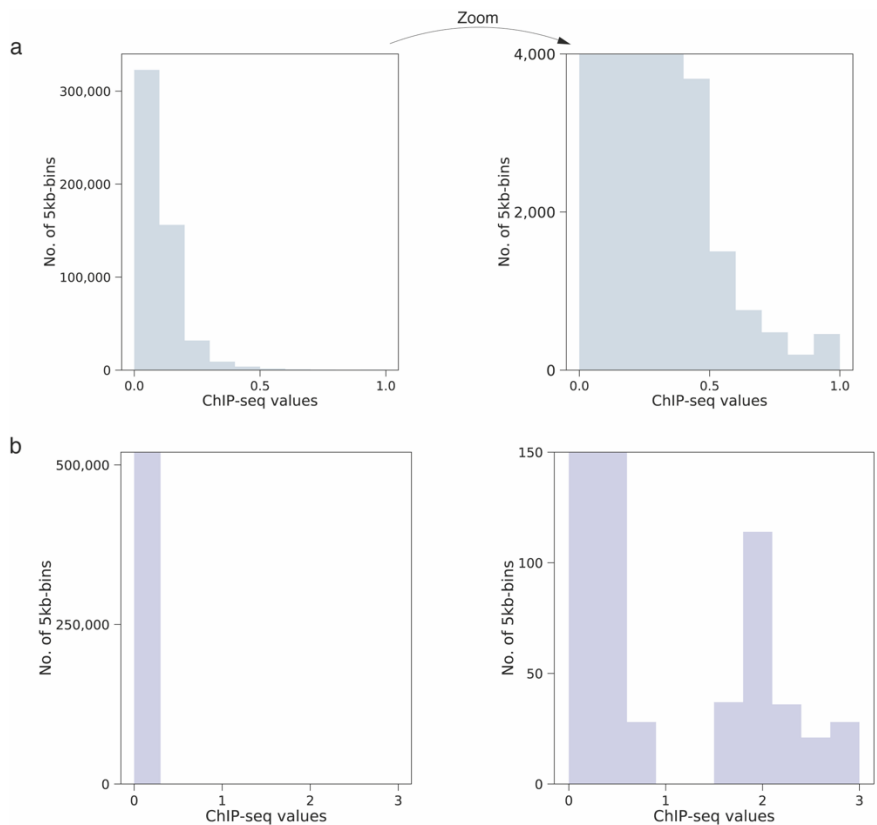
- Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., . . . Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381-385. doi:10.1038/nature11049
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, *44*(D1), D733-745. doi:10.1093/nar/gkv1189
- Ogiyama, Y., Schuettengruber, B., Papadopoulos, G. L., Chang, J. M., & Cavalli, G. (2018). Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Mol Cell*, *71*(1), 73-88 e75. doi:10.1016/j.molcel.2018.05.032
- Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., . . . van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*, *38*(4), 603-613. doi:S1097-2765(10)00321-7 [pii]
10.1016/j.molcel.2010.03.016
- Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., . . . Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281-1295. doi:10.1016/j.cell.2013.04.053
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. doi:10.1093/bioinformatics/btq033

- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665-1680. doi:10.1016/j.cell.2014.11.021
- Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nat Rev Genet*, 19(12), 789-800. doi:10.1038/s41576-018-0060-8
- Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., . . . Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 112(47), E6456-6465. doi:10.1073/pnas.1518552112
- Serra, F., Bau, D., Goodstadt, M., Castillo, D., Filion, G. J., & Marti-Renom, M. A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol*, 13(7), e1005665. doi:10.1371/journal.pcbi.1005665
- Shin, Y., & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science*, 357(6357). doi:10.1126/science.aaf4382
- Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., . . . Graf, T. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet*, 50(2), 238-249. doi:10.1038/s41588-017-0030-7
- Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., . . . Laue, E. D. (2017). 3D structures of individual mammalian

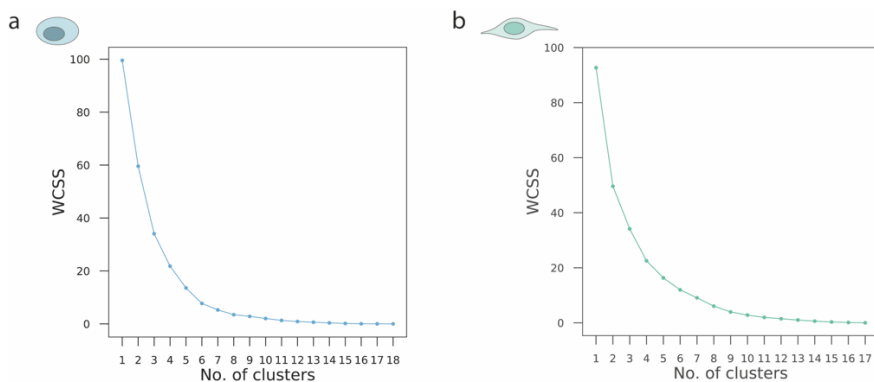
- genomes studied by single-cell Hi-C. *Nature*, 544(7648), 59-64.
doi:10.1038/nature21429
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111.
doi:10.1093/bioinformatics/btp120
- Vidal, E., le Dily, F., Quilez, J., Stadhouders, R., Cuartero, Y., Graf, T., . . . Filion, G. J. (2018). OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res*, 46(8), e49. doi:10.1093/nar/gky064
- Wei, H., Dong, X., You, Y., Hai, B., Duran, R. C., Wu, X., . . . Wu, J. Q. (2021). OLIG2 regulates lncRNAs and its own expression during oligodendrocyte lineage formation. *BMC Biol*, 19(1), 132.
doi:10.1186/s12915-021-01057-6
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., . . . Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2), 307-319.
doi:10.1016/j.cell.2013.03.035
- Wood, H. B., & Episkopou, V. (1999). Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mech Dev*, 86(1-2), 197-201. doi:10.1016/s0925-4773(99)00116-1
- Yeo, J. C., & Ng, H. H. (2013). The transcriptional regulation of pluripotency. *Cell Res*, 23(1), 20-32. doi:10.1038/cr.2012.172
- Yesudhas D, Anwar MA, & S, C. (2019). Structural mechanism of DNA-mediated Nanog–Sox2 cooperative interaction. *The Royal Society of Chemistry*, 9, 8121-8130.

- Ying, Q. L., Stavridis, M., Griffiths, D., Li, M., & Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol*, 21(2), 183-186. doi:10.1038/nbt780
- Zhang, T., Zhang, Z., Dong, Q., Xiong, J., & Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol*, 21(1), 45. doi:10.1186/s13059-020-01957-w
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9), R137. doi:10.1186/gb-2008-9-9-r137
- Zheng, H., & Xie, W. (2019). The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol*. doi:10.1038/s41580-019-0132-4

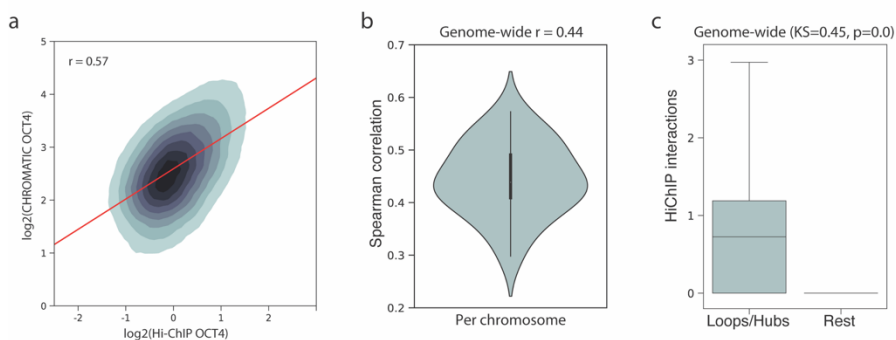
SUPPLEMENTARY FIGURES



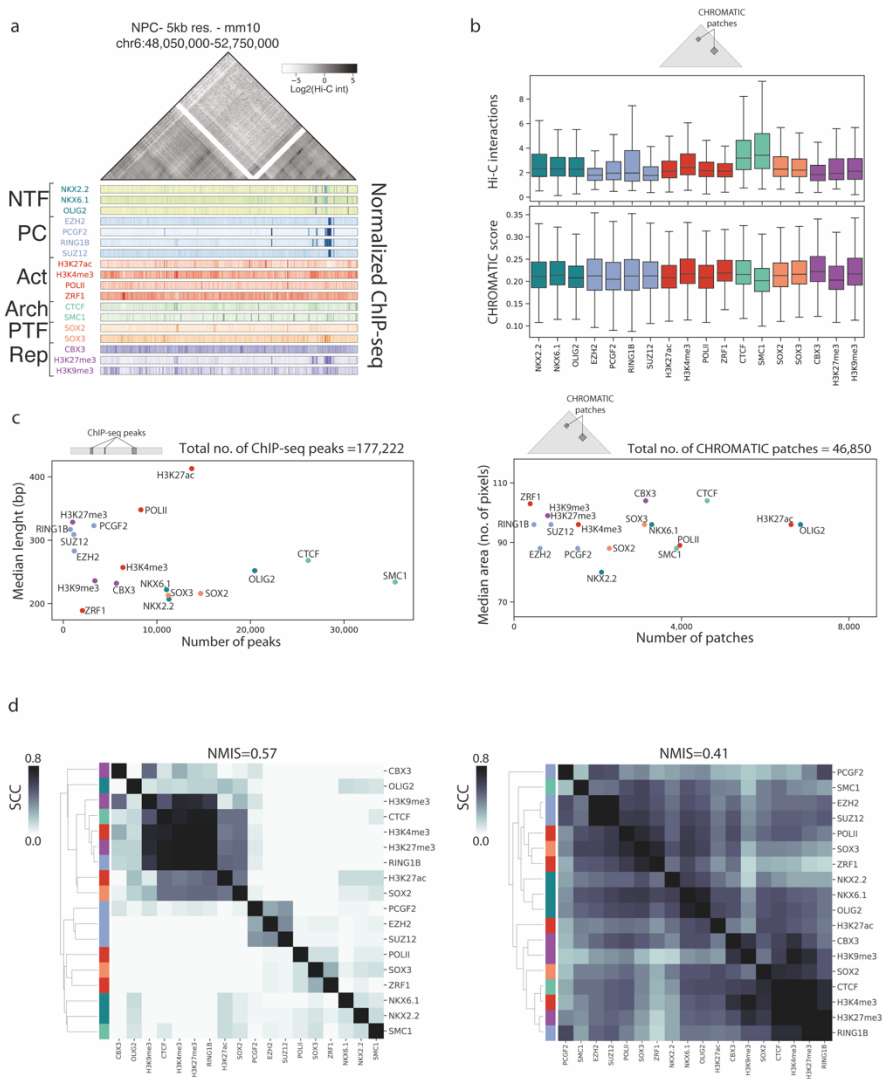
Supplementary Figure 1. Re-scaling of ChIP-seq values. **a** ChIP-seq values before re-scaling. Values are distributed from 0 to 1. **b** ChIP-seq values after re-scaling. Re-scaled values are separated in two groups, one of low ChIP-seq values and one of high ChIP-seq values.



Supplementary Figure 2. Determination of the number of clusters for major types of 3D interactions. **a** K-means algorithm was run multiple times with a different number of clusters, from 1 to 18 in ESC, where 18 3D-types were identified. For each solution, the Within Cluster Sum of Squares (WCSS) is shown. The elbow point appeared in correspondence of 4 clusters of major types of interactions. **b** Same as a, for NPC. K-means algorithm was run multiple times with a different number of clusters, from 1 to 17 in NPC, where 17 3D-types were classified.

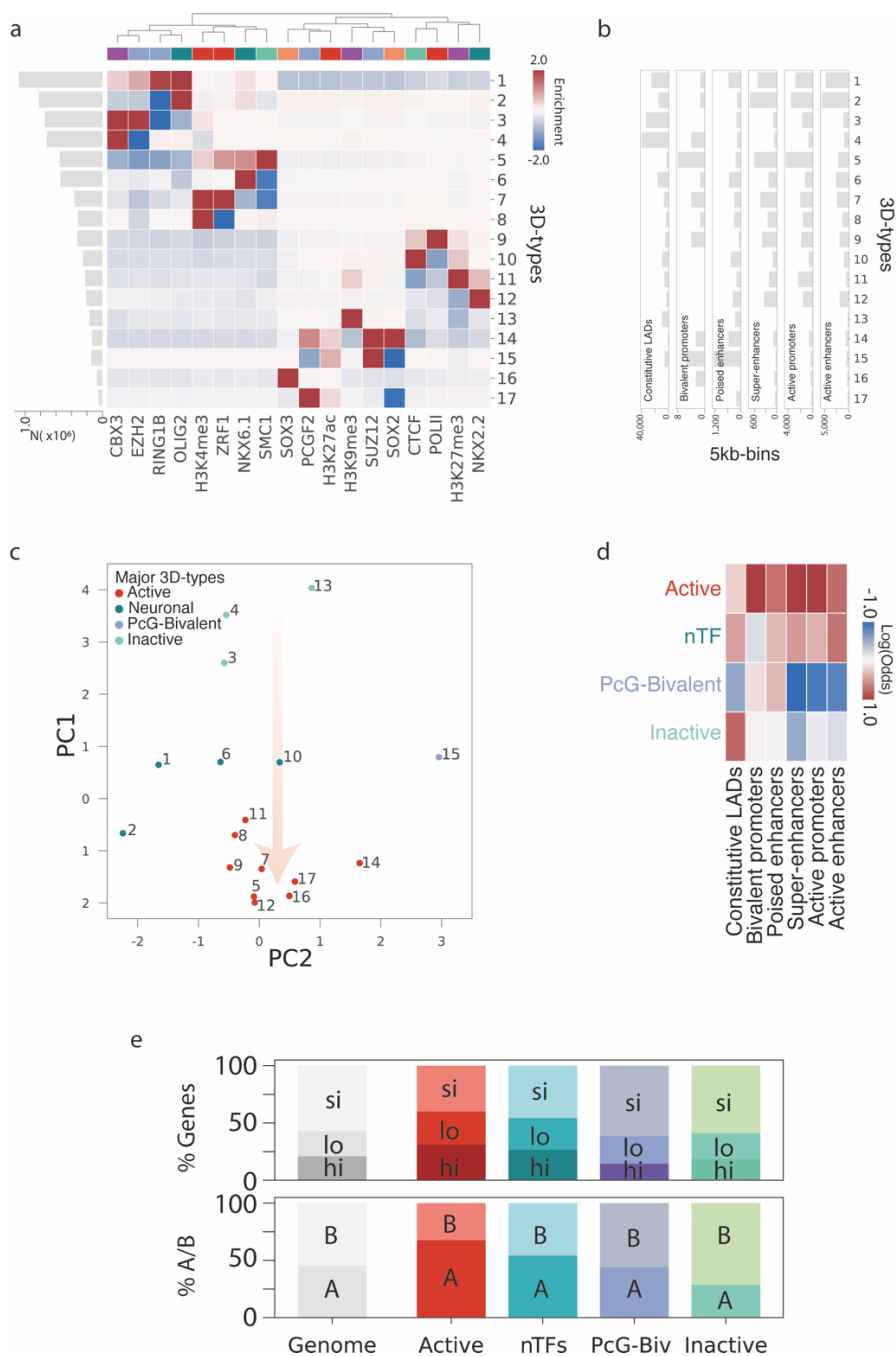


Supplementary Figure 3. Validation with HiChIP data of OCT4. **a** Correlation between CHROMATIC coefficients and HiChIP values for OCT4 in chromosome 6. Spearman correlation coefficient $r = 0.57$ ($p\text{-value} = 0$). **b** Spearman correlation coefficients per chromosome. Genome-wide median $r = 0.44$ genome-wide. **c** Boxplots of HiChIP values of OCT4 from detected CHROMATIC OCT4 patches compared HiChIP interactions elsewhere in the matrix (statistically different distributions as for Kolmogorov-Smirnov statistical test $= 0.45, p\text{-val} = 0$).



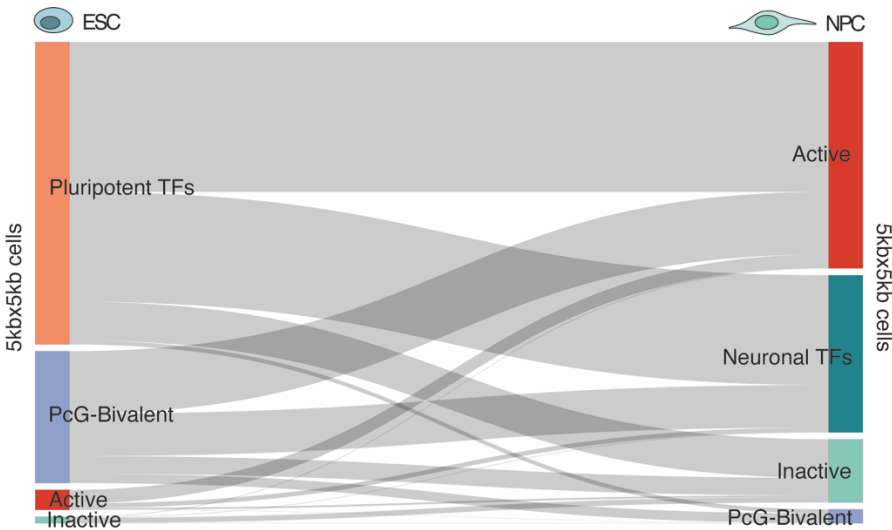
Supplementary Figure 4. CHROMATIC applied to NPC. **a** Example of CHROMATIC applied to Hi-C interaction maps and ChIP-seq profiles at the resolution of 5kb, for 18 factors in NPC. Factors are colored according to their factional role. **b** Top, value distributions of original Hi-C interactions corrected by decay and median filter in NPC, before CHROMATIC processing, in correspondence of the patches detected by CHROMATIC. For each patch, the average of the corresponding Hi-C values is considered. Bottom, CHROMATIC coefficient distributions in NPC, in correspondence of the detected patches. **c** Left, number of ChIP-seq peaks for each factor with respect to their median length (base pairs), in NPC. Right, number of patches detected genome-wide by CHROMATIC for each factor with respect to their median area (number of 5kbX5kb pixels). **d** Unsupervised

hierarchical clustering of factors studied in NPC based on their genome-wide pair-wise correlation, of ChIP-seq tracks on the left and of CHROMATIC maps on the right.

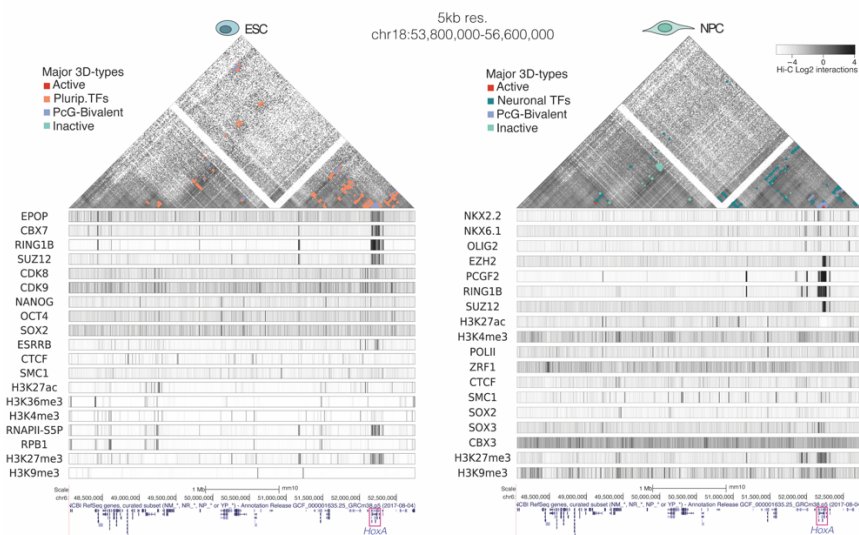


Supplementary Figure 5. 3D-types identified in NPC. **a** Resulting emissions of LSA in NPC defining sets of types of 3D interactions (3D-types) in terms of enrichment (in red) or depletion (in blue) of factors. Factors are colored according to their functional role as in **Suppl. Fig. 4**. Left, bar plot indicates the number of

5kbX5kb pixels associated to each 3D-type. **b** Overlap in number of 5kb-bins between 1D loci corresponding to each 3D-type and chromatin types, in NPC. **c** LogOdds of the overlap between 1D loci corresponding to each 3D-type and the functional genomic features in **b** was used as input for principal component analysis. Plots depict the values of principal components 1 and 2 (PC1, PC2) for the different 3D-types, which were further classified by K-means unsupervised clustering, in NPC. The arrow shows the direction from inactive to active for the identified 3D-types. **d** LogOdds of the overlap between 1D loci corresponding to each major 3D-type and the functional genomic features in ESC. **e** Percentage of silent (si), lowly-expressed (lo) and highly-expressed (hi) genes, and percentage of A and B compartments for the whole genome (in gray) and for the 4 major 3D-types, in ESC.



Supplementary Figure 6. Changes in 3D interaction types during mouse neural development. Sankey plot describing the transitions between the different types of 3D interactions, between ESC (left) and NPC (right). Here, “Unclassified” cells are excluded.



Supplementary Figure 7. Changes in complex functional 3D hubs during mouse neural development at the *HoxA* locus. Interactions classified into major 3D-types, in ESC (left) and NPC (right), in the *HoxA* locus in chromosome 6. Hi-C maps (top) and ChIP-seq tracks (below) are in grey. Major 3D-types classified by CHROMATIC are in colors. The *HoxA* gene cluster is highlighted. In ESC, *HoxA* cluster is not expressed and is kept in a bivalent domain by Polycomb proteins (in blue). In NPC, a small active domain appears, in agreement with the already known activation of a portion of *HoxA* genes.

SUPPLEMENTARY TABLES

Supplementary Table 1.

Raw counts and mapped statistics of RNA-seq experiments.

Sample	Raw reads	Mapped reads
ESC rep1	48463796	46557499 (96%)
ESC rep2	40148534	38336872 (95%)
NPC rep1	41686101	40318090 (97%)
NPC rep2	50350011	48441325 (96%)

Supplementary Table 2.

Hi-C experimental statistics for merged replicates of mESCs.

Filtered artifacts	
Duplicated	104,577,419
Too short	63,891,817
Error	26,144,108
Extra dangling-end	323,042,784
Too large	19,470
Dangling-end	163,593,222
Over-represented	31,102,973
Too close from RES	497,700,422
Self-circle	2,116,071
Random breaks	4,220,714

Valid reads		
Total	Valid	% valid
7,260,480,082	1,537,751,681	21.18

Supplementary Table 3.

Hi-C experimental statistics for merged replicates of NPCs.

Filtered artifacts

Duplicated	2,906,553,357
Too short	191,364,911
Error	30,930,472
Extra dangling-end	626,655,991
Too large	78,448
Dangling-end	301,911,787
Over-represented	98,650,960
Too close from RES	1,563,662,772
Self-circle	4,702,795
Random breaks	20,843,082

Valid reads

Total	Valid	% valid
8,677,570,910	3,974,901,849	45.81

Supplementary Table 4.

Hi-ChIP experimental statistics for merged replicates of mESCs.

SMC1a

Filtered artifacts

Duplicated	142,222,792
Too short	19,258,610
Error	590,117
Extra dangling-end	111,269,935
Too large	4,460
Dangling-end	92,950,614
Over-represented	11,885,254
Too close from RES	80,263,176
Self-circle	3,015,462
Random breaks	3,398,952

Valid reads

Total	Valid	% valid
585,071,818	219,998,058	37.6

OCT4

Filtered artifacts

Duplicated
Too short
Error
Extra dangling-end
Too large
Dangling-end
Over-represented
Too close from RES
Self-circle
Random breaks

189,316,364
20,572,571
314,290
106,603,963
4,670
69,916,554
6,648,950
87,065,455
2,104,720
2,175,642

Valid reads

Total	Valid	% valid
657,261,466	252,920,123	38.48

CONCLUSIONS

This thesis is focused on the characterization of the role of chromatin-associated factors in genome topology, which in turn is important for proper spatiotemporal regulation of gene expression and cell fate decisions.

In Chapter 1, we studied the transcriptional and architectural consequences of histone H1 variants depletion in human breast cancer cells. From this chapter, we can specifically conclude that:

1. Despite the small changes in H1 variants distribution, knock-down of H1 translated into more isolated but de-compacted chromatin structures at the scale of Topologically Associating Domains (TADs).
2. Such changes in TAD structure correlated with a coordinated gene expression response of their resident genes.

In Chapter 2, we presented CHROMATIC, a novel and generalized computational method that integrates chromatin interactions and factor occupancy data with genome structural data to reveal the contribution of chromatin-associated factors to genome topology.

From the first part of this chapter, dedicated to the description of the computational tool and of its utility, we can specifically conclude that:

1. CHROMATIC integrates Hi-C and ChIP-seq data in a single map of *in silico* HiChIP, representing chromatin interactions associated to any factor of interest.
2. CHROMATIC interactions correlate with HiChIP data, while being much faster and less expensive than real HiChIP experiments.
3. By deconvolving the Hi-C data into factor-specific interactions otherwise hidden by background levels, CHROMATIC allows to discern the role of each studied factor in the global genome structure and to better identify factors participating in genome architecture in a cell-type specific manner.
4. Compared to the analysis of data mapped exclusively on linear chromatin (1D) such as ChIP-seq, CHROMATIC output is more informative of the functional role performed by factors in the nucleus.
5. The study of 3D co-localization patterns of factors allows to identify types of functional 3D interactions, which may reflect already known interactions between different chromatin factors, or may help discover new associations between molecules with specific functional roles. Such types of 3D interactions can be regarded as 3D chromatin states and represent a functional annotation of chromatin 3D interactions.

From the second part of Chapter 2, dedicated to the application of CHROMATIC to embryonic stem cells (ESCs) and neural progenitor cells (NPCs) data, we can specifically conclude that:

1. ES cells transition from a plastic state to a more specialized one when differentiating to NPCs,

2. Stem cell differentiation involves substantial changes in chromatin 3D conformation and factor occupancy, even though a subgroup of NPC interactions associated to an inactive state are already established in ESCs.

ANNEX 1

In vivo temporal resolution of acute promyelocytic leukemia progression reveals a role of *Klf4* in suppressing early leukemic transformation

Candidate's contribution: Analysis of the Hi-C experiments.

Mas G, Santoro F, Blanco E, Gamarra Figueroa GP, Le Dily F, Frigè G, Vidal E, Mugianesi F, Ballaré C, Gutierrez A, Sparavier A, Marti-Renom MA, Minucci S, Di Croce L. *In vivo* temporal resolution of acute promyelocytic leukemia progression reveals a role of *Klf4* in suppressing early leukemic transformation. *Genes Dev.* 2022 Apr 1;36(7-8):451-467. doi: 10.1101/gad.349115.121. Epub 2022 Apr 21. PMID: 35450883.

In vivo temporal resolution of acute promyelocytic leukemia progression reveals a role of *Klf4* in suppressing early leukemic transformation

Glòria Mas,^{1,7} Fabio Santoro,^{2,3} Enrique Blanco,¹ Gianni Paolo Gamarra Figueroa,¹ François Le Dily,¹ Gianmaria Frige,^{2,3} Enrique Vidal,¹ Francesca Mugianesi,^{1,4} Cecilia Ballaré,¹ Arantxa Gutierrez,¹ Aleksandra Sparavier,^{1,4} Marc A. Marti-Renom,^{1,4,5,6} Saverio Minucci,^{2,3} and Luciano Di Croce^{1,5,6}

¹Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain;

²Department of Experimental Oncology, European Institute of Oncology (IEO), Milan 20139, Italy; ³Department of Oncology and Hemato-oncology, University of Milan, Milan 20139, Italy; ⁴Centro Nacional de Análisis Genómico (CNAG), Centre for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology, Barcelona 08028, Spain; ⁵Universitat Pompeu Fabra (UPF), Barcelona, Spain; ⁶Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain

Genome organization plays a pivotal role in transcription, but how transcription factors (TFs) rewire the structure of the genome to initiate and maintain the programs that lead to oncogenic transformation remains poorly understood. Acute promyelocytic leukemia (APL) is a fatal subtype of leukemia driven by a chromosomal translocation between the promyelocytic leukemia (PML) and retinoic acid receptor α (RARA) genes. We used primary hematopoietic stem and progenitor cells (HSPCs) and leukemic blasts that express the fusion protein PML-RARA as a paradigm to temporally dissect the dynamic changes in the epigenome, transcriptome, and genome architecture induced during oncogenic transformation. We found that PML-RARA initiates a continuum of topologic alterations, including switches from A to B compartments, transcriptional repression, loss of active histone marks, and gain of repressive histone marks. Our multiomics-integrated analysis identifies *Klf4* as an early down-regulated gene in PML-RARA-driven leukemogenesis. Furthermore, we characterized the dynamic alterations in the *Klf4* cis-regulatory network during APL progression and demonstrated that ectopic *Klf4* overexpression can suppress self-renewal and reverse the differentiation block induced by PML-RARA. Our study provides a comprehensive in vivo temporal dissection of the epigenomic and topological reprogramming induced by an oncogenic TF and illustrates how topological architecture can be used to identify new drivers of malignant transformation.

[**Keywords:** chromatin; chromatin topology; gene regulation; leukemia]

Supplemental material is available for this article.

Received October 15, 2021; revised version accepted March 25, 2022.

The 3D organization of the genome, ranging from nucleosomes to heterochromatin/euchromatin compartments and chromosome territories, provides a fundamental mechanism for genome regulation (Schoenfelder and Fraser 2019; Zheng and Xie 2019). Transcriptional regulatory elements, including enhancers and promoters, are in physical contact to fine-tune the timing and magnitude of gene expression, and perturbation of this contact can profoundly affect cell identity, differentiation, and tumorigenesis (Gröschel et al. 2014; Northcott et al. 2014; Lupiáñez et al. 2015; Flavahan et al. 2016; Hnisz et al. 2016; Akdemir

et al. 2020). Epigenetic changes often drive the initiation, maintenance, and progression of cancer, and their reversibility and plasticity make them attractive targets in the clinical field (Dawson 2017). However, little is known about how the genome structure is rewired during the acquisition of oncogenic features, or whether structural changes are functionally linked to epigenome and transcriptome alterations during oncogenic transformation.

Acute promyelocytic leukemias (APLs) represent 10%–15% of acute myeloid leukemias (AMLs) and are characterized by the presence of the t(15;17) chromosomal

⁷Present address: University of Miami Miller School of Medicine, Miami, FL 33136, USA.

Corresponding authors: luciano.dicroce@crg.eu, gxm578@miami.edu, saverio.minucci@ieo.it

Article published online ahead of print. Article and publication date are online at <http://www.genesdev.org/cgi/doi/10.1101/gad.349115.121>.

© 2022 Mas et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genesdev.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Mas et al.

translocation between *PML* and *RAR α* (de Thé et al. 1990; Goddard et al. 1991). Expression of the oncofusion protein PML-RAR α in hematopoietic stem/progenitor cells (HSPCs) results in a differentiation block at the promyelocytic stage and in malignant transformation, as these cells are able to recapitulate most clinical and morphological features of human APL in transplantation mouse models (Brown et al. 1997; Grisolan et al. 1997; He et al. 1997; Grignani et al. 2000; Westervelt et al. 2003; Guibal et al. 2009; Wojiski et al. 2009). Mechanistic studies using APL cell lines and primary blasts have shown that the PML-RAR α oncofusion protein competes with normal RAR α functions and recruits histone deacetylases (HDACs), the NuRD chromatin remodeling complex, and Polycomb-repressive complexes (PRCs) to constitutively repress target genes of the TFs RAR α and PU.1 (Pandolfi 2001; Villa et al. 2007; Morey et al. 2008; Martens et al. 2010; Wang et al. 2010; Mas and Di Croce 2016). These epigenetic complexes mediate long-range interactions to instruct gene expression programs during development and in tumor cells (Denholtz et al. 2013; Schoenfelder et al. 2015; Mas et al. 2018; Oksuz et al. 2018; Basu et al. 2020). In addition to its repressive functions, PML-RAR α binds superenhancer regions to directly transactivate genes that encode key myeloid-determining TFs or enzymes, including *GFI1*, *MPO*, *WT1*, and *MYC* (Tan et al. 2021). PML-RAR α also disrupts PML nuclear bodies, which are structures involved in the control of cell cycle, apoptosis, senescence, DNA damage, and antiviral immunity (Bernardi and Pandolfi 2007; Scherer and Stamminger 2016; Chang et al. 2018). Induction of DNA damage by PML-RAR α results in increased mutability, favoring the occurrence of cooperating secondary mutations and development of full-blown leukemia (di Masi et al. 2016; Voisset et al. 2018). Recent reports using APL cell lines and primary blasts showed that PML-RAR α mediates the formation of long-range interactions to repress the expression of genes controlling myeloid differentiation and maturation (Li et al. 2018; Wang et al. 2020) and activate the *GFI1* superenhancer (Tan et al. 2021). However, these studies did not provide a dynamic perspective of how PML-RAR α remodels the genome to impair the function and differentiation of normal primary HPSCs to generate fully transformed leukemic blasts. Here, we used the PML-RAR α model system as a paradigm to temporally dissect the dynamics of epigenomic and transcriptomic reprogramming occurring at the onset, during progression, and in full-blown APL leukemias in animal models that faithfully recapitulate human APL clinical features. Our global profiling identified the *Klf4* locus as one of the most extensively reorganized genes during PML-RAR α -driven APL progression. *Klf4* encodes a TF with important roles in myeloid differentiation (Feinberg et al. 2007; Park et al. 2016, 2019a). *Klf4* expression has been shown to be lower in samples from AML patients than in those from healthy individuals (Faber et al. 2013b; Morris et al. 2016). However, the function of *Klf4* in APL has remained controversial, with a few studies reporting that *Klf4* overexpression induces differentiation using the APL cell line HL-60 (Feinberg et al. 2007; Alder et al.

2008; Morris et al. 2016), and others showing that *Klf4* expression supports cell growth and survival of the APL cell line NB4 (Lewis et al. 2021). Using our integrative multiomics analysis, we temporally resolved the genomic alterations induced by PML-RAR α and showed that the *Klf4* locus undergoes extensive reprogramming of enhancer-promoter interactions, transcriptional down-regulation, and gain of repressive histone modifications. We further showed that ectopic overexpression of *Klf4* partially restored the phenotypic defects induced by the expression of PML-RAR α . This work provides a dynamic model of the genomic reprogramming triggered by an oncogenic TF in vivo and highlights the use of topological information to identify new drivers of malignant transformation.

Results

PML-RAR α induces a progressive reorganization of genome architecture

To dissect the dynamic changes induced by PML-RAR α in genome architecture and transcription during leukemia progression, we infected primary mouse bone marrow hematopoietic stem/progenitor cells (lineage negative [Lin⁻]) with lentiviruses carrying an empty vector control or a Flag-tagged human PML-RAR α and harvested cells at different stages of APL transformation (Fig. 1A). Stage 0 and stage I corresponded to sorted GFP⁺ cells transformed with empty vector or PML-RAR α -3xFlag vector, respectively (Supplemental Fig. S1A,B). We followed the progressive transformation of cells carrying PML-RAR α by culturing them in semisolid media and harvesting after 2 or 4 wk of serial replating (equivalent to stage II or III, respectively). The final stage of APL transformation (stage IV) corresponded to blasts isolated from mice that were transplanted with cells carrying PML-RAR α ; mice developed leukemia after ~6 mo (Fig. 1A). We verified that cells expressing PML-RAR α -3xFlag showed increased serial replating capacity, impaired differentiation, and promyelocytic morphology, as compared with cells expressing empty vector control (Supplemental Fig. S1C,D). We then used multiple biological replicates of cells from stages 0–IV to generate in situ Hi-C libraries (Rao et al. 2014), RNA-seq libraries, and ChIP-seq libraries in order to comprehensively characterize the genome architecture, the transcriptome, and the epigenome, respectively, during the process of leukemic transformation.

We obtained high-quality maps of the 3D genome organization across all stages (Supplemental Fig. S1E), which allowed us to examine the segregation of active (A, gene-rich) and inactive (B, gene-poor) compartments (Lieberman-Aiden et al. 2009; Imakaev et al. 2012). Principal component analysis (PCA) of the eigenvectors of all autosomes revealed that PML-RAR α induced genome-wide, cumulative changes in A/B compartments (Fig. 1B,C). Overall, 7.1% of the genome changed compartment at some point during APL transformation, with 3% of the genome stably switching from the A to B compartment, and 1.1% switching from B to A (Fig. 1D,E). A greater proportion of switching events from stage 0 to III occurred from the A to B

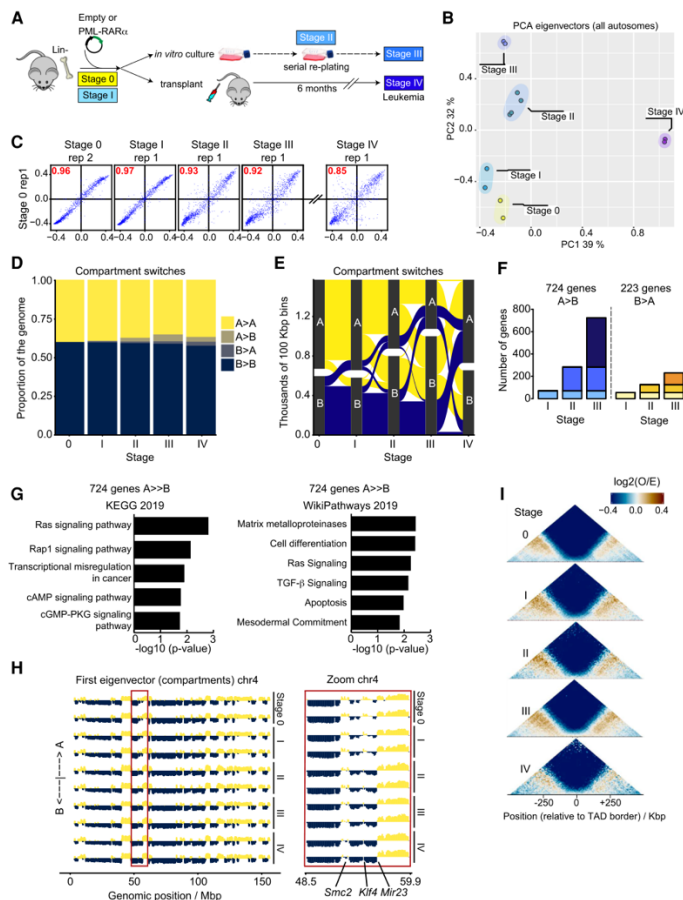


Figure 1. PML-RARα induces a continuum of A-to-B compartment switches. (A) Experimental strategy and definition of stages. Lineage-negative cells (Lin⁻) were isolated from bone marrow of adult mice (6 to 8 wk old) and infected with lentiviruses carrying an empty vector (Empty) or a vector containing a Flag-tagged PML-RARα fusion gene (PML-RARα-3xFlag). Successfully infected Lin⁻ cells (GFP⁺) were sorted and corresponded to stage 0 (carrying empty vector) or stage I (carrying PML-RARα-3xFlag vector). Stage I cells were then cultured in methylcellulose media and harvested at the second replating (for stage II) or fourth replating (for stage III). Stage I cells were also transplanted into lethally irradiated recipient mice. Blast cells were harvested from bone marrow of mice developing leukemias, at ~6 mo after transplant (stage IV). (B) Principal component analysis (PCA) based on Hi-C eigenvectors for all autosomes. (C) Scatter plots of first eigenvectors along the time course for chromosome 7. The Y-axis represents the first eigenvector associated with replicate 1 of stage 0, and the X-axis represents the first eigenvector of the different stages (replicate 2 for stage 0, and replicate 1 for the remaining stages). Pearson correlations are highlighted in red. (D) Proportion of the genome that changed compartment during the time course. We assumed that a region was A (or B) if all replicates at the same stage were flagged as A (or B). Regions not consistent between replicates were considered ambiguous and represented 2% of the genome. About 5% of the genome was excluded due to low mappability. (E) Alluvial plot showing the dynamic A-to-B compartment switching of bins during the time course. Stages are represented along the X-axis, and the genomic size is represented on the Y-axis as well as by the width of the ribbons. Bins that did not switch compartments or that were flagged as “ambiguous” at any point were excluded. (F) Stacked bar plots of the number of genes in bins that were (1) in compartment A at stage 0 and switched to compartment B at another stage (left), or (2) in compartment B at stage 0 and switched to compartment A at another stage (right). Only genes that switched compartments from one stage to the next and were stably maintained in the new compartment were considered (i.e., genes in bins that switched compartments more than once during the time course were excluded). (G) KEGG analysis and WikiPathways analysis of 724 genes that switched from A to B compartments. (H) Example of A-to-B compartment switching of chromosome 4. The left panel shows the first eigenvector (compartments) chr4 along the genomic position in megabases (X-axis). Each row corresponds to one independent biological replicate of the indicated stage. A compartments are depicted in yellow, and B compartments are shown in blue. The right panel corresponds to an 11.4-Mb zoomed region of chromosome 4 that contains the *Klf4* locus. (I) Aggregate genome-wide contact profiles centered on TAD borders defined in stage 0. Data are the log₂ ratio of observed and expected contacts in 10-kb bins, pooling biological replicates.

Mas et al.

compartment (Fig. 1D,E; Supplemental Fig. S1F); this is in agreement with the known role of PML-RAR α as a transcription repressor (Di Croce et al. 2002; Segalla et al. 2003; Villa et al. 2004, 2006, 2007; Carbone et al. 2006; Morey et al. 2008; Saumet et al. 2009; Martens et al. 2010; Subramanyam et al. 2010; Saeed et al. 2011, 2012; Cole et al. 2016). We found a cumulative total of 724 genes in bins that stably switched from compartment A to B, and 223 genes in bins that switched from B to A (Fig. 1F). The gene set that switched from A to B was enriched for genes associated with MAPK signaling (including those encoding Ras, Rap1, and cAMP), immune signaling via TGF- β , cellular differentiation, apoptosis, and transcriptional misregulation in cancer, as shown by KEGG analysis (Fig. 1G). In addition, this A-to-B gene set was significantly enriched for SMAD4 targets, as shown by ChEA analysis (adjusted $P=0.00045$) and Polycomb targets (enriched in H3K27me₃, adjusted $P=0.00034$); specific genes included *Mef2c*, *Flt3*, *Hmga2*, *Maf*, *Pax7*, *Met*, *Igf1*, *Wnt16*, *Aff1*, *Ptk2*, *Runx2*, *Rel*, and *Prom1*. Of note, the A-to-B gene set also included several genes with known roles in leukemia development at compartment boundaries, such as *Klf4*, *Setbp1*, *Efl1*, and *Hhip* (Fig. 1H; Supplemental Fig. S1H; Alder et al. 2008; Kobune et al. 2012; Faber et al. 2013a; Schoenhals et al. 2013; Huang et al. 2014; Filarsky et al. 2016; Morris et al. 2016; Seipel et al. 2016; Makishima 2017; Park et al. 2019b; Tan et al. 2019). In contrast, the set of genes that switched from the B to A compartment was enriched for immune system processes, as shown by KEGG pathway analysis (Supplemental Fig. S1G); this included *Il33*, *Il9r*, *Klf5*, and *Mcm10*. When examining Hi-C interactions with intra-TAD regions, we observed minimal changes in TAD border strength (Fig. 1I). Overall, our data showed that PML-RAR α expression induced a dynamic reorganization of the genome, affecting a large set of actively transcribed regions of the genome and causing their interaction patterns to switch toward those in the inactive chromatin compartment.

PML-RAR α promotes dynamic changes in gene expression that are linked to changes in genome topology

Our in situ Hi-C data indicated that PML-RAR α reorganized long-range interactions in a cumulative manner across the genome and was potentially accompanied by dynamic transcription alterations. To confirm this hypothesis, we performed RNA-seq on independent biological replicates harvested in duplicate at all stages. Based on PCA of the RNA-seq data sets, we observed a trajectory of transcriptome alterations concurrent with PML-RAR α expression; of note, full-blown leukemias (stage IV) showed extensive transcriptome reprogramming as compared with the other stages (Supplemental Fig. S2A). We used two differential gene expression analyses to identify (1) genes significantly deregulated during APL transformation with respect to stage 0 (control) cells, and (2) genes uniquely deregulated (i.e., excluding genes that were also deregulated at other stages) at each stage of APL progression as compared with stage 0, which identified genes that are “transiently” altered during the kinetic analysis (Supplemental Table S1). The

first analysis revealed an increasing number of significantly deregulated genes (Q -value <0.05) from stage 0 during leukemic transformation (Fig. 2A). In addition, these differentially regulated genes progressively increased or decreased in expression along the time course (Fig. 2B), suggesting that PML-RAR α expression induced early alterations in expression (e.g., at stage I) that were maintained during APL transformation. The second analysis revealed that a relatively small subset of genes was uniquely up-regulated or down-regulated at early stages of APL transformation, and that a larger number of genes was transcriptionally deregulated specifically at stages III and IV (Supplemental Fig. S2B; Supplemental Table S1). These results put forward the hypothesis that early expression of PML-RAR α impaired the expression of a relatively few genes encoding for key hematopoietic TFs, which subsequently altered the transcriptional landscape genome-wide. Indeed, several genes encoding TFs or enzymes were either significantly up-regulated (e.g., *Bcl2*, *Bmp2*, *Hes1*, *Mycn*, *Twist1*, and *Id2*) or down-regulated (e.g., *Cdh1*, *Lef1*, *Rarb*, and *Rarg*) at stages I and II. Overall, more genes were found to be down-regulated than up-regulated (Fig. 2A; Supplemental Fig. S2B), in line with previous reports of PML-RAR α driving transcriptional repression (Morey et al. 2008; Gaillard et al. 2015; Li et al. 2018; Wang et al. 2020).

We next performed gene ontology and KEGG pathway analyses to dissect the pathways perturbed by PML-RAR α . Genes with an increased expression in early stages showed enrichment in pathways related to MAPK signaling, regulation of cell proliferation and/or adhesion, or pathways in cancer. In turn, genes with a reduced expression in early stages were mostly related to hematopoietic cell lineage or immune response (Supplemental Table S1; Supplemental Fig. S2C). GSEAs of genes during APL progression revealed increased expression of genes involved in pathways related to cell cycle (E2F targets, G2M checkpoint, and mitotic spindle) or DNA repair, and decreased expression of genes involved in apoptosis and immune signaling pathways (Fig. 2C). In the leukemic stage (IV), a large number of genes was deregulated with respect to their status in stage 0 (Supplemental Fig. S2D); however, a large proportion of these genes already showed altered expression at stage III (Supplemental Fig. S2E). Included in the top transcriptionally deregulated genes were genes that encode TFs or enzymes that play fundamental roles in myeloid differentiation (Rosenbauer and Tenen 2007) and HSPC function, including *Gata2*, *Cebpa*, *Bcl2*, *Hoxa10*, *Irf8*, *Myc*, *Spi1*, and *Klf4* (Fig. 2D). These results were validated in independent biological samples using qRT-PCR (Supplemental Fig. S2F).

Overall, our Hi-C and transcriptomic data indicated that cells expressing PML-RAR α undergo progressive and profound alterations in genome architecture that may be correlated to changes in gene transcription. To confirm this hypothesis, we examined the transcriptional status of genes located in bins that switch compartments during APL transformation. Indeed, expression of genes in regions that switched from the A compartment at stage 0 to the B compartment at any later stage was significantly down-regulated (Fig. 2E). Although not statistically significant,

Temporal dissection of leukemia transformation

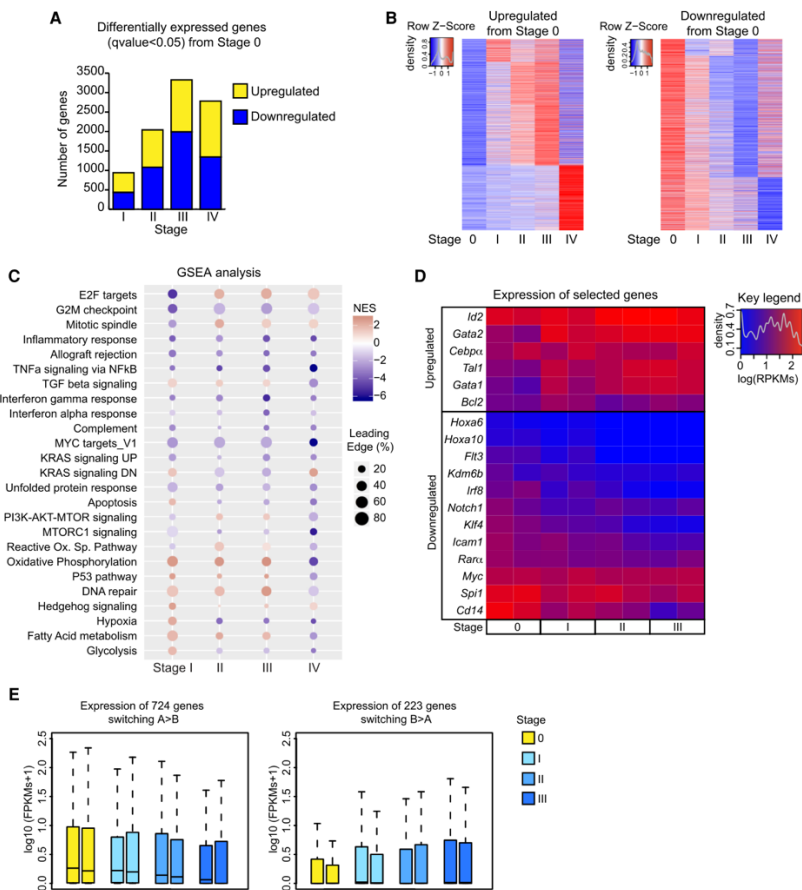


Figure 2. PML-RARA promotes dynamic changes in expression of gene pathways that control cell cycle progression, immune signaling, and DNA repair. (A) Number of differentially expressed genes (DEGs) obtained by DESeq2 (Q-value > 0.05) at each of the indicated stages, using stage 0 as baseline. (B) Heat maps showing unsupervised clustering of expression levels (Z-score values) of genes at each stage, using stage 0 as reference. The left heat map depicts all genes that were up-regulated with respect to stage 0, and the right heat map depicts all down-regulated genes. (C) GSEA signatures of DEGs at each indicated stage with respect to stage 0. Gene expression signatures related to cell cycle control and p53/DNA damage repair were positively enriched during the time course, while signatures related to immune signaling were negatively enriched. (D) Heat map depicting the dynamic gene expression alterations (log RPKM values) of key hematopoietic transcription factors and leukemia-associated genes at each stage. (E, left panel) Expression of genes in regions that were in compartment A at stage 0 and switched to compartment B at another stage (724 genes). $P = 2 \times 10^{-4}$ between stages 0 and II; $P = 1.1 \times 10^{-8}$ between stages 0 and III. (Right panel) Same as the left panel, but for genes in the B compartment at stage 0 that switched to the A compartment at another stage (223 genes). $P = 0.79$ between stages 0 and II; $P = 0.11$ between stages 0 and III. Genes in bins that switched compartments more than once during the kinetic assay were excluded. P-values were computed using the Wilcoxon test (two-sided).

the opposite trend was observed for genes that switched from the B to A compartment. Genes that stably switched from one compartment to the other were frequently found

at pre-existing boundaries between A and B compartments at stage 0 (22.28% of the total genes are located at ± 100 kb from A/B boundaries, whereas this proportion increases to

Mas et al.

66.58% and 66.97% for stable A-to-B and stable B-to-A genes, respectively; P -value $< 2.2 \times 10^{-16}$). Together, our data indicated that PML-RAR α induced early chromatin topological alterations, and in particular switched the interaction patterns of active regions of the genome to inactive chromatin compartments, which correlated with transcription repression.

PML-RAR α induces epigenomic alterations at enhancers, which correlate with changes in expression of nearby genes

PML-RAR α induces important expression changes of key genes involved in hematopoietic stem cell function and differentiation (Fig. 2D; Supplemental Fig. S2F; Tan et al. 2021). Given that transcriptional output is controlled by the activity of distal regulatory enhancers, we hypothesized that PML-RAR α influences transcription of these genes by modulating enhancer activation. To comprehensively examine alterations of enhancer activity during APL progression, we collected samples at all experimental stages and performed ChIP-seq to map the genome-wide distribution of active enhancers (H3K4me1 and H3K27ac), active promoters (H3K4me3 and H3K27ac), and Polycomb-mediated repression (H3K27me3). These experiments revealed interesting patterns in both the number of peaks and their genomic distribution (Supplemental Fig. S3A,B). First, while the global number of H3K4me1-enriched regions was very similar between stages, the number of regions enriched in H3K4me3 and H3K27ac—hallmarks of active promoters—substantially decreased during APL progression (stages III and IV) (Supplemental Fig. S3A). Second, regions decorated by H3K27me3 increased along the four stages (Supplemental Fig. S3A), suggesting that PML-RAR α led to a cumulative repression of the epigenome. Third, the reduced H3K27ac and increased H3K27me3 levels mostly occurred outside promoters of coding genes (i.e., in intergenic and intragenic regions), suggesting that PML-RAR α had a primary role in epigenetic repression of putative enhancers (Supplemental Fig. S3B). Following these results, we next mapped the dynamic loss of enhancer activity during leukemic transformation (Fig. 3A). We identified 27,341 active enhancers at stage 0, of which 5%, 17%, 20%, and 21% lost H3K27ac at stages I, II, III, and IV, respectively (Fig. 3A; Supplemental Fig. S3C). We observed a striking progressive reduction of H3K27ac levels at enhancers with reduced levels in one stage during the subsequent stages. For example, enhancers with reduced H3K27ac levels at stage II continued to lose H3K27ac levels at stages III and IV (Fig. 3A). Importantly, loss of H3K27ac was accompanied by a gain in the repressive histone mark H3K27me3 (Fig. 3B). Interestingly, motif analysis of sequences of these enhancers revealed significant hits for the PU.1/SPI1 and the myeloid-determining transcription factor GFI1B (Supplemental Fig. S3D). These data align with previous literature (Wang et al. 2010; Tan et al. 2021) and confirm the role of the PML-RAR α -PU.1 and PML-RAR α -GFI1B axes during APL progression. Moreover, the KLF4 motif was enriched at enhancers that are inactivated during leukemia progression, suggesting that KLF4 down-regulation might be one of the

key events that could induce decommissioning of enhancers at a later time point, although some of the observed changes might be indirect. Together, our results indicate that PML-RAR α induced a vast reprogramming of the epigenome that involved the repression of active enhancers concomitant with a gain of Polycomb-mediated repression.

To closely examine the dynamic alterations of the epigenome during leukemic transformation, we next subtracted the normalized signal intensity of H3K27ac and H3K27me3 at each stage from the baseline signal at stage 0; we then inspected the regulatory landscape near key hematopoietic transcription factors. We observed that PML-RAR α expression induced a progressive loss of H3K27ac at enhancers near *Klf4* and *Spi1*, which became transcriptionally repressed during APL transformation (Figs. 2D, 3C). In contrast, *Gata2* and its putative enhancers showed progressively increased H3K27ac and reduced H3K27me3 levels (Fig. 3C), in line with its increased expression (Fig. 2D). These examples suggested that the epigenetic alterations induced by PML-RAR α at enhancers were associated with changes in nearby gene expression. To address this question genome-wide, we examined the levels of expression of genes located within 5 kb from enhancers that presented a significant decrease in the H3K27ac levels at each stage (Supplemental Fig. S3C). Our data confirmed that loss of enhancer activity correlated with a significant decrease in expression of nearby genes (Fig. 3D).

Next, we used Hi-C to examine whether the overall physical contacts within the same TAD (topologically associating domain; i.e., contacts with other promoters or enhancers) were affected in promoters that lost or gained H3K27ac during APL transformation. Notably, promoters that had decreased levels of H3K27ac—and thus had become repressed—showed decreased contacts during the early phases of leukemic transformation (Fig. 3E). In contrast, activated promoters with increased H3K27ac levels showed the opposite trend, whereas the contacts of stably active or inactive promoters were maintained (Fig. 3E). Examples of intra-TAD reorganizations for a repressed gene (*KLF4*) and an activated gene (*GATA2*) are shown in Figure 4 and Supplemental Figure S4, respectively. These results suggest that changes in intra-TAD interactions may be required for transcriptional activation but not for transcriptional deactivation. To generalize those observations, we ranked TADs according to their changes in domain score, which reflect internal reorganization and compartmentalization of TADs (Krijger et al. 2016; Stadhouers et al. 2018) between stage 0 and stage III. The 10% of TADs with a higher increase in domain score at stage III showed increased levels of H3K27ac (Fig. 3F, left panel), reflecting intra-TAD reorganization and establishment of regulatory contacts in TADs that become active, confirming previous observations (Krijger et al. 2016; Stadhouers et al. 2018). Changes in domain score did not correlate with changes in H3K27me3 levels, and both TADs with a higher increase or decrease in gene expression showed an increased domain score at stage III, indicating a link between TAD reorganization and gene expression changes and suggesting complex reorganization of TADs upon PML-RAR α expression. Collectively, our data indicated that PML-RAR α

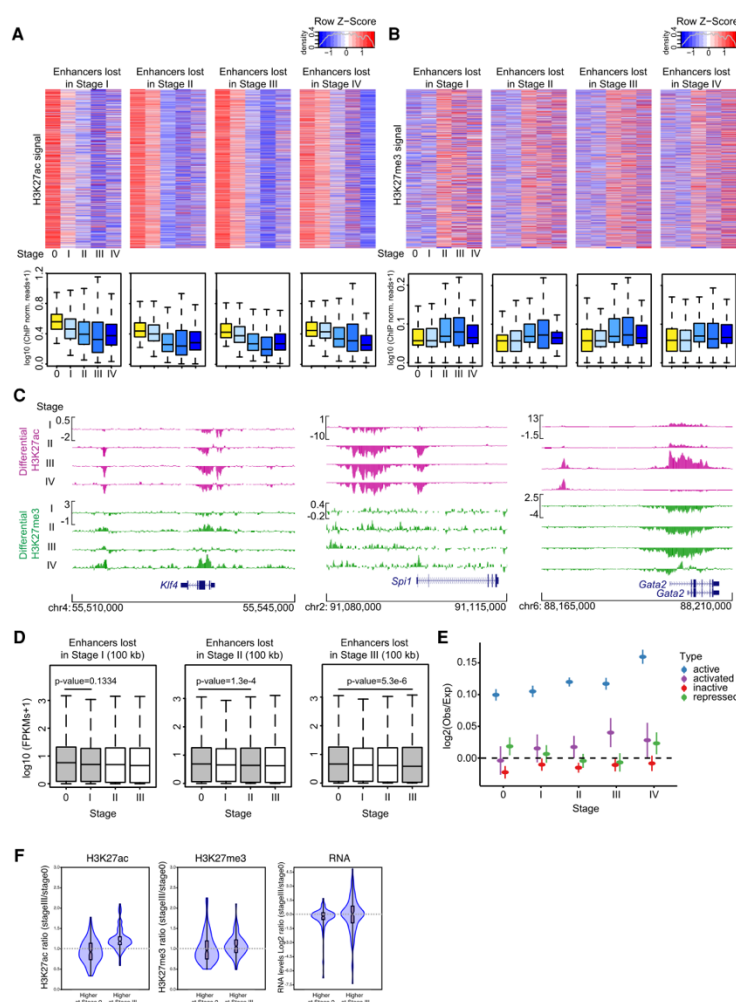


Figure 3. PML-RAR α induces epigenomic alterations at distal regulatory elements correlated with changes in expression of nearby genes. (A) Intensity of H3K27ac signal at 27,341 stage 0 active enhancers (non-TSS regions with overlapping H3K4me1 and H3K27ac peaks) that lost H3K27ac during the time course. Box plots below correspond to the normalized H3K27ac ChIP-seq signal intensity of enhancers lost at each stage. For H3K27ac, $P = 2.2 \times 10^{-16}$ between stages 0 and I, between stages 0 and II, between stages 0 and III, between stages 0 and IV. For H3K27me3, $P = 0.33$ between stages 0 and I, $P = 2.2 \times 10^{-16}$ between stages 0 and II, $P = 2.2 \times 10^{-16}$ between stages 0 and III, and $P = 2.2 \times 10^{-16}$ between stages 0 and IV. (B) Intensity of H3K27me3 signal at the same enhancers shown in A. Box plots show the normalized H3K27me3 ChIP-seq signal intensity of enhancers lost at each stage. (C) UCSC genome browser snapshots of differential H3K27ac (purple) and H3K27me3 (green) ChIP-seq profiles at promoters and putative distal regulatory elements of the indicated genes. Each row corresponds to the ChIP-seq signal intensity at each indicated stage subtracted from the signal at stage 0 as baseline. (D) Expression of genes within 5 kb around active enhancers at stage 0 that are lost in stage I (1464 enhancers; left graph), stage II (4668 enhancers; middle graph), or stage III (5406 enhancers; right graph). P-values were computed using Wilcoxon test (two-sided). (E) Dynamic changes of overall contact enrichment (intra-TAD) of promoters depending on activation status from stage 0 to stage III are as follows: Active (blue dots) maintained H3K27ac in both stages, inactive (red dots) were not marked by H3K27ac in either stage, gain (purple dots) gained H3K27ac at stage III, and loss (green dots) lost H3K27ac at stage III. Contact enrichments were measured as \log_2 of observed contacts over expected ($\log_2(\text{Obs}/\text{Exp})$) and were corrected against background. (F) Box plots showing the changes in H3K27ac, H3K27me3, and RNA levels per TAD for the top and bottom 10% of TADs with higher changes in domain score.

Mas et al.

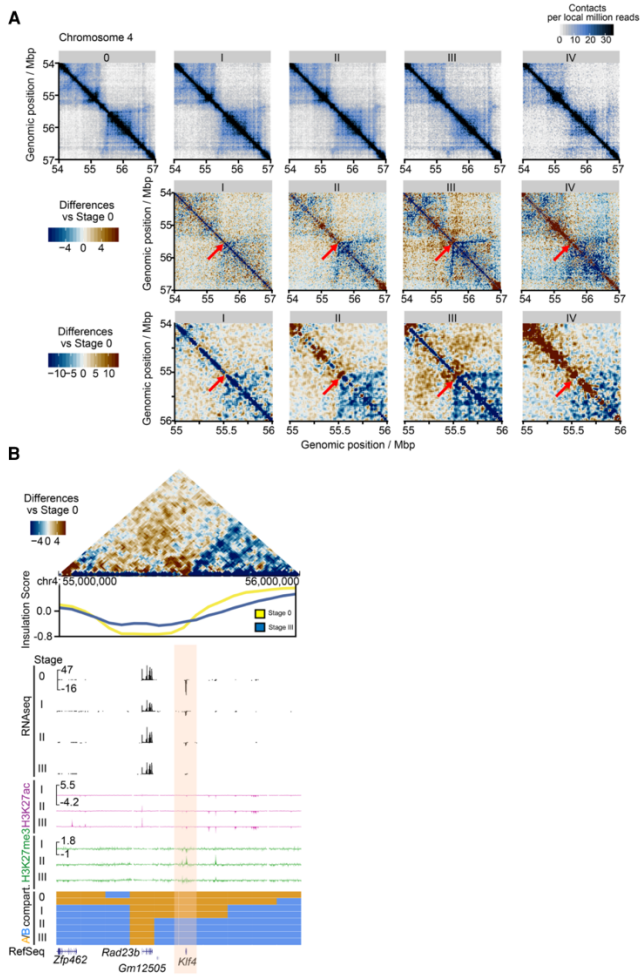


Figure 4. The *Klf4* locus undergoes progressive changes in long-range interactions driven by PML-RARA expression. (*A*, top panels) Hi-C interaction matrices showing normalized interaction counts at the region between 54 and 57 Mb of chromosome 4. Each stage is indicated by the top gray bar. (*Middle panels*) Hi-C interaction matrices following subtraction of the signal at stage 0. Blue depicts interactions that decreased after stage 0, and brown depicts interactions that increased after stage 0. Red arrows indicate the location of the *Klf4* locus. (*Bottom panels*) Same as the middle panels, but zooming in at the region between 55 and 56 Mb of chromosome 4, depicting a progressive loss of interactions of *Klf4* with downstream genomic elements. (*B*) Differential matrix at 5-kb resolution, showing normalized interaction counts at the region between 55 and 56 Mb of chromosome 4 at stage III after subtraction of stage 0 signal. Blue depicts interactions that decreased at stage III, and brown depicts interactions that increased at stage III. The insulation score track is shown below for stages 0 and III. RNA-seq tracks (black) show a progressive decrease of *Klf4* gene expression. Differential H3K27ac (purple) and H3K27me3 (green) ChIP-seq tracks show a sequential loss of H3K27ac signal and gain of H3K27me3. A/B compartment bins showed progressive compartment switching of the *Klf4* locus from A (orange) to B (blue). RefSeq genes of this genomic region are shown at the bottom.

promoted extensive epigenomic reprogramming of enhancer regions and induced transcriptional changes by re-wiring promoter–promoter and promoter–enhancer contacts, thus affecting genes that encode for critical regulators of hematopoietic differentiation.

The Klf4 genomic locus undergoes progressive rearrangement of long-range interactions during PML-RARA-induced transformation

KLF4 is a master hematopoietic transcription factor that acts as a tumor suppressor in leukemia by activating the expression of genes that promote myeloid differentiation,

apoptosis, and cell cycle arrest (Feinberg et al. 2007; Alder et al. 2008; Huang et al. 2014; Filarsky et al. 2016; Morris et al. 2016). Our data showed that the *Klf4* locus underwent an extensive regulatory reprogramming during APL transformation with an A-to-B compartment switch (Fig. 1H) and transcriptional and epigenomic repression (Figs. 2D, 3C; Supplemental Fig. S2F). To investigate whether this reprogramming was accompanied by alterations in long-range interactions, we inspected the temporal changes in interactions centered around the *Klf4* gene. The *Klf4* locus is located at the boundary between two well-defined TADs (Fig. 4A). During APL transformation, we observed a progressive loss of long-range interactions of the *Klf4* locus

with the downstream TAD, with multiple interaction loops profoundly decreased at stage III as compared with stage 0 (Fig. 4A, middle Hi-C map). Simultaneously, the insulation between the two TADs (Crane et al. 2015) decreased during the kinetic (Fig. 4B), leading to increased interactions with the upstream TAD linked to the change of compartment of the *Klf4* locus. Interestingly, such changes in long-range interactions were accompanied by epigenetic and transcriptional reprogramming, as shown by decreases in H3K27ac levels and gene expression toward the downstream TAD (Fig. 4B).

Similar topological alterations were observed in the *Etv1* locus, a gene that is recurrently transposed in acute leukemia (Sacchi et al. 1986) and that was also repressed during APL progression (Supplemental Fig. S4A, left panel). The *Etv1* gene progressively lost contacts, H3K27ac signal strength, and transcriptional output, concomitant with a gain of H3K27me3 levels. We also inspected the pattern of interactions around the *Gata2* locus as an example of a gene encoding a master regulator of myeloid differentiation that is up-regulated in APL (Fig. 2D; Zhang et al. 2008a; Li et al. 2018). The *Gata2* gene showed conspicuously increased contacts with neighboring genes in a region of ~0.5 Mb (Supplemental Fig. S4A, right panel). In addition, we observed an enrichment in H3K27ac levels and transcriptional output at the *Gata2* locus during APL transformation (Supplemental Fig. S4A, right panel).

Given the remarkable topological rearrangements observed at the *Klf4* locus, we next sought to identify potential *cis*-regulatory elements that interacted with the *Klf4* locus and to examine their contact profiles during APL transformation. To this end, we generated virtual 4C-seq maps at stage 0 and stage III that were centered at the *Klf4* locus (Fig. 5A). These maps revealed that, in normal hematopoietic stem/progenitor cells, the *Klf4* promoter had strong interactions with potential enhancers located at ~119, 198, and 274 kb upstream of the promoter (Fig. 5A). These interactions were markedly reduced at stage III of the time course, while interactions downstream from the *Klf4* promoter tended to increase. Notably, the *Klf4* putative enhancers identified at stage 0 were enriched in H3K4me1 and H3K27ac in normal cells, and these marks were reduced in stage III. In addition, the region spanning the +119-kb enhancer showed a conspicuous increase in the levels of the Polycomb-repressive mark H3K27me3 (Fig. 5A; Di Carlo et al. 2019). The virtual 4C-seq map around the *Etv1* locus also confirmed that transcriptional repression was accompanied by loss of contacts between the *Etv1* promoter and its downstream enhancers, which decreased their activity from stage 0 to stage III, as shown by the loss of active histone modifications (Supplemental Fig. S4B). In contrast, the *Gata2* locus (which is up-regulated during APL progression) showed a marked gain in interactions both upstream of and downstream from the gene, including at *Gata2* putative enhancers (Supplemental Fig. S4C). Furthermore, we found a marked decrease in the Polycomb-mediated repressive mark H3K27me3 around *Gata2*. Altogether, our data showed that PML-RARα expression induced exten-

sive rearrangements in long-range interactivity at loci encoding for master hematopoietic transcription factors.

Klf4 overexpression inhibits self-renewal and promotes differentiation of PML-RARα-expressing cells

Our data indicated that PML-RARα progressively down-regulated *Klf4* expression by remodeling long-range interactions at the *Klf4* locus. Both tumor suppressor and oncogenic roles have been reported for *Klf4* in the context of APL (Feinberg et al. 2007; Alder et al. 2008; Morris et al. 2016; Lewis et al. 2021). *Klf4* expression appears to be lower in APL patient samples carrying the t(15;17) translocation as compared with other AML subtypes or healthy bone marrow (Supplemental Fig. S5A). To determine whether down-regulation of this TF contributes to leukemic phenotypes, we examined whether ectopic *Klf4* expression was able to reverse the phenotypic alterations driven by PML-RARα. To this end, we generated lineage-negative (Lin[−]) cells expressing PML-RARα only, *Klf4* only, or both using a retroviral strategy (Supplemental Fig. S5B). Expression of PML-RARα arrested cellular differentiation, as shown by the decreased frequency of CD11b⁺ cells and the increased frequency of cKit⁺ cells, as compared with an empty vector control (Supplemental Fig. S5C). *Klf4* overexpression alone did not significantly change the frequency of cKit⁺ or CD11b⁺ cells as compared with empty vector control (Supplemental Fig. S5C). However, ectopic *Klf4* expression in cells simultaneously expressing PML-RARα substantially increased differentiation (Fig. 5B; Supplemental Fig. S5C). In addition, the self-renewal capacity of PML-RARα-expressing cells was completely abrogated when *Klf4* was overexpressed (Fig. 5C; Supplemental Fig. S5D,E). These results were further supported by RNA-seq analysis of cells overexpressing KLF4 in the absence and presence of PML-RARα (Supplemental Fig. S6A,B). By comparing the transcriptome of cells co-overexpressing PML-RARα and KLF4 with the one from cells overexpressing PML-RARα only, we observed alterations in different processes that are essential for cell growth and leukocyte function. Interestingly, KLF4 expression in PML-RARα cells results in up-regulation of cellular senescence programs (Supplemental Fig. S6C). Together, these results suggested that *Klf4* down-regulation, which is induced by PML-RARα, is indeed a leukemia-promoting event that can be reversed by ectopic *Klf4* expression.

Discussion

We have shown that the expression of the chimeric protein PML-RARα in primary HPSCs induces a rapid and extensive remodeling of contacts genome-wide as well as reprogramming of both the epigenome and the transcriptome. We showed that this process is, at least partially, dynamic and continuous, impacting transcription and the enhancer landscape around genes that encode for key transcription factors, which control the differentiation and function of HPSCs. Among these alterations, we identified major changes and transcriptional repression at the *Klf4* gene and neighboring enhancers and showed that

Mas et al.

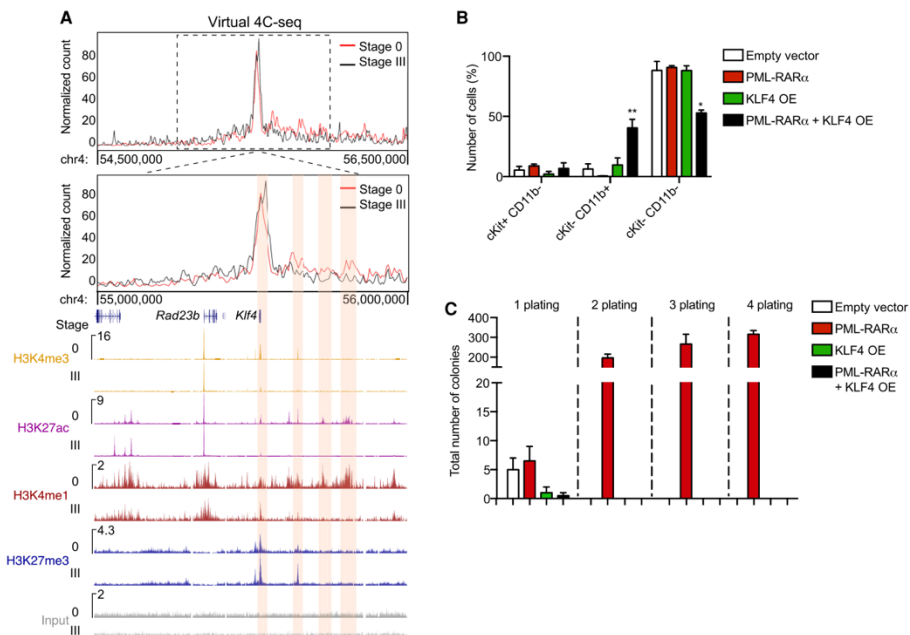


Figure 5. *Klf4* overexpression restores the normal function of PML-RAR α -expressing HSCs. (A) Virtual 4C-seq signal around the *Klf4* genomic region. The top panel illustrates overall contact profiles at the gene promoter at stage 0 (red line) and stage III (black line) for the region between 54.5 and 56.5 Mb of chromosome 4. The zoomed-in panel corresponds to the region between 55 and 56 Mb of chromosome 4. ChIP-seq tracks of the indicated histone modifications at stage 0 and stage III are shown. Shaded regions highlight the regions spanning the *Klf4* gene and the putative enhancers upstream of the *Klf4* promoter (+119 kb, +198 kb, and +274 kb from the promoter). (B) Immunophenotyping analyses of lineage-negative bone marrow cells overexpressing KLF4-GFP, PML-RAR α -hCD4, or both (KLF4 OE + PML-RAR α). Cells were sorted and analyzed by flow cytometry using the indicated cell surface markers. *P*-values were calculated using a Student's *t*-test between PML-RAR α and KLF4 OE + PML-RAR α conditions. (***) *P* = 0.016, (*) *P* = 0.003. (C) Quantification of colony-forming units (CFUs) during four consecutive replatings of sorted cells overexpressing KLF4-GFP, PML-RAR α -hCD4, or both (KLF4 OE + PML-RAR α).

ectopic overexpression of *Klf4* restored the differentiation capacity of HPSCs that expressed PML-RAR α . Our findings add to recent studies addressing the role of PML-RAR α (Li et al. 2018; Wang et al. 2020; Tan et al. 2021) and of other oncoproteins, such as RUNX1-ETO (Ptasinska et al. 2019), in genome architecture by (1) using a primary cellular and animal model system that closely recapitulates clinical and morphological features of human APL, (2) providing the first temporal multiomics analysis of the alterations driven by the chimeric protein at both promoter and enhancer regions, and (3) identifying specific changes that occur at *Klf4* putative enhancers and demonstrating a tumor suppressor role of this transcription factor in promyelocyte leukemic transformation. While our experimental system has been validated extensively and is known to maintain two key functional aspects of PML-RAR α function (inhibiting differentiation and enhancing self-renewal), it is important to acknowledge that using in vitro cultured cells to characterize early

stages of PML-RAR α function may not provide the complete view of the molecular events or genetic mutations required to develop leukemogenesis in vivo.

A few studies have mapped PML-RAR α occupancy genome-wide using human APL cell lines or APL blasts (Hoemme et al. 2008; Martens et al. 2010; Mikesch et al. 2010; Wang et al. 2010, 2020; Singh et al. 2018; Tan et al. 2021). Despite multiple attempts, we were unable to map the fusion protein in primary mouse HSPCs by ChIP-seq; however, our temporal analysis allowed us to identify early (and potentially direct) alterations in the topology, epigenome, and transcriptome driven by PML-RAR α expression. Such earlier changes are more likely driven by the direct effect of the fusion protein and its primary targets, while stage IV alterations might reflect the occurrence of additional genetic alterations and other potentially cooperative effects that promote the fully transformed leukemic phenotype. Among the early alterations, we focused on those occurring at the *Klf4* locus,

which we characterized in depth. Furthermore, we cross-validated our RNA-seq data by overlapping our differentially expressed gene lists with known PML-RAR α target genes in the NB4 APL cell line (Tan et al. 2021), and with known target genes of the TF PU.1 in HSCs (Pundhir et al. 2018), which are reported to be coregulated by PML-RAR α (Wang et al. 2010; Yang et al. 2014). These analyses showed that ~30% of differentially expressed genes between stages 0 and II, and between stages 0 and III, are bona fide PML-RAR targets in the patient-derived NB4 cells. We also found that >52% or 60% of up-regulated or down-regulated genes, respectively, were PU.1 targets in HSCs. In addition, genes down-regulated during the kinetic analysis showed significant enrichment in PRC2 (Suz12) as well as SMAD4 targets (adjusted *P*-values 1×2.5^{-8} and 1×1.05^{-6} , respectively), confirming previous reports (Lin et al. 2004; Villa et al. 2007; Morey et al. 2008) and further validating our analyses. Interestingly, up-regulated genes were enriched not only for Suz12 and SMAD4 targets (adjusted *P*-values 10×4.88^{-8} and 10×9.4^{-3} , respectively), but also for Gata2 targets (adjusted *P*-value 10×4.88^{-8}). Expression of *Gata2* increased very early during the kinetic analysis (Fig. 2D; Supplemental Fig. S4A,C) and in human APL patient samples (Sukhai et al. 2008; Katerndahl et al. 2021). Notably, *Gata2* up-regulation was recently reported to suppress PML-RAR α -induced leukemic transformation, indicating that, in addition to *Klf4*, *Gata2* modulation might contribute to suppressing malignant transformation (Katerndahl et al. 2021).

The temporal resolution of our data revealed that the previously reported repressive functions of PML-RAR α occurred in a progressive manner as cells underwent transformation (Figs. 2B,D, 3A). We found that the most pronounced alterations in long-range interactions, the epigenome, and the transcriptome occurred from stage III to stage IV (full-blown leukemia). These observations integrated other studies (Gaillard et al. 2015) that have shown that the initial changes in gene expression driven by PML-RAR α are relatively subtle and related to metabolism, cell cycle, and DNA damage response signatures (Fig. 2C), which may be insufficient to terminally arrest differentiation. This model is further reinforced by DNA methylation analyses that report only modest epigenome alterations at early stages of APL development (Schoofs et al. 2013; Gaillard et al. 2015). It is important to note that stage IV leukemic blasts were isolated from live animals and thus were likely to be influenced by microenvironmental cues in the bone marrow. Furthermore, the development of full-blown APL blasts requires secondary mutations, which could further contribute to the divergence in our stage IV samples. Future studies are warranted to identify the contribution of secondary lesions and the bone marrow microenvironment in the cellular phenotypes observed during later stages of APL development.

Here, we focused on uncovering early alterations occurring at regulatory enhancers encoding for key hematopoietic TFs, including *Klf4*, that can subsequently have major impacts in the transcriptome, genome architecture, and methylome of APL blasts. The role of *Klf4* in hematopoietic malignancies has remained controversial. Our study sheds

new light on this issue by demonstrating a tumor suppressor role of *Klf4* in the context of APL, showing that the gene is progressively down-regulated during APL progression and that its ectopic overexpression counteracts oncogenic transformation. Given that *Klf4* is required for mesoderm lineage commitment (Aksoy et al. 2014), we speculate that *Klf4* down-regulation rewires gene regulatory networks that promote HSPC differentiation, thus contributing to leukemogenesis. We found that the *cis*-regulatory landscape within the *Klf4* locus substantially changed its pattern of long-range interactions and histone modifications concomitant with a reduction in *Klf4* expression. The increase in long-range interactions observed around the *Klf4* locus could be triggered by the switch from the A to B compartment: As *Klf4* is progressively embedded into a larger B compartment, B-to-B interactions might be facilitated. Although treatment of APL with all-*trans* retinoic acid in combination with chemotherapy results in remission in >90% of patients, our data suggest that ATRA combined with enhanced *KLF4* expression may open a novel avenue of therapeutic intervention. Indeed, we have observed synergistic effects of ATRA in inducing apoptosis, enhanced G1-phase arrest, and differentiation of cells coexpressing PML-RAR α and *KLF4* (data not shown), confirming the tumor suppressor role of *KLF4* overexpression and its molecular effects in the context of APL.

Our work delineates the dynamic mechanisms whereby the oncogenic TF PML-RAR α builds a network of chromosome interactions that repress transcription of master hematopoietic regulators. We propose that the dynamic changes in the genome architecture mediated by PML-RAR α may serve as a general paradigm for other oncogenic proteins that act as transcriptional repressors, bringing new light to the molecular mechanisms by which these transcriptional repressors drive malignant transformation, and possibly leading to the identification of novel transformative therapeutic strategies.

Materials and methods

Murine APL model, bone marrow harvest, and cell culture

Bone marrow lineage-negative hematopoietic stem/progenitor cells from 8- to 10-wk-old female 129SvEv mice were harvested and infected with high-titer retroviruses expressing either an empty PINCO-3xFlag vector or a PINCO-PML-RAR α -3xFlag vector carrying human PML-RAR α . PINCO plasmids expressing human PML-RAR α from the 5' viral long terminal repeat (LTR) and GFP from an internal promoter (cytomegalovirus [CMV]) were described previously (Grignani et al. 1998; Minucci et al. 2002) and were modified by cloning three copies of a Flag tag (69 bp) at the C-terminal of the human PML-RAR α sequence. GFP⁺ cells transformed with empty vector or PML-RAR α -3xFlag vector were sorted by FACS and correspond to stage 0 and stage I, respectively. Stage I cells were then plated in methylcellulose supplemented with cytokines and stem cell factor and serially replated for 2 wk (stage II) and 4 wk (stage III). The GFP⁺ cells that were passaged on methylcellulose were not resorted at each passage. In parallel, ~1 million GFP⁺ PML-RAR α -3xFlag transduced lineage-negative cells (stage I) were transplanted via tail vein injection into lethally irradiated (9 Gy) syngeneic mice (129SvEv) as previously described (Minucci et al. 2002). The animals were

Mas et al.

monitored periodically for signs of disease and the presence of blasts as evaluated by complete blood counts (CBC) and peripheral blood smears. Leukemic mice were humanely euthanized, and leukemic blasts were isolated from the spleen (with >95% of leukemic cell infiltration) for subsequent experiments (stage IV).

Bone marrow lineage-negative cells were obtained and transduced as described previously (Minucci et al. 2002). Serial replating assays of GFP⁺ cells were performed by seeding 10,000 cells/well in methylcellulose medium (Methocult, Stem Cell Technologies M3434) and replating every 7 d. Flow cytometry analyses were performed by staining cells with antimouse CD11b (eBioscience 25-0112-82) and using BD FACSCalibur 2.

Animal handling was performed following Italian laws (D.L. vo 116/92 and subsequent additions), which enforce EU Council Directive 86/609/EEC of November 24, 1986, on the approximation of laws, regulations, and administrative provisions of the Member States regarding the protection of animals used for scientific purposes. Mice were housed according to guidelines from the Commission Recommendation 2007/526/EC, June 18, 2007. The protocol was approved by the Italian Ministry of Health (authorization October 2013).

Western blotting

Whole-cell lysates of 293T cells infected with empty PINCO-3xFlag vector, PINCO-PML-RAR α (Minucci et al. 2002), or PINCO-PML-RAR α -3xFlag were obtained using RIPA buffer containing protease inhibitors (Roche). Sixty micrograms of total protein was loaded per lane on an 8% SDS-PAGE. After blocking in 5% milk-TBST-1X, the following antibodies were incubated overnight at 4°C: anti-Flag (mouse monoclonal, 1:500; Sigma F1804) and anti-Tubulin (mouse monoclonal, 1:5000; Abcam ab7291). Immunodetection was performed using ECL.

In situ Hi-C experiments

For in situ Hi-C experiments, 5 million to 10 million cells of each stage were harvested at two independent biological replicates per stage. Cells were cross-linked for 10 min at room temperature with 1% formaldehyde and quenched during a 5-min incubation at room temperature with 125 mM glycine, followed by a 15-min incubation on ice and two washes with cold PBS; samples were then pelleted and frozen at -80°C .

In situ Hi-C libraries were generated as previously described (Rao et al. 2014) with minor modifications (Mas et al. 2018). Two biological replicates were sequenced for all stages, with one additional technical replicate for stage II, giving between 70 million and 400 million valid reads per replicate. Supplemental Table S2 summarizes the statistics and reads obtained for all in situ Hi-C samples.

RNA-seq and quantitative real-time PCR (qRT-PCR)

For RNA-seq experiments, 1 million to 3 million cells at each stage were resuspended in 350 μL of Qiazol (Qiagen) and frozen at -80°C . RNA was obtained by thawing the samples, adding an additional 350 μL of Qiazol, and using the miRNEasy mini kit as recommended (Qiagen). After RNA extraction, contaminating genomic DNA was eliminated with DNase I digestion. Two independent biological replicates per stage were used to generate RNA-seq libraries.

RNA samples were quantified using Nanodrop, and RNA quality was evaluated with an Agilent Bioanalyser (RIN > 9.9). Total RNA (1 μL) was used to generate RNA-seq libraries with rRNA depletion using TruSeq stranded total RNA library preparation

kit (Illumina RS-122-2201). Libraries were sequenced in a HiSeq 2000 (75-bp, paired-end reads) to obtain ~300 million raw reads per sample.

For qRT-PCR, 0.5–1 μg of total RNA obtained from independent biological samples at each stage was converted to cDNA, and qRT-PCR was conducted using SYBR Green (LightCycler Roche) and the following primer sequences (5' to 3'): *Klf4* (Fwd-CGGGAAGGGAGAAGACA, Rev-GAGTTCCTCACGCCA AC), *Spi1* (Fwd-GCGTGCAAAATGGAAGGGTT, Rev-GTGTG CGGAGAAATCCCAGT), *Irf8* (Fwd-CAATCAGGAGGTGGA TGCTT, Rev-AGCACAGCGTAACCTCGTCT), *Myc* (Fwd-CC TAGTGCTGCATGAGGAGA, Rev-TCCACAGACACCACAT-CAATT), *Flt3* (Fwd-ATCTCCGAGGGTGTTCAGA, Rev-T GAACAGCTTGGTGCATTTCG), *Gata2* (Fwd-GCTTCACCCC TAAGCAGAGA, Rev-TGGCACCACAGTTGACACA), *Gata1* (Fwd-ACGACCACTACAACACTCTGGC, Rev-TGGCGGTTT CTCGCTCGATTTC), *c-Kit* (Fwd-GATCTGCTCTGCGTCCT GTT, Rev-CTTGACAGATGGCTGAGACG), and *Bcl2* (Fwd-GAACTGGGGGAGGATTGTGG, Rev-GGCCATATAGTTCC ACAAGGC). *Rplp0* (Fwd-TTCATTGTGGGAGCAGAC, Rev-CAGCAGTTTCTCCAGAGC) was used as housekeeping control for *Klf4*, *Gata1*, *c-Kit*, and *Bcl2*. For those genes, final values were multiplied by 1000. For the rest of genes, β -actin (Fwd-GGCCCA GAGCAAGAGAGGTATCC, Rev-ACGCACGATTTCCTCTC CAGC) was used as housekeeping control.

ChIP-seq experiments

For ChIP-seq experiments, 5 million to 10 million cells were cross-linked as described above. Experiments were performed as previously published (Mas et al. 2018). Chromatin complexes were immunoprecipitated using anti-H3K27me3 (Millipore 07-449), anti-H3K4me1 (Abcam ab8895), anti-H3K4me3 (Diagenode C15410003), and anti-H3K27ac (Millipore 07-360). A small aliquot of ChIP DNA was used for ChIP-qPCR validations using primers of transcriptionally active and repressed genes (*Nucleolin*, *Sox2*, and *Gapdh*) to verify enrichment of the histone modifications. About 2–10 ng of ChIP or input DNA material was used to prepare ChIP-seq libraries following the NEBNext Ultra DNA library preparation kit for Illumina (NEB E7370L) as per the manufacturer's instructions. Final ChIP-seq libraries were size-selected to remove fragments <100 bp and then amplified for 10 PCR cycles. Libraries were sequenced on a HiSeq 2000 platform (Illumina) to obtain ~30 million reads per library (50 bp, single end).

RNA-seq and ChIP-seq data analyses

RNA-seq replicate samples were mapped against the mm10 mouse genome assembly using TopHat (Trapnell et al. 2009) with the option -g 1 to discard reads that could not be uniquely mapped to just one region. DESeq2 (Love et al. 2014) was run to quantify the expression of every annotated transcript using the RefSeq catalog of exons and to identify each set of differentially expressed genes. Expression values shown in the box plots correspond to the averaged FPKMs across the two replicates in each stage. The rows of the heat maps of gene expression were scaled to have mean 0 and a standard deviation of 1 (Z-score). To define the set of unique differentially expressed genes (up or down), only genes reported to significantly change expression in a single stage as compared with stage 0 were included in the heat maps. Gene set enrichment analysis of the preranked lists of genes by DESeq2 stat value was performed with the GSEA software (Subramanian et al. 2005).

ChIP-seq raw reads were mapped against the mm10 mouse genome assembly using Bowtie (Langmead et al. 2009) with the option -m 1 to discard reads that did not map uniquely to one

region. MACS (Zhang et al. 2008b) was run with the default parameters but with the shift size adjusted to 100 bp to perform the peak calling of ChIP-seq experiments. The genome distribution of each set of peaks was calculated by counting the number of peaks fitted on each class of region according to RefSeq annotations (O'Leary et al. 2016). "Promoter" was the region within ± 2.5 kb of the transcription start site (TSS), intragenic regions corresponded to the rest of the gene not classified as promoter, and the rest of the genome was considered to be intergenic. Peaks that overlapped with more than one genomic feature were counted in multiple categories. Active enhancers were defined by the presence of overlapping peaks of H3K4me1 and H3K27ac at stage 0 within intronic and intergenic regions. To define the set of active enhancers at stage 0 that were lost along the rest of stages (I, II, III, and IV), the ChIP-seq signal of H3K27ac was subtracted in stage 0 from the rest of H3K27ac profiles, and enhancers were identified in which the final value was less than one normalized read. To identify examples of enhancers gaining H3K27ac signal, the subtraction was inversely performed. The same procedure was used to determine the gain or loss of H3K27me3 in the same enhancer collection. Heat maps displaying the density of H3K27ac and H3K27me3 ChIP-seq reads around the center of each enhancer were generated by counting the number of reads for each individual enhancer and normalizing this value with the total number of mapped reads of the sample. The rows of the heat maps were scaled to have mean 0 and standard deviation 1 (Z-score), and plots were generated using SeqCode (Blanco et al. 2021).

In all analyses, we used release 68 of the RefSeq annotations (O'Leary et al. 2016) as provided by the UCSC genome browser on the refGene.txt file (Tyner et al. 2017). This RefSeq version contains 34,904 transcripts corresponding to 24,338 mouse genes. No preprocessing filtering steps were performed on this file. The UCSC genome browser was used to generate screenshots of the genomic landscape of selected genes (Tyner et al. 2017). Enrichr (Kuleshov et al. 2016) was used to perform gene ontology (GO), KEGG, and other functional analysis (such as ChEA) of the gene sets obtained from RNA-seq and genes in bins that switched A/B compartments. Supplemental Table S1 lists all differentially expressed genes and Enrichr results in each comparison. Graphical treatment and quantification of the ChIP-seq and the RNA-seq experiments was performed using SeqCode (Blanco et al. 2021).

Hi-C data analysis

Hi-C data were processed using an in-house pipeline based on TADbit (Serra et al. 2017). Reads were mapped according to a fragment-based strategy: Each side of the sequenced read was mapped in full length to the reference genome mouse December 2011 (GRCm38/mm10). TADbit filtering module was used to remove noninformative contacts and to create contact matrices as previously described (Serra et al. 2017). PCR duplicates were removed, and the Hi-C filters applied corresponded to potential nondigested fragments (extradangling ends), nonligated fragments (dangling ends), self-circles, and random breaks. Contact matrices were normalized for sequencing depth and genomic biases using OneD (Vidal et al. 2018). A and B chromatin compartment analysis was performed at 100-kb resolution as previously described (Lieberman-Aiden et al. 2009; Serra et al. 2017). Differential Hi-C matrices were computed from normalized Hi-C matrices at 5-kb resolution. Matrices of the specific regions were corrected for read coverage, a Gaussian filter was applied for noise reduction, and the difference between maps at stage III and stage 0 was plotted. Virtual 4C-seq profiles were generated from local

coverage-normalized Hi-C matrices at 5-kb resolution, and Gaussian filter was applied for smoothing. The 5-kb bin containing the TSS of the gene of interest was used as the viewpoint. The domain score of consensus TADs was computed as previously described (Krieger et al. 2016; Stadhouers et al. 2018). TADs were ranked according to the ratio stage III/stage 0 of this score. The normalized level of H3K27ac, H3K27me3, and RNA per TAD was obtained using respective ChIP-seq and RNA-seq data sets. The ratio of the levels of these marks between stage III and stage 0 was compared between the 10% of TADs with higher changes (higher at stage 0 or higher at stage III).

Klf4 overexpression experiments

The *KLF4* (mouse) and *PML-RAR α* (human) cDNAs were cloned into MSCV-GFP and MSCV-hCD4 vectors (Addgene vector 35712) under the control of the EV promoter. The ecotropic phoenix packaging cell line was transiently transduced with the retroviral vectors cited above, and the retroviral supernatant was collected and filtered. Bone marrow lineage-negative cells were obtained from C57/BL6 wild-type mice and transduced with retroviruses carrying either MSCV-GFP-KLF4, MSCV-hCD4-PE-PML-RAR α , or both by two rounds of spinfection in nontissue culture-treated plates (Corning 351147) coated with retronectin (Takara T100A). Transduced Lin[−] cells were sorted and serially replated in methylcellulose medium (Methocult, Stem Cell Technologies M3434) by seeding 10,000 cells/well and replating every 7 d. Flow cytometry analyses were performed by staining cells with PE anti-hCD4 (BD Pharmingen 555347), APC antimouse Cd11b (BD 53312), and APC-fluo780 antimouse c-kit (Invitrogen 47-1171-82) using FACSAria (BD).

Data availability

All sequencing data sets are available at GEO under accession number GSE151837.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We are indebted to L. Morey and members of the Di Croce laboratory for insightful discussions and critical reading of the manuscript. We thank V.A. Raker for scientific editing, and the Centre for Genomic Regulation Genomics Unit for their help in genomic experiments. G.M. conducted this work with support from "Becas Leonardo a Investigadores y Creadores Culturales" from the Banco Bilbao Vizcaya Argentaria Foundation. The work in the Di Croce laboratory is supported by grants from the Spanish Ministry of Science and Innovation (PID2019-108322GB-I00), "Fundación Vencer El Cáncer" (VEC), the European Regional Development Fund (ERDF), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (Programa Operatiu FEDER de Catalunya 2014-2020; AGAUR, 2017 SGR and 2019 FLB 00426), the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement number 713673 "ChromDesign," and the Fondo Social Europeo (FSE). A.S. is supported by a fellowship from "la Caixa" Foundation (ID 100010434). The work was partially supported by awards from the European Research Council under the 7th Framework Program (FP7/2007-2013 609989), the European Union's Horizon 2020 Research

Mas et al.

and Innovation Program (676556), and the Spanish Ministerio de Ciencia, Innovación y Universidades (BFU2017-85926-P) to M.A.M.-R. We acknowledge support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the EMBL partnership, and the cofinancing with funds from the European Regional Development Fund (FEDER), Centro de Excelencia Severo Ochoa, Centres de Recerca de Catalunya Programme/Generalitat de Catalunya. Work in S.M.'s laboratory was funded by the Association for Cancer Research (IG20).

Author contributions: G.M., S.M., and L.D.C. designed the study. F.L.D., E.V., and F.M. performed *in situ* Hi-C data analyses and interpretation, with guidance from M.A.M.-R. F.S. and A.G. prepared mouse primary samples for all NGS experiments. G.P.G.F. and A.S. conducted the *Klf4* rescue experiments. G.M., C.B., and G.F. carried out ChIP-seq experiments. G.M. and C.B. conducted RNA-seq and *in situ* Hi-C experiments. E.B. performed bioinformatic analyses for ChIP-seq and RNA-seq. All authors contributed to the discussion and interpretation of the results. G.M. and L.D.C. wrote the manuscript with input from all authors.

References

- Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhim R, Campbell PJ, Chin L, Dixon JR, Futreal PA, et al. 2020. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet* **52**: 294–305. doi:10.1038/s41588-019-0564-y
- Aksoy I, Giudice V, Delahaye E, Wianny F, Aubry M, Mure M, Chen J, Jauch R, Bogu GK, Nolden T, et al. 2014. *Klf4* and *Klf5* differentially inhibit mesoderm and endoderm differentiation in embryonic stem cells. *Nat Commun* **5**: 3719. doi:10.1038/ncomms4719
- Alder JK, Georgantas RW III, Hildreth RL, Kaplan IM, Morisot S, Yu X, McDevitt M, Civin CI. 2008. Krüppel-like factor 4 is essential for inflammatory monocyte differentiation *in vivo*. *J Immunol* **180**: 5645–5652. doi:10.4049/jimmunol.180.8.5645
- Basu S, Shukron O, Ponjavic A, Parruto P, Boucher W, Zhang W, Reynolds N, Lando D, Shah D, Sober L, et al. 2020. Live-cell 3D single-molecule tracking reveals how NuRD modulates enhancer dynamics. *bioRxiv* doi:10.1101/2020.04.03.003178
- Bernardi R, Pandolfi PP. 2007. Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nat Rev Mol Cell Biol* **8**: 1006–1016. doi:10.1038/nrm2277
- Blanco E, Gonzalez-Ramirez M, Di Croce L. 2021. Productive visualization of high-throughput sequencing data using the Seq-Code open portable platform. *Sci Rep* **11**: 19545. doi:10.1038/s41598-021-98889-7
- Brown D, Kogan S, Lagasse E, Weissman I, Alcalay M, Pelicci PG, Atwater S, Bishop JM. 1997. A PML/RARA transgene initiates murine acute promyelocytic leukemia. *Proc Natl Acad Sci* **94**: 2551–2556. doi:10.1073/pnas.94.6.2551
- Carbone R, Botrugno OA, Ronzoni S, Insinga A, Di Croce L, Pelicci PG, Minucci S. 2006. Recruitment of the histone methyltransferase SUV39H1 and its role in the oncogenic properties of the leukemia-associated PML-retinoic acid receptor fusion protein. *Mol Cell Biol* **26**: 1288–1296. doi:10.1128/MCB.26.4.1288-1296.2006
- Chang HR, Munkhjargal A, Kim MJ, Park SY, Jung E, Ryu JH, Yang Y, Lim JS, Kim Y. 2018. The functional roles of PML nuclear bodies in genome maintenance. *Mutat Res* **809**: 99–107. doi:10.1016/j.mrfmmm.2017.05.002
- Cole CB, Verdoni AM, Ketkar S, Leight ER, Russler-Germain DA, Lamprecht TL, Demeter RT, Magrini V, Ley TJ. 2016. PML-RARA requires DNA methyltransferase 3A to initiate acute promyelocytic leukemia. *J Clin Invest* **126**: 85–98. doi:10.1172/JCI82897
- Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**: 240–244. doi:10.1038/nature14450
- Dawson MA. 2017. The cancer epigenome: concepts, challenges, and therapeutic opportunities. *Science* **355**: 1147–1152. doi:10.1126/science.aam7304
- Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, Pellegrini M, Plath K. 2013. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**: 602–616. doi:10.1016/j.stem.2013.08.013
- de Thé H, Chomienne C, Lanotte M, Degos L, Dejean A. 1990. The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor to a novel transcribed locus. *Nature* **347**: 558–561. doi:10.1038/347558a0
- Di Carlo V, Mocavini I, Di Croce L. 2019. Polycomb complexes in normal and malignant hematopoiesis. *J Cell Biol* **218**: 55–69. doi:10.1083/jcb.201808028
- Di Croce L, Raker VA, Corsaro M, Fazi F, Fanelli M, Faretta M, Fuks F, Lo Coco F, Kouzarides T, Nervi C, et al. 2002. Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. *Science* **295**: 1079–1082. doi:10.1126/science.1065173
- di Masi A, Cilli D, Berardinelli F, Talarico A, Pallavicini I, Pennisi R, Leone S, Antoccia A, Noguera NI, Lo-Coco F, et al. 2016. PML nuclear body disruption impairs DNA double-strand break sensing and repair in APL. *Cell Death Dis* **7**: e2308. doi:10.1038/cddis.2016.115
- Faber K, Bullinger L, Ragu C, Garding A, Mertens D, Miller C, Martin D, Walcher D, Döhner K, Döhner H, et al. 2013a. CDX2-driven leukemogenesis involves KLF4 repression and deregulated PPAR γ signaling. *J Clin Invest* **123**: 299–314. doi:10.1172/JCI64745
- Faber K, Bullinger L, Ragu C, Garding A, Mertens D, Miller C, Martin D, Walcher D, Döhner K, Döhner H, et al. 2013b. CDX2-driven leukemogenesis involves KLF4 repression and deregulated PPAR γ signaling. *J Clin Invest* **123**: 299–314. doi:10.1172/JCI64745
- Feinberg MW, Wara AK, Cao Z, Lebedeva MA, Rosenbauer F, Iwasaki H, Hirai H, Katz JP, Haspel RL, Gray S, et al. 2007. The Krüppel-like factor KLF4 is a critical regulator of monocyte differentiation. *EMBO J* **26**: 4138–4148. doi:10.1038/sj.emboj.7601824
- Filarsky K, Garding A, Becker N, Wolf C, Zucknick M, Claus R, Weichenhan D, Plass C, Döhner H, Stilgenbauer S, et al. 2016. Krüppel-like factor 4 (KLF4) inactivation in chronic lymphocytic leukemia correlates with promoter DNA-methylation and can be reversed by inhibition of NOTCH signaling. *Haematologica* **101**: e249–e253. doi:10.3324/haematol.2015.138172
- Flavahan WA, Drier Y, Liao BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, Bernstein BE. 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**: 110–114. doi:10.1038/nature16490
- Gaillard C, Tokuyasu TA, Rosen G, Sothen J, Vitaliano-Prunier A, Roy R, Passegue E, de Thé H, Figueroa ME, Kogan SC. 2015. Transcription and methylation analyses of preleukemic promyelocytes indicate a dual role for PML/RARA in leukemia initiation. *Haematologica* **100**: 1064–1075.
- Goddard AD, Borrow J, Freemont PS, Solomon E. 1991. Characterization of a zinc finger gene disrupted by the t(15;17) in

- acute promyelocytic leukemia. *Science* **254**: 1371–1374. doi:10.1126/science.1720570
- Grignani F, Kinsella T, Mencarelli A, Valtieri M, Riganelli D, Grignani F, Lanfrancone L, Peschle C, Nolan GP, Pelicci PG. 1998. High-efficiency gene transfer and selection of human hematopoietic progenitor cells with a hybrid EBV/retroviral vector expressing the green fluorescence protein. *Cancer Res* **58**: 14–19.
- Grignani F, Valtieri M, Gabbianelli M, Gelmetti V, Botta R, Luchetti L, Masella B, Morsilli O, Pelosi E, Samoggia P, et al. 2000. PML/RAR α fusion protein expression in normal human hematopoietic progenitors dictates myeloid commitment and the promyelocytic phenotype. *Blood* **96**: 1531–1537. doi:10.1182/blood.V96.4.1531
- Grisolano JL, Wesselschmidt RL, Pelicci PG, Ley TJ. 1997. Altered myeloid development and acute leukemia in transgenic mice expressing PML-RAR α under control of cathepsin G regulatory sequences. *Blood* **89**: 376–387. doi:10.1182/blood.V89.2.376
- Gröschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, van der Velden VHJ, Havermans M, Avelino R, van Lom K, et al. 2014. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**: 369–381. doi:10.1016/j.cell.2014.02.019
- Guibal FC, Alberich-Jorda M, Hirai H, Ebralidze A, Levantini E, Di Ruscio A, Zhang P, Santana-Lemos BA, Neuberg D, Wagers AJ, et al. 2009. Identification of a myeloid committed progenitor as the cancer-initiating cell in acute promyelocytic leukemia. *Blood* **114**: 5415–5425. doi:10.1182/blood-2008-10-182071
- He LZ, Tribioli C, Rivi R, Peruzzi D, Pelicci PG, Soares V, Cattorretti G, Pandolfi PP. 1997. Acute leukemia with promyelocytic features in PML/RAR α transgenic mice. *Proc Natl Acad Sci* **94**: 5302–5307. doi:10.1073/pnas.94.10.5302
- Hnisz D, Weintraub AS, Day DS, Valtion AL, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–1458. doi:10.1126/science.1249024
- Hoemme C, Peerzada A, Behre G, Wang Y, McClelland M, Nieselt K, Zschunke M, Disselhoff C, Agrawal S, Isken F, et al. 2008. Chromatin modifications induced by PML-RAR α repress critical targets in leukemogenesis as analyzed by ChIP-chip. *Blood* **111**: 2887–2895. doi:10.1182/blood-2007-03-079921
- Huang Y, Chen J, Lu C, Han J, Wang G, Song C, Zhu S, Wang C, Li G, Kang J, et al. 2014. HDAC1 and Klf4 interplay critically regulates human myeloid leukemia cell proliferation. *Cell Death Dis* **5**: e1491. doi:10.1038/cddis.2014.433
- Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003. doi:10.1038/nmeth.2148
- Katerndahl CDS, Rogers ORS, Day RB, Cai MA, Rooney TP, Helton NM, Hoock M, Ramakrishnan SM, Srivatsan SN, Wartman LD, et al. 2021. Tumor suppressor function of Gata2 in acute promyelocytic leukemia. *Blood* **138**: 1148–1161. doi:10.1182/blood.2021011758
- Kobune M, Iyama S, Kikuchi S, Horiguchi H, Sato T, Murase K, Kawano Y, Takada K, Ono K, Kamihara Y, et al. 2012. Stromal cells expressing hedgehog-interacting protein regulate the proliferation of myeloid neoplasms. *Blood Cancer J* **2**: e87. doi:10.1038/bcj.2012.36
- Krijger PH, Di Stefano B, de Wit E, Limone F, van Oevelen C, de Laat W, Graf T. 2016. Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell* **18**: 597–610. doi:10.1016/j.stem.2016.01.007
- Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**: W90–W97. doi:10.1093/nar/gkw377
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25. doi:10.1186/gb-2009-10-3-r25
- Lewis AH, Bridges CS, Punia VS, Cooper AFJ, Puppi M, Lacorazza HD. 2021. Krüppel-like factor 4 promotes survival and expansion in acute myeloid leukemia cells. *Oncotarget* **12**: 255–267. doi:10.18632/oncotarget.27878
- Li Y, He Y, Liang Z, Wang Y, Chen F, Djekidel MN, Li G, Zhang X, Xiang S, Wang Z, et al. 2018. Alterations of specific chromatin conformation affect ATRA-induced leukemia cell differentiation. *Cell Death Dis* **9**: 200. doi:10.1038/s41419-017-0173-6
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293. doi:10.1126/science.1181369
- Lin HK, Bergmann S, Pandolfi PP. 2004. Cytoplasmic PML function in TGF- β signalling. *Nature* **431**: 205–211. doi:10.1038/nature02783
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. doi:10.1186/s13059-014-0550-8
- Lupiañez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserli H, Opitz JM, Laxova R, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012–1025. doi:10.1016/j.cell.2015.04.004
- Makishima H. 2017. Somatic SETBP1 mutations in myeloid neoplasms. *Int J Hematol* **105**: 732–742. doi:10.1007/s12185-017-2241-1
- Martens JH, Brinkman AB, Simmer F, Francois KJ, Nebbioso A, Ferrara F, Altucci L, Stunnenberg HG. 2010. PML-RAR α /RXR alters the epigenetic landscape in acute promyelocytic leukemia. *Cancer Cell* **17**: 173–185. doi:10.1016/j.ccr.2009.12.042
- Mas G, Di Croce L. 2016. The role of Polycomb in stem cell genome architecture. *Curr Opin Cell Biol* **43**: 87–95. doi:10.1016/j.ccb.2016.09.006
- Mas G, Blanco E, Ballare C, Sanso M, Spill YG, Hu D, Aoi Y, Le Dily F, Shilatifard A, Marti-Renom MA, et al. 2018. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet* **50**: 1452–1462. doi:10.1038/s41588-018-0218-5
- Mikesch J-H, Gronemeyer H, So CWE. 2010. Discovery of novel transcriptional and epigenetic targets in APL by global ChIP analyses: emerging opportunity and challenge. *Cancer Cell* **17**: 112–114. doi:10.1016/j.ccr.2010.01.012
- Minucci S, Monestiroli S, Giavara S, Ronzoni S, Marchesi F, Ininga A, Diverio D, Gasparini P, Capillo M, Colombo E, et al. 2002. PML-RAR induces promyelocytic leukemias with high efficiency following retroviral gene transfer into purified murine hematopoietic progenitors. *Blood* **100**: 2989–2995. doi:10.1182/blood-2001-11-0089
- Morey L, Brenner C, Fazi F, Villa R, Gutierrez A, Buschbeck M, Nervi C, Minucci S, Fuks F, Di Croce L. 2008. MBD3, a

Mas et al.

- component of the NuRD complex, facilitates chromatin alteration and deposition of epigenetic marks. *Mol Cell Biol* **28**: 5912–5923. doi:10.1128/MCB.00467-08
- Morris VA, Cummings CL, Korb B, Boaglio S, Oehler VG. 2016. Deregulated KLF4 expression in myeloid leukemias alters cell proliferation and differentiation through microRNA and gene targets. *Mol Cell Biol* **36**: 559–573. doi:10.1128/MCB.00712-15
- Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, Shih DJ, Hovestadt V, Zapata M, Sturm D, et al. 2014. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**: 428–434. doi:10.1038/nature13379
- Oksuz O, Narendra V, Lee CH, Descostes N, LeRoy G, Raviram R, Blumenberg L, Karch K, Rocha PP, Garcia BA, et al. 2018. Capturing the onset of PRC2-mediated repressive domain formation. *Mol Cell* **70**: 1149–1162.e1145. doi:10.1016/j.molcel.2018.05.023
- O'Leary NA, Wright MW, Brister JR, Ciufio S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–D745. doi:10.1093/nar/gkv1189
- Pandolfi PP. 2001. Oncogenes and tumor suppressors in the molecular pathogenesis of acute promyelocytic leukemia. *Hum Mol Genet* **10**: 769–775. doi:10.1093/hmg/10.7.769
- Park CS, Shen Y, Lewis A, Lacorazza HD. 2016. Role of the reprogramming factor KLF4 in blood formation. *J Leukoc Biol* **99**: 673–685. doi:10.1189/jlb.1RU1215-539R
- Park CS, Lewis A, Chen T, Lacorazza D. 2019a. Concise review: regulation of self-renewal in normal and malignant hematopoietic stem cells by Krüppel-like factor 4. *Stem Cells Transl Med* **8**: 568–574. doi:10.1002/sctm.18-0249
- Park CS, Lewis AH, Chen TJ, Bridges CS, Shen Y, Suppipat K, Puppi M, Tomolonis JA, Pang PD, Mistretta TA, et al. 2019b. A KLF4-DYRK2-mediated pathway regulating self-renewal in CML stem cells. *Blood* **134**: 1960–1972. doi:10.1182/blood.2018875922
- Ptasinska A, Pickin A, Assi SA, Chin PS, Ames L, Avellino R, Gröschel S, Delwel R, Cockerill PN, Osborne CS, et al. 2019. RUNX1-ETO depletion in t(8;21) AML leads to C/EBP α - and AP-1-mediated alterations in enhancer-promoter interaction. *Cell Rep* **28**: 3022–3031.e3027. doi:10.1016/j.celrep.2019.08.040
- Pundhir S, Bratt Lauridsen FK, Schuster MB, Jakobsen JS, Ge Y, Schoof EM, Rapin N, Waage J, Hasemann MS, Porse BT. 2018. Enhancer and transcription factor dynamics during myeloid differentiation reveal an early differentiation block in Cebpa null progenitors. *Cell Rep* **23**: 2744–2757. doi:10.1016/j.celrep.2018.05.012
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680. doi:10.1016/j.cell.2014.11.021
- Rosenbauer F, Tenen DG. 2007. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat Rev Immunol* **7**: 105–117. doi:10.1038/nri2024
- Sacchi N, Watson DK, Guerts van Kessel AH, Hagemeijer A, Kersey J, Drabkin HD, Patterson D, Papas TS. 1986. Hu-ets-1 and Hu-ets-2 genes are transposed in acute leukemias with (4;11) and (8;21) translocations. *Science* **231**: 379–382. doi:10.1126/science.3941901
- Saeed S, Logie C, Stunnenberg HG, Martens JH. 2011. Genome-wide functions of PML-RAR α in acute promyelocytic leukemia. *Br J Cancer* **104**: 554–558. doi:10.1038/sj.bjc.6606095
- Saeed S, Logie C, Francoijs KJ, Frige G, Romanenghi M, Nielsen FG, Raats L, Shahhoseini M, Huynen M, Altucci L, et al. 2012. Chromatin accessibility, p300, and histone acetylation define PML-RAR α and AML1-ETO binding sites in acute myeloid leukemia. *Blood* **120**: 3058–3068. doi:10.1182/blood-2011-10-386086
- Saumet A, Vetter G, Bouttier M, Portales-Casamar E, Wasserman WW, Maurin T, Mari B, Barbry P, Vallar L, Friederich E, et al. 2009. Transcriptional repression of microRNA genes by PML-RAR α increases expression of key cancer proteins in acute promyelocytic leukemia. *Blood* **113**: 412–421. doi:10.1182/blood-2008-05-158139
- Scherer M, Stamminger T. 2016. Emerging role of PML nuclear bodies in innate immune signaling. *J Virol* **90**: 5850–5854. doi:10.1128/JVI.01979-15
- Schoenfelder S, Fraser P. 2019. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* **20**: 437–455. doi:10.1038/s41576-019-0128-0
- Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, Dimitrova E, Matheson L, Tavares-Cadete F, Furlan-Magaril M, et al. 2015. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet* **47**: 1179–1186. doi:10.1038/ng.3393
- Schoenhals M, Kassambara A, Veyrune JL, Moreaux J, Goldschmidt H, Hose D, Klein B. 2013. Krüppel-like factor 4 blocks tumor cell proliferation and promotes drug resistance in multiple myeloma. *Haematologica* **98**: 1442–1449. doi:10.3324/haematol.2012.066944
- Schoofs T, Rohde C, Hebestreit K, Klein HU, Göllner S, Schulze I, Lerdrup M, Dietrich N, Agrawal-Singh S, Witten A, et al. 2013. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood* **121**: 178–187. doi:10.1182/blood-2012-08-448860
- Segalla S, Rinaldi L, Kilstup-Nielsen C, Badaracco G, Minucci S, Pelicci PG, Landsberger N. 2003. Retinoic acid receptor a fusion to PML affects its transcriptional and chromatin-remodeling properties. *Mol Cell Biol* **23**: 8795–8808. doi:10.1128/MCB.23.23.8795-8808.2003
- Seipel K, Marques MT, Bozzini MA, Meinken C, Mueller BU, Pabst T. 2016. Inactivation of the p53-KLF4-CEBPA axis in acute myeloid leukemia. *Clin Cancer Res* **22**: 746–756. doi:10.1158/1078-0432.CCR-15-1054
- Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. 2017. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**: e1005665. doi:10.1371/journal.pcbi.1005665
- Singh AA, Petraglia F, Nebbioso A, Yi G, Conte M, Valente S, Mandoli A, Scisciola L, Lindeboom R, Kerstens H, et al. 2018. Multi-omics profiling reveals a distinctive epigenome signature for high-risk acute promyelocytic leukemia. *Oncotarget* **9**: 25647–25660. doi:10.18632/oncotarget.25429
- Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y, et al. 2018. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet* **50**: 238–249. doi:10.1038/s41588-017-0030-7
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102**: 15545–15550. doi:10.1073/pnas.0506580102

- Subramanyam D, Belair CD, Barry-Holson KQ, Lin H, Kogan SC, Passegue E, Blueloch R. 2010. PML-RAR α and Dnmt3a1 cooperate in vivo to promote acute promyelocytic leukemia. *Cancer Res* **70**: 8792–8801. doi:10.1158/0008-5472.CAN-08-4481
- Sukhai M, Thomas M, Goswami R, Xuan Y, Reis PP, Kamel-Reid S. 2008. Deregulation of transcription factors GATA-1, GATA-2 and C/EBP α in acute promyelocytic leukemia. *Blood* **112**: 2241–2241. doi:10.1182/blood.V112.11.2241.2241
- Tan S, Kermasson L, Hoslin A, Jaako P, Faille A, Acevedo-Aroza A, Lengline E, Ranta D, Poirée M, Fenneteau O, et al. 2019. EFL1 mutations impair eIF6 release to cause Shwachman-Diamond syndrome. *Blood* **134**: 277–290. doi:10.1182/blood.2018893404
- Tan Y, Wang X, Song H, Zhang Y, Zhang R, Li S, Jin W, Chen S, Fang H, Chen Z, et al. 2021. A PML/RAR α direct target atlas redefines transcriptional deregulation in acute promyelocytic leukemia. *Blood* **137**: 1503–1516. doi:10.1182/blood.2020005698
- Trapnell C, Pachter L, Salzberg SL. 2009. Tophat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**: 1105–1111. doi:10.1093/bioinformatics/btp120
- Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, et al. 2017. The UCSC genome browser database: 2017 update. *Nucleic Acids Res* **45**: D626–D634.
- Vidal E, le Dily F, Quilez J, Stadhouers R, Cuartero Y, Graf T, Marti-Renom MA, Beato M, Filion GJ. 2018. Oned: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res* **46**: e49. doi:10.1093/nar/gky064
- Villa R, De Santis F, Gutierrez A, Minucci S, Pelicci PG, Di Croce L. 2004. Epigenetic gene silencing in acute promyelocytic leukemia. *Biochem Pharmacol* **68**: 1247–1254. doi:10.1016/j.bcp.2004.05.041
- Villa R, Morey L, Raker VA, Buschbeck M, Gutierrez A, De Santis F, Corsaro M, Varas F, Bossi D, Minucci S, et al. 2006. The methyl-CpG binding protein MBD1 is required for PML-RAR α function. *Proc Natl Acad Sci* **103**: 1400–1405. doi:10.1073/pnas.0509343103
- Villa R, Pasini D, Gutierrez A, Morey L, Occhionorelli M, Vire E, Nomdedeu JF, Jenuwein T, Pelicci PG, Minucci S, et al. 2007. Role of the Polycomb repressive complex 2 in acute promyelocytic leukemia. *Cancer Cell* **11**: 513–525. doi:10.1016/j.ccr.2007.04.009
- Voisset E, Moravcsik E, Stratford EW, Jaye A, Palgrave CJ, Hills RK, Salomoni P, Kogan SC, Solomon E, Grimwade D. 2018. Pml nuclear body disruption cooperates in APL pathogenesis and impairs DNA damage repair pathways in mice. *Blood* **131**: 636–648. doi:10.1182/blood-2017-07-794784
- Wang K, Wang P, Shi J, Zhu X, He M, Jia X, Yang X, Qiu F, Jin W, Qian M, et al. 2010. PML/RAR α targets promoter regions containing PU.1 consensus and RARE half sites in acute promyelocytic leukemia. *Cancer Cell* **17**: 186–197. doi:10.1016/j.ccr.2009.12.045
- Wang P, Tang Z, Lee B, Zhu JJ, Cai L, Szalaj P, Tian SZ, Zheng M, Plewczynski D, Ruan X, et al. 2020. Chromatin topology reorganization and transcription repression by PML-RAR α in acute promyeloid leukemia. *Genome Biol* **21**: 110. doi:10.1186/s13059-020-02030-2
- Westervelt P, Lane AA, Pollock JL, Oldfather K, Holt MS, Zimonjic DB, Popescu NC, DiPersio JF, Ley TJ. 2003. High-penetrance mouse model of acute promyelocytic leukemia with very low levels of PML-RAR α expression. *Blood* **102**: 1857–1865. doi:10.1182/blood-2002-12-3779
- Wojiski S, Guibal FC, Kindler T, Lee BH, Jesneck JL, Fabian A, Tenen DG, Gilliland DG. 2009. PML-RAR α initiates leukemia by conferring properties of self-renewal to committed promyelocytic progenitors. *Leukemia* **23**: 1462–1471. doi:10.1038/leu.2009.63
- Yang XW, Wang P, Liu JQ, Zhang H, Xi WD, Jia XH, Wang KK. 2014. Coordinated regulation of the immunoproteasome subunits by PML/RAR α and PU.1 in acute promyelocytic leukemia. *Oncogene* **33**: 2700–2708. doi:10.1038/onc.2013.224
- Zhang SJ, Ma LY, Huang QH, Li G, Gu BW, Gao XD, Shi JY, Wang YY, Gao L, Cai X, et al. 2008a. Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc Natl Acad Sci* **105**: 2076–2081. doi:10.1073/pnas.0711824105
- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008b. Model-based analysis of ChIP-seq (MACS). *Genome Biol* **9**: R137. doi:10.1186/gb-2008-9-9-r137
- Zheng H, Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* **20**: 535–550. doi:10.1038/s41580-019-0132-4



ANNEX 2

TADs enriched in histone H1.2 strongly overlap with the B compartment, inaccessible chromatin, and AT-rich Giemsa bands

Candidate's contribution: Analysis of the Hi-C experiments.

Serna-Pujol N, Salinas-Pena M, Mugianesi F, Lopez-Anguita N, Torrent-Llagostera F, Izquierdo-Bouldstridge A, Marti-Renom MA, Jordan A. *TADs enriched in histone H1.2 strongly overlap with the B compartment, inaccessible chromatin, and AT-rich Giemsa bands*. FEBS J. 2021 Mar;288(6):1989-2013. doi: 10.1111/febs.15549. Epub 2020 Sep 24. PMID: 32896099.

TADs enriched in histone H1.2 strongly overlap with the B compartment, inaccessible chromatin, and AT-rich Giemsa bands

Núria Serna-Pujol¹, Mónica Salinas-Pena¹, Francesca Mugianesi², Natalia Lopez-Anguita^{1,*}, Francesc Torrent-Llagostera¹, Andrea Izquierdo-Bouldstridge¹, Marc A. Marti-Renom^{2,3,4,5}  and Albert Jordan¹ 

¹ Molecular Biology Institute of Barcelona (IBMB-CSIC), Spain

² CNAG-CRG, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Spain

³ Centre for Genomic Regulation, The Barcelona Institute for Science and Technology, Spain

⁴ Pompeu Fabra University, Barcelona, Spain

⁵ ICREA, Barcelona, Spain

Keywords

genome compartments; Giemsa bands; histone H1 variants; linker histone; TADs

Correspondence

A. Jordan, IBMB-CSIC, Baldiri Reixac 4, Barcelona 08028, Spain

Tel: +34 93 402 0487

E-mail: albert.jordan@ibmb.csic.es

*Present address

Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, Berlin, 14195, Germany

(Received 4 May 2020, revised 22 July 2020, accepted 1 September 2020)

doi:10.1111/febs.15549

Giemsa staining of metaphase chromosomes results in a characteristic banding useful for identification of chromosomes and its alterations. We have investigated *in silico* whether Giemsa bands (G bands) correlate with epigenetic and topological features of the interphase genome. Staining of G-positive bands decreases with GC content; nonetheless, G-negative bands are GC heterogeneous. High GC bands are enriched in active histone marks, RNA polymerase II, and SINEs and associate with gene richness, gene expression, and early replication. Low GC bands are enriched in repressive marks, lamina-associated domains, and LINEs. Histone H1 variants distribute heterogeneously among G bands: H1X is enriched at high GC bands and H1.2 is abundant at low GC, compacted bands. According to epigenetic features and H1 content, G bands can be organized in clusters useful to compartmentalize the genome. Indeed, we have obtained Hi-C chromosome interaction maps and compared topologically associating domains (TADs) and A/B compartments to G banding. TADs with high H1.2/H1X ratio strongly overlap with B compartment, late replicating, and inaccessible chromatin and low GC bands. We propose that GC content is a strong driver of chromatin compaction and 3D genome organization, that Giemsa staining recapitulates this organization denoted by high-throughput techniques, and that H1 variants distribute at distinct chromatin domains.

Databases

Hi-C data on T47D breast cancer cells have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number [GSE147627](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147627).

Abbreviations

bphs, bands per haploid sequence; G band, Giemsa band; Gneg, negative (unstained) Giemsa bands; Gpos, Giemsa-positive (stained) bands; LAD, lamina-associated domain; mESCs, mouse embryonic stem cells; NAD, nucleolus-associated domain; PTM, post-translational modification; RNAPII, RNA-polymerase II; S/MAR, scaffold or matrix attachment region; TAD, topologically associating domain; TSS, transcription start site.

Introduction

Eukaryotic DNA is packaged into chromatin, whose repeating structural unit is the nucleosome. Each nucleosome consists of an octamer of core histones (H2A, H2B, H3, and H4) around which ~ 147 base pairs (bp) of DNA are wrapped. Histone H1 binds at both entry/exit sites to the linker DNA at the nucleosome, participating in the formation of higher-order chromatin structures [1]. Unlike core histones, H1 proteins are more evolutionary diverse. The human histone H1 family includes seven somatic subtypes (or variants) (H1.1 to H1.5, H1.0, and H1X), three testis-specific (H1t, H1T2, and H1LS1), and one oocyte-specific variant (H1oo) [2–4]. Among somatic variants, H1.1–H1.5 variants are expressed in a replication-dependent manner while H1.0 and H1X are replication-independent. Regarding their patterns of expression, H1.2 to H1.5 and H1X are ubiquitously expressed, H1.1 is restricted to certain tissues, and H1.0 accumulates in terminally differentiated cells.

This large repertoire of H1 variants leads to wonder whether somatic H1 variants are redundant or show specific properties in terms of functionality and genomic distribution. Classically, H1 has been seen as a structural component associated with chromatin compaction, but in recent years, several evidences support the idea of H1 playing a more dynamic role in chromatin regulation [4,5]. Previous studies have shown that histone H1 variants are involved in several nuclear processes including transcription, replication, genome stability, splicing, or heterochromatin maintenance, among others [6–10].

To fully characterize H1 variants specific functionality, it is important to address their genomic distribution, due to the growing evidence that chromatin organization is crucial to genome function. Reports point to a variant-specific genomic distribution among cell types. In mouse embryonic stem cells (ESCs), H1c and H1d (H1.2 and H1.3 orthologs) were found to be depleted from high GC, gene-rich regions, and abundant at major satellites [11]. By using DamID technology in human IMR90 cells, results showed that H1.2–H1.5 was depleted from CpG-dense and regulatory regions, whereas H1.1 had a distinct profile [12]. Besides, H1.5 was enriched in genic and intergenic regions in IMR90 cells but not in ESCs, suggesting that its genomic distribution depends on the differentiation state [13]. In human fibroblasts, mapping of H1.0 revealed its correlation with GC content and abundance at gene-rich chromosomes [14]. In breast cancer cells, H1.2 was the variant that showed the most specific pattern. H1.2 was found enriched in low

GC domains and lamina-associated domains (LADs) [15]. Moreover, combined depletion of H1.2 and H1.4 leads to the activation of heterochromatic repeats, supporting the role of H1.2 in heterochromatin organization [10]. Regarding replication-independent variants, H1.0 and H1X were more abundant at high GC, gene-rich chromosomes. H1.0 was also found enriched at nucleolus-associated domains (NADs) while H1X was more associated with coding regions and RNA polymerase II binding sites [16]. A general feature for all H1 variants, in all species, is its depletion from the transcription start site (TSS) of active genes, meaning that upon transcriptional activation H1 is removed from the TSS of genes.

Nevertheless, although uncovering specific features for H1 variants, data support that all H1 variants are distributed across the whole genome [15]. For this reason, methods to compartmentalize the genome could be useful to study and compare H1 variants genomic distribution. Due to the complex paradigm of chromatin organization, this compartmentalization has to be addressed by a multi-omics approach.

From a functional point of view, the genome has classically segregated into euchromatin and heterochromatin. Transcriptionally active euchromatin present an open state to facilitate accession of transcription machinery, replicates early within S-phase, and is abundant in SINE repetitive elements and active histone modifications. On the contrary, closed and transcriptionally silent heterochromatin is characterized by late replication timing, LINEs and inactive histone modifications [17]. Moreover, it is well established that chromosomes occupy a nonrandom regions in the nucleus (chromosome territories), where gene-poor regions are placed at the heterochromatic nuclear periphery and gene-rich ones to the euchromatic interior. Chromosome conformation capture techniques (such as Hi-C) have revealed the existence of topologically associating domains (TADs), self-organized chromatin domains in spatial proximity that interact more frequently within themselves than with the rest of the genome [18–20]. These structures are conserved across species and are relatively stable in different cell types [18,21]. Hi-C data also lead to the discovery of the so-called A and B genomic compartments, comprising active and inactive regions, respectively [22]. Independently, other chromatin domains participating in nucleus organization have been described, such as aforementioned LADs or NADs [23,24].

Other layers of chromosome architecture have also been studied for years. In 1970s, several staining methods of metaphase chromosomes arised, that is, Giemsa

staining [25]. Although the precise molecular basis of Giemsa has remained unknown for decades, it is widely accepted that staining correlates with AT-rich sequences and chromatin compaction [26,27]. Giemsa bands (G bands) have been useful in cytogenetics allowing detection of chromosomes rearrangements in diseased cells. However, they have not been much explored in relation to functional genomics. Staining of G-positive (Gpos) bands correlates with AT content; nonetheless, unstained or G-negative (Gneg) bands, expected to be GC-rich, are as heterogeneous in its GC or AT content as Gpos.

Here, we used G bands as epigenetic units to investigate the differential distribution of linker histones. We have *in silico* investigated how G bands correlate with epigenetic, accessibility and topological features of the interphase genome, taking advantage of previously published ChIP-seq, ATAC-seq, and newly generated Hi-C data in breast cancer cells. Our results show a heterogeneous and opposite distribution of histones H1.2 and H1X within G bands, being H1.2 associated with low GC bands and H1X with high GC bands. We have found a strong correlation between B compartment, TADs presenting a high H1.2/H1X ratio, low GC bands, and compact chromatin. To our knowledge, this is the first report including an extensive characterization of G bands based on a wide repertoire of genome-wide data, including H1 variants abundance or Hi-C experiments, among others. Moreover, the balance between two H1 variants has never been considered as an epigenetic feature nor related to genome topology before. Overall, this work represents a comprehensive attempt to further investigate how chromatin is organized within the nucleus, integrating histone H1 variants as putative chromatin organizers.

Results

Characterization of Giemsa bands with epigenetic features and GC content dependency

Giemsa staining of metaphase chromosomes results in an alternating dark and light banding pattern that became useful for identifying individual chromosomes and their abnormalities in diseased cells (Fig. 1A). After the sequencing of the human genome and with the help of a dynamic programming algorithm employing data from thousands of fluorescence *in situ* hybridization experiments, the boundaries of each of the bands were estimated [28]. The estimated starting and ending position of each of the Giemsa-positive bands, classified into four groups according to its increasing staining intensity (Gpos25-Gpos100), and

intergenic bands (Gneg), was obtained from the UCSC human genome database. The number of Gpos bands ranged from 81 to 121. Gpos bands occupied from 7.6% to 17.6% of the genome and Gneg bands a 46% (Fig. 1B).

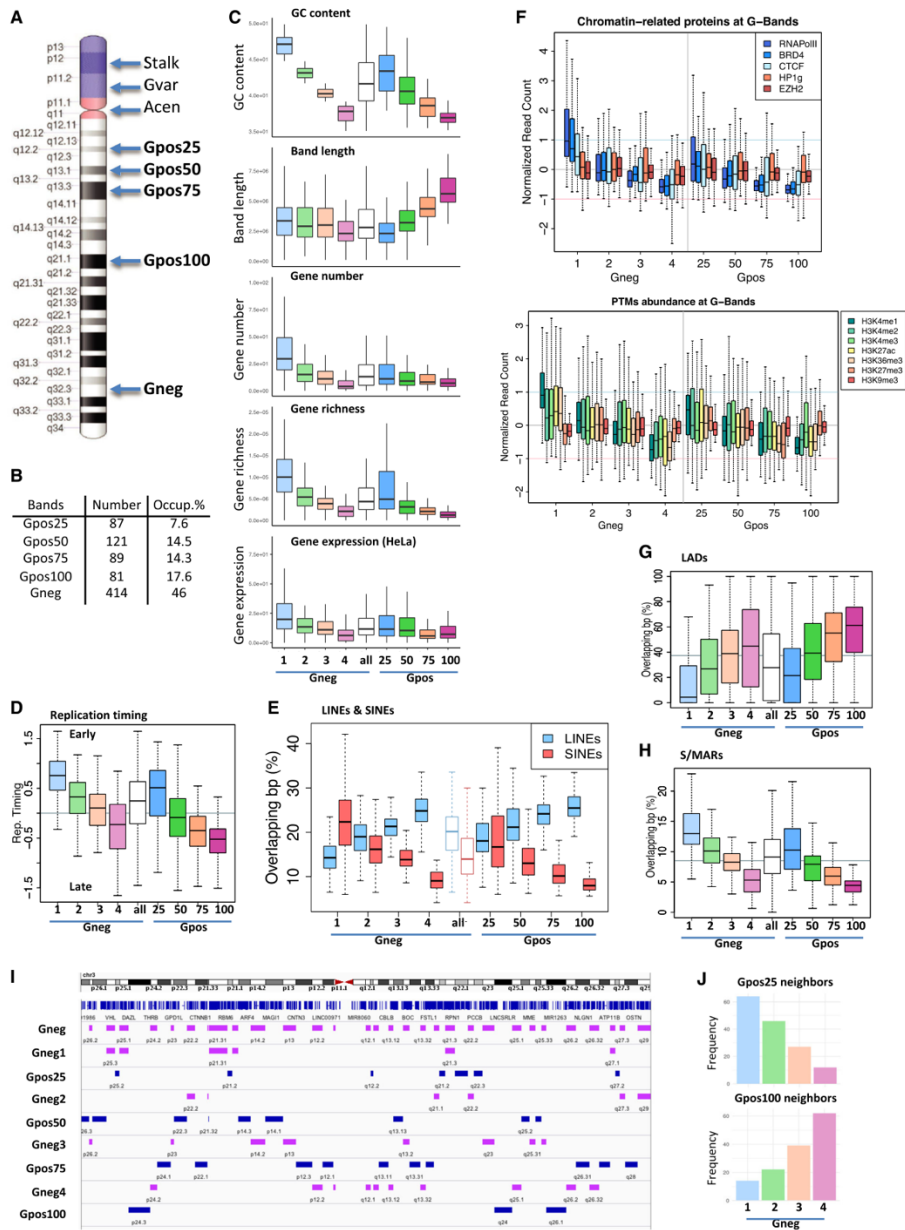
The molecular basis of cytogenetic bands is not well understood. Banding was thought to correspond to GC-poor (dark bands) and GC-rich (light bands) regions. However, Gpos100 bands were consistently AT-rich, but Gpos25 and particularly Gneg bands were highly heterogeneous in its GC content (Fig. 1C). Gneg bands presented a mean GC content intermediate between Gpos25 and Gpos50, indicating that banding could not be explained only by the base composition. Therefore, we wanted to investigate whether banding could be explained by epigenetic features such as core histone marks or linker histone variants.

Darker bands (Gpos100) were longer on average and have been associated with chromatin condensation. Accordingly, they contained the lowest gene content, gene richness, and average gene expression of all the bands (Fig. 1C), as well as longer introns (data not shown). Gneg bands presented intermediate features. As a consequence, we decided to split the Gneg bands in four equivalent groups according to their GC content (Gneg1–4). Gene richness and gene expression correlated positively with GC content (Fig. 1C).

Replication occurs first at active/open chromatin and later at compact chromatin. Data on replication timing for HeLa cells are available, and we used it to calculate the average replication timing at each G band. As expected, within G-positive bands, replication timing was lower (late) at Gpos100 (Fig. 1D). Within G-negative bands, replication timing correlated with the GC content; high GC bands replicated the earliest.

It was previously reported that darker bands are enriched in LINEs and G-negative bands are enriched in SINEs [28]. We have calculated, per chromosome, the percentage of bases in each of the eight band types that is contained within LINEs and SINEs (Fig. 1E). The abundance of SINEs correlated with the GC content, whereas abundance of LINEs correlated with AT content, more than with the darkness of G bands.

Next, we explored the abundance of core histone post-translational modifications (PTMs) and transcription or chromatin-related proteins (from breast cancer T47D cells publically available data) at Gneg and Gpos bands (Fig. 1F). On the one hand, the abundance of PTMs related to gene activation and factors such as RNA polymerase II (RNAPII), BRD4, or CTCF decreased accordingly to the GC content, that is, being high within Gneg1 and Gpos25 and low at



Gneg4 and Gpos100 bands. On the other hand, repressive marks such as H3K27me3 or H3K9me3, as well as EZH2 methyl transferase and heterochromatin protein HP1 gamma, did not follow this pattern following GC content. Instead, they were more abundant than active marks at Gpos75 and Gpos100 bands, but also at Gneg4. The overlap between LADs and G bands also increased at low GC bands, particularly at Gpos100 (Fig. 1G), coinciding with H3K9me3 enrichment over active marks.

Scaffold (metaphase) or matrix (interphase) attachment regions (S/MARs) are involved in control of gene expression, replication, DNA repair, and chromatin to chromosome transition. By linking DNA to the nuclear scaffold, they generate structural and functional loops that span ≈ 20 –100 kb. S/MARs are relatively short sequences (100–1000 bp long) containing one or several of these features: AT richness ($\approx 70\%$), OriC, kinked or curved DNA, TG richness, and topoisomerase-II sites [29]. Because of their AT richness, it was initially proposed that S/MARs were present densely within dark G bands [30]. Mapping of human S/MARs using ChIP-seq data of 14 S/MAR binding proteins was recently achieved [31]. These sites were confirmed to contain the previously described features including AT richness. Nonetheless, we found that they were enriched within high GC bands, both Gneg and Gpos bands (Fig. 1H), as expected for elements involved in the control of gene expression and replication. Accordingly, S/MAR density was found to correlate with gene density [31]. Moreover, S/MARs also correlated with retrovirus integration sites [31]. Accordingly, we found that hotspots for retroviral integration were enriched within high GC bands (data not shown).

As a consequence of this analysis, Gneg interbands were seen epigenetically heterogeneous, being its GC content an important predictive factor of its

characteristics, but not of its lack of Giemsa staining. We then hypothesized that Gneg bands surrounding Gpos bands with a particular GC content could have similar GC values, forming patches of bands with similar features, as shown in Fig. 1I. In fact, we determined that the most abundant neighbors of Gpos25 bands were Gneg1 bands, and Gpos100 bands were preferably surrounded by Gneg4 bands (Fig. 1J). In conclusion, Giemsa-stained bands were surrounded by unstained bands of similar GC content, gene content, and other features, except that were shorter.

Correlation of Giemsa staining with AT content is enhanced along chromosome condensation

Along chromosome condensation, the initially observed banding of 850 bands per haploid sequence (bphs) (prometaphase) gets condensed down to 400 bphs (metaphase) (Fig. 2A) [32]. We predicted that neighbor Gpos and Gneg bands (at 850 bphs) with similar GC content would become either dark (stained) or white (unstained) bands at 400 bphs, depending on its GC content, as shown in Fig. 2B. As an example, bands p25.1, p25.3 (Gneg) and p25.2 (Gpos25) of chromosome 3 became band p25 (white, high GC) at 400 bphs, while bands p14.1, p14.3 (Gpos50), and p14.2 (Gneg) became band p14 (dark, low GC) (Fig. 2B). As a consequence, the difference between GC content at stained versus not-stained bands was increased at 400 bphs compared with 850 bphs, that is upon chromosome condensation (Fig. 2C).

Analyzing what proportion of each of the G bands at 850 bphs became dark or white at 400 bphs, we obtained that a big proportion of Gneg4 became dark, and a big amount of Gpos25 and Gpos50 became white (Fig. 2D). A circular permutation of G bands at 850 bphs confirmed that this observation depends on

Fig. 1. Characterization of Giemsa bands. (A) Ideogram of a human metaphase chromosome showing banding after Giemsa staining. G-positive bands are classified into four types (Gpos25 to Gpos100) according to increasing staining intensity. Unstained bands or interbands are called G-negative (Gneg). Ideograms are from NCBI's Genome Decoration Page. (B) Table indicating the number of bands of each type existing in human chromosomes and the percentage of base pair occupancy in the genome. (C) Box plots showing the GC content, base pair length, gene number content, gene richness (gene number/base pair length), and average gene expression (from HeLa cells) of each G band for each band type. Gneg bands were divided into four equal groups according to GC content. (D) Box plot showing replication timing at each G band (normalized by band length), for each band type. HeLa-S3 public Repli-seq data were used. (E) Box plot showing the proportion of overlapping base pairs between LINEs or SINEs and each G band, for each band type. (F) Box plots showing abundance of chromatin-related proteins or histone PTMs at each G band, for each band type. Enrichment was calculated by computing the average normalized read count of the peaks mapped at each G band. Publicly available data from T47D cells were used, except for EZH2 and H3K27me3 that correspond to HeLa cells. (G) Box plot showing overlapping base pairs between LADs and each G band, for each band type. (H) Box plot showing overlapping base pairs between S/MARs and each G band, for each band type. (I) Browser snapshot of human chromosome 3 showing the position of Gpos and Gneg bands. (J) Bar plots showing the frequency of Gneg band groups that are neighbors of Gpos25 or Gpos100 bands.

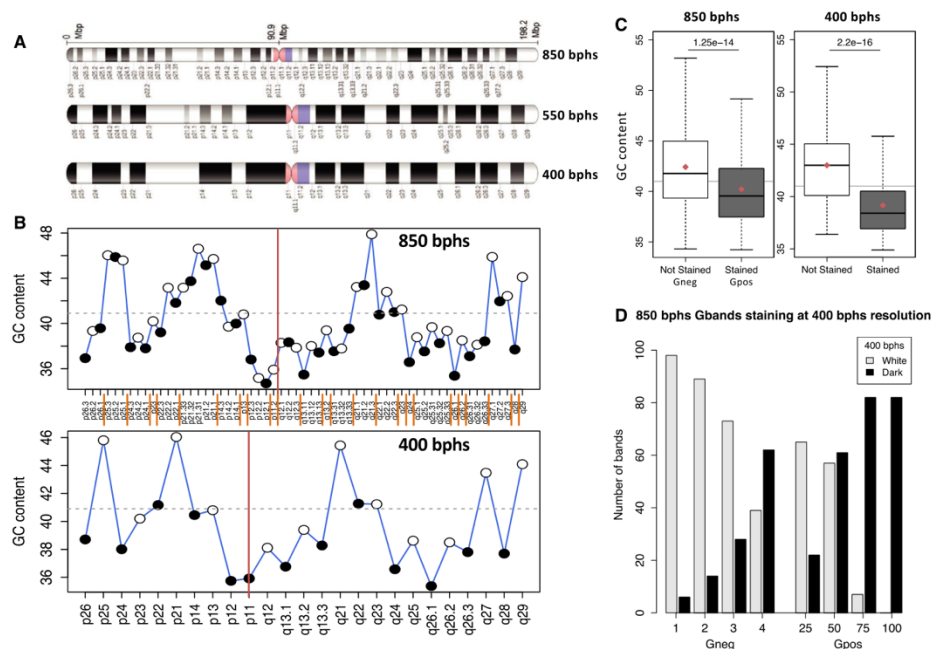
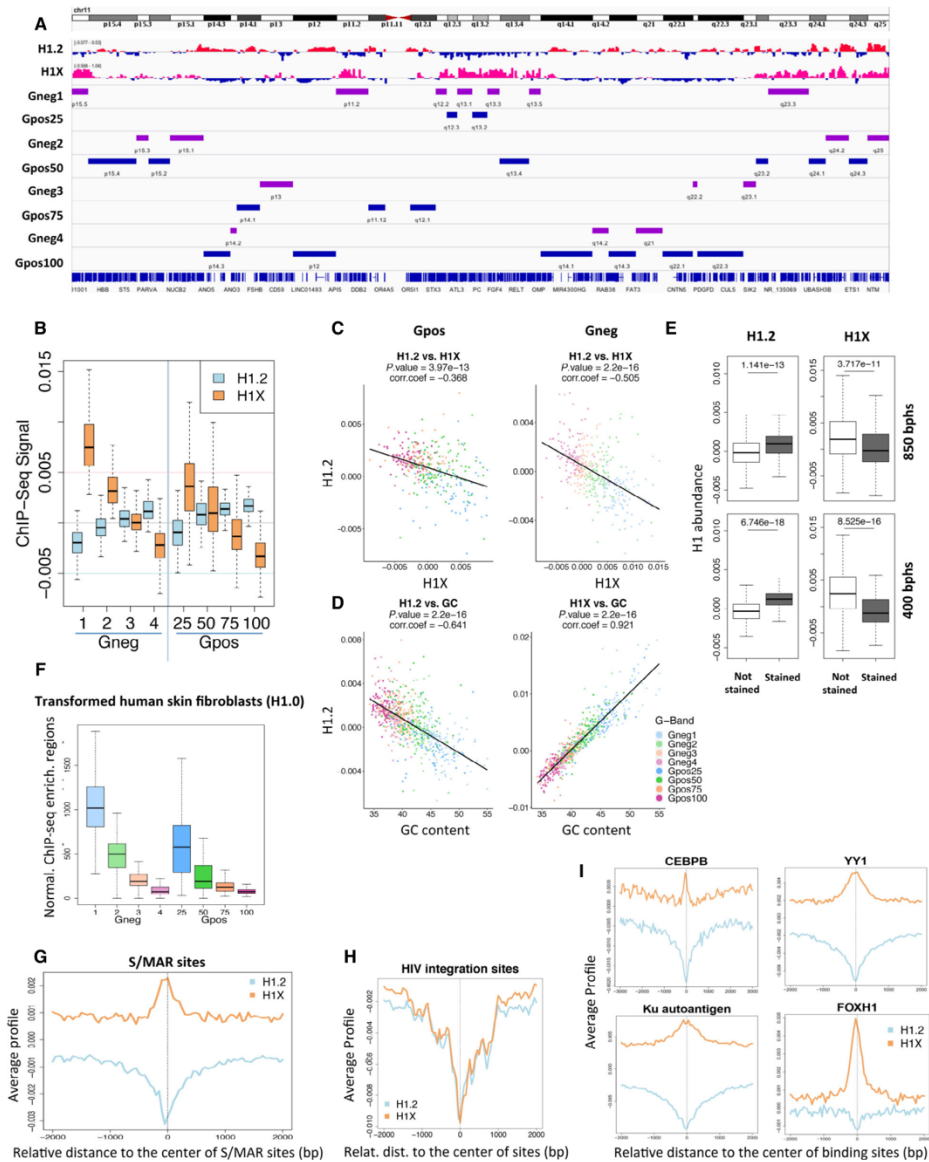


Fig. 2. Correlation of Giemsa staining with AT content is enhanced along chromosome condensation. (A) Ideograms of human chromosome 3 at 850, 550, and 400 bands per haploid sequence (bphs) resolution. Along metaphase condensation, the number of G bands (resolution) decreases, and bands are classified just as stained (dark) or unstained (light). Ideograms are from NCBI's Genome Decoration Page. (B) Representation of GC content of Gpos/stained (dark circle) and Gneg/unstained (light circle) bands along chromosome 3, at 850 and 400 bphs resolution. Clusters of bands at 850 bphs that are merged to a single band at 400 bphs are separated by orange lines. (C) Box plots showing the GC content of Gneg and Gpos bands at 850 and 400 bphs. The Wilcoxon test was used to evaluate the significance of the differences in GC content. (D) Bar plot showing the frequency of bands of each type at 850 bphs that end up stained (dark) or unstained (white) at 400 bphs.

the actual position of the bands (data not shown). Bands that changed their staining status along condensation and ended stained or not as expected according

to its GC content (Gneg4 and Gpos25, respectively) are the shortest bands on average (Fig. 1C). Then, these bands could be seen as short interbands

Fig. 3. Histone H1 variants distribute heterogeneously among G bands. (A) Browser snapshot of human chromosome 11 showing H1.2 and H1X input-subtracted ChIP-seq signal from T47D cells and the position of Gpos and Gneg bands. (B) Box plots showing H1.2 and H1X input-subtracted ChIP-seq abundance within G bands, for each band type. (C) Scatter plots of H1.2 and H1X input-subtracted ChIP-seq abundance at each Gpos (left) or Gneg (right) band. Pearson's correlation coefficient is shown as well as *P*-value. (D) Scatter plots of H1.2 or H1X input-subtracted ChIP-seq abundance against GC content at each Gneg and Gpos band. Pearson's correlation coefficient is shown as well as *P*-value. (E) Box plots showing H1.2 and H1X input-subtracted ChIP-seq abundance within unstained (light) or stained (dark) G bands at 850 or 400 bphs. The Wilcoxon test was used to evaluate the significance of the differences in H1.2 and H1X enrichment. (F) Box plots showing the normalized number of H1.0 ChIP-seq enrichment regions from *in vitro* transformed human skin fibroblasts (GSE66169) within G bands. (G–I) Abundance of H1 variants at retroviral integration sites and S/MAR protein binding sites. Average, input-subtracted ChIP-seq signal of H1.2 and H1X around the center of S/MARs sites (mapped in [31]) (G), HIV-1 integration sites (H), or around the center of the S/MAR binding protein sites indicated (I).



inadequately stained initially (850 bphs), that mimic surrounding, larger bands later (400 bphs), forming larger patches stained or not according to their GC content. These results reinforced the notion that Giemsa staining depends on AT richness, but this is better seen in highly condensed chromosomes. Still, correlation is not perfect because, even at 400 bphs, some stained bands have higher GC content than some unstained bands (Fig. 2B). Nonetheless, locally, stained bands always have lower GC than neighbor unstained bands. This suggests that chromosomes are partitioned into a small number of large domains of high or low GC content and, within them, relative differences in GC dictate band staining.

Another possibility to explain the lack of correlation between staining and GC content at 850 bphs could be that staining was more sensitive to the existence of long AT tracks than to the average AT content. Nonetheless, we obtained that AT content and abundance of AT tracks correlated (correlation coefficient = 0.994, P -value < 0.001), and the number of AT tracks of different lengths was not more different between Gpos and Gneg bands than its average AT content, discarding this hypothesis (data not shown).

In summary, AT content is partially responsible for the intensity of Giemsa banding of metaphase chromosomes and correlates with epigenetic features of chromatin already present at interphase chromosomes.

Histone H1 variants in breast cancer cells distribute heterogeneously among G bands

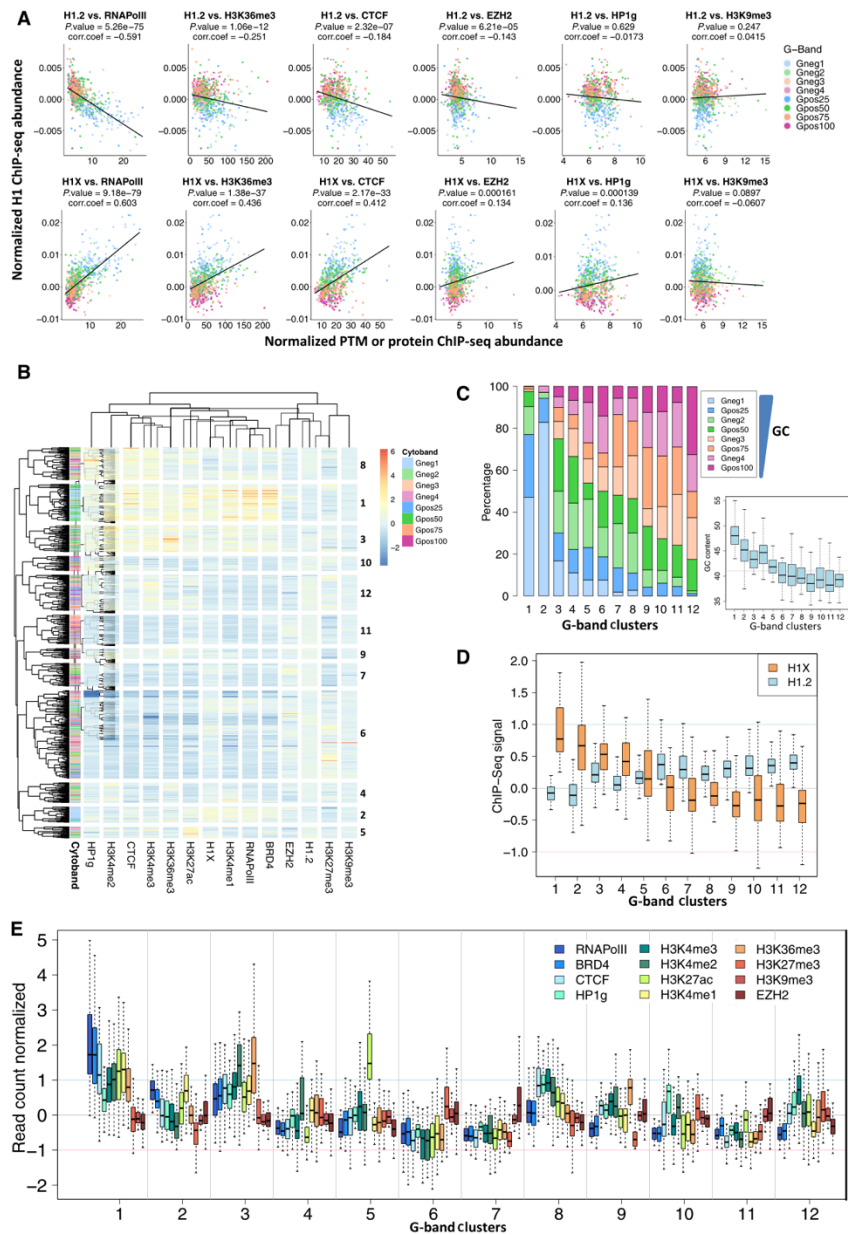
We have previously reported that histone H1 variants distribute heterogeneously along the human genome in T47D breast cancer cells, being H1.2 the variant that is more abundant within closed and intergenic regions, and H1X the most abundant within RNA polymerase II-enriched regions [15,16]. Then, we interrogated whether the abundance of these two H1 variants differed among G bands. In a genome browser, it was apparent that H1X was enriched at Gneg1 and Gpos25, while H1.2 was more abundant at Gneg4 and Gpos100, suggesting a relation with GC content

(Fig. 3A). Indeed, H1X was enriched at G bands with high GC content, while H1.2 was rich at low GC bands, both G-positive and G-negative (Fig. 3B). As a consequence, H1.2 and H1X abundance at both types of G bands correlated inversely (Fig. 3C). The positive correlation between H1X and the GC content of bands was stronger than the inverse correlation between H1.2 and GC content (correlation coefficient 0.92 versus -0.64) (Fig. 3D). Differences in H1 variants abundance at stained versus not-stained G bands were enhanced at 400 bphs compared with 850 bphs (Fig. 3E), as it occurred with GC content (Fig. 2C). H1.2 was significantly enriched at stained bands, and H1X was more abundant at nonstained bands.

Parallel to profiling the distribution of endogenous H1.2 and H1X with variant-specific antibodies, we had profiled H1.0 and H1.4 C-terminally tagged with the hemagglutinin (HA) peptide, stably expressed in T47D cells, with anti-HA antibodies. H1.4-HA and H1.0-HA distribution was similar to H1X and different to H1.2 [15,16]. We calculated the abundance of these two HA-tagged variants into G bands. We obtained that both were enriched toward high GC bands, being H1.0-HA the one that was more similar to H1X, in agreement with our previous reports (data not shown). Moreover, using published data on H1.0 profiling in human skin fibroblast [14], we determined that H1.0 was enriched at high GC Gpos and Gneg bands as well (Fig. 3F). Therefore, we decided to focus on endogenous H1.2 and H1X for further studies, as representatives of the different H1 profiles observed.

Because S/MARs and H1X were enriched at high GC bands in a very similar way (Figs 1H and 3B), we compared the abundance of H1.2 and H1X around the center of mapped S/MAR sites. H1X was enriched at S/MAR sites while H1.2 was clearly depleted (Fig. 3G). As mentioned above, S/MARs correlated with retrovirus integration sites [31], which were enriched within high GC bands. Instead, both H1.2 and H1X were found depleted from putative HIV-1 and HTLV-1 integration sites, suggesting that retroviruses integrate at H1-depleted loci (Fig. 3H and data not shown). This showed that H1X was not enriched at all features that are enriched

Fig. 4. Clustering of G bands according to H1 variants and epigenetic features. (A) Scatter plots of H1.2 or H1X input-subtracted ChIP-seq abundance against abundance of the indicated histone marks or chromatin-associated proteins at each Gpos and Gneg band. Pearson's correlation coefficient is shown as well as P -value. (B) Heat map and dendrogram of the abundance of H1 variants, histone marks, and chromatin-associated proteins at Gpos and Gneg bands. Twelve clusters of G bands are shown, ordered from high to low proportion of high GC content bands (Gneg1 + Gpos25 + Gneg2 + Gpos50). (C) Bar plot showing the proportion of each G band type within the 12 clusters of bands generated in (B). GC content at clusters is also shown. (D) Box plots showing H1.2 and H1X input-subtracted ChIP-seq abundance within G bands, for each G bands cluster. (E) Box plots showing abundance of the indicated histone marks or chromatin-associated proteins within G bands, for each G bands cluster.



within high GC bands, but functional selectivity exists. Next, we calculated the abundance of H1s around the binding sites of the proteins that were used to define S/MAR [31]. H1X was more abundant than H1.2 at all S/MAR protein binding sites, but different profiles were observed (Fig. 3I and data not shown). For CEBPB, YY1, Ku antigen, and FOXH1, H1X was locally enriched around the center of the binding site, but not for the others (CTCF, NMP4, BRIGHT, BRCA1, SAF-A/hnRNP-U, SATB1, SMAR1). H1.2 was depleted from all tested sites. This was previously observed for RNA polymerase II [15,16]. All together, these data confirmed that H1X is present at places where transcription and replication initiate.

Clustering of G bands according to H1 variants and epigenetic features

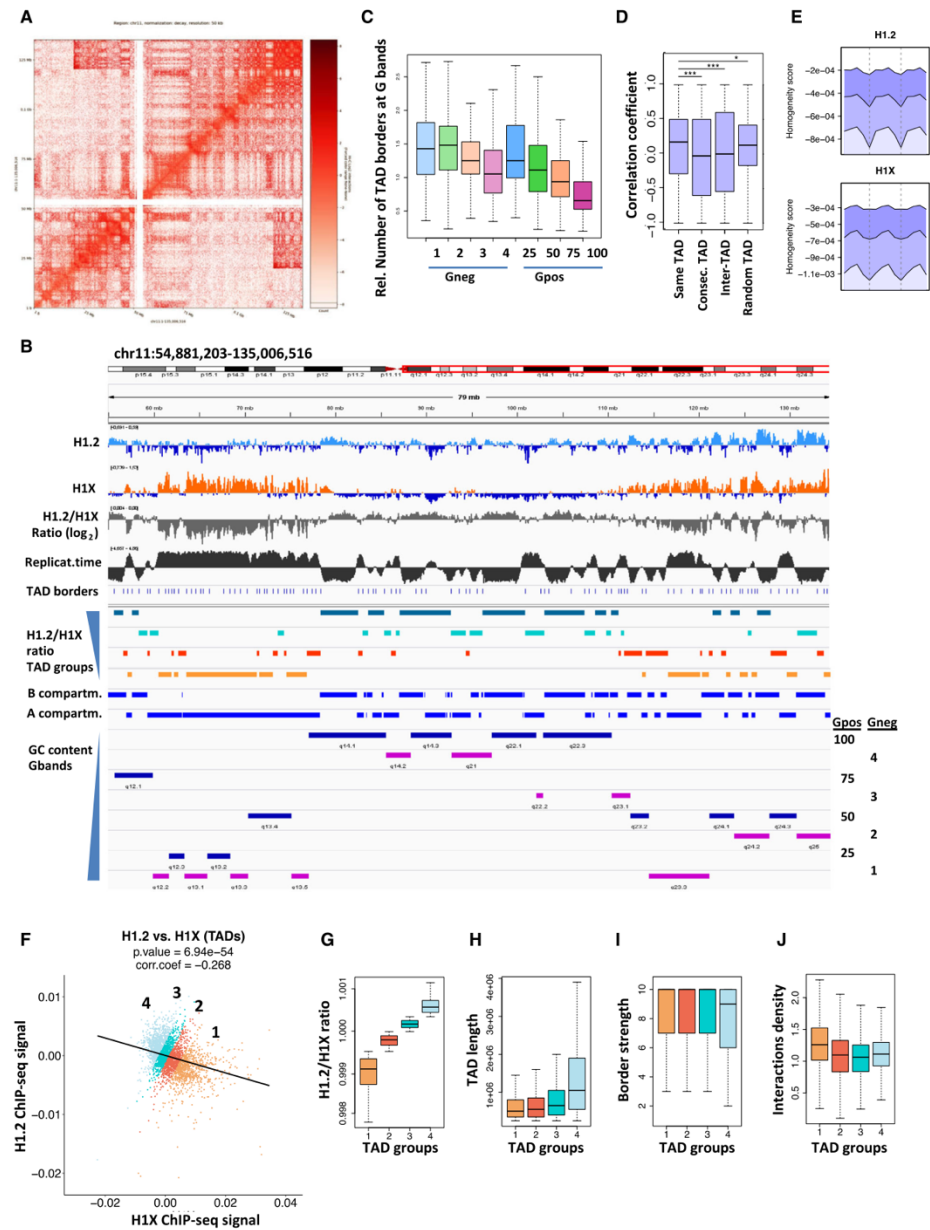
To study the colocalization of the different H1 variants with epigenetic factors within the G bands, the abundance of H1.2 and H1X at Gpos and Gneg bands in T47D cells was compared with the abundance of PTMs and chromatin-associated factors (Fig. 4A and Fig. S1A,B). H1.2 correlated negatively with active histone marks, such as H3K27ac, H3K4me1, or H3K36me3, and RNAPII, BRD4, or CTCF. Significant negative correlation was also observed between H1.2 and EZH2, related to transcriptional repression. The repressive marks H3K9me3 or H3K27me3 showed no correlation with H1.2 abundance, nor HP1 gamma. Instead, H1X correlated positively with all histone marks and chromatin-associated factors tested, except H3K9me3 and H3K27me3. All these results were

similar when the abundance of H1s and PTMs at Gpos or Gneg bands was used separately (Fig. S1C–F) and confirms that H1.2 localizes at inactive G bands whereas H1X is more abundant at high GC content bands enriched in active chromatin.

Next, the calculated abundance of H1 variants, core histone marks, and chromatin-associated factors at Gpos and Gneg bands was used to cluster the G bands and, consequently, compartmentalize the human genome according to epigenetically relevant features (Fig. 4B). Active marks, RNAPII, CTCF, and H1X clustered together, as did H3K9me3, H3K27me3, and EZH2 with H1.2. Next, G bands were clustered into 12 groups with 6 clusters enriched in active epigenetic features and 6 in repressive marks. Each cluster contained a different proportion of G band types; clusters were named from 1 to 12 according to decreasing proportion of high GC content bands (Fig. 4C). As expected, GC content decreased along the defined clusters (see insert in Fig. 4C).

Next, the abundance of H1 variants and epigenetic features at G bands contained in each of the 12 clusters was calculated (Fig. 4D,E). H1 variants increased or decreased progressively according to the GC content of the bands included in each cluster, particularly H1X, as H1.2 was similarly abundant at clusters 6 to 12. Clusters 1 to 4 were enriched in H1X while clusters 6 to 12 were enriched in H1.2. RNAPII or active histone marks were enriched toward the high GC content clusters, in particular clusters 1 to 3, but also cluster 8. Repressive marks or EZH2 was enriched in clusters 6, 7, and 9 to 12. Interestingly, cluster 2 contained predominantly Gneg1 bands and was enriched in H1X,

Fig. 5. Clustering of TADs according to its content in histone H1 variants. (A) Hi-C interaction map of chromosome 11 in T47D cells, at the resolution of 50 kb. The map is normalized, corrected by decay, and in Log2 scale. (B) Representative IGV snapshot of human chromosome 11 (partial). Tracks refer as follows (from top to bottom): H1.2 and H1X input-subtracted ChIP-seq signal from T47D cells; the calculated H1.2/H1X ratio (log2) over 100-kb bins; replication timing of the genome from T47D cells (smoothed signal of early/late S-phase read counts in 5 kb windows); TAD borders obtained by Hi-C in T47D cells; the extension of TADs classified into four groups according to H1.2/H1X ratio as described in (F–G); the extent of A/B compartments obtained by Hi-C; and the position of Gpos and Gneg bands. (C) Box plot showing the number of TAD borders within each G band, corrected by band length, for each G band type. (D, E) TADs as homogeneous units of H1 variants abundance. (D) Distributions of pairwise correlation coefficients of H1 profiles between 100-kb genome bins located within the same TAD, within consecutive or randomly picked TADs (inter-TADs), or within a similar randomly defined domain ($***P < 0.001$; $*P < 0.05$; Wilcoxon test). (E) Homogeneity score of linker histones enrichment between consecutive subsegments over three successive TADs. For this analysis, TADs were divided into five subsegments of equal size. The opposite of the absolute difference of the H1 variants ChIP-Seq signal was calculated for two consecutive subsegments on three consecutive TADs. Higher scores indicate higher similarity between the consecutive subsegments. The 25th, 50th, and 75th percentiles (black lines from top to bottom, respectively) of the 14 consecutive values were computed genome-wide. Dashed lines correspond to the TADs borders. (F) Scatter plot of H1.2 and H1X ChIP-seq abundance at each individual TAD. Pearson's correlation coefficient is shown as well as P -value. TADs corresponding to the four groups defined in (G) according to H1.2/H1X ratio are differentially colored. (G) Box plot showing the ChIP-seq H1.2/H1X ratio within TADs in the four groups generated with equal count of TADs in each. (H–J) Box plots showing the base pairs length (H), border strength (I), and interactions density (J) of TADs belonging to the four groups defined according to H1.2/H1X ratio.



RNAPII, H3K4me1, also in BRD4 and H3K27ac, but not other active marks. Cluster 1 was clearly the most active one with absence of repressive features, while cluster 6 showed the highest abundance of repressive features and absence of active ones. Cluster 8 was enriched in particular features that formed a cluster in the dendrogram: CTCF, HP1 gamma, and H3K4me2/3. Cluster 5 was highly enriched in H3K27ac, whereas it was quite neutral on the rest of active/inactive features, including H1.2 and H1X.

In summary, clustering of G bands according to epigenetic features and H1 variants content compartmentalized the human genome and identified different types of chromatin units. Interestingly, when clusters were ordered according to the abundance of high GC bands or GC content, H1 variants decreased or increased progressively, something not clearly seen for the other epigenetic features or proteins, indicating that histone H1 best correlates with the GC content of the genome.

Overlap between TADs defined by the abundance of H1 variants, G banding, A/B compartments, replication timing, and ATAC-seq accessibility regions

Chromosome conformation capture techniques, such as Hi-C [22], allow to detect local and distal contacts within the genome and to establish the position of borders flanking the so-called TADs. We performed Hi-C experiments in T47D cells, and we calculated the position of TADs within the genome, obtaining a total of 3247 TADs. Figure 5A shows the normalized Hi-C interaction map of chromosome 11 at the resolution of 50 kb as an example.

The comparison of the positions of TADs and G bands denoted that often the limits of G bands were in close proximity to TAD borders (Fig. 5B); therefore, we further investigated the coincidences between

these two features and in relation to H1 variants abundance. First, we calculated the frequency of TAD borders that fell into each of the G bands normalized by their length. Gneg bands and, in general, high GC bands, showed a higher relative frequency of TAD borders than Gpos100 (Fig. 5C). Moreover, Gpos100 is longer on average than other G bands (Fig. 1C). As a consequence, Gneg and high GC Gpos bands are shorter and contain several short TADs, whereas Gpos100 (and Gpos75) contains one or a few long TADs (Fig. 5B).

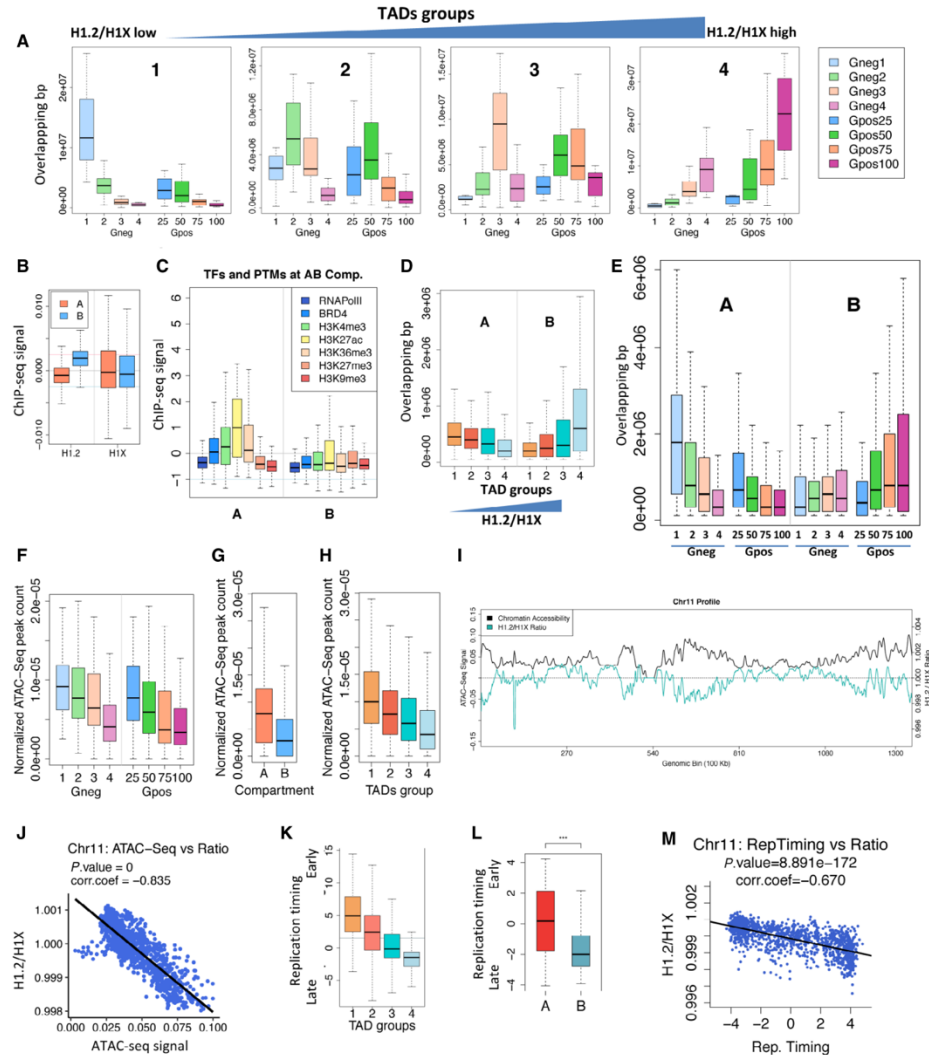
We observed that shifts on the distribution of H1 variants often coincided also with TAD borders (Fig. 5B). Before using TADs as units to compare the distribution of H1.2 and H1X variants, we asked whether this distribution (calculated within 100-kb bins) was more homogeneous within the same TAD than between consecutive TADs, randomly picked TADs or randomly defined domains. Correlation coefficient between the two H1 variants was significantly higher within the same TAD than any other comparison, suggesting that H1 variants were more homogeneous within than between TADs and that transitions between variants occurred preferentially at the borders (Fig. 5D). Besides, we performed 5000 randomizations of TAD borders to further confirm whether the relationship between H1s occupancy depends on these genomic units. Our results showed that the average correlation coefficient between the histones was significantly higher within the real TAD borders compared with the distribution of the average correlation coefficients calculated for the random domains (data not shown). This hypothesis was additionally tested by dividing TADs into subsegments and computing a homogeneity score of linker histones enrichment, which was higher between intra-TAD subsegments (Fig. 5E). Given that TADs and G bands tend to overlap (Fig. 5B), we also performed this analysis for G bands. We found that linker histones distribution

Fig. 6. Overlap between TAD groups defined by H1.2/H1X ratio, G bands, A/B compartments, ATAC-seq accessibility regions, and replication timing. (A) Box plot showing overlapping base pairs between TADs classified according to H1.2/H1X ratio (from low, Group 1; to high, Group 4) and the G bands. (B) Box plot showing the occupancy of H1.2 and H1X variants (input-subtracted ChIP-seq signal) within A/B compartments. (C) Boxplots showing the average normalized read count of the peaks mapped at A or B compartments of each histone PTM or chromatin-associated protein indicated. (D) Box plot showing overlapping base pairs between TADs classified according to H1.2/H1X ratio (Groups 1 to 4) and the A/B compartments ($N_A = 1098$, $N_B = 1098$). (E) Box plot showing overlapping base pairs between G bands and the A/B compartments. (F–H) Box plots showing the relative number of ATAC-seq peaks within G bands (F), A/B compartments (G), or TADs classified according to H1.2/H1X ratio (Groups 1 to 4), normalized by TAD length (H). (I) Profiles of ATAC-seq accessibility and H1.2/H1X abundance ratio along chromosome 11, calculated within 100 kb bins. (J, M) Scatter plots between ChIP-seq H1.2/H1X abundance ratio and ATAC-seq accessibility (J) or replication timing (M) within 100-kb bins along chromosome 11. Pearson's correlation coefficient is shown as well as *P*-value. (K, L) Box plot showing the T47D replication timing (ENCODE) (normalized by TAD length) within TADs classified according to H1.2/H1X ratio (Groups 1 to 4) (K), or within A/B compartments (L).

was more homogeneous within the same G band than between consecutive, alternate, or within similar random genomic regions (data not shown).

Next, we calculated the abundance of H1.2 and H1X within each TAD (Fig. 5F). As expected, an

inverse correlation was observed. The ratio between H1.2 and H1X abundance was calculated for each TAD and used to generate four equal groups of TADs, from low to high H1.2/H1X ratio (Fig. 5G). TADs with a high H1.2/H1X ratio, presumably more



compacted, were much larger in average (Fig. 5H) and were enriched in Gpos bands, especially Gpos100 (Fig. 6A). Instead, TADs with the lowest H1.2/H1X ratio were enriched in high GC bands, mainly Gneg1. In fact, TADs with similar H1.2/H1X ratios are seen as clusters that resemble the G bands (Fig. 5B). Long stretches of TADs with a high H1.2/H1X ratio greatly overlap with Gpos100 bands and so on. This allows us to propose that G bands extension and staining correlate with the relative abundance of two histone H1 variants with opposite genomic distribution and are related to the topology of the genome, which has been proposed to be highly conserved between cell types, as occurs for G banding.

Hi-C data allow to compute the relative strength of each TAD border and the relative intra-TAD interactions density in which each TAD is involved. TAD border strength was slightly higher in TADs with the lowest H1.2/H1X ratio (Fig. 5I). Those TADs also presented a major abundance of TADs with a high interaction density (Fig. 5J). Border strength and interactions density within TADs correlated positively (correlation coefficient = 0.274, *P*-value < 0.001). In conclusion, TADs with low H1.2/H1X ratio, the GC-rich ones, are better defined according to their border strength and present a higher relative number of interactions given their size, as expected from open chromatin genome regions. TADs with high H1.2/H1X ratio, within AT-rich G bands, are not defined as well and present less interactions, probably because they are immersed in closed chromatin regions, as shown below.

Hi-C experiments also allow to establish a division of the genome into two compartments, A (active) and B (repressive). We hypothesized that A/B compartments could also reflect differences in the abundance of H1.2 and H1X and maybe greatly overlap with the stretches of TADs defined by the H1.2/H1X ratio or even with the G bands staining (Fig. 5B). From our Hi-C data, we established the A/B compartments and calculated the abundance of the H1 variants in each A or B compartment fragment. B compartment was greatly enriched in H1.2, whereas H1X was only slightly increased in A compartment (Fig. 6B). Instead, A compartment was enriched in active histone H3 marks and transcription factors (Fig. 6C). B compartment was highly enriched in the group of TADs containing a high H1.2/H1X ratio. Instead, A compartment was enriched in TADs with low H1.2/H1X ratio (Fig. 6D). As expected, B compartment greatly overlapped with the Gpos bands (Gpos75 and Gpos100), whereas A compartment overlapped with high GC bands (Gneg1, Gneg2, and Gpos25)

(Fig. 6E). Moreover, G bands that present a higher base pair overlap with the B compartment showed a higher AT content (correlation coefficient = 0.56, *P*-value < 0.001).

Next, we used accessibility data of T47D cells previously obtained by ATAC-seq [10] to calculate its overlap with G bands, A/B compartments, and TADs classified according to H1.2/H1X ratio. High GC content G bands showed a major density of accessibility peaks (Fig. 6F). Interestingly, Gpos50, 75, and 100 were particularly deprived of accessibility peaks. As predicted, A compartment was also enriched in high accessibility regions compared with B compartment (Fig. 6G). Moreover, accessibility peaks were enriched within TADs presenting a low H1.2/H1X ratio, denoting that H1.2-rich TADs are more compact (Fig. 6H). This was further confirmed by profiling along chromosomes the ATAC-seq accessibility and ChIP-seq H1 variants abundance within 100-kb bins; it was evident that H1X correlates strongly with accessibility, while H1.2 or the H1.2/H1X ratio correlated negatively with accessibility (Fig. 6I,J and data not shown). This reinforces the relationship of H1.2 and H1X with repressed and active genomic regions, respectively.

Table 1. Summary of chromatin and topology features of high and low GC cytobands.

	Gneg1	Gpos25	Gneg4	Gpos100
GC content	High		Low	
Giemsa staining	Unstained	Positive (Light)	Unstained	Positive (Dark)
Repetitive elements	SINEs		LINES	
Replication Timing	Early		Late	
Histone modifications	Active		Repressive	
Chromatin Domains/sites	RNApol II binding sites, S/MARs		LADs	
Gene density	Dense		Poor	
Gene expression	High		Low	
Chromatin accessibility	Accessible		Compact	
Histone H1 variants (T47D)	H1X		H1.2	
Genome compartment	A		B	
TADs	Low H1.2/H1X Ratio		High H1.2/H1X Ratio	
TAD length, num. TADs per Gband	Short, High		Long, Low	
TAD border strength, interactions density	High, High		Low, Low	

Finally, we studied whether replication timing correlated with all features described here. We showed above that high GC cytobands replicated earlier than low GC bands (Fig. 1D). Replication timing represented in a browser formed clusters that clearly overlapped the TAD clusters defined by H1.2/H1X ratio, genome compartments, and G bands (Fig. 5B). Late replicating regions overlapped with TADs enriched in H1.2 and the B compartment (Fig. 6K,L). A strong inverse correlation existed between replication timing and the H1.2/H1X ratio within 100-kb bins along chromosomes (Fig. 6M).

In summary, topological domains enriched in H1.2 compared with H1X or other variants correspond to poorly accessible, late replicating regions that overlap with the B compartment of the 3D genome and with the low GC Giemsa bands of the metaphase chromosomes (Table 1).

Correlation between epigenetic scores and chromatin accessibility within G bands clustered according to H1 variants and epigenetic features

Taking advantage of the topological and accessibility data available, we further analyzed the 12 G bands clusters generated to compartmentalize the genome using histone PTMs, H1 variants, and chromatin proteins (Fig. 4B). First, we calculated the base pair overlap between A/B compartments and the G bands included in each of the 12 clusters. Bands within clusters 1 to 5 with high GC content (abundant Gneg1, Gneg2, and Gpos25) and high abundance of H1X and active marks were located mainly in A compartment. Bands within clusters 8 to 12 with low GC content (abundant Gpos100, Gpos75, and Gneg4) and high H1.2 abundance showed major overlap with the B compartment (Fig. 7A).

Further, we represented in a 3D plot the twelve G bands clusters according to their H1.2/H1X ratio, a calculated compartment B/A ratio, and a repressive or heterochromatic 'epigenetic score' obtained from the ratio between the average abundance of repressive versus active histone marks or chromatin factors (Fig. 7B). As expected, clusters with high H1.2/H1X ratio also showed high B/A compartments ratio and repressive epigenetic score, that is, clusters 9 to 12. Nonetheless, this representation denoted particularities of several clusters that have been described above. Cluster 6 presented the highest repressive epigenetic score, and cluster 8 an epigenetic score lower than expected according to its H1.2/H1X and B/A ratios. Among the clusters with low H1.2/H1X and B/A ratios, that is, 1 to 4, there is some heterogeneity on

the epigenetic score, being clusters 1 and 3 those showing the highest proportion of active marks (Figs 4E and 7B,D).

When computing the ATAC-seq accessibility within the 12 G bands clusters described above, it was observed that initial clusters enriched in H1X and located within the A compartment were more accessible than clusters of bands enriched in H1.2 (Fig. 7C). Still, the best correlation of accessibility occurred with the calculated epigenetic score; clusters 1, 3, and 8 presented the lowest repressive epigenetic score (or the highest active/euchromatic epigenetic score) and the highest accessibility (Fig. 7C,D). Pairwise correlations between the different parameters studied in the 12 clusters of G bands confirmed that the active/euchromatic epigenetic score correlated the best with ATAC-seq accessibility (Spearman's correlation = 0.76), but not as well with GC content, A compartment or H1X (Fig. 7E).

In conclusion, by first dividing the heterogeneous Gneg bands in four groups according to GC content and, later, all Giemsa bands into 12 clusters according to the abundance of H1 variants and other epigenetic features, we functionally compartmentalized the genome in a way that allowed to search for correlations with accessibility and topological data. Each cluster contained G bands of different types that presented common features. The GC content within each cluster was not more homogeneous than inside each of the five original Giemsa bands categories (Gneg, Gpos25–100), indicating that GC content was not the main parameter dictating clustering once epigenetic features were used. All together, we propose that the clustering made here including H1 variants may be useful to identify and characterize different functional chromatin units inside the human genome.

The overlap between H1.2-rich TADs, the B compartment, and AT-rich G bands is extensive to mouse ESCs

Finally, we asked whether the correlations described here were extensive to other cell types or species. Different H1 variants correlate with high or low GC content in different studies [4,11–16]. Genomic localization data on H1.2 are not available elsewhere, except for DamID studies of H1 variants in IMR90 human fibroblasts and ChIP-seq of tagged variants in knock-in mouse ESCs [11,12]. In both cases, H1.2 was abundant at low GC DNA as in T47D cells. We obtained available data on mouse ESCs H1 ChIP-seq and Hi-C [11,33], together with the coordinates of mouse Giemsa bands from UCSC server, to test the

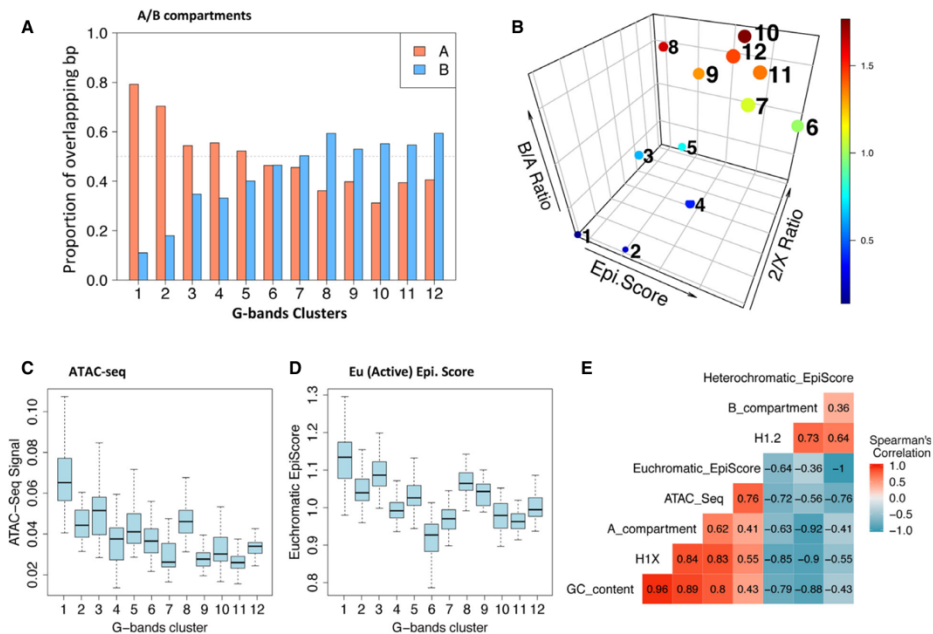
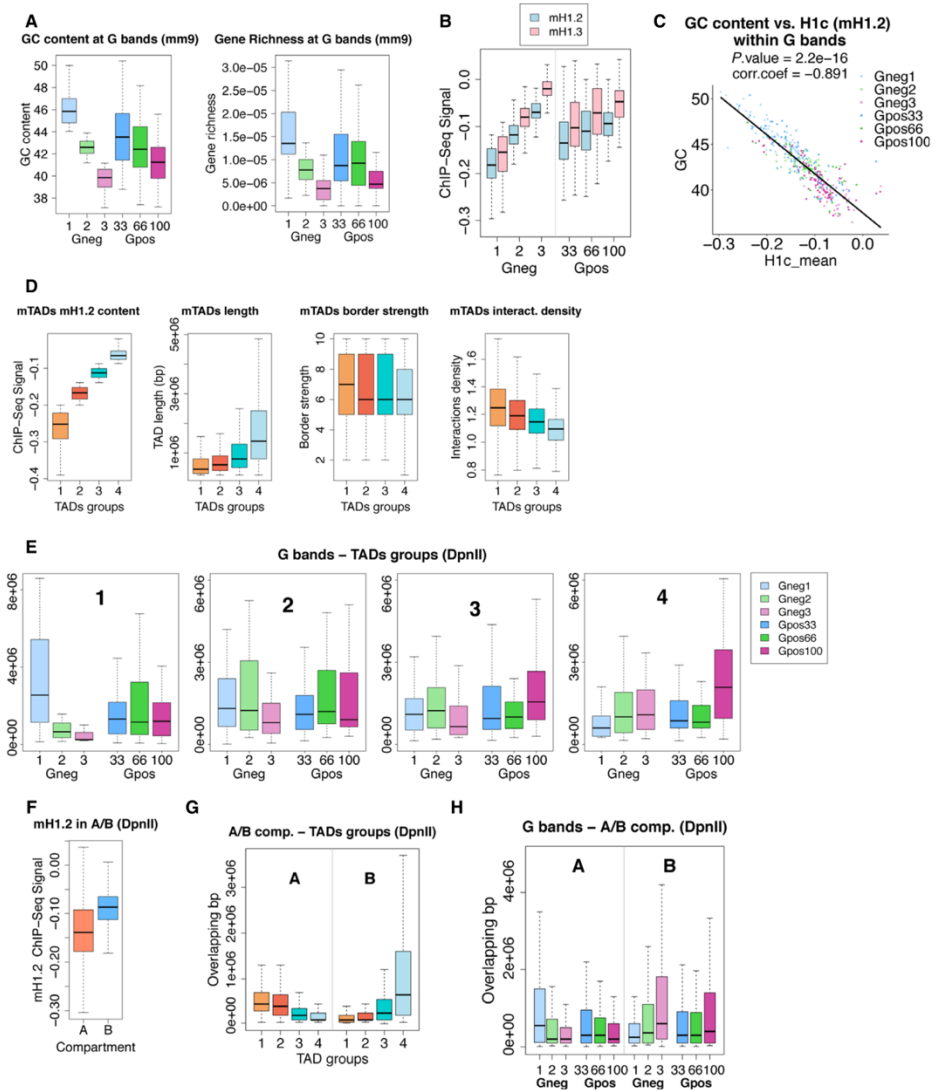


Fig. 7. Correlations between epigenetic scores, H1 variants abundance, and chromatin accessibility within G bands clusters. (A) Proportion of overlapping base pairs between A/B compartments and the G bands clustered according to histone marks, H1 variants, and chromatin proteins (Clusters 1 to 12; Fig. 4). (B) 3D plot of G band clusters according to its H1.2/H1X ratio, a calculated B/A compartments ratio, and a repressive 'epigenetic score' obtained from the ratio between the average abundance of repressive versus active histone marks or chromatin factors. Color scale refers to the B/A compartments ratio, and size of dots refers to the H1.2/H1X ratio. (C) ATAC-seq accessibility within the twelve G bands clusters. (D) Active/euchromatic epigenetic score within the twelve G bands clusters, calculated as the inverse of the repressive epigenetic score defined for B, for better comparison with ATAC-seq accessibility. (E) Correlation matrix of the different parameters studied in the 12 clusters of G bands. The graph shows the pairwise correlation coefficient between the average within the clusters of the following variables: GC content, ATAC-seq signal, euchromatic and heterochromatic epigenetic scores, H1.2 and H1X abundances, and A and B compartments overlapping.

Fig. 8. Overlap between TAD groups defined by H1.2 content, G bands, and A/B compartments from mouse ESCs. (A) Box plots showing the GC content and gene richness (gene number/base pair length) of each mouse G band type. Mouse G-positive bands are classified into three types (Gpos33 to Gpos100) according to increasing staining intensity. Gneg bands were divided into three equal groups according to GC content. (B) Box plots showing mouse Myc-H1.2 (H1c) and Flag-H1.3 (H1d) input-subtracted ChIP-seq abundance from mESCs (GSE46134) within G bands, for each band type. (C) Scatter plot of mouse H1.2 ChIP-seq abundance and GC content at each individual G band. Pearson's correlation coefficient is shown as well as *P*-value. (D) Box plots showing mouse H1.2 ChIP-seq abundance, TAD length, border strength, and interactions density of TADs (*N* = 2460) from mouse ESCs (GSE75426) divided into four groups according to their H1.2 content. (E) Box plot showing overlapping base pairs between TADs classified according to mouse H1.2 content (from low, Group 1; to high, Group 4) and the mouse G bands. (F) Box plot showing the occupancy of mouse H1.2 (input-subtracted ChIP-seq signal) within A/B compartments obtained by Hi-C in mESCs cells (GSE75426). (G) Box plot showing overlapping base pairs between TADs classified according to mouse H1.2 content (Groups 1 to 4) and the A/B compartments (*N*_A = 1367, *N*_B = 1418). (H) Box plot showing overlapping base pairs between G bands and the A/B compartments.

described correlations. Mouse G-positive bands are classified into four groups: 85 Gpos33, 44 Gpos66, 1 Gpos75, and 83 Gpos100 bands. We divided G-negative 190 bands into three equal groups according to

their GC content. Gpos and Gneg bands with the lowest GC content (Gneg3 and Gpos100) presented the lowest gene richness (Fig. 8A) and the highest abundance of mouse H1c (H1.2) and H1d (H1.3) (Fig. 8B).



H1.2 correlated negatively with GC content within G bands (Fig. 8C). Because H1X was not profiled in mESCs, we were unable to calculate the H1.2/H1X ratio. From the wild-type mESCs Hi-C data, we calculated the location of TADs and A/B compartments using the same protocol used in T47D cells. Abundance of H1s within individual TADs was calculated and four groups of TADs were generated according to the H1.2 content (Fig. 8D). TADs enriched in H1.2 were longer and presented low border strength and interactions density (Fig. 8D). TADs with the highest H1.2 content (group 4) were enriched at low GC bands, particularly Gpos100, whereas TADs with the lowest H1.2 content (group 1) were enriched at Gneg1 bands (Fig. 8E). Abundance of H1s within compartments was also calculated. H1.2 was enriched at the B compartment (Fig. 8F). Moreover, TADs with a high H1.2 content were enriched at the B compartment (Fig. 8G), and this compartment was enriched on low GC Gpos bands (Fig. 8H). Altogether, these results confirmed that the overlap between TADs enriched in histone H1.2 (among others), the B compartment, and gene-poor, AT-rich Giemsa bands is also observed in mouse ESCs and we anticipate that it might be, at least, widely extended. A remaining issue is which mammalian H1 variants accumulate at high and low GC compartments, in different cell types, to establish whether the variant preference is universal or depends on cell type or differentiation stage, or on H1 variants content. From the data available up to date, H1.2 is preferentially located at low GC, compacted or inactive regions. Whether H1X or other variants prefer high GC, active regions, extensively, needs further studies.

Discussion

It is well established that the eukaryotic genome is topologically compartmentalized inside the nucleus at several levels including chromosome territories, active and inactive compartments, TADs, and loops [34]. Initial evidences of the nonhomogeneous nature of the interphase genome came from different physico-chemical techniques that identified two major forms of chromatin, euchromatin and heterochromatin, with distinct compaction properties and location inside the nucleus, back to the 1960s. In the 1970s, several staining methods of metaphase chromosomes identified characteristic and well-conserved bands that later were associated with different features or sequences of DNA, including GC content. Here, we have combined available data on mapping of Giemsa bands and ChIP-seq data on epigenetic features with our histone H1 variants ChIP-

seq, Hi-C, and ATAC-seq data in breast cancer cells to fully characterize the overlap between genome compartments defined by these classical and state-of-the-art high-throughput methodologies. By comparing the location of G bands to chromatin accessibility maps (ATAC-seq) and the location of the A and B compartments and TADs (Hi-C) classified according to the relative abundance of different H1 variants, we have found strong correlations that support the biological relevance of these techniques to establish different compaction/activity states of the genome compartments. Besides, we demonstrate that genomic properties of compartments established in the interphase genome are in agreement with those shown by the characteristic banding of metaphase chromosomes, and vice versa. This supports the reversibility of chromosome architecture through the cell cycle, which may be sustained by the retention of architectural proteins (CTCF, cohesins) allowing the recovery of the original interphase chromatin loop structure at the end of mitosis [35].

In our previous studies, we mapped somatic H1 variants in breast cancer cells to study their specific genomic distribution. To date, specific ChIP-grade antibodies were only available for human H1.2 and H1X variants, so, for the remaining variants, HA-tagged H1 variants were overexpressed in the cells [15]. Regarding endogenous H1.2 and H1X, data uncovered some specific features for both variants. More recently, we realized that patches of enrichment of H1.2 and H1X greatly overlap with the classical chromosomal bands resulting from Giemsa staining (G bands). In this work, we have characterized G bands at several epigenetic levels to use them as genomic units to compartmentalize the genome and evaluate histone H1 variants genomic distribution (Table 1). High GC bands are enriched in active histone marks, RNA polymerase II and SINEs, and associate with gene richness, gene expression, and early replication. Low GC bands are enriched in repressive histone marks, LADs, LINEs, and late replication domains. Our results support a heterogeneous distribution of histones H1.2 and H1X within G bands that is reinforced at highly condensed chromosomes. Thus, H1.2 was found enriched in low GC bands whereas H1X was more abundant at high GC bands. From our data on HA-tagged H1 variants or elsewhere data available, we have shown that H1.0 and H1.4 are also enriched at high GC bands. Consequently, evaluating the abundance of H1 variants within G bands allows to easily compare the genomic preferences of different variants within a cell type, or to compare a variant between cell types.

We rapidly realized that both G-positive and G-negative bands were heterogeneous and not highly differentiated among them in all features investigated initially, including GC content, gene richness, replication timing, epigenetic marks, and histone H1 variants content. Gpos bands were already categorized according to staining intensity (Gpos25–Gpos100), and this was inversely correlated to GC content, gene richness, replication timing, SINES, S/MARs, active core histone marks, transcription factors, and histone H1X. When Gneg bands were classified into four groups according to GC content, we realized they also showed the same correlations, indicating that all features obey to the GC content of regional domains of the genome (Fig. 1). These observations opened a question mark, as Gneg and Gpos bands with similar GC content and epigenetic features stained differently, at least at 850 bphs resolution, while historically it was suggested that Giemsa was staining AT-rich regions [26,27]. To solve this paradox, others suggested that the banding pattern may be related to the differences in GC content between neighboring regions [30]. We observed that Gneg and Gpos bands that were located close to each other presented similar GC contents and, upon chromosome compaction (400 bphs resolution), became stained or remained unstained more consistently with their GC content, that is, neighbor Gpos100 and Gneg4 became stained, and neighbor Gpos25 and Gneg1 did not. In other words, most of low GC Gneg bands (Gneg4) become stained at 400 bphs, while most of high GC Gpos bands (Gpos25) remain unstained. Thus, the correlation of staining with AT content is reinforced at 400 bphs compared with 850 bphs, upon chromosome compaction (Fig. 2C,D). Still, Giemsa banding cannot be explained only by the difference in base composition, especially within the Gneg bands. Instead, GC content correlates with almost every epigenetic and topological feature studied here, specially H1 variants abundance (discussed below).

One difference between Gpos and Gneg bands having a similar GC content was the average band length (Fig. 1C). Gpos100 and Gpos75 bands were longer than any Gneg band. Besides, they contained a reduced number of TAD borders within them, and those TADs with a high proportion of H1.2 were also longer than others (Fig. 5C,H). As a consequence, there was a relatively good overlap between Gpos100 bands and TADs with high H1.2 abundance. In addition, TADs within the B compartment were longer on average than TADs within the A compartment (data not shown). From all these observations, we can conclude that the domains of repressed or compacted

chromatin tend to form longer patches than active or open chromatin. Therefore, heterochromatin is less compartmentalized than euchromatin, and probably, compartmentalization (TAD borders) is needed for the proper regulation of active chromatin and gene expression occurring inside.

In general, differences between Gpos and Gneg bands with a similar GC content increased notably when topological features from the Hi-C data were analyzed. For instance, Gpos100 bands, but not Gneg4, highly overlapped with the B compartment and with TADs enriched in H1.2 (Fig. 6A,E). On the contrary, Gneg1 bands, but not Gpos25, overlapped with the A compartment and with TADs enriched in H1X (low H1.2/H1X ratio). As a consequence, Giemsa staining seems to better correspond to topological and compaction properties of genome domains.

Still, within Gneg or Gpos bands, topological features correlated to some extent with their GC content. For instance, within Gneg bands that were classified entirely based on GC content herein, their overlap with the A compartment, or with the different TAD groups based on H1.2/H1X ratio, depended greatly on GC content. Whether GC content is a prior determinant of the epigenetic and topological features of genomes, or the base pair composition of the genome has evolved as a consequence of the existence of compartments with high or low activity/accessibility, is an interesting issue that would need further debate. Assuming that low GC content is favorable for compaction, if a region is under functional constraint to maintain a compact chromatin structure, an increase in GC content would be selectively disadvantageous or an increase in AT content would be advantageous. Alternatively, GC to AT derive through evolution may occur spontaneously more often at inactive/compact regions.

We have also described that S/MARs, which in general are AT-rich sequences, are densely present in both Gpos and Gneg high GC bands. DNA molecules that are rich in AT stretches are flexible and prone to strand separation, properties needed for S/MAR functions, but these elements do not need to be immersed in AT-rich bands or domains. Apparently, S/MARs are short AT-rich stretches within GC-rich environments such as the high GC cytobands, where gene expression occurs and replication starts. S/MARs and H1X follow a similar distribution within G bands, so it would be interesting to further investigate which is the involvement of histone H1X in the function of S/MARs and, in general, in controlling gene expression and replication. We already reported that H1X is enriched at RNAPII binding sites [16]. Now, we have

found that H1X is enriched around S/MAR proteins binding sites, while H1.2 is deprived.

Another interesting observation we made was that histones H1.2 and, especially, H1X correlate with the GC content of G bands, both Gpos and Gneg, more consistently than any other epigenetic feature we investigated (i.e., core histone marks, transcription factors, etc.). This is still clearer when we generated 12 clusters of G bands according to epigenetic features including H1 variants. Upon classifying them according to the decreasing proportion of high GC bands and, consequently, decreasing GC content, H1X also decreased proportionally and H1.2 increased, but the other features did not follow a clear pattern across the 12 clusters although there was a tendency. Active marks and transcription-related proteins accumulated over repressive ones at the initial clusters, and the opposite occurred toward the final clusters. This behavior may be due to the fact that histones H1 distribute uniformly along chromatin as every nucleosome may contain one linker histone and, consequently, each variant may paint a particular G band or chromosome domain uniformly according to its characteristics and GC content. Transcription factors and most of core histone marks occupy better defined positions at promoters, enhancers, coding regions, etc., and some variability may exist within a G band despite having some general behavior dictated by GC content and location within chromosome territories, among others. Obviously not all genes within a G band may be in the same state, especially because their transcriptional activity depends on the expression program of each cell type at every moment of the development or in response to diverse stimuli. Instead, the nature of G bands and even chromosome territories seems to be widely conserved across cell types.

Clustering of G bands according to epigenetic features and H1 content was a useful method to compartmentalize the genome, similar to previous initiatives based on epigenetic profiling of the genome divided in size-defined bins, resulting in defined clusters that were named the 'colors' of chromatin [36,37]. Here, genome segments (G bands) are much longer but the compartmentalization method proposed, although based on Giemsa staining, indirectly underlies multiple functional properties, including GC content. Further, this is the first time that H1 variants with different distribution have been used as an epigenetic feature. This method gave rise to several clusters with particular combinations of epigenetic features that might be functionally relevant and would need further investigation.

Moreover, we represented in a 3-axis diagram the characteristics of these 12 clusters based on a

repressive epigenetic score, its H1.2/H1X ratio, and a calculated compartment B/A ratio that was useful to identify clusters where the three parameters correlate, and clusters where some of the parameter deviates from the expected result, allowing to further identify and characterize particular regions of the genome. Therefore, the methods described here allow combining epigenetic data with topological information to better investigate the diversity that may be found within genome compartments. Notably, GC content, H1 variants content, and overlap with A/B compartments showed a strong correlation among clusters, whereas the epigenetic score (calculated from the abundance of histone marks and chromatin factors) presented the best correlation with ATAC-seq accessibility. This suggests that the first parameters may be related to the division of the genome in the classical euchromatin and heterochromatin compartments, and the second group of parameters may be occurring due to local changes in chromatin related to genome functions including gene expression.

Our previous studies showed that combined H1 depletion in breast cancer cells causes induction of repetitive elements, such as satellites [10]. In this last study, one of the variants depleted was H1.2 that here, we have found to be enriched in B compartment and compact TADs, characteristics presumably associated with heterochromatin. Moreover, Hi-C data in H1 triple knockout ES mouse cells revealed that reduced levels of histone H1 result in altered epigenetic and topological organization at the most active chromosomal domains [33]. Altogether, these data suggest that histone H1 levels are crucial for maintenance of the global genome topological organization, both at active and at inactive compartments. Indeed, our data show that H1.2 and H1X inversely correlate with genome topology parameters, so it is reasonable to hypothesize that altering H1 variants homeostasis could have different consequences on genome topology, in a H1-variant-dependent manner. This work supports the notion of H1 variants functional specificity, not only at the linear level but also in correspondence with the 3D genome.

We have found that H1.2/H1X ratio is closely related to G bands and genome topology. Both G banding and genome topology are expected to be highly conserved among different cell types, but this is not happening with H1 variants distribution. Several studies point to a cell type-specific distribution of H1 variants [4,11–16], so further research will be needed to elucidate if H1.2/H1X ratio correlation with G bands and topology found in breast cancer cells is maintained across cell types. If not conserved, other

H1 variants could be responsible for the mentioned correlation, in a cell type-specific manner. We believe that an extensive study of the abundance and genome distribution of all H1 variants in different cell types would be of great interest to understand H1 function and specificity in genome organization. In mouse ESCs, H1.2 and H1.3 present a similar distribution, enriched at low GC regions [11]. We have shown that TADs enriched in H1.2 are longer, present low interactions density, and correlate with the B compartment and AT-rich cytobands, indicating that the model exposed here is extensive to other cell types and species.

In conclusion, our study shows that linker histones are involved in compartmentalization of the genome. We have detected differences between H1 variants distribution within G bands, TADs, and A/B compartments that correlate with the epigenetic landscape as well as with genome sequence properties, such as GC content or the abundance of repetitive elements. Therefore, we hypothesize that H1 variants are organized according to a nonrandom clustering of the genome required to physically delineate regions with distinct functionalities.

Materials and methods

Cells culturing conditions

Breast cancer T47D-MTVL (carrying one stably integrated copy of luciferase reporter gene driven by the MMTV promoter) derivative cells were grown at 37 °C with 5% CO₂ in RPMI 1640 medium, supplemented with 10% FBS, 2 mM L-glutamine, 100 U·mL⁻¹ penicillin, and 100 µg·mL⁻¹ streptomycin, as described previously [38]. These cell lines are a model to study gene expression regulation by steroid hormones and the interplay of chromatin components and states including histone H1.

G bands characterization

Genome-wide GC content and G bands coordinates at 850 bands per haploid sequence (bphs) resolution were obtained from the UCSC human genome database. G bands average GC content was calculated with BEDTools Map to subsequently split Gneg bands into four subgroups according to their decreasing GC content. We used in-house scripts to calculate the G bands percentage of genomic occupancy as well as their average gene content, band length, gene richness, and gene expression.

LINES, SINES, and LADs coordinates were retrieved from the UCSC server. HeLa-S3 and T47D replication timing data, S/MARs coordinates and HIV-1 and HTLV-1 integration sites were obtained from the ENCODE,

MARome [31], and RID [39] databases, respectively. The overlapping coordinates between G bands and these regions were calculated with BEDTools Intersect and subsequently analyzed with in-house R scripts.

Since G bands coordinates at 400 bphs resolution are not available, we computed their expected starting and ending positions merging the bands at 850 bphs that give rise to each 400-bphs band according to the available ideograms (as an example, bands p24.1, p24.2, and p24.3 give rise to band p24). The properties of 400-bphs G bands, such as average GC content and H1 variants enrichment, were calculated with BEDTools as described previously for the 850 bphs bands. Next, in order to calculate the proportion of consecutive A or T nucleotides per G band, the DNA sequences of the human chromosomes were obtained from the NCBI database. We designed an R script which iterates along chromosome sequences and subtracts the fragment corresponding to each G band. It finally calculates the proportion of 1 to 5 or more consecutive A/T nucleotides at G bands as well as their total average AT content.

H1 variants ChIP-Seq analysis

Histone H1 ChIP-Seq data from T47D included in the Gene Expression Omnibus (GEO) dataset GSE49334 has been reprocessed for this study. Single-end reads were quality-checked via FASTQC v0.11.9 (S. Andrews, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and aligned to the human GRCh37/hg19 reference genome using bowtie2 v2.3.5.1 [40] with default options. Next, SAMTOOLS v1.9 [41] utilities were used to sort the alignments and filter out the low-quality ones with the flag 3844. Input and H1 variant genome coverage was calculated with BEDTOOLS v2.28.0 [42]. Genome coverage was normalized by reads per million and regions with zero coverage were also reported in the ChIP-Seq annotation (*genomcov -ibam -bga -scale* options). MACS2 (Model-based Analysis of ChIP-Seq) v2.1.2 [43] was used to subtract input coverage from H1 variants and to generate signal tracks (*bdgcmp -m subtract* option). We used BEDTools Map to determine the enrichment of histone H1 variants within the eight groups of G bands. ChIP signals around the center of S/MARs and HIV-1/HTLV-1 integration sites were calculated by using 'Sitepro' script of CEAS package [44] with normalized input-subtracted-average tags in 50-bp bins in a set window.

PTMs and chromatin-associated proteins analysis

We conducted our epigenetic analysis for T47D cells by downloading and reprocessing PTMs and chromatin-associated proteins raw data from the GEO database. GEO accession numbers are GSE109229 (RNAPolII, BRD4), GSE41617 (H3K4me1, H3K4me3), GSE120162 (CTCF, H3K9ac, H3K27ac), GSE63109 (H3K4me2, H3K9me2,

H3K36me3), GSE64467 (HP1 γ), and GSE29611 (EZH2, H3K27me3). ChIP-Seq reads were processed as described [45] with minor modifications. Briefly, reads were aligned to the reference human genome (GRCh37/hg19) using BOWTIE2 v2.3.5.1 with default parameters. Mapped reads were sorted and filtered to discard the low-quality ones with SAMTOOLS. HOMER (Hypergeometric Optimization of Motif EnRichment) v4.11 [46] was used to call peaks using an input from T47D cells as a control. The '-style histone' option was specified for PTMs and the '-style factor' option for transcription factors and some specific histone marks which are known to develop narrow peaks (e.g., H3K4me3 or H3K9ac). The enrichment of PTMs and chromatin-associated proteins within G bands was calculated by mapping the normalized read count onto G bands with BEDTools Map.

Clustering of G bands

We designed an R script to calculate the Pearson's correlation between H1 variants and the analyzed epigenetic factors within G bands, to establish the 12 clusters of bands and to finally characterize them. Specifically, we computed the clusters' Gpos and Gneg bands proportion and we used the previously generated files to study the distribution of the GC content, the H1 variants, and the epigenetic factors. The packages pheamap, ggplot2, and plot3D were used to visualize the results.

In situ Hi-C analysis

Hi-C libraries were generated from untreated derivative T47D cells as previously described [47,48]. In brief, adherent cells were cross-linked with 1% formaldehyde in PBS for 10 min at room temperature and glycine 0.125 M was added for 5 min at room temperature and for 15 min at 4 °C to stop the crosslink reaction. Before permeabilization, cells were treated for 5 min with trypsin. Nuclei digestion was performed with 400 units of MboI restriction enzyme. The ends of restriction fragments were labeled using biotinylated nucleotides and ligated with T4 DNA ligase. After reversal of crosslinks, DNA was purified and sheared (Diagenode BioruptorPico, Seraing, Belgium) to obtain 300–500 bp fragments and ligation junctions were pull down with streptavidin beads. Hi-C libraries were finally amplified, controlled for quality, and sequenced on an Illumina HiSeq 2500 sequencer (Illumina, Inc., San Diego, CA, USA).

Hi-C data preprocessing, normalization, and generation of interaction matrices

The analysis of Hi-C data, from FASTQ files mapping to genome segmentation into A/B compartments and TADs, was performed using TADBIT software [49]. TADbit pipeline starts by performing a quality control on the raw data in FASTQ format. Next, sequencing reads were mapped to the

reference genome (GRCh37/hg19) applying a fragment-based iterative strategy and using the GEM MAPPER [50]. Mapped reads were filtered to remove those resulting from unspecified ligations, errors, or experimental artifacts. Specifically, nine different filters were applied using the default parameters in TADbit: self-circles, dangling ends, errors, extra dangling ends, over-represented, too short, too long, duplicated, and random breaks [49]. Hi-C data were normalized with OneD correction [51] at the resolutions of 1 Mb, 500 kb, 100 kb, and 10 kb, to remove Hi-C biases and artifacts. Filtered read-pairs were binned at the resolutions of 1 Mb, 500 kb, 100 kb, and 10 kb, applying biases from the normalization step and decay correction to generate interaction matrices.

Hi-C data on T47D breast cancer cells have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE147627.

Genome segmentation into topologically associating domains

We identified TADs at the resolution of 50 kb using TADbit with default parameters. TADbit segments the genome into constitutive TADs after analyzing contact distribution along the genome. TADbit employs a BIC-penalized breakpoint detection algorithm based on probabilistic interaction frequency model that returns the optimal segmentation of the chromosome [52]. This algorithm leads to a ~99% average genome coverage. In the output, TADbit also describes TADs border strength and TADs density. TADs border strength is the algorithm likelihood corresponding to each border (the higher the strength, the higher the algorithm confidence). TADs density represents the number of interactions within each TAD compared with the others (the higher the density, the higher the number of interactions within the TAD).

Genome segmentation into A/B compartments

We segmented the genome into A/B compartments at 100 kb resolution on OneD-normalized and decay-corrected matrices, using HOMER software [46]. Briefly, HOMER calculates correlation between the contact profiles of each bin against each other and performs principal component analysis (PCA) on chromosome-wide matrices. Normally, A compartment is assigned to genomic bins with positive first principal component (PC1), and B compartment is assigned to genomic bins with negative PC1.

Computing the overlap between G bands, TADs, A/B compartments, and ATAC-Seq regions

BEDTools Map was used to calculate the average H1.2 and H1X enrichment within TADs and A/B compartments while the overlapping coordinates between G bands, TADs, and A/B compartments were computed with BEDTools

Intersect. We also computed the H1.2 and H1X abundance within 100-kb bins to confirm that those located within the same TAD are more homogeneous in their H1 variants content than bins located within consecutive or alternate TADs or within similar random domains. Then, we used R to define four groups of TADs according to their increasing H1.2/H1X ratio and calculate their average length, border strength, and interactions density. We also developed a function to calculate the overlapping base pairs between two sets of intersected coordinates and therefore calculated the total overlapping nucleotides between G bands, the four groups of TADs, and the A/B compartments. This function was also used for calculating the overlapping base pairs between A/B compartments and the G bands included in each of the 12 clusters. The overlapping coordinates between the ATAC-Seq peaks and the four groups of TADs were calculated with BEDTools Intersect to subsequently compute the average number of peaks per TAD.

ATAC-Seq analysis

We reprocessed our ATAC-Seq data identified by the accession number GSE100762 as described [53] with slight modifications. Paired-end sequencing reads were quality-checked via FASTQC v0.11.9, trimmed, and subsequently aligned to the human GRCh37/hg19 reference genome using BOWTIE2 v2.3.5.1. SAMTOOLS v1.9 was used to sort and filter out the low-quality alignments with the flag 1796, remove reads mapped in the mitochondrial chromosome, and discard those reads with a MAPQ score below 30. The peak calling was performed with MACS2 v2.1.2 by specifying the *-BAMPE* mode. Filtered BAM files were also used to compute the ATAC-Seq genome coverage, which was normalized by reads per million (*bedtools genomecov -ibam -bga -scale* options). BEDTools Map was used to compute the average ATAC-Seq signal within 100-kb genomic bins as well as within G bands.

Analysis of data on mouse ESCs

mESCs GC content, G bands coordinates, and transcript annotation were obtained from the UCSC database while data on genome 3D organization and H1 variants distribution were downloaded from the GEO server. FASTQ files from Hi-C experiments performed in mESCs (GSE75426) were processed as described before for human T47D cells to compute TADs and A/B compartments coordinates. Processed input-subtracted ChIP-Seq files (GSE46134) were used to calculate the average abundance of histones H1c and H1d within G bands, TADs, and A/B compartments by using BEDTools utilities.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation [BFU2017-82805-C2-1-P to

AJ, BFU2017-85926-P to MAM-R (AEI/FEDER, UE)]. This research was partially funded by the European Union's Seventh Framework Programme ERC grant agreement 609989 to MAM-R, European Union's Horizon 2020 research and innovation program grant agreement 676556 to MAM-R. We also acknowledge the Generalitat de Catalunya Suport Grups de Recerca AGAUR 2017-SGR-597 to AJ and 2017-SGR-468 to MAM-R. CRG acknowledges support from 'Centro de Excelencia Severo Ochoa 2013–2017', SEV-2012-0208, and the CERCA Programme/Generalitat de Catalunya. We acknowledge Generalitat de Catalunya for an AGAUR-FI predoctoral fellowship [to MS-P and to FM].

Conflict of interest

The authors declare no conflict of interest.

Author contributions

NS-P and MS-P designed research, performed the experiments, analyzed data, and wrote the paper; NL-A, FT-L, FM, and MAM-R analyzed data; AI-B performed the experiments; AJ designed research, analyzed data, and wrote the paper; and NS-P, MS-P, MAM-R, AI-B, and AJ contributed to discussion.

References

- 1 Bednar J, Horowitz RA, Grigoryev SA, Carruthers LM, Hansen JC, Koster AJ & Woodcock CL (1998) Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc Natl Acad Sci USA* **95**, 14173–14178.
- 2 Happel N & Doenecke D (2009) Histone H1 and its isoforms: contribution to chromatin structure and function. *Gene* **431**, 1–12.
- 3 Izzo A, Kamieniarz K & Schneider R (2008) The histone H1 family: specific members, specific functions? *Biol Chem* **389**, 333–343.
- 4 Millán-Ariño L, Izquierdo-Bouldstridge A & Jordan A (2016) Specificities and genomic distribution of somatic mammalian histone H1 subtypes. *Biochim Biophys Acta* **1859**, 510–519.
- 5 Fyodorov DV, Zhou B-R, Skoultschi AI & Bai Y (2018) Emerging roles of linker histones in regulating chromatin structure and function. *Nat Rev Mol Cell Biol* **19**, 192–206.
- 6 Laybourn PJ & Kadonaga JT (1991) Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. *Science* **254**, 238–245.

- 7 Almeida R, Fernández-Justel JM, Santa-María C, Cadoret JC, Cano-Aroca L, Lombrana R, Herranz G, Agresti A & Gómez M (2018) Chromatin conformation regulates the coordination between DNA replication and transcription. *Nat Commun* **9**, 1590.
- 8 Bayona-Feliu A, Casas-Lamesa A, Reina O, Bernués J & Azorín F (2017) Linker histone H1 prevents R-loop accumulation and genome instability in heterochromatin. *Nat Commun* **8**, 283.
- 9 Glaich O, Leader Y, Lev Maor G & Ast G (2019) Histone H1.5 binds over splice sites in chromatin and regulates alternative splicing. *Nucleic Acids Res* **47**, 6145–6159.
- 10 Izquierdo-Bouldstridge A, Bustillos A, Bonet-Costa C, Aribau-Miralbés P, García-Gomis D, Dabad M, Esteve-Codina A, Pascual-Reguant L, Peiró S, Esteller M *et al.* (2017) Histone H1 depletion triggers an interferon response in cancer cells via activation of heterochromatic repeats. *Nucleic Acids Res* **45**, 11622–11642.
- 11 Cao K, Lailler N, Zhang Y, Kumar A, Uppal K, Liu Z, Lee EK, Wu H, Medrzycki M, Pan C *et al.* (2013) High-resolution mapping of H1 linker histone variants in embryonic stem cells. *PLoS Genet* **9**, e1003417.
- 12 Izzo A, Kamieniarz-Gdula K, Ramírez F, Noureen N, Kind J, Manke T, van Steensel B & Schneider R (2013) The genomic landscape of the somatic linker histone subtypes H1.1 to H1.5 in human cells. *Cell Rep* **3**, 2142–2154.
- 13 Li JY, Patterson M, Mikkola HKA, Lowry WE & Kurdistan SK (2012) Dynamic distribution of linker histone H1.5 in cellular differentiation. *PLoS Genet* **8**, e1002879.
- 14 Torres CM, Biran A, Burney MJ, Patel H, Henser-Brownhill T, Cohen AHS, Li Y, Ben-Hamo R, Nye E, Spencer-Dene B *et al.* (2016) The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity. *Science* **353**: aaf1644. <https://doi.org/10.1126/science.aaf1644>.
- 15 Millán-Ariño L, Islam ABMMK, Izquierdo-Bouldstridge A, Mayor R, Terme JM, Luque N, Sancho M, López-Bigas N & Jordan A (2014) Mapping of six somatic linker histone H1 variants in human breast cancer cells uncovers specific features of H1.2. *Nucleic Acids Res* **42**, 4474–4493.
- 16 Mayor R, Izquierdo-Bouldstridge A, Millán-Ariño L, Bustillos A, Sampaio C, Luque N & Jordan A (2015) Genome distribution of replication-independent histone H1 variants shows H1.0 associated with nucleolar domains and H1X associated with RNA polymerase II-enriched regions. *J Biol Chem* **290**, 7474–7491.
- 17 Solovei I, Thanisch K & Feodorova Y (2016) How to rule the nucleus: divide et impera. *Curr Opin Cell Biol* **40**, 47–59.
- 18 Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS & Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380.
- 19 Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Van Berkum NL, Meisig J, Sedat J *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385.
- 20 Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A & Cavalli G (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472.
- 21 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680.
- 22 Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293.
- 23 Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Peterfia B, Solovei I, Cremer T, Dopazo J & Langst G (2010) Initial genomics of the human nucleolus. *PLoS Genet* **6**, e1000889.
- 24 Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951.
- 25 Caspersson T, Lomakka G & Zech L (1971) The 24 fluorescence patterns of the human metaphase chromosomes – distinguishing characters and variability. *Hereditas* **67**, 89–102.
- 26 Comings DE (1978) Mechanisms of chromosome banding and implications for chromosome structure. *Annu Rev Genet* **12**, 25–46.
- 27 Holmquist G, Gray M, Porter T & Jordan J (1982) Characterization of Giemsa dark- and light-band DNA. *Cell* **31**, 121–129.
- 28 Furey TS & Haussler D (2003) Integration of the cytogenetic map with the draft human genome sequence. *Hum Mol Genet* **12**, 1037–1044.
- 29 Singh GB, Kramer JA & Krawetz SA (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res* **25**, 1419–1425.
- 30 Niimura Y & Gojobori T (2002) In silico chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence. *Proc Natl Acad Sci USA* **99**, 797–802.

- 31 Narwade N, Patel S, Alam A, Chattopadhyay S, Mittal S & Kulkarni A (2019) Mapping of scaffold/matrix attachment regions in human genome: a data mining exercise. *Nucleic Acids Res* **47**, 7247–7261.
- 32 Costantini M, Clay O, Federico C, Saccone S, Auletta F & Bernardi G (2007) Human chromosomal bands: nested structure, high-definition map and molecular basis. *Chromosoma* **116**, 29–40.
- 33 Geeven G, Zhu Y, Kim BJ, Bartholdy BA, Yang SM, Macfarlan TS, Gifford WD, Pfaff SL, Versteegen MJAM, Pinto H *et al.* (2015) Local compartment changes and regulatory landscape alterations in histone H1-depleted cells. *Genome Biol* **16**: 289. <https://doi.org/10.1186/s13059-015-0857-0>.
- 34 Cavalli G & Misteli T (2013) Functional implications of genome topology. *Nat Struct Mol Biol* **20**, 290–299.
- 35 Bernardi G (2015) Chromosome architecture and genome organization. *PLoS One* **10**, e0143739.
- 36 Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ *et al.* (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224.
- 37 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330.
- 38 Sancho M, Diani E, Beato M & Jordan A (2008) Depletion of human histone H1 variants uncovers specific roles in gene expression and cell growth. *PLoS Genet* **4**, e1000227.
- 39 Shao W, Shan J, Kearney MF, Wu X, Maldarelli F, Mellors JW, Luke B, Coffin JM & Hughes SH (2016) Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology* **13**, 47.
- 40 Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359.
- 41 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G & Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
- 42 Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.
- 43 Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137.
- 44 Shin H, Liu T, Manrai AK & Liu SX (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605–2606.
- 45 Zhang G, Zhao Y, Liu Y, Kao LP, Wang X, Skerry B & Li Z (2016) Foxa1 defines cancer cell specificity. *Sci Adv* **2**, e1501473.
- 46 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H & Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–589.
- 47 Vara C, Paytuví-Gallart A, Cuartero Y, Le Dily F, García F, Salva-Castro J, Gómez HL, Julia E, Moutinho C, Aiese Cigliano R *et al.* (2019) Three-dimensional genomic structure and cohesin occupancy correlate with transcriptional activity during spermatogenesis. *Cell Rep* **28**, 352–367.e9.
- 48 Pascual-Reguant L, Blanco E, Galán S, Le Dily F, Cuartero Y, Serra-Bardénys G, Di Carlo V, Iturbide A, Cebrià-Costa JP, Nonell L *et al.* (2018) Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. *Nat Commun* **9**, 3420.
- 49 Serra F, Baù D, Goodstadt M, Castillo D, Filion G & Martí-Renom MA (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**, e1005665.
- 50 Marco-Sola S, Sammeth M, Guigó R & Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**, 1185–1188.
- 51 Vidal E, le Dily F, Quilez J, Stadhouders R, Cuartero Y, Graf T, Martí-Renom MA, Beato M & Filion GJ (2018) OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res* **46**, e49.
- 52 Le Dily FL, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, Filion G *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* **28**, 2151–2162.
- 53 Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B *et al.* (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959–962.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Correlations between H1 variants and epigenetic features within G bands.

ANNEX 3

Differential contribution to gene expression prediction of histone modifications at enhancers or promoters

Candidate's contribution: Analysis of the Hi-C experiments.

Mar González-Ramírez, Cecilia Ballaré, Francesca Mugianesi, Malte Beringer, Alexandra Santanach, Enrique Blanco, Luciano Di Croce.
Differential contribution to gene expression prediction of histone modifications at enhancers or promoters. PLoS Comput Biol. 2021 Sep 2;17(9):e1009368.
doi: 10.1371/journal.pcbi.1009368.

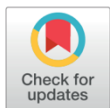
RESEARCH ARTICLE

Differential contribution to gene expression prediction of histone modifications at enhancers or promoters

Mar González-Ramírez¹, Cecilia Ballaré¹, Francesca Mugianesi^{1,2}, Malte Beringer¹, Alexandra Santanach¹, Enrique Blanco¹, Luciano Di Croce^{1,3,4*}

1 Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, **2** CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain, **3** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **4** ICREA, Pg. Barcelona, Spain

* luciano.dicroce@crgeu



Abstract

The ChIP-seq signal of histone modifications at promoters is a good predictor of gene expression in different cellular contexts, but whether this is also true at enhancers is not clear. To address this issue, we develop quantitative models to characterize the relationship of gene expression with histone modifications at enhancers or promoters. We use embryonic stem cells (ESCs), which contain a full spectrum of active and repressed (poised) enhancers, to train predictive models. As many poised enhancers in ESCs switch towards an active state during differentiation, predictive models can also be trained on poised enhancers throughout differentiation and in development. Remarkably, we determine that histone modifications at enhancers, as well as promoters, are predictive of gene expression in ESCs and throughout differentiation and development. Importantly, we demonstrate that their contribution to the predictive models varies depending on their location in enhancers or promoters. Moreover, we use a local regression (LOESS) to normalize sequencing data from different sources, which allows us to apply predictive models trained in a specific cellular context to a different one. We conclude that the relationship between gene expression and histone modifications at enhancers is universal and different from promoters. Our study provides new insight into how histone modifications relate to gene expression based on their location in enhancers or promoters.

OPEN ACCESS

Citation: González-Ramírez M, Ballaré C, Mugianesi F, Beringer M, Santanach A, Blanco E, et al. (2021) Differential contribution to gene expression prediction of histone modifications at enhancers or promoters. *PLoS Comput Biol* 17(9): e1009368. <https://doi.org/10.1371/journal.pcbi.1009368>

Editor: Chongzhi Zang, University of Virginia, UNITED STATES

Received: June 11, 2021

Accepted: August 21, 2021

Published: September 2, 2021

Copyright: © 2021 González-Ramírez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting Information files](#).

Funding: The work in the Di Croce laboratory is supported by grants from the Spanish Ministry of Science and Innovation (BFU2016-75008-P, and PID2019-108322GB-I00), "Fundación Vencer El Cáncer" (VEC), the European Regional Development Fund (FEDER), and from AGAUR (SGR 2017-2019, AGAUR 2019 FI_B 00426). We

Author summary

Gene expression can be properly predicted by the ChIP-seq signal of histone modifications at promoters, but whether this is also true at enhancers is unclear. In this study we develop predictive models of gene expression that demonstrate the predictive power of histone modifications at enhancers in the context of mouse embryonic stem cells, during differentiation, and in animal development. Moreover, by assessing the contribution of each histone modification, we found that enhancer predictive models and promoter

acknowledge support from the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

predictive models have different histone modification requirement. Therefore, different histone modifications relate better to enhancer or promoter function(s). Finally, by applying predictive models trained in a specific cellular context to a different one, we concluded that the relationship between gene expression and histone modifications at enhancers is universal.

Introduction

Appropriate regulation of gene expression is necessary for correct development and homeostasis of organisms. Different classes of regulatory genomic regions are coordinated to establish the appropriate gene transcriptional programs in every cell. These regulatory elements include, among others, promoters and enhancers [1]. Promoters are non-coding DNA fragments located in the surroundings of a transcriptional start site (TSS) that initiate gene transcription, whereas enhancers are distal non-coding DNA fragments that amplify gene expression [1]. DNA is wrapped around histones to form nucleosomes, which are the basic structural unit of chromatin. Post-translational modifications at histones can affect chromatin function by altering its structure, for example by facilitating or preventing the accessibility of transcription factors (TFs) to certain genomic regions [2]. Distinct histone modifications at regulatory elements are associated with gene activation, such as trimethylation of histone H3 at lysine 4 (H3K4me3) [3–5], and acetylation of histone H3 at lysine 27 (H3K27ac) [6], or with gene repression, such as trimethylation of histone H3 at lysine 27 (H3K27me3) [7]. In contrast, monomethylation of histone H3 at lysine 4 (H3K4me1) is a histone modification associated with enhancers [8]. Combinations of histone modifications can have synergistic or antagonist effects on gene regulation. Promoters and enhancers are in fact decorated by a particular combination of different histone modifications according to the transcriptional state of their target gene.

Nowadays, RNA-seq is the main technique to assess gene expression levels, while ChIP-seq experiments allow to map histone modifications genome-wide. Indeed, much effort has been made to understand the quantitative relationship between ChIP-seq levels of histone modifications and gene expression in different cellular contexts [9–15]. However, none of these studies have introduced epigenetic information of enhancers into the modelling for predicting gene expression, but rather have focused only on promoters or gene bodies. Indeed, gene expression has been alternatively modelled using data on chromatin accessibility at promoters in combination to enhancers, together with information about TFs and chromatin remodelers [16]. In this regard, it has been recently shown that including information to the promoter predictive models about chromatin accessibility and the predicted affinity of TF for enhancers can improve the model performance significantly [17]. However, the independent contribution of enhancer and promoter information separately has not been evaluated yet. Although ChIP-seq levels of H3K27ac at enhancers have been modelled with gene expression to (i) obtain predictive models of differential gene expression across tissues and conditions [18], and to (ii) identify enhancer-promoter associations [19], yet to our knowledge, multiple histone modifications exclusively at enhancers have not been used to generate predictive models of gene expression. Therefore, we consider that modelling gene expression from enhancer epigenetic information might help to understand how the contribution to gene expression differs between promoters and enhancers and, more broadly, how enhancers function.

Here, we set out to explore the quantitative relationship between histone modifications and gene expression, focusing on enhancer regions. Our main goal is to decipher which histone

modifications correlate with enhancer function. To do so, we asked the following questions: (i) are histone modifications at enhancers predictive of gene expression? (ii) Which histone modifications are more predictive in the enhancer models? (iii) Are the same histone modifications also important for the promoter predictive models? (iv) Is an enhancer predictive model learned in a specific cell type useful to predict gene expression in another one? To address these issues, we developed a novel computational approach based on the combination of chromatin segmentation and linear regression to infer gene expression using ChIP-seq data from histone modifications at enhancers and promoters. To construct proper predictive models, the full spectrum of active and repressed regions is needed. Therefore, we took advantage of mouse embryonic stem cells (ESCs) for which active and repressed regulatory regions can be identified. ESCs contain active enhancers (AEs) and poised (repressed) enhancers (PEs), which respectively coordinate with active promoters (APs) and bivalent (repressed) promoters (BPs) to regulate gene expression [20].

We first performed ChIP-seq experiments of several histone modifications to identify all four types of regulatory regions and to model gene expression in ESCs. Next, as BPs and PEs can either be activated or remain repressed during later stages of differentiation [20], we have applied our framework to predict gene expression in “in vitro” and “in vivo” differentiated cells. Further, we successfully predicted gene expression in a differentiation time point different from the one in which the model was built. To overcome potential pitfalls of using information from different sources (e.g. cell types, labs, etc.), we applied a normalization approach based on a local regression (LOESS) method. LOESS normalization has shown to be useful for normalizing RNA-seq and ChIP-seq data coming from different sources. We found that histone modification levels at enhancers, as well as at promoters, can predict gene expression of their target genes. Remarkably, we determined that BPs and also PEs are good predictors of gene expression at later stages of differentiation and development. Notably, we also observed that histone modifications have different contributions depending on their location at enhancers or promoters. We propose that the relationship of gene expression and histone modifications at PEs is universal, as we have successfully predicted gene expression in a specific cell type using a model previously trained in another one.

Results

Identification of promoters and enhancers in ESCs

We performed ChIP-seq experiments of H3K4me3, H3K27me3, H3K27ac, and H3K4me1 in mouse ESCs to identify the different types of regulatory regions. This set of histone modifications has been previously used to distinguish between AEs, PEs, APs, and BPs in mouse ESCs [21,22]. We next generated a 9-state chromatin segmentation model of ESCs using ChIP-seq data (Fig 1A and 1B). As expected, active states mark transcriptionally active regions, while repressed states denote transcriptionally repressed regions (Fig 1B; see also our previously published RNA-seq data [23]).

To understand the resulting map of states, we calculated the matrix of transition enrichments between all states in the model. The transition value between two different states, x and y , is defined as the number of times that a segment of state y is found after a segment of state x , as measured from left to right in the linear genome. The enrichment score is defined as the ratio between the observed number of transitions and the expected number of transitions by chance. Two groups of states (active, 1–4, and repressed, 6–8) clearly emerged from the global picture (Fig 1C) (note that state 5 represents a distinct category; see Discussion). The high enrichment of transitions between states belonging to the same category suggests that they might mark the same functional regulatory regions, rather than be caused by different

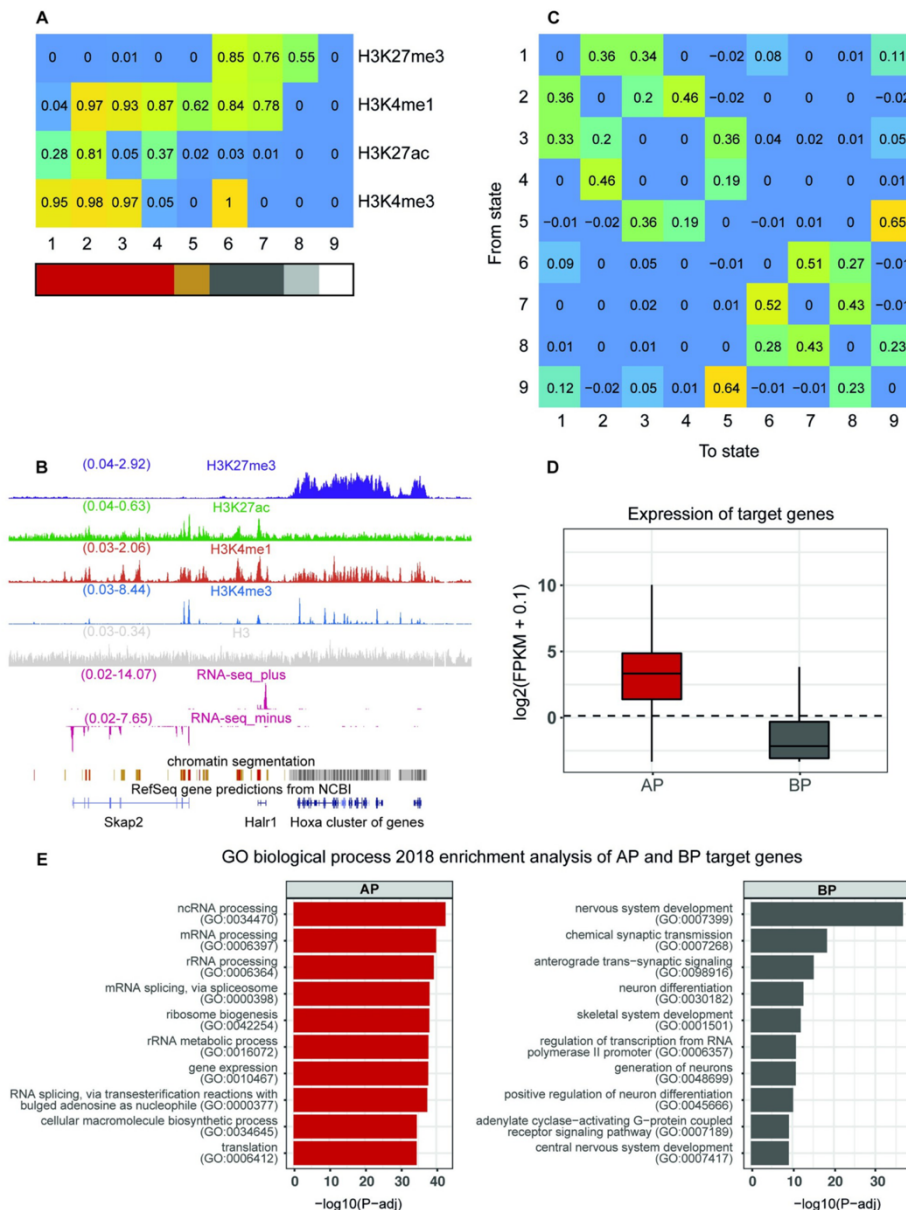


Fig 1. Identification of repressed and active functional regions in ESCs. (A) State definition of the chromatin segmentation model in ESCs. The values represent the probability (from 0 to 1) of finding each histone modification (vertical) in genomic segments of states 1 to 9 (horizontal). The cells of the matrix are colored according to the value of probability they contain inside. Red: states with histone modifications associated to activation (active, 1–4); dark yellow, H3K4me1-only state (Intermediate, 5); grey, states in which H3K27me3 was present (repressed, 6–8); dark grey, poised states, in which H3K27me3 colocalized with H3K4me3 and/or H3K4me1 (states 6 and 7); light grey, H3K27me3-only regions (state 8); and white, unmarked state (9). (B) Example of a genomic region containing two expressed genes (*Skap2* and *Halr1*), which are covered by active states (in red), and a cluster of repressed genes (*HoxA*), which are covered by repressed states (in grey). Active chromatin segments integrate the signal of H3K27ac, H3K4me3, and H3K4me1 and lack H3K27me3. Repressed chromatin segments integrate the signal of H3K27me3, H3K4me3, and H3K4me1 and lack H3K27ac. Expression of *Skap2* and *Halr1*, and silencing of *HoxA* genes, were confirmed by the RNA-seq profiles [23]. Y-axis represents normalized count of reads by total reads. The screenshot was taken from the UCSC Genome Browser [62]. (C) Enrichment of state transitions (e.g., number of observed transitions divided by the number of expected transitions by chance) from the segments of one state (vertical) towards the segments of another state (horizontal) in the linear chromatin. The cells of the matrix are colored according to the value of enrichment they contain inside. (D) Expression of genes associated to active promoters (AP; 10,786 genes) or bivalent promoters (BP; 3,459 genes). The dotted line represents 1 FPKM. (E) Top GO biological process (2018 categories) for each list of genes in D.

<https://doi.org/10.1371/journal.pcbi.1009368.g001>

functional regions separated by the unmarked state. Visual inspection confirmed that states 1 to 4 marked the same active regulatory regions, revealing that differences in the state definition are due to differences in the shape of the ChIP-seq peaks (S1A Fig). We also observed that the repressed states 6 and 7 decorated poised or bivalent regulatory regions (e.g., marked with H3K27me3, in combination with H3K4me3 and/or H3K4me1), whereas state 8 was generated by the tail-end of broad peaks of H3K27me3 (S1B Fig). Similarly, state 5 was associated with the tail-end of broad peaks of H3K4me1 in active regions, but it also associated with single peaks of H3K4me1 near active regions (S1A Fig).

Based on these results, we decided to merge the contiguous segments of states 1 to 4 as a list of potential active regulatory regions, and the segments of states 6 and 7 as a list of potential poised or bivalent regulatory regions. We reasoned that functional regions should have a minimum length of 600 bp and thus discarded shorter regions, which we considered as background signal. We also discarded those cases in which an active region and a poised region were contiguous, as this was ambiguous. Promoters were defined as those regions that overlapped by at least 1 bp to a region ± 500 bp around a TSS according to RefSeq [24]. Enhancers were defined as regions that were not classified as promoters and overlapped by at least 1 bp with a peak of the enhancer mark p300 [21]. As H3K4me3 can be present in enhancers [25–28], we did not discard enhancers containing H3K4me3, although this histone modification has been traditionally only associated with promoters. In total, we found 9,421 APs, 3,344 BPs, 16,904 AEs, and 2,699 PEs (S1–S4 Tables).

Next, we matched our set of promoters to their target genes, using the same parameters as before (overlap by at least 1 bp to a region ± 500 bp around a TSS according to RefSeq [24]). Using RNA-seq data [23], we confirmed that genes associated with APs are expressed, while genes associated with BPs are not (Fig 1D). Gene ontology (GO) term enrichment analysis performed with Enrichr [29] confirmed that genes with APs are involved in housekeeping roles, while genes with BPs are mostly related to development and differentiation (Fig 1E), as is expected in mouse ESCs. On the other hand, we used available high-throughput chromosome conformation capture (3C) data of Hi-C [30], to link AEs and PEs with target genes. As interacting enhancers and promoters have been shown to match their chromatin state [31], we associated an enhancer with the target gene of a promoter when both enhancer and promoter are in the same category (e.g., both active, or both repressed) and each one overlaps with one of the two sides of the same Hi-C significant interaction (total of 43,892,155 significant Hi-C interactions). We confirmed that PEs were significantly enriched in interactions with BPs over APs ($p < 2.2e-16$ Exact Binomial Test, observed probability: 0.29, expected probability by chance: 0.24), whereas AEs were significantly enriched in interactions with APs over BPs ($p < 2.2e-16$ Exact Binomial Test, observed probability: 0.80, expected probability by chance:

0.76). In total, we found 10,786 genes associated to APs, 3,459 genes to BPs, 10,206 genes to AEs, and 2,526 genes to PEs (S1, S2, S5 and S6 Tables). Likewise, 15,841 AEs and 2,466 PEs were associated to at least one gene, and 8,931 APs and 2,443 BPs, to at least one enhancer (S5 and S6 Tables).

Development of a predictive model of gene expression using histone modifications at enhancers

After identifying the set of enhancers and promoters in ESCs, we built the gene expression predictive models. We first performed additional ChIP-seq experiments for other histone modifications, in order to have additional variables to predict gene expression. Specifically, we performed ChIP-seq experiments for trimethylation of histone H3 at lysine 36 (H3K36me3), ubiquitination of histone H2B (H2Bub), monomethylation of histone H3 at lysine 27 (H3K27me1), dimethylation of histone H3 at lysine 27 (H3K27me2), trimethylation of histone H4 at lysine 20 (H4K20me3), and dimethylation of histone H3 at lysine 79 (H3K79me2). The input of our predictive models consisted of the ChIP-seq data of these six histone marks as well as the four histone marks previously used to define promoters and enhancers (H3K4me3, H3K4me1, H3K27ac and H3K27me3), together with our previously-published RNA-seq expression data [23].

Initial studies on gene expression prediction revealed that using two to three histone modifications (rather than larger set) are sufficient to accurately predict gene expression, and do not find substantial improvements with the addition of other histone modifications into the models [9–11]. Indeed, follow up publications directly utilize three to four histone modifications [12, 13]. However, our objective is not only to predict gene expression, but also to assess histone modification contribution in enhancer predictive models in comparison to promoter predictive models. Therefore, we used a set of ten histone modifications, each one with different properties (associated to activation, associated to repression, broad marks, sharp marks, etc.). A total of 11,387 protein coding genes previously associated to a promoter (active or bivalent) and at least one enhancer (active or poised) entered the modelling. We divided the set of genes into two subsets: training and test. The Pearson's correlation coefficient (r) between the measured expression in the test subset and the predicted one was used to assess performance.

As one of our aims is to compare enhancer predictive models to promoter predictive models, we needed to build them using the same approach to identify enhancers and promoters so they are comparable. Therefore, we first generated a predictive model for the promoters identified using our approach (named Hi-C-all promoter model) to confirm that histone modifications at these elements are predictive of gene expression. The predictive capacity of promoters has been previously shown in several cell types from different model organisms, including ESCs, where promoters were defined as a pre-set distance from a TSS [9–15]. As a control, we repeated the predictive model learning in a training subset in which expression values were randomized. We obtained an r -value of 0.81 for the promoter model, and an r -value of -0.07 for the random promoter model (S2A Fig). The low performance of the random promoter model strongly indicated that the high predictive power of the Hi-C-all promoter model was not due to random structures in the data. Importantly, the performance of our promoter model was comparable to previously described predictive models, in which r -values around 0.8 were reported [9–15]. Indeed, Karlic and colleagues obtained an r -value of 0.77 in CD4 + T-cells in the seminal paper on gene expression prediction from histone modification levels [9]. Coefficients and p -values of the predictors for all the predictive models generated in this study can be found in S7 Table.

Next, we trained a second predictive model of gene expression (the Hi-C–all enhancer model) using the levels of histone modifications at the previously-defined enhancers as predictors. We obtained a performance in the test subset of $r = 0.38$ (Fig 2B). Although the r -value of this model is still modest, this is the first time to our knowledge that enhancers have been shown to be predictors of gene expression through their histone modification levels. Critically, when the expression data for learning the model were randomized, the performance was poor ($r = -0.15$, Fig 2B).

We hypothesized that the modest performance of the Hi-C–all enhancer model could be due to some enhancer–promoter associations that were simply regions in close 3D proximity but not functionally linked. To enrich the set of interactions for functional promoter–enhancer loops, we applied a more restrictive threshold on the Hi-C significant interactions ($\text{FDR} = 0$ and $\ln(p\text{-value}) \leq -100$). We obtained a total of 5,555,844 interactions (8% of the total interactions). We then recalculated the enhancer–promoter–gene associations and obtained 1,846 PEs associated to 1,382 BPs and to 1,434 target genes, and 11,777 AEs associated to 7,211 APs and to 8,254 target genes (S8 and S9 Tables). We selected the protein-coding genes included in the new associations (a total of 8,639 genes) to recalculate the predictive models (hereon in termed Hi-C–top models) and random models for promoters (Fig 2A) and enhancers (Fig 2B). While the Hi-C–top promoter model performed similarly to the previous model ($r = 0.79$ vs. $r = 0.81$, respectively), the Hi-C–top enhancer model was significantly improved ($r = 0.49$ vs. $r = 0.38$). This result further confirmed that enhancers, as well as promoters, possess a quantitative relationship with gene expression.

We now know that an enhancer preferentially interacts with promoters located in the same topologically associating domain (TAD) rather than those located in neighboring domains [32]. Moreover, TADs have an average size of around 1 Mb. Therefore, assigning genes to enhancers located in a distance lower than 1 Mb might seem appropriate. Indeed, when evaluating a new ESC predictive model that associates enhancers to promoters of the same chromatin state that are closer than 1 Mb (1 Mb model; 11,986 protein-coding genes), we achieved a performance of $r = 0.34$ (S2C Fig). Nonetheless, that performance is lower than the ESC models based on Hi-C data that we have built previously ($r = 0.38$ and $r = 0.49$ for Hi-C–all and Hi-C–top enhancer models, respectively). This suggests that matching genes to regulatory elements by 1 Mb distance leads to some false-positive associations, yet maintaining its predictive capacity.

Finally, from the Hi-C–top interactions we selected those involving a distal enhancer (> 5 Kb from a TSS, a total of 5,235 AEs and 696 PEs) to confirm that the predictive capacity was not exclusive of proximal enhancers. We generated a new distal enhancer model (Hi-C–top_distal; 7,925 protein-coding genes) that properly predicted gene expression with a $r = 0.41$ (S2D Fig).

Enhancers and promoters exhibit similar histone modification patterns. We therefore wondered whether the histone modifications mostly contributing to the prediction of expression were the same ones as well, or whether there were differences in the contributions between the enhancer and the promoter predictive models. To address this, we assessed variable importance in the Hi-C–top model for promoters and enhancers. Notably, H3K27me3—a histone modification associated with transcriptional gene repression—was the prevalent mark in both classes of regulatory elements (Fig 2C and 2D). In contrast to promoters, in which H3K27me3 has a relatively similar importance as other marks (e.g., H2Bub, H3K4me3, and H3K36me3, Fig 2C), H3K27me3 in enhancers represented up to 55% of the total importance (Fig 2D). Therefore, even though promoters and enhancers contribute to predict gene expression mostly through H3K27me3, this contribution seems to be uniquely driven by H3K27me3 in the

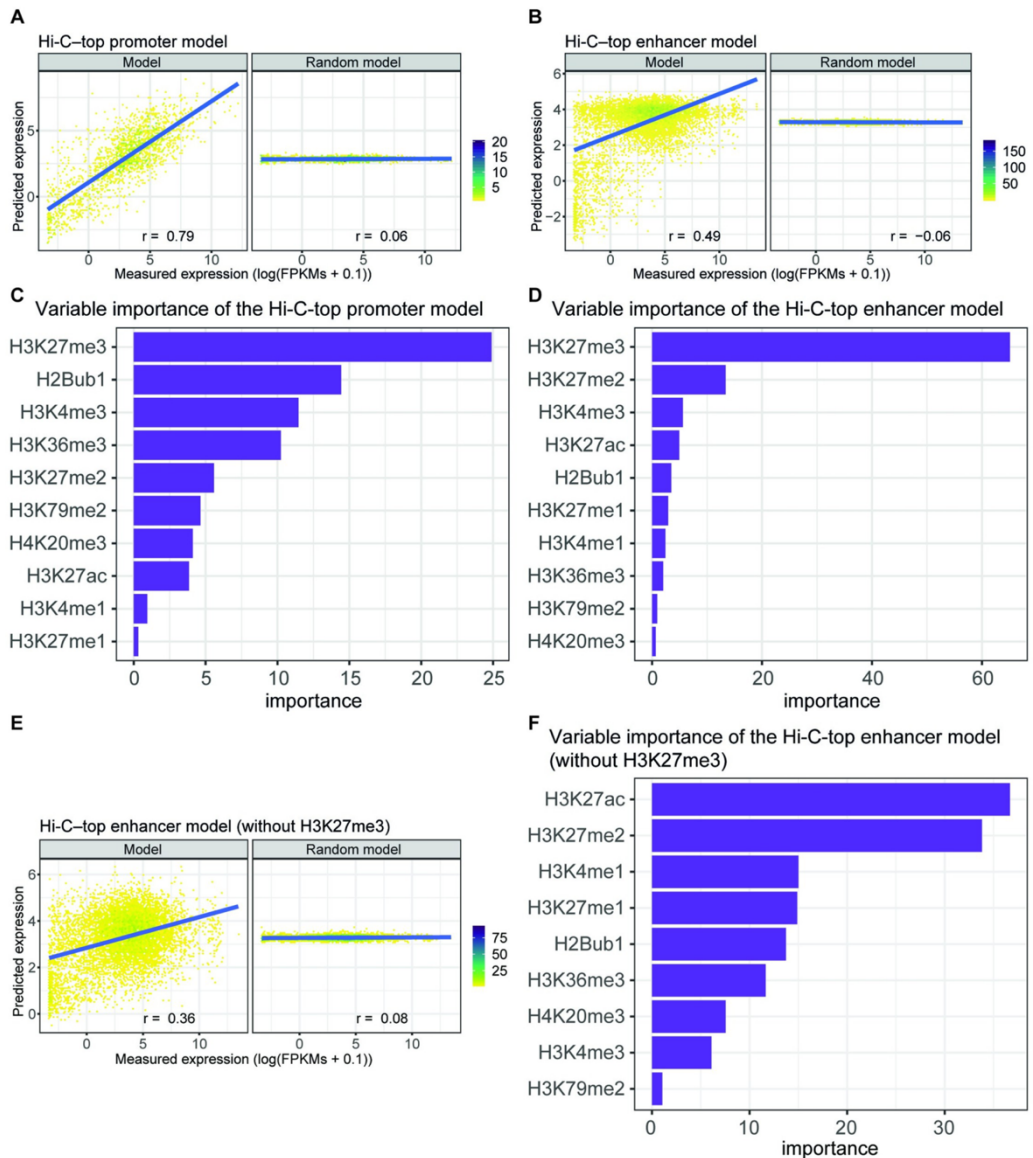


Fig 2. Performance and variable importance of enhancer and promoter Hi-C-top predictive models in ESCs. Predicted expression of the test subset of genes calculated by the models versus their measured expression by RNA-seq. Model performances are represented by the Pearson's correlation (r) between predicted and measured expression values. (A) Left, the model trained on the promoter regions associated to at least one enhancer using the top significant interactions of Hi-C (Hi-

C-top promoter model). Right, the performance of the same model after randomizing the expression of the training subset of genes. The color bar represents the density of dots. (B) Left, the model trained on the enhancer regions associated to at least one promoter using the top significant interactions of Hi-C (Hi-C-top enhancer model). Right, the performance of the same model after randomizing the expression of the training subset of genes. The color bar represents the density of dots. (C) Importance of each histone modification used to train the Hi-C-top promoter predictive model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t-statistic for each model parameter. (D) As for C, but for the Hi-C-top enhancer predictive model. (E) As for B, but the model is trained without H3K27me3 as predictive variable. (F) As for D, but the model is trained without H3K27me3 as predictive variable.

<https://doi.org/10.1371/journal.pcbi.1009368.g002>

enhancer predictive model and shared by other histone marks in the promoter predictive model.

Interestingly, H3K27ac—considered the canonical marker of enhancer activation [33]—had little importance in the enhancer model. H3K27me3 and H3K27ac have antagonistic effects and are generally mutually exclusive marks since they occur on the same lysine and are chemically prohibited [34]. Thus, we reasoned that H3K27ac importance was masked by H3K27me3 presence in the predictive model, as both histone modifications are not independent variables. Indeed, when excluding H3K27me3 as a predictive variable, the enhancer model maintains its predictive capacity ($r = 0.36$; Fig 2E), and the most important variable is in fact H3K27ac (Fig 2F). However, in this case, H3K27ac has a relatively similar importance as H3K27me2. Therefore, H3K27me3 seems more predictive than H3K27ac at enhancers in relationship with the rest of the histone modifications.

LOESS normalization of ChIP-seq and RNA-seq data from heterogeneous sources

We next wanted to determine whether enhancers are predictive of gene expression in other cellular contexts besides ESCs, and whether a predictive model learned in one cell type could predict gene expression in another. As true colocalization of H3K27me3 with H3K4me1 or p300 in the same DNA fragment has been only studied in ESCs, it is still not clear whether PEs exist in other developmental scenarios [20]. To address this issue, we took advantage of the capacity of PEs and BPs to switch into an active state for certain cell types, in a lineage-specific manner during differentiation from ESCs, while remaining inactive in others [20]. We hypothesized that differentiation data could be used to obtain predictive models exclusively from PEs and BPs. We focused on several time points for two cell differentiation mouse models: i) cardiac lineage: mesoderm, cardio precursors, and cardiomyocytes [35]; and ii) neural lineage: neural precursors and cortical neurons [30].

We downloaded RNA-seq and ChIP-seq data of five histone modifications (H3K27me3, H3K4me3, H3K27ac, H3K4me1, and H3K36me3) that were available in the literature for both differentiation models. To remove potential biases (e.g., due to the source of data generation or to a batch effect), we normalized the sequencing samples of the same feature at all the available time points. For this, we applied a normalization based on a local regression (LOESS) that was originally proposed for the pairwise normalization of expression microarrays [36] but generalized for multiple arrays [37]. LOESS normalization is based on a MA methodology, where M is the \log_2 ratio of the intensities of the samples, and A is the \log_2 of the average intensity. It assumes that the intensities of the two samples should be equal, therefore $M = 0$. Finally, corrections based on a LOESS are applied to obtain a MA plot in where the regression line approximates $M = 0$.

We first applied this normalization method over the expression of a set of 20,706 protein-coding genes in the mouse genome, at each differentiation time point, and ESCs (MA plots of each differentiation time point against ESCs before and after LOESS normalization are shown in S3A Fig). As LOESS normalization assumes that expression is equal in all samples, a general balance in global expression distribution of all time points is expected (S3B Fig). To further

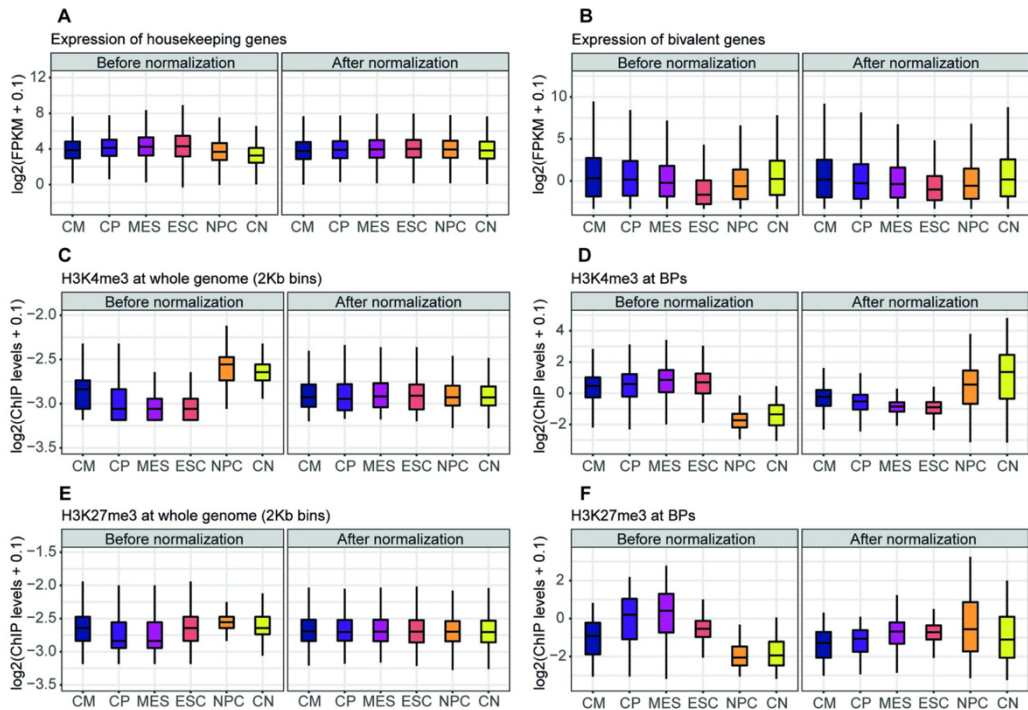


Fig 3. RNA-seq and ChIP-seq data before and after LOESS normalization. (A) Raw and normalized expression of 3,277 housekeeping genes along cardiac and neural differentiation from ESCs. (B) Raw and normalized expression of 3,459 bivalent genes along the same time points as A. (C) Raw and normalized H3K4me3 ChIP-seq signal levels at all 2-Kb bins of the genome. (D) Raw and normalized H3K4me3 ChIP-seq signal levels at 3,344 BPs. (E, F), same as C and D, respectively, but for H3K27me3. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors.

<https://doi.org/10.1371/journal.pcbi.1009368.g003>

confirm the normalization efficiency, we tested its performance on two different subsets of genes: housekeeping genes and bivalent genes. We hypothesized that housekeeping genes would show a balanced distribution of expression after normalization, while bivalent genes would increase their expression globally during differentiation (as some of them are activated). For this, we extracted a list of mouse housekeeping genes across 14 mouse tissues from the literature [38] to check their expression. We also evaluated the normalization on our list of bivalent genes (e.g., those associated to BPs). Indeed, after LOESS normalization, the expression of housekeeping genes was correctly balanced (Fig 3A), whereas the expression of bivalent genes maintained the characteristic pattern of increased expression across time (Fig 3B).

We then ran the same normalization method on the ChIP-seq samples for H3K27me3, H3K4me3, H3K4me1, and H3K36me3 at the same time points (MA plots before and after LOESS normalization over the full set of bins of 2 Kb at Chr19 are shown in S4 Fig). Similar to expression data, we evaluated our normalization method for the H3K4me3 and H3K27me3 ChIP-seq levels across differentiation on two different sets of genomic regions: the whole collection of bins of 2 Kb in which the genome is segmented, and the coordinates of our collection

of BPs. We hypothesized that global ChIP-seq levels of the whole genome would become balanced irrespectively of the particular histone modification analyzed, while BPs should present a different pattern for H3K4me3 and H3K27me3 (e.g., increase and decrease of signal along differentiation time points respectively, as a subset of the bivalent genes becomes activated during time). Indeed, after LOESS normalization, the levels of H3K4me3 along the whole genome were balanced (Fig 3C), whereas the same histone mark at BPs presented a clear pattern of increase during differentiation (Fig 3D). Notably, for H3K27me3, we observed the same balance in the levels along the whole genome after LOESS normalization (Fig 3E), while BPs exhibited a pattern of decreased signal across differentiation, in contrast to that observed for H3K4me3 (Fig 3F). In all cases, therefore, there is a substantial improvement after applying this normalization approach, while the analysis solely based on data before normalization would be misleading.

Poised enhancers and bivalent promoters are good predictors of gene expression during differentiation

After normalizing the data on expression and histone modifications across differentiation, we next generated predictive models using PEs and BPs for each differentiation time point. We used the Hi-C-top interactions involving PEs, BPs, and target genes in ESCs (1,846 PEs and 1,382 BPs associated with 1,434 target genes). From this dataset, a total of 1,063 protein-coding genes were used in the analysis. As the number of genes is smaller than in the previous gene sets, we decided to build the models at each cell type from the full set of genes to be evaluated in the rest of the differentiation time points. This approach has the advantage of allowing us to check whether the relationship between gene expression and histone marks in PEs is universal. This would be true if a model trained in a specific cellular context has a good performance in predicting gene expression in another one. We hypothesized that, as shown previously for promoters and gene bodies [9–11,14], there is a universal relationship between gene expression and histone modifications at PEs.

As a control, we randomized expression data and calculated predictive models for each time point. Next, we evaluated the performance of the randomized models on each differentiation dataset. We observed that predictive models for PEs and BPs obtained a significantly higher performance than randomized controls (Figs 4A, S5 and S6A). Surprisingly, all PE models achieved the best performance in cardiomyocytes (Table 1), suggesting that cardiomyocyte gene expression is easier to predict than the gene expression of the other time points. Moreover, all PE models had similar performances at each time point (Table 1). These observations are also true for the BP models (Table 2). Taken together, our results indicate that there is a universal quantitative relationship between gene expression and histone modifications at PEs and BPs across cardiac and neural differentiation.

In order to confirm the predictive capacity of distal PEs (>5 Kb from a TSS, a total of 696 PEs), we generated new distal PE models (S7 Fig). Indeed, distal PEs maintained the predictive capacity. In this case, 486 protein-coding genes were included in the modelling.

Finally, we assessed the variable importance of the PE models for identifying differences in the contribution of each histone modification to the predictive models in different cellular contexts (Fig 4B). Strikingly, we observed that, in general, the two most important variables are H3K27me3 and H3K36me3. H3K36me3 was the most important histone modification for cardiac differentiation, whereas H3K27me3 was the most important for neural differentiation. In general, H3K27ac followed the above-mentioned histone modifications. H3K4me1 had a relatively low relevance to the predictive models, which suggests that it is involved in delimitating the enhancer regions rather than in contributing to its function. H3K4me3, which was

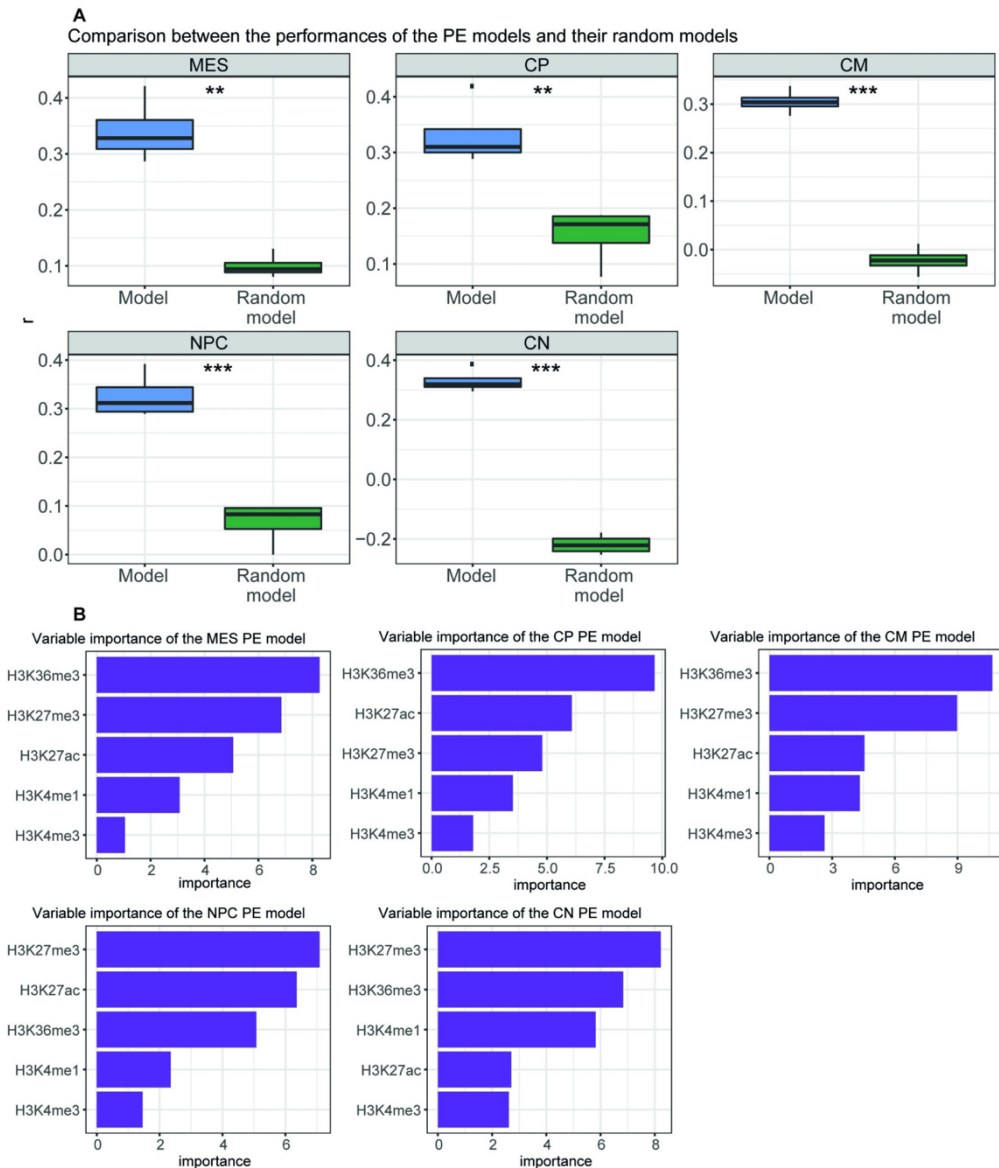


Fig 4. PE models trained in differentiation time points. (A) Performance of each differentiation enhancer model on the rest of the differentiation time points as compared to performance over random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression.

Significance was assessed using a paired Student's *t*-test between the performance of the models and the performance of the random models paired by the differentiation test set (*****p* < 0.0001, ****p* < 0.001, ***p* < 0.01, **p* < 0.05). CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (B) Importance of the histone modifications for each differentiation enhancer model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the *t*-statistic for each model parameter.

<https://doi.org/10.1371/journal.pcbi.1009368.g004>

vastly associated with promoter activity, is accordingly the least informative mark for the prediction of gene expression using PEs, suggesting that H3K4me3 is not associated to enhancer activity. The variable importance in the BPs showed that H3K27me3, H3K27ac and, importantly, H3K4me3 were the most informative variables (S6B Fig). Our results suggest that the quantitative relationship between histone modifications varies according to their location in PEs or BPs. Critically, even though there is a universal quantitative relationship between histone modifications and gene expression, this relationship can vary depending on the cellular context.

As for ESCs, we also assessed the effect of H3K27me3 absence over H3K27ac importance in the current scenario. Therefore, we generated new PE predictive models without H3K27me3 (S8A Fig). Interestingly, H3K27ac was generally the most predictive variable for these predictive models, followed by H3K36me3 (S8B Fig). Therefore, we wondered whether in absence of H3K27ac, H3K27me3 would be also more important than H3K36me3. Indeed, when generating new PE models in absence of H3K27ac (S9A Fig), H3K27me3 had the highest importance in all the cases (S9B Fig). Therefore, H3K27me3 seems the most informative variable for predicting gene expression from PEs.

The importance of H3K36me3 seen in Fig 4B agrees with the fact that almost 60% of the PEs that entered the modelling are located within gene bodies. As H3K36me3 is located in the gene body of active genes [39,40], intragenic enhancers also become marked when the genes start to be expressed during differentiation. Therefore, we divided PEs into two groups, intragenic or intergenic, and built new PE predictive models. Both, intragenic and intergenic models were capable of predicting gene expression (S10A and S11A Figs). When assessing for variable importance, we observed that, as expected, H3K36me3 maintained its high contribution in the intragenic predictive models (S10B Fig). However, H3K36me3 importance was reduced in the intergenic predictive models (S11B Fig). On the contrary, H3K27me3 maintained its importance in both, intergenic and intragenic models. Thus, H3K27me3, and not H3K36me3, behaves as a truly universal predictor of PE activity.

Poised enhancers and bivalent promoters are good predictors of gene expression in mouse embryonic tissues

In order to extend our findings from *in vitro* differentiation to *in vivo*, we learnt predictive PE and BP models from mouse developmental stages at different tissues. We downloaded ChIP-

Table 1. Performance of each PE differentiation model at every differentiation time point.

	MES	CP	CM	NPC	CN
MES model	-	0.34	0.42	0.32	0.29
CP model	0.3	-	0.42	0.32	0.29
CM model	0.3	0.34	-	0.31	0.28
NPC model	0.3	0.33	0.39	-	0.29
CN model	0.3	0.32	0.39	0.32	-

For each time point of cardiac (MES/CP/CM) and neural (NPC/CN) PE models (rows), the performance of the PE predictive models is shown for each cell type (columns). The performance values are represented as Pearson's correlation (*r*) between the measured expression and the predicted one. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors.

<https://doi.org/10.1371/journal.pcbi.1009368.t001>

Table 2. Performance of each BP differentiation model at every differentiation time point.

	MES	CP	CM	NPC	CN
MES model	-	0.78	0.77	0.66	0.77
CP model	0.74	-	0.79	0.67	0.72
CM model	0.72	0.78	-	0.65	0.71
NPC model	0.71	0.77	0.77	-	0.72
CN model	0.73	0.77	0.78	0.69	-

For each time point of cardiac (MES/CP/CM) and neural (NPC/CN) BP models (rows), the performance of the BP predictive models is shown for each cell type (columns). The performance values are represented as Pearson's correlation (r) between the measured expression and the predicted one. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors.

<https://doi.org/10.1371/journal.pcbi.1009368.t002>

seq data of H3K27me3, H3K27ac, H3K36me3, H3K4me3 and H3K4me1 and RNA-seq on mouse embryos (heart tissue from 10.5 embryonic day, liver tissue from 11.5 embryonic day, neural tube tissue from 12.5 embryonic day, kidney tissue from 14.5 embryonic day, and lung tissue from 15.5 embryonic day) from ENCODE [41]. We first normalized the ChIP-seq and expression data following the LOESS approach. A total of 1,087 protein-coding genes entered the analysis. We observed that PEs were also predicting gene expression during mouse embryo development (Fig 5A). Again, when assessing for variable importance, we found that H3K27me3 was contributing the most, followed by H3K27ac and H3K36me3 (Fig 5B).

Next, we confirmed that BPs were also predicting gene expression during mouse embryo development (S12A Fig). In this case, the most predictive variable was H3K27ac followed by H3K27me3 and H3K4me3 (S12B Fig). Therefore, we further confirmed that differences in variable importance between PE models and BP models exist, which suggests that different histone modifications relate better to PE and BP function respectively.

Discussion

To study the full spectrum of active and repressed enhancers, we used ESCs as a model system. Nevertheless, a third class of enhancers, termed intermediate or primed enhancers, exists in this cellular context; intermediate enhancers are decorated with H3K4me1 but lack both H3K27ac and H3K27me3 [21,42]. Originally, intermediate enhancers were classified as PEs [33]; however, more recent publications now use the term PE only for H3K27me3-marked enhancers [21,22,42–44]. The intermediate enhancer signature was found in our chromatin state model (state 5). We decided to focus only in AEs and PEs, though, as intermediate enhancers remain poorly understood, and their target promoters have not been unambiguously identified. In the near future, intermediate enhancers could be introduced in the modeling to explore their impact on the performance of the predictions and to discover new relationships between histone modifications at enhancers and gene expression. However, this is not a limitation for our differentiation predictive models. Here, enhancers and promoters are required to be in either a poised or a bivalent state in ESCs, but many will transition towards an active, or even intermediate state, along cardiac and neural *in vitro* differentiation, and along embryo development. Therefore, in the subset of PEs during differentiation, we have assessed the dynamics of a complete spectrum of enhancers in cardiac (mesoderm/cardio precursors/cardiomyocytes) and neural (neural precursors/cortical neurons) cells, and in developmental tissues (heart, liver, neural tube, kidney and lung). However, it is worth mentioning that our conclusions likely only apply to this system where PEs have been described.

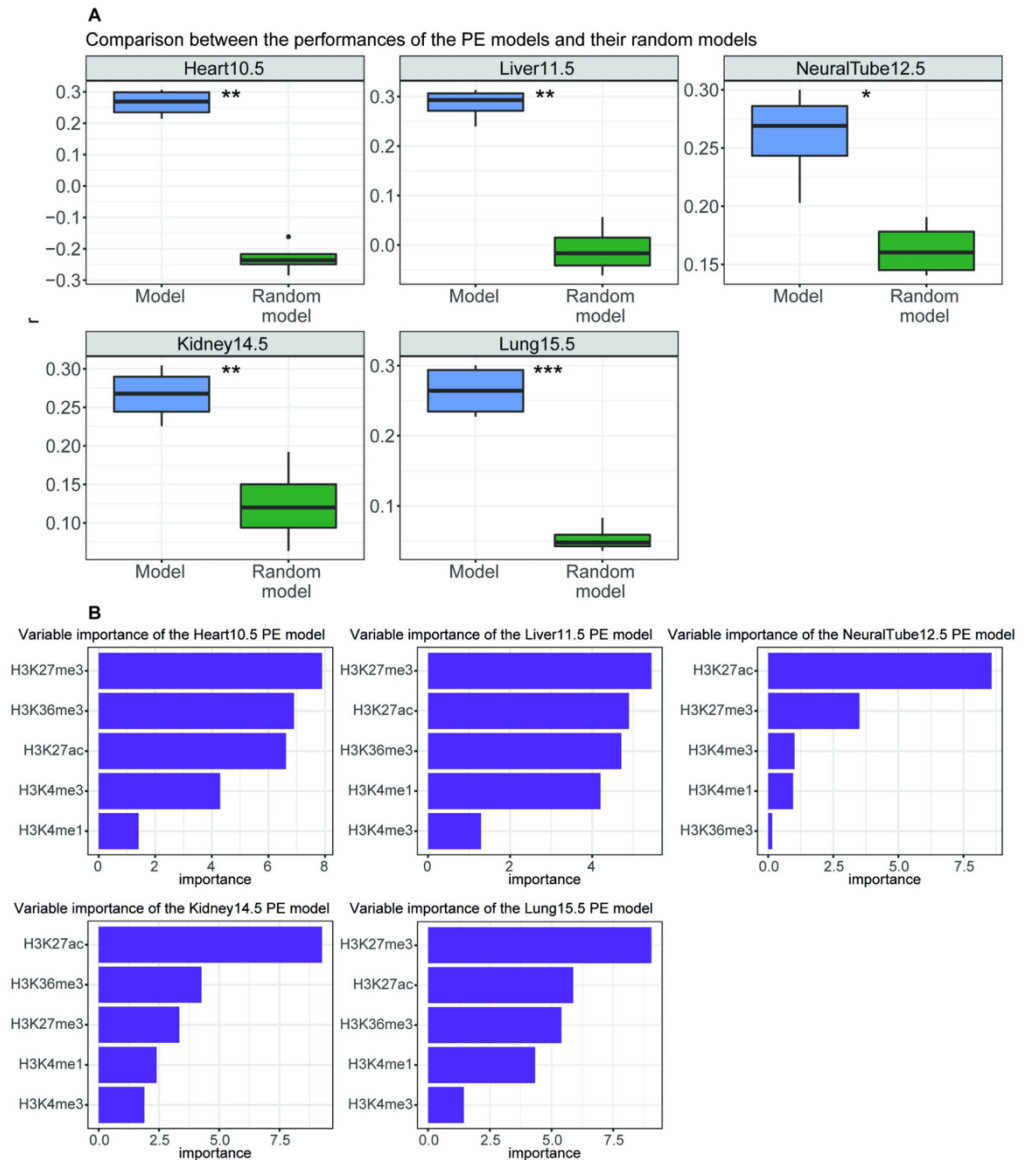


Fig 5. PE models trained using developmental stages. (A) Performance of each differentiation PE model on the rest of the developmental stages as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set (**** $p < 0.0001$).

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). (B) Importance of histone modifications for each development PE model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t-statistics for each model parameter. Heart10.5, heart tissue from 10.5 embryonic day; Kidney14.5, kidney tissue from 14.5 embryonic day; Liver11.5, liver tissue from 11.5 embryonic day; Lung15.5, lung tissue from 15.5 embryonic day; NeuralTube12.5, neural tube tissue from 12.5 embryonic day.

<https://doi.org/10.1371/journal.pcbi.1009368.g005>

Therefore, further work on PEs in other cellular contexts will be necessary in order to generalize our findings.

One could speculate that the ESC enhancer model performance is an artefact of matching the chromatin state of promoters and enhancers (both active or both repressed). However, as no expression information is introduced in this step, further conclusions are not affected by this matching procedure. Moreover, we have confirmed the predictive capacity of PEs in the differentiation PE models, in which we do not require a coordinated activation of PEs and BPs. One could argue that the correlation coefficients obtained in the ESC enhancer models are a mere consequence of the bimodality observed in our scatterplots of gene expression prediction. However, in absence of H3K27me3 as a predictive variable, this bimodality disappears, and importantly, the correlation between predicted and measured expression is maintained ($r = 0.36$). Indeed, this correlation denotes the good performance of the model, and therefore, confirms the predictive power of histone modifications at enhancers. Thus, H3K27me3 could be the cause of the bimodality, probably because it is the histone modification that better differentiates between active and repressed enhancers in ESCs. Interestingly, no bimodality is observed in the scatterplots of the differentiation predictive models, which further confirms the predictive power of histone modifications at enhancers.

At the promoter level, the performance of ESCs and differentiation models was very similar ($r = 0.79$ vs. $r \approx 0.75$). Our limitation of being only able to use a reduced number of histone modifications in the differentiation models (an issue that will be easily overcome when more ChIP-seq datasets are available) could be the reason for the minimal difference in the promoter models' performance. Indeed, at the enhancer level, the difference in performance between ESCs and the differentiation models was higher ($r = 0.49$ vs. $r > 0.3$). Other factors besides the number of ChIP-seq datasets used could explain such a difference: (i) as mentioned before, enhancers and promoters in the ESC model were required to match their chromatin state, which could lead to overrating the performance of the ESC enhancer model; (ii) the enhancer-promoter Hi-C interactions were taken from data published on ESCs. Nevertheless, we predicted gene expression in cellular contexts distinct from ESCs; (iii) some genes might be specific for a cell lineage, and their interactions with enhancers might be lost in the other cell lineages. In these cases, the enhancer and the gene would no longer be related; (iv) enhancer-promoter interactions relevant for early stages of differentiation might be lost once they have served their function. This would imply that the enhancer and its target gene are no longer coordinated. In fact, it has been shown that intensive rearrangement of promoter-enhancer interactions occurs during differentiation, and that these loops become disrupted when their target genes are repressed [30,31].

Moreover, one could argue that the linear regression approach used in this study might be a too generic model, which could be the reason for lower performance of the enhancer models when compared to promoter models. Thus, we re-analyzed our collection of enhancer-gene associations in the Hi-C-top dataset with other methodologies, besides linear regression, which are conceptually more complex. In all cases, although more time-consuming, the performance of these methods did not improve our initial result (Table 3). We believe that the difference in performance between promoter (including BP) and enhancer (including PE) models could be due to: (i) the difficulties of assigning enhancers to their target genes genome-

Table 3. Comparative analysis of other methodologies to predict gene expression in the Hi-C-top dataset.

Model	method value in caret R package [45]	Performance (r)	Most important variable
Neural Network	neuralnet	0.48	H3K27me3
Lasso	lasso	0.48	H3K27me3
Random Forest	rf	0.44	H3K27me3
Support Vector Machines with Linear Kernel	svmLinear	0.48	H3K27me3
Principal Component Analysis	pcr	0.46	H3K27me3
Linear regression (used in this work)	lm	0.49	H3K27me3

Predictive models were obtained using default parameters. Variable importance was assessed with varImp function from caret R package [45].

<https://doi.org/10.1371/journal.pcbi.1009368.t003>

wide, which can lead to incorrect associations; and (ii) the more complex gene expression regulation by enhancers, when more than one enhancer—with different levels (or type) of histone marks—can regulate the same gene.

How to assign enhancers to target genes is still under debate. In this study, we used Hi-C and matched chromatin states to link enhancers to genes and promoters. A recent study evaluated distinct ways of linking genes to enhancers by modelling gene expression and DNase-seq data [17]. They showed that expression predictive models using chromatin conformation data, such as Hi-C, performed better than those using other traditional ways of assigning target genes, such as the closest-gene method or by distance. The closest-gene method consists of assigning each enhancer to the nearest TSS. This prevents enhancers from being assigned to two or more genes but does not take into account that one enhancer can regulate the expression of more than one gene [46,47]. The distance method consists of assigning an enhancer to all the genes that are closer than a pre-set number of base pairs. We achieved a performance of $r = 0.34$ by using 1 Mb distance to assign enhancers to promoters. Although this predictive model had lower performance than the models based on Hi-C data ($r = 0.38$ and $r = 0.49$ for Hi-C-all and Hi-C-top enhancer models, respectively), it maintained the predictive capacity. This suggests that in absence of Hi-C data, using 1 Mb distance to assign target genes to enhancers performs well. Moreover, data from other chromatin capture techniques could be useful to associate enhancers to promoters as well. Indeed, preliminary results using promoter capture Hi-C to associate enhancers to promoters improved the performance of our enhancer predictive models in comparison to those in which Hi-C data was used. We argue that this gain in the performance of gene expression prediction is likely due to promoter capture Hi-C enriching for the best interactions.

Apart from 3C techniques, two novel computational methods have been developed to properly identify enhancer–gene associations using chromatin capture data (such as Hi-C and Hi-ChIP) and enhancer activity data (such as H3K27ac ChIP-seq and DHS-seq) [19,48]. For instance, the so-called FOCS inference method provides a map of active enhancer–promoter associations consistent across several cellular contexts, although no cell-type specific associations could be detected [48]. Further, the activity-by-contact method identified cell-type specific associations of active enhancers and genes [19]. However, neither methodology can be applied to PEs due to the lack of enhancer activity. Indeed, the capacity of PEs to dynamically predict variable gene expression during differentiation suggests that our approach of assigning target genes to PEs performs properly in this context.

We have reported differences in the histone modification contribution to the expression predictive models depending on their location in enhancers or promoters. The different contribution of each histone modification suggests that the epigenetic landscape is different in enhancers and promoters. For example, although H3K4me3 has been previously shown to be

located in enhancers [25–28], our results suggest that its presence in enhancers has little association to gene expression. Therefore, H3K4me3 does not seem to be a good indicator of enhancer activity. In contrast, H3K4me3 proved to be key in predicting gene expression from the differentiation BP models, confirming its relevance in establishing promoter activity. Moreover, whereas H3K36me3 proved to be important for the differentiation PE models—mainly those intragenic—, it showed little contribution to the BP ones. Even though there is a universal relationship between histone modifications and gene expression, we observed that H3K36me3 is more informative in the cardiac PE models than in the neural PE models. We reached this conclusion thanks to our LOESS normalization approach, which allowed us to reduce biases in all datasets used (RNA-seq and ChIP-seq), such that our results were not influenced by the different origin of data. Without such a normalization, the conclusions reached would be wrong. However, it is worth mentioning that we assume constancy of ChIP-seq signal and expression, although they might change in their abundance during differentiation. This problem will be solved in the future with spike-in normalization. Strikingly, H3K27me3 was found to be the most important histone modification in the majority of predictive models for enhancers and promoters. This suggests that H3K27me3 plays a key role in gene regulation, as it is important for both types of regulatory regions. Our results show that mainly H3K27me3, and also in combination with H3K36me3 and H3K27ac, are sufficient to predict future gene expression from PEs. In any case, the predictive power of our models will benefit in the future from the introduction of other histone modifications into the modelling, which can be extremely useful for identifying unknown quantitative relationships between histone modifications at enhancers and gene expression.

Finally, other types of information could also be introduced in the modelling in the future. In fact, previous work has modelled gene expression using accessibility data (e.g. DHS-seq) [14,16,17], and other types of ChIP-seq samples (e.g. TFs or RNA polymerase II) [10,11,13,49]. It would be also interesting to use enhancer RNA (eRNA) data to predict gene expression of target genes. Promising results have been obtained in predicting eRNA transcription by modelling GRO-seq and histone modification ChIP-seq at enhancers [50]. Indeed, Pearson's correlation between PRO-seq [51] signal and H3K27ac ChIP-seq signal at intergenic enhancers in ESC is 0.41, which further supports that eRNA expression might be a good predictor, probably similar to H3K27ac. All this information at enhancers could be integrated into the modelling to improve the power and, most importantly, to discover new quantitative relationships between gene expression and multiple epigenetic features.

Materials and methods

Cell culture

E14Tg2A ESCs were cultured feeder-free on 15-cm plates coated with 0.1% gelatin. Plates were coated with gelatin for 15 min at 37°C, and then non-bound gelatin was removed. ESCs were cultured with Glasgow minimum essential medium (Sigma) supplemented with β -mercaptoethanol, sodium pyruvate, penicillin–streptomycin, non-essential amino acids, GlutaMAX, 20% fetal bovine serum (Hyclone), and leukemia inhibitory factor (LIF).

Chromatin immunoprecipitation

Cells were grown in 15-cm plates until 70% confluency and crosslinked in 1% formaldehyde in growth medium for 10 min at room temperature in a shaker. To stop fixation, glycine was added to a final concentration of 0.125 M and incubated for 5 min at room temperature. Cells were then washed twice with ice-cold PBS and harvested by gently scrapping plates (on ice) in

PBS plus protease inhibitors. Cells from two 15-cm plates were pooled together and centrifuged at $3,400 \times g$ at 4°C for 5 min. Cell pellets were frozen at -80°C until use.

Chromatin was prepared by resuspending the crosslinked pellet in 1.3 ml ice cold ChIP buffer [$1 \times$ volume SDS buffer (100 mM NaCl, 50 mM Tris-HCl pH 8.1, 5 mM EDTA pH 8.0, and 0.5% SDS) and $0.5 \times$ volume Triton dilution buffer (100 mM NaCl, 100 mM Tris-HCl pH 8.6, 5 mM EDTA pH 8.0, and 5% Triton X-100)] plus proteinase inhibitors. Samples were sonicated 40 cycles (30 seconds on/30 seconds off) in a Bioruptor Pico (Diagenode) and centrifuged at $16,000 \times g$ at 4°C for 20 min to remove the cell debris. To check chromatin size, a 25- μl aliquot was mixed with 175 μl of PBS plus 5 μl of 20 mg/ml proteinase K, and de-crosslinked for 5 h at 65°C . DNA was purified using the QIAquick PCR purification kit (Qiagen), quantified in Nanodrop, and checked by electrophoresis on a 1.2% agarose gel.

ChIP experiments were performed using 30 μg of chromatin (DNA) and 5 μg of antibody in a final volume of 500 μl ChIP buffer. Aliquots of 5 μl were removed as input material (1%). ChIP samples were incubated overnight at 4°C on rotation, and then Protein A agarose beads (Diagenode) (42 μl per ChIP) were blocked 30 min with 0.05% BSA, washed, and added to the ChIP reaction. Samples were incubated for 2 h at 4°C with rotation. After incubation, beads were washed three times with 1 ml of low-salt buffer (140 mM NaCl, 50 mM HEPES pH 7.5, and 1% Triton X-100) and once with 1 ml high-salt buffer (500 mM NaCl, 50 mM HEPES pH 7.5, and 1% Triton X-100). ChIPed material was eluted from the beads in 200 μl freshly prepared elution buffer (1% SDS, 100 mM NaHCO_3) at 65°C in a shaker (1000 rpm) for 1 h. Input samples were also brought to 200 μl with elution buffer. After addition of 8 μl of 5 M NaCl to the eluted chromatin and input samples, samples were de-crosslinked overnight at 65°C . The next day, samples were treated with proteinase K [1 μl of 20 mg/ml Proteinase K, plus 4 μl 0.5 M EDTA, and 8 μl Tris-HCl pH 6.5] for 1 h at 45°C . ChIPed DNA and inputs were purified using the QIAquick PCR purification kit (Qiagen) and eluted in 60 μl . The following antibodies were used in the ChIP experiments: H3K27me3 (Millipore, #07–449); H3K4me3 (Diagenode, C15410003); H3K4me1 (Abcam, ab8895); H3K27Ac (Millipore, #07–360); H3 (Abcam, Ab1791); H3K36me3 (Abcam, ab9050); H3K27me1 (Active Motif, #61015); H3K27me2 (Cell Signaling, #9728); H3K79me2 (Abcam, ab3594); H2Bub (Cell Signaling, #5546); and H4K20me3 (Abcam, ab9053). Library preparation for ChIP-seq experiments was performed at the UPF/CRG Genomics Unit. Libraries were sequenced using Illumina HiSeq2000 sequencer.

Input datasets

Raw files and processed data from experiments performed in this study are available at the Gene Expression Omnibus (GEO) under the accession number GSE150633. Raw data of multiple samples from the literature was downloaded and reanalyzed to be included in the study. RNA-seq data of mouse ESCs was extracted from a previous publication from our lab (GEO accession number: GSE79606) [23]. ChIP-seq data of p300 in ESCs was obtained via GEO (GEO accession number: GSE89211) [21]. ChIP-seq data of H3K27me3, H3K4me3, H3K27ac, H3K4me1, and H3K36me3, and RNA-seq data of cardiac differentiation (mesoderm, cardio-precursors and cardiomyocytes), were obtained from <https://b2b.hci.utah.edu/gnomex/> (accession numbers: 44R and 7R2) [35]. ChIP-seq data of H3K27me3, H3K4me3, H3K27ac, H3K4me1, and H3K36me3, RNA-seq data of neural differentiation (neural precursors and cortical neurons), and Hi-C data of ESCs were retrieved from GEO (GEO accession number: GSE96107) [30]. PRO-seq data of mouse ESC was obtained from a previous publication from our lab (GEO accession number: GSE99530) [51]. ChIP-seq data of H3K27me3, H3K4me3, H3K27ac, H3K4me1, and H3K36me3, and RNA-seq of mouse developmental stages (heart tissue from 10.5 embryonic day, liver tissue from 11.5 embryonic day, neural tube tissue from

Table 4. Information on ChIP-seq experiments produced in this study.

ChIP-seq	Total reads	Mapped reads	Multilocus reads
H2Bub	42268079	32671442 (77.30%)	7633345 (18.06%)
H3K4me1	43515038	35408178 (81.37%)	6365630 (14.63%)
H3K4me3	45129019	34342716 (76.10%)	9253789 (20.51%)
H3K27me1	45349251	26830739 (59.16%)	15003526 (33.08%)
H3K27me2	60609374	44651269 (73.67%)	12963849 (21.39%)
H3K27me3	37386877	26823175 (71.74%)	8234058 (22.02%)
H3K27ac	35210421	25105518 (71.30%)	8368750 (23.77%)
H3K36me3	42562036	28957647 (68.04%)	11123530 (26.13%)
H3K79me2	60011275	43270733 (72.10%)	13458082 (22.43%)
H4K20me3	34180446	16175574 (47.32%)	15452266 (45.21%)
H3	53276040	35997064 (67.57%)	14569719 (27.35%)
Input	41583841	28857996 (69.40%)	10598866 (25.49%)

<https://doi.org/10.1371/journal.pcbi.1009368.t004>

12.5 embryonic day, kidney tissue from 14.5 embryonic day, lung tissue from 15.5 embryonic day) were obtained from ENCODE project [41]. The list of ENCODE accession numbers can be found in S10 Table. When replicates were available, pooling was done except for the ChIP-seq samples of H3K4me3 of neural precursors (replicate 1 was used) and H3K27ac of neural precursors (replicate 2 was used).

ChIP-seq analysis

The sequence reads of ChIP-seq data were mapped to the mm10 version of the mouse genome with the BOWTIE software [52], setting the option—m 1, which eliminates reads that align in more than one region. The ChIP-seq profiles were obtained using the function buildChIPprofile from SeqCode (<https://github.com/eblancoga/seqcode>). For the p300 ChIP-seq, peak calling against input was performed using MACS [53] with the option—shiftsize 100, which shifts tags to their midpoint. Information about the total number of reads and read mapping of each ChIP-seq experiment produced in this study can be found in Table 4.

Chromatin segmentation

ChromHMM [54] was used to obtain a chromatin segmentation model for ESCs using the default parameters. The input data were ChIP-seq experiments of H3K4me3, H3K27me3, H3K27ac, and H3K4me1, using ChIP-seq of H3 as control. First, the function BinarizeBam was used to binarize the input mapped data. Next, the LearnModel function was ran to learn different chromatin segmentation models of ESCs, using from 4 to 16 states; the 9-state model was selected because it showed the higher number of states with no redundancy.

RNA-seq analysis

The pair-end sequence reads of RNA-seq data were mapped to the mm10 version of the mouse genome with TopHat [55], setting the options—mate-inner-dist 100, which is the expected mean distance between mate pairs, and -g 1, which eliminates those reads which align in more than one region. The RNA-seq profiles were obtained using the function buildChIPprofile from SeqCode. The FPKMs (fragments per kilobase of transcript per million mapped reads) of each gene in the RefSeq catalogue [24] of the mouse genome were calculated using Cufflinks [56], setting the option—max-bundle-frags 5,000,000, which specifies the maximum genomic length for the bundles.

PRO-seq analysis

The single-end sequence reads of PRO-seq data was mapped to the mm10 version of the mouse genome with TopHat [55], setting the options—mate-inner-dist 100, which is the expected mean distance between mate pairs, and -g 1, which eliminates those reads which align in more than one region. The normalized count of reads of PRO-seq at intergenic enhancers averaged by the length of the region was calculated by recoverChIPlevels from SeqCode.

Hi-C analysis

Hi-C data were processed with TADbit [57]. Briefly, sequencing reads were mapped to the reference genome (mm10) by applying a fragment-based strategy, which is dependent on the GEM mapper [58]. Mapped reads were filtered to remove those resulting from unspecified ligations, errors, or experimental artefacts. Specifically, seven different filters were applied using the default parameters in TADbit: self-circles, dangling ends, errors, extra dangling ends, over-represented, duplicated, and random breaks [57]. After pooling replicates, Hi-C data were normalized with OneD correction [59] at 5 kb of resolution to remove known biases. Significant Hi-C interactions were called with the analyzeHiC function of HOMER software suit [60], binned at 5 kb of resolution, and with the default p -value threshold of 0.001.

Gene expression predictive model

The regression linear models were built to predict gene expression by adjusting the following formula:

$$y_i \sim \beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in} + \epsilon$$

where y_i is the \log_2 of the FPKMs of gene i , with a pseudo count of 0.1. x_{i1} to x_{in} are the \log_2 -normalized count of reads of each ChIP-seq signal at the defined promoters or enhancers averaged by the length of the region calculated by recoverChIPlevels from SeqCode, plus a pseudo count of 0.1. β_0 to β_n are the coefficients that we would like to calculate and ϵ is the error. The predictive models were trained on protein-coding genes. The set of data was randomly divided into two subsets, a training subset with the 80% of entries, and a test subset with the remaining 20% of entries. In the case of differentiation, each of the time points was used as training subsets and then the predictive models were evaluated in the rest. A 10-fold cross-validation was repeat three times to verify that the quantitative relationship between expression and histone modifications was not specific for a subset of the data. The following functions were used: trainControl to perform the 10-fold cross-validation, train to train the models, and varImp to calculate the variable importance, from the R package caret [45]. For models trained on enhancers, genes were introduced into the dataset as many times as the number of associated enhancers they had. To evaluate the specificity of our predictive models, we randomly shuffled the expression values of all the genes in the mouse genome. Thus, for each initial predictive model a random model was also obtained, where all the values y_i were shuffled, maintaining the values of x_{i1} to x_{in} intact. This operation generates a new table of gene expression assignments in which the putative relationship between histone marking and expression of genes (if any) is completely lost. The random models were next generated following the same procedure as the models.

LOESS normalization

The FPKMs of all protein-coding genes and ChIP-seq levels of PEs and BPs were normalized for ESCs, cardiomyocytes, cortical neurons, cardio precursors, mesoderm, and neural

precursors; and also, for heart tissue from 10.5 embryonic day, liver tissue from 11.5 embryonic day, neural tube tissue from 12.5 embryonic day, kidney tissue from 14.5 embryonic day, and lung tissue from 15.5 embryonic day. To normalize the ChIP-seq levels of H3K27me3, H3K4me3, H3K27ac, H3K4me1, and H3K36me3 on the PEs and BPs, the genome was first divided into 2 Kb bins (note that bin size reflects the average size of PEs and BPs). Next, the count of reads, normalized by total number of reads and averaged by the length, was calculated with `recoverChIPlevels` function from `SeqCode`. Finally, the normalization parameters were calculated in those bins and applied to the count of reads normalized by the total number of reads and averaged by the length of PEs and BPs. The `normalize.loess` function of the R package `affy` [61] was used to normalize ChIP-seq data and expression data. Genes and bins with a 0 in any columns were discarded, as it was not possible to determine whether it was due to a sequencing error or a real absence of signal.

Supporting information

S1 Fig. Functional regions are covered by more than one class of state. (A) Segments of active states 1–4 cover the same functional regions delimited by peaks of H3K4me3, H3K27ac and H3K4me1. Differences in the definition of active states are due to the shape of the peaks over the same functional elements. The screenshot was taken from the UCSC Genome Browser [62]. (B) Segments of repressed states 6 and 7 denote the sharp peaks of H3K4me3 and H3K4me1 found inside broad regions covered by H3K27me3. State 8 corresponds to the fraction of H3K27me3 peaks that does not overlap with the other two marks. Differences in the definition of repressed states 6 and 7 are due to the shape of the peaks over the same functional elements. The screenshot was taken from the UCSC Genome Browser [62]. (TIF)

S2 Fig. Performance of enhancer and promoter predictive models in ESCs. Predicted expression of the test subset of genes calculated by the models versus their measured expression by RNA-seq. Model performances are represented by the Pearson's correlation (r) between predicted and measured expression values. (A) Left, the model trained on the promoter regions associated to at least one enhancer using all significant interactions of Hi-C (Hi-C-all promoter model). Right, the performance of the same model after randomizing the expression of the training subset of genes. The color bar represents the density of dots. (B) Left, the model trained on the enhancer regions associated to at least one promoter using all the significant interactions of Hi-C (Hi-C-all enhancer model). Right, the performance of the same model after randomizing the expression of the training subset of genes. The color bar represents the density of dots. (C) As for B, but using 1 Mb distance to connect enhancers to promoters. (F) As for B, but using from the Hi-C-top interactions, only distal enhancers (> 5 Kb from a TSS) to generate the model. (TIF)

S3 Fig. Performance of LOESS normalization in RNA-seq data. (A) MA plot before and after normalization of expression data at each differentiation time point against ESCs. M represents the \log_2 ratio of the intensities of the two samples and A is the \log_2 of the average intensity. Intensity is determined in FPKMs. After normalization, the regression line tends to $M = 0$. The color bar represents the density of dots. (B) Boxplot of expression of 15,065 protein-coding genes before and after LOESS normalization. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S4 Fig. Performance of LOESS normalization in the ChIP-seq data. MA plots before and after normalization of each differentiation time point against ESCs. M represents the \log_2 ratio of the intensities of the two samples, and A is the \log_2 of the average intensity. Intensity corresponds to normalized count of reads by total number of reads of the ChIP-seq samples of (A) H3K27me3, (B) H3K4me3, (C) H3K27ac, (D) H3K4me1, and (E) H3K36me3. The color bars represent the density of dots. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S5 Fig. Performance of PE differentiation models. Predicted expression of the test differentiation time points calculated by the models versus their measured expression by RNA-seq. Model performances are represented by the Pearson's correlation (r) between predicted and measured expression values. The color bars represent the density of dots. (A) Model trained in mesoderm. (B) Model trained in cardio precursors. (C) Model trained in cardiomyocytes. (D) Model trained in neural precursors. (E) Model trained in cortical neurons. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S6 Fig. BP models trained using differentiation time points. (A) Performance of each differentiation BP model on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set ($****p < 0.0001$, $***p < 0.001$, $**p < 0.01$, $*p < 0.05$). (B) Importance of histone modifications for each differentiation BP model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S7 Fig. Distal PE models trained using differentiation time points. Performance of each differentiation BP model on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set ($****p < 0.0001$, $***p < 0.001$, $**p < 0.01$, $*p < 0.05$). CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S8 Fig. PE models trained using differentiation time points without H3K27me3 as predictive variable. (A) Performance of each differentiation model without H3K27me3 as predictive variable on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set ($****p < 0.0001$, $***p < 0.001$, $**p < 0.01$, $*p < 0.05$). (B) Importance of histone modifications for each differentiation intragenic model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S9 Fig. PE models trained using differentiation time points without H3K27ac as predictive variable. (A) Performance of each differentiation model without H3K27ac as predictive variable on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). (B) Importance of histone modifications for each differentiation intragenic model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S10 Fig. Intragenic PE models trained using differentiation time points. (A) Performance of each differentiation intragenic model on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). (B) Importance of histone modifications for each differentiation intragenic model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S11 Fig. Intergenic PE models trained using differentiation time points. (A) Performance of each differentiation intergenic model on the rest of the differentiation time points as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). (B) Importance of histone modifications for each differentiation intergenic model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. CM, cardiomyocytes; CN, cortical neurons; CP, cardio precursors; MES, mesoderm; NPC, neural precursors. (TIF)

S12 Fig. BP models trained using developmental stages. (A) Performance of each differentiation BP model on the rest of the developmental stages as compared to the performance over the random models. Performance is represented as Pearson's correlation (r) between predicted expression and measured expression. Significance was assessed using a paired Student's t -test of the performance of the models or of the random models paired by a differentiation test set (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). (B) Importance of histone modifications for each development BP model. Importance is defined as the contribution of each variable in the linear regression predictive model and corresponds to the absolute value of the t -statistics for each model parameter. Heart10.5, heart tissue from 10.5 embryonic day; Kidney14.5, kidney tissue from 14.5 embryonic day; Liver11.5, liver tissue from 11.5 embryonic day;

Lung15.5, lung tissue from 15.5 embryonic day; NeuralTube12.5, neural tube tissue from 12.5 embryonic day.
(TIF)

S1 Table. List of active promoters and target genes. Coordinates of the identified active promoters and their target genes (genome assembly mm10).
(XLSX)

S2 Table. List of bivalent promoters and target genes. Coordinates of the identified bivalent promoters and their target genes (genome assembly mm10).
(XLSX)

S3 Table. List of active enhancers. Coordinates of the identified active enhancers (genome assembly mm10).
(XLSX)

S4 Table. List of poised enhancers. Coordinates of the identified poised enhancers (genome assembly mm10).
(XLSX)

S5 Table. List of active enhancers, associated active promoters, and target genes (Hi-C-all). Coordinates of the identified active enhancers (*_e), associated active promoters (*_p), and target genes (genome assembly mm10). The association was done using all significant Hi-C interactions (Hi-C-all).
(XLSX)

S6 Table. List of poised enhancers, associated bivalent promoters, and target genes (Hi-C-all). Coordinates of the identified poised enhancers (*_e), associated bivalent promoters (*_p), and target genes (genome assembly mm10). The association was done using all significant Hi-C interactions (Hi-C-all).
(XLSX)

S7 Table. Model predictors. Coefficient and *p*-value of every predictor in each predictive model generated in this study.
(XLSX)

S8 Table. List of active enhancers, associated active promoters, and target genes (Hi-C-top). Coordinates of the identified active enhancers (*_e), associated active promoters (*_p), and target genes (genome assembly mm10). The association was done using the top significant Hi-C interactions (Hi-C-top).
(XLSX)

S9 Table. List of poised enhancers, associated bivalent promoters, and target genes (Hi-C-top). Coordinates of the identified poised enhancers (*_e), associated bivalent promoters (*_p) and target genes (genome assembly mm10). The association was done using the top significant Hi-C interactions (Hi-C-top).
(XLSX)

S10 Table. List of ENCODE accession numbers. Accession numbers of the mouse embryo development data used in this study. Heart10.5, heart tissue from 10.5 embryonic day; Kidney14.5, kidney tissue from 14.5 embryonic day; Liver11.5, liver tissue from 11.5 embryonic day; Lung15.5, lung tissue from 15.5 embryonic day; NeuralTube12.5, neural tube tissue from 12.5 embryonic day.
(XLSX)

Acknowledgments

We thank all members of the Di Croce lab for valuable discussion, and VA Raker for scientific editing.

Author Contributions

Conceptualization: Mar González-Ramírez, Enrique Blanco, Luciano Di Croce.

Data curation: Mar González-Ramírez, Enrique Blanco, Luciano Di Croce.

Formal analysis: Mar González-Ramírez, Enrique Blanco, Luciano Di Croce.

Funding acquisition: Luciano Di Croce.

Investigation: Mar González-Ramírez, Cecilia Ballaré, Francesca Mugianesi, Malte Beringer, Alexandra Santanach, Enrique Blanco.

Methodology: Mar González-Ramírez, Enrique Blanco.

Project administration: Enrique Blanco.

Supervision: Enrique Blanco, Luciano Di Croce.

Validation: Mar González-Ramírez.

Writing – original draft: Mar González-Ramírez, Enrique Blanco, Luciano Di Croce.

Writing – review & editing: Mar González-Ramírez, Cecilia Ballaré, Francesca Mugianesi, Malte Beringer, Enrique Blanco, Luciano Di Croce.

References

1. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet.* 2020. <https://doi.org/10.1038/s41576-019-0209-0> PMID: 31988385.
2. Kouzarides T. Chromatin modifications and their function. *Cell.* 2007; 128(4):693–705. <https://doi.org/10.1016/j.cell.2007.02.005> PMID: 17320507.
3. Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, et al. Active genes are tri-methylated at K4 of histone H3. *Nature.* 2002; 419(6905):407–11. <https://doi.org/10.1038/nature01080> PMID: 12353038.
4. Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, et al. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A.* 2002; 99(13):8695–700. <https://doi.org/10.1073/pnas.082249499> PMID: 12060701; PubMed Central PMCID: PMC124361.
5. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol.* 2004; 6(1):73–7. <https://doi.org/10.1038/ncb1076> PMID: 14661024.
6. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet.* 2008; 40(7):897–903. <https://doi.org/10.1038/ng.154> PMID: 18552846; PubMed Central PMCID: PMC2769248.
7. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, et al. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science.* 2002; 298(5595):1039–43. <https://doi.org/10.1126/science.1076997> PMID: 12351676.
8. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet.* 2007; 39(3):311–8. <https://doi.org/10.1038/ng1966> PMID: 17277777.
9. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A.* 2010; 107(7):2926–31. <https://doi.org/10.1073/pnas.0909344107> PMID: 20133639; PubMed Central PMCID: PMC2814872.
10. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.*

- 2011; 12(2):R15. <https://doi.org/10.1186/gb-2011-12-2-r15> PMID: 21324173; PubMed Central PMCID: PMC3188797.
11. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* 2012; 40(2):553–68. <https://doi.org/10.1093/nar/gkr752> PMID: 21926158; PubMed Central PMCID: PMC3258143.
 12. Wang C, Tian R, Zhao Q, Xu H, Meyer CA, Li C, et al. Computational inference of mRNA stability from histone modification and transcriptome profiles. *Nucleic Acids Res.* 2012; 40(14):6414–23. <https://doi.org/10.1093/nar/gks304> PMID: 22495509; PubMed Central PMCID: PMC3413115.
 13. Tippmann SC, Ivanek R, Gaidatzis D, Scholer A, Hoerner L, van Nimwegen E, et al. Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol Syst Biol.* 2012; 8:593. <https://doi.org/10.1038/msb.2012.23> PMID: 22806141; PubMed Central PMCID: PMC3421439.
 14. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 2012; 13(9):R53. <https://doi.org/10.1186/gb-2012-13-9-r53> PMID: 22950368; PubMed Central PMCID: PMC3491397.
 15. Read DF, Cook K, Lu YY, Le Roch KG, Noble WS. Predicting gene expression in the human malaria parasite *Plasmodium falciparum* using histone modification, nucleosome positioning, and 3D localization features. *PLoS Comput Biol.* 2019; 15(9):e1007329. <https://doi.org/10.1371/journal.pcbi.1007329> PMID: 31509524; PubMed Central PMCID: PMC6756558.
 16. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. *Epigenetics Chromatin.* 2020; 13(1):4. <https://doi.org/10.1073/pnas.1704553114> PMID: 28576882; PubMed Central PMCID: PMC5488952.
 17. Schmidt F, Kern F, Schulz MH. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin.* 2020; 13(1):4. <https://doi.org/10.1186/s13072-020-0327-0> PMID: 32029002; PubMed Central PMCID: PMC7003490.
 18. Wang S, Zang C, Xiao T, Fan J, Mei S, Qin Q, et al. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Res.* 2016; 26(10):1417–29. <https://doi.org/10.1101/gr.201574.115> PMID: 27466232; PubMed Central PMCID: PMC5052056.
 19. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet.* 2019; 51(12):1664–9. <https://doi.org/10.1038/s41588-019-0538-0> PMID: 31784727; PubMed Central PMCID: PMC6886585.
 20. Blanco E, Gonzalez-Ramirez M, Alcaine-Colet A, Aranda S, Di Croce L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends Genet.* 2020; 36(2):118–31. <https://doi.org/10.1016/j.tig.2019.11.004> PMID: 31818514.
 21. Cruz-Molina S, Respuela P, Tebartz C, Kolovos P, Nikolic M, Fueyo R, et al. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell.* 2017; 20(5):689–705 e9. <https://doi.org/10.1016/j.stem.2017.02.004> PMID: 28285903.
 22. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* 2011; 21(8):1273–83. <https://doi.org/10.1101/gr.122382.111> PMID: 21632746; PubMed Central PMCID: PMC3149494.
 23. Beringer M, Pisano P, Di Carlo V, Blanco E, Chammas P, Vizan P, et al. EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell.* 2016; 64(4):645–58. <https://doi.org/10.1016/j.molcel.2016.10.018> PMID: 27863225.
 24. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016; 44(D1):D733–45. <https://doi.org/10.1093/nar/gkv1189> PMID: 26553804; PubMed Central PMCID: PMC4702849.
 25. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell.* 2007; 129(4):823–37. <https://doi.org/10.1016/j.cell.2007.05.009> PMID: 17512414.
 26. Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, et al. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* 2011; 30(20):4198–210. <https://doi.org/10.1038/emboj.2011.295> PMID: 21847099; PubMed Central PMCID: PMC3199384.
 27. Koch F, Andrau JC. Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity. *Transcription.* 2011; 2(6):263–8. <https://doi.org/10.4161/tms.2.6.18747> PMID: 22223044; PubMed Central PMCID: PMC3265787.
 28. Russ BE, Olshansky M, Li J, Nguyen MLT, Gearing LJ, Nguyen THO, et al. Regulation of H3K4me3 at Transcriptional Enhancers Characterizes Acquisition of Virus-Specific CD8(+) T Cell-Lineage-Specific

- Function. *Cell Rep.* 2017; 21(12):3624–36. <https://doi.org/10.1016/j.celrep.2017.11.097> PMID: 29262339.
29. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016; 44(W1):W90–7. <https://doi.org/10.1093/nar/gkw377> PMID: 27141961; PubMed Central PMCID: PMC4987924.
 30. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell.* 2017; 171(3):557–72 e24. <https://doi.org/10.1016/j.cell.2017.09.043> PMID: 29053968; PubMed Central PMCID: PMC5651218.
 31. Freire-Pritchett P, Schoenfelder S, Varnai C, Wingett SW, Cairns J, Collier AJ, et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife.* 2017; 6. <https://doi.org/10.7554/eLife.21926> PMID: 28332981; PubMed Central PMCID: PMC5407860.
 32. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 2014; 24(3):390–400. <https://doi.org/10.1101/gr.163519.113> PMID: 24398455; PubMed Central PMCID: PMC3941104.
 33. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A.* 2010; 107(50):21931–6. <https://doi.org/10.1073/pnas.1016071107> PMID: 21106759; PubMed Central PMCID: PMC3003124.
 34. Shema E, Jones D, Shores N, Donohue L, Ram O, Bernstein BE. Single-molecule decoding of combinatorially modified nucleosomes. *Science.* 2016; 352(6286):717–21. Epub 2016/05/07. <https://doi.org/10.1126/science.1253823> PMID: 27151869; PubMed Central PMCID: PMC4904710.
 35. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell.* 2012; 151(1):206–20. <https://doi.org/10.1016/j.cell.2012.07.035> PMID: 22981692; PubMed Central PMCID: PMC3462286.
 36. Dudoit SY Y. H.; Callow M. J.; Speed T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin.* 2002; 12(1):111–39.
 37. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003; 19(2):185–93. <https://doi.org/10.1093/bioinformatics/19.2.185> PMID: 12538238.
 38. Hounkpe BWC F.; Lima F.; de Paula E. V. HT Atlas v1.0 database: redefining human and mouse house-keeping genes and candidate reference transcripts by mining massive RNA-seq datasets bioRxiv. 2019. <https://doi.org/https://doi.org/10.1101/787150>
 39. Krogan NJ, Kim M, Tong A, Golshani A, Cagney G, Canadien V, et al. Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol.* 2003; 23(12):4207–18. <https://doi.org/10.1128/MCB.23.12.4207-4218.2003> PMID: 12773564; PubMed Central PMCID: PMC427527.
 40. Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem.* 2005; 280(18):17732–6. <https://doi.org/10.1074/jbc.M500796200> PMID: 15760899.
 41. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shores N, et al. Expanded encyclopedia of DNA elements in the human and mouse genomes. *Nature.* 2020; 583(7818):699–710. Epub 2020/07/31. <https://doi.org/10.1038/s41586-020-2493-4> PMID: 32728249.
 42. Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet.* 2015; 47(10):1179–86. <https://doi.org/10.1038/ng.3393> PMID: 26323060; PubMed Central PMCID: PMC4847639.
 43. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature.* 2011; 470(7333):279–83. <https://doi.org/10.1038/nature09692> PMID: 21160473; PubMed Central PMCID: PMC4445674.
 44. Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell.* 2012; 11(5):633–48. <https://doi.org/10.1016/j.stem.2012.07.006> PMID: 22981823; PubMed Central PMCID: PMC3751405.
 45. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles.* 2008; 28(5):1–26. <https://doi.org/10.18637/jss.v028.i05>
 46. Chang TH, Primig M, Hadchouel J, Tajbakhsh S, Rocancourt D, Fernandez A, et al. An enhancer directs differential expression of the linked Mrf4 and Myf5 myogenic regulatory genes in the mouse. *Dev Biol.* 2004; 269(2):595–608. <https://doi.org/10.1016/j.ydbio.2004.02.013> PMID: 15110722.

47. Link N, Kurtz P, O'Neal M, Garcia-Hughes G, Abrams JM. A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans. *Genes Dev.* 2013; 27(22):2433–8. <https://doi.org/10.1101/gad.225565.113> PMID: 24240233; PubMed Central PMCID: PMC3841732.
48. Hait TA, Amar D, Shamir R, Elkon R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol.* 2018; 19(1):56. <https://doi.org/10.1186/s13059-018-1432-2> PMID: 29716618; PubMed Central PMCID: PMC5930446.
49. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A.* 2009; 106(51):21521–6. <https://doi.org/10.1073/pnas.0904863106> PMID: 19995984; PubMed Central PMCID: PMC2789751.
50. Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* 2013; 41(22):10032–43. <https://doi.org/10.1093/nar/gkt826> PMID: 24038352; PubMed Central PMCID: PMC3905895.
51. Mas G, Blanco E, Ballare C, Sanso M, Spill YG, Hu D, et al. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet.* 2018; 50(10):1452–62. Epub 2018/09/19. <https://doi.org/10.1038/s41588-018-0218-5> PMID: 30224650.
52. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10(3):R25. Epub 2009/03/06. <https://doi.org/10.1186/gb-2009-10-3-r25> PMID: 19261174; PubMed Central PMCID: PMC2690996.
53. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008; 9(9):R137. <https://doi.org/10.1186/gb-2008-9-9-r137> PMID: 18798982; PubMed Central PMCID: PMC2592715.
54. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012; 9(3):215–6. <https://doi.org/10.1038/nmeth.1906> PMID: 22373907; PubMed Central PMCID: PMC3577932.
55. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25(9):1105–11. <https://doi.org/10.1093/bioinformatics/btp120> PMID: 19289445; PubMed Central PMCID: PMC2672628.
56. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010; 28(5):511–5. <https://doi.org/10.1038/nbt.1621> PMID: 20436464; PubMed Central PMCID: PMC3146043.
57. Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol.* 2017; 13(7):e1005665. <https://doi.org/10.1371/journal.pcbi.1005665> PMID: 28723903; PubMed Central PMCID: PMC5540598.
58. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods.* 2012; 9(12):1185–8. <https://doi.org/10.1038/nmeth.2221> PMID: 23103880.
59. Vidal E, le Dily F, Quilez J, Stadhouders R, Cuartero Y, Graf T, et al. OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.* 2018; 46(8):e49. <https://doi.org/10.1093/nar/gky064> PMID: 29394371; PubMed Central PMCID: PMC5934634.
60. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010; 38(4):576–89. <https://doi.org/10.1016/j.molcel.2010.05.004> PMID: 20513432; PubMed Central PMCID: PMC2898526.
61. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* 2004; 20(3):307–15. <https://doi.org/10.1093/bioinformatics/btg405> PMID: 14960456.
62. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. *Curr Protoc Bioinformatics.* 2007; Chapter 1:Unit 1 4. <https://doi.org/10.1002/0471250953.bi0104s17> PMID: 18428780.

REFERENCES

- Akdemir, K. C., Le, V. T., Kim, J. M., Killcoyne, S., King, D. A., Lin, Y. P., . . . Andrew Futreal, P. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat Genet*, 52(11), 1178-1188. doi:10.1038/s41588-020-0708-0
- Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res*, 40(22), 11202-11212. doi:10.1093/nar/gks925
- Allahyar, A., Vermeulen, C., Bouwman, B. A. M., Krijger, P. H. L., Verstegen, M., Geeven, G., . . . de Laat, W. (2018). Enhancer hubs and loop collisions identified from single-allele topologies. *Nat Genet*, 50(8), 1151-1160. doi:10.1038/s41588-018-0161-5
- Aloia, L., Di Stefano, B., Sessa, A., Morey, L., Santanach, A., Gutierrez, A., . . . Di Croce, L. (2014). Zrf1 is required to establish and maintain neural progenitor identity. *Genes Dev*, 28(2), 182-197. doi:10.1101/gad.228510.113
- Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep*, 9(1), 9354. doi:10.1038/s41598-019-45839-z
- Andrey, G., & Mundlos, S. (2017). The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 144(20), 3646-3658. doi:10.1242/dev.148304
- Apostolou, E., Ferrari, F., Walsh, R. M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., . . . Hochedlinger, K. (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell*, 12(6), 699-712. doi:10.1016/j.stem.2013.04.013
- Aranda, S., Mas, G., & Di Croce, L. (2015). Regulation of gene transcription by Polycomb proteins. *Sci Adv*, 1(11), e1500737. doi:10.1126/sciadv.1500737

- Arrastia, M. V., Jachowicz, J. W., Ollikainen, N., Curtis, M. S., Lai, C., Quinodoz, S. A., . . . Guttman, M. (2021). Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nat Biotechnol*. doi:10.1038/s41587-021-00998-1
- Ay, F., Bailey, T. L., & Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res*, 24(6), 999-1011. doi:10.1101/gr.160374.113
- Ay, F., Vu, T. H., Zeitz, M. J., Varoquaux, N., Carette, J. E., Vert, J. P., . . . Noble, W. S. (2015). Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C. *BMC Genomics*, 16, 121. doi:10.1186/s12864-015-1236-7
- Azagra, A., Marina-Zarate, E., Ramiro, A. R., Javierre, B. M., & Parra, M. (2020). From Loops to Looks: Transcription Factors and Chromatin Organization Shaping Terminal B Cell Differentiation. *Trends Immunol*, 41(1), 46-60. doi:10.1016/j.it.2019.11.006
- Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*, 18(5), 285-298. doi:10.1038/nrm.2017.7
- Bantignies, F., Grimaud, C., Lavrov, S., Gabut, M., & Cavalli, G. (2003). Inheritance of Polycomb-dependent chromosomal interactions in *Drosophila*. *Genes Dev*, 17(19), 2406-2420. doi:10.1101/gad.269503
- Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., . . . Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell*, 144(2), 214-226. doi:10.1016/j.cell.2010.12.026
- Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823-837. doi:S0092-8674(07)00600-9 [pii] 10.1016/j.cell.2007.05.009
- Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M., & Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*, 143(3), 470-484. doi:10.1016/j.cell.2010.10.012

- Beagan, J. A., & Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nat Genet*, 52(1), 8-16. doi:10.1038/s41588-019-0561-1
- Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C., Chotalia, M., Xie, S. Q., . . . Pombo, A. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646), 519-524. doi:10.1038/nature21411
- Beliveau, B. J., Boettiger, A. N., Avendano, M. S., Jungmann, R., McCole, R. B., Joyce, E. F., . . . Wu, C. T. (2015). Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat Commun*, 6, 7147. doi:10.1038/ncomms8147
- Beliveau, B. J., Boettiger, A. N., Nir, G., Bintu, B., Yin, P., Zhuang, X., & Wu, C. T. (2017). In Situ Super-Resolution Imaging of Genomic DNA with OligoSTORM and OligoDNA-PAINT. *Methods Mol Biol*, 1663, 231-252. doi:10.1007/978-1-4939-7265-4_19
- Beliveau, B. J., Joyce, E. F., Apostolopoulos, N., Yilmaz, F., Fonseka, C. Y., McCole, R. B., . . . Wu, C. T. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci U S A*, 109(52), 21301-21306. doi:10.1073/pnas.1213818110
- Belmont, A. S. (2014). Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr Opin Cell Biol*, 26, 69-78. doi:10.1016/j.ceb.2013.10.002
- Bendandi, A., Dante, S., Zia, S. R., Diaspro, A., & Rocchia, W. (2020). Chromatin Compaction Multiscale Modeling: A Complex Synergy Between Theory, Simulation, and Experiment. *Front Mol Biosci*, 7, 15. doi:10.3389/fmolb.2020.00015
- Beringer, M., Pisano, P., Di Carlo, V., Blanco, E., Chammas, P., Vizan, P., . . . Di Croce, L. (2016). EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell*, 64(4), 645-658. doi:10.1016/j.molcel.2016.10.018
- Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA

- elements in the human genome. *Nature*, 489(7414), 57-74. doi:10.1038/nature11247
- nature11247 [pii]
- Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4), 669-681. doi:10.1016/j.cell.2007.01.033
- Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S., . . . Hess, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793), 1642-1645. doi:1127344 [pii]
- 10.1126/science.1127344
- Bickmore, W. A., & van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6), 1270-1284. doi:10.1016/j.cell.2013.02.001
- S0092-8674(13)00146-3 [pii]
- Bintu, B., Mateo, L. J., Su, J. H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., . . . Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413). doi:10.1126/science.aau1783
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., . . . de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), 799-816. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17571346
- Blanco, E., Gonzalez-Ramirez, M., & Di Croce, L. (2021). Productive visualization of high-throughput sequencing data using the SeqCode open portable platform. *Sci Rep*, 11(1), 19545. doi:10.1038/s41598-021-98889-7
- Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., . . . Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat Struct Mol Biol*, 25(9), 833-840. doi:10.1038/s41594-018-0112-y
- Boettiger, A. N., Bintu, B., Moffitt, J. R., Wang, S., Beliveau, B. J., Fudenberg, G., . . . Zhuang, X. (2016). Super-resolution imaging

- reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586), 418-422. doi:10.1038/nature16496
- Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., . . . Young, R. A. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, 175(7), 1842-1855 e1816. doi:10.1016/j.cell.2018.10.042
- Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nat Rev Genet*, 17(11), 661-678. doi:10.1038/nrg.2016.112
- Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., . . . Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3), 557-572 e524. doi:10.1016/j.cell.2017.09.043
- Bowman, G. D., & Poirier, M. G. (2015). Post-translational modifications of histones that influence nucleosome dynamics. *Chem Rev*, 115(6), 2274-2295. doi:10.1021/cr500350x
- Brandao, H. B., Gabriele, M., & Hansen, A. S. (2021). Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Curr Opin Cell Biol*, 70, 18-26. doi:10.1016/j.ccb.2020.11.002
- Bruneau, B. G., & Nora, E. P. (2018). Chromatin Domains Go on Repeat in Disease. *Cell*, 175(1), 38-40. doi:10.1016/j.cell.2018.08.068
- Bunting, K. L., Soong, T. D., Singh, R., Jiang, Y., Beguelin, W., Poloway, D. W., . . . Melnick, A. M. (2016). Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity*, 45(3), 497-512. doi:10.1016/j.immuni.2016.08.012
- Cai, S., Chen, C., Tan, Z. Y., Huang, Y., Shi, J., & Gan, L. (2018). Cryo-ET reveals the macromolecular reorganization of *S. pombe* mitotic chromosomes in vivo. *Proc Natl Acad Sci U S A*, 115(43), 10977-10982. doi:10.1073/pnas.1720476115
- Cao, K., Lailier, N., Zhang, Y., Kumar, A., Uppal, K., Liu, Z., . . . Fan, Y. (2013). High-resolution mapping of h1 linker histone

- variants in embryonic stem cells. *PLoS Genet*, 9(4), e1003417. doi:10.1371/journal.pgen.1003417
- Cardozo Gizzi, A. M., Cattoni, D. I., Fiche, J. B., Espinola, S. M., Gurgo, J., Messina, O., . . . Nollmann, M. (2019). Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol Cell*, 74(1), 212-222 e215. doi:10.1016/j.molcel.2019.01.011
- Cattoni, D. I., Cardozo Gizzi, A. M., Georgieva, M., Di Stefano, M., Valeri, A., Chamousset, D., . . . Nollmann, M. (2017). Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat Commun*, 8(1), 1753. doi:10.1038/s41467-017-01962-x
- Cavalli, G., & Misteli, T. (2013). Functional implications of genome topology. *Nat Struct Mol Biol*, 20(3), 290-299. doi:10.1038/nsmb.2474
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., . . . Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6), 1106-1117. doi:10.1016/j.cell.2008.04.043
- Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E. K., Adam, S. A., . . . Belmont, A. S. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol*, 217(11), 4025-4048. doi:10.1083/jcb.201807108
- Cheutin, T., & Cavalli, G. (2019). The multiscale effects of polycomb mechanisms on 3D chromatin folding. *Crit Rev Biochem Mol Biol*, 54(5), 399-417. doi:10.1080/10409238.2019.1679082
- Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P. Y., Stroud, H., Yu, Y., . . . Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304), 388-392. doi:10.1038/nature09147
- Choy, M. K., Javierre, B. M., Williams, S. G., Baross, S. L., Liu, Y., Wingett, S. W., . . . Keavney, B. D. (2018). Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene

- networks. *Nat Commun*, 9(1), 2526. doi:10.1038/s41467-018-04931-0
- Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., . . . Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*, 168(3), 442-459 e420. doi:10.1016/j.cell.2016.12.016
- Clapier, C. R., & Cairns, B. R. (2009). The biology of chromatin remodeling complexes. *Annu Rev Biochem*, 78, 273-304. doi:10.1146/annurev.biochem.77.062706.153223
- Consortium, E. P., Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shores, N., . . . Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818), 699-710. doi:10.1038/s41586-020-2493-4
- Cremer, T., & Cremer, M. (2010). Chromosome territories. *Cold Spring Harb Perspect Biol*, 2(3), a003889. doi:10.1101/cshperspect.a003889
- Csank, A. K., & Henikoff, S. (1996). Genetic modification of heterochromatic association and nuclear organization in *Drosophila*. *Nature*, 381(6582), 529-531. doi:10.1038/381529a0
- Cullen, K. E., Kladde, M. P., & Seyfred, M. A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. *Science*, 261(5118), 203-206. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8327891
- Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G., & Peters, J. M. (2019). DNA loop extrusion by human cohesin. *Science*, 366(6471), 1338-1345. doi:10.1126/science.aaz3418
- Day, N., Hemmaphard, A., Thurman, R. E., Stamatoyannopoulos, J. A., & Noble, W. S. (2007). Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11), 1424-1426. doi:10.1093/bioinformatics/btm096
- de Wit, E., Bouwman, B. A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J., . . . de Laat, W. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, 501(7466), 227-231. doi:10.1038/nature12420

- de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, 26(1), 11-24. doi:10.1101/gad.179804.111
- 26/1/11 [pii]
- de Wit, E., Vos, E. S., Holwerda, S. J., Valdes-Quezada, C., Verstegen, M. J., Teunissen, H., . . . de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell*, 60(4), 676-684. doi:10.1016/j.molcel.2015.09.023
- Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558), 1306-1311. doi:10.1126/science.1067799
- 295/5558/1306 [pii]
- Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., . . . Blobel, G. A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6), 1233-1244. doi:10.1016/j.cell.2012.03.051
- Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., . . . Disteche, C. M. (2015). Bipartite structure of the inactive mouse X chromosome. *Genome Biol*, 16, 152. doi:10.1186/s13059-015-0728-8
- Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., . . . Plath, K. (2013). Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell*, 13(5), 602-616. doi:10.1016/j.stem.2013.08.013
- Denker, A., & de Laat, W. (2016). The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev*, 30(12), 1357-1382. doi:10.1101/gad.281964.116
- Despang, A., Schopflin, R., Franke, M., Ali, S., Jerkovic, I., Paliou, C., . . . Ibrahim, D. M. (2019). Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat Genet*, 51(8), 1263-1271. doi:10.1038/s41588-019-0466-z
- Di Croce, L., & Helin, K. (2013). Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol*, 20(10), 1147-1155. doi:10.1038/nsmb.2669

- Di Giammartino, D. C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., . . . Apostolou, E. (2019). KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nat Cell Biol*, 21(10), 1179-1190. doi:10.1038/s41556-019-0390-6
- Di Pierro, M., Cheng, R. R., Lieberman Aiden, E., Wolynes, P. G., & Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc Natl Acad Sci U S A*, 114(46), 12126-12131. doi:10.1073/pnas.1714980114
- Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proc Natl Acad Sci U S A*, 113(43), 12168-12173. doi:10.1073/pnas.1613607113
- Di Stefano, M., Stadhouders, R., Farabella, I., Castillo, D., Serra, F., Graf, T., & Marti-Renom, M. A. (2020). Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. *Nat Commun*, 11(1), 2564. doi:10.1038/s41467-020-16396-1
- Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., . . . Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), 331-336. doi:10.1038/nature14222
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376-380. doi:10.1038/nature11082
- nature11082 [pii]
- Dogan, E. S., & Liu, C. (2018). Three-dimensional chromatin packing and positioning of plant genomes. *Nat Plants*, 4(8), 521-529. doi:10.1038/s41477-018-0199-5
- Dostie, J., & Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc*, 2(4), 988-1002. doi:nprot.2007.116 [pii]
- 10.1038/nprot.2007.116

- Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., . . . Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2), 374-387. doi:10.1016/j.cell.2014.09.030
- Doyle, M. J., & Sussel, L. (2007). Nkx2.2 regulates beta-cell function in the mature islet. *Diabetes*, 56(8), 1999-2007. doi:10.2337/db06-1766
- Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188-230. doi:10.1002/aris.1440380105
- Eagen, K. P., Aiden, E. L., & Kornberg, R. D. (2017). Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc Natl Acad Sci U S A*, 114(33), 8764-8769. doi:10.1073/pnas.1701291114
- Eltsov, M., Maclellan, K. M., Maeshima, K., Frangakis, A. S., & Dubochet, J. (2008). Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proc Natl Acad Sci U S A*, 105(50), 19732-19737. doi:10.1073/pnas.0810057105
- Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., . . . Guttman, M. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, 341(6147), 1237973. doi:10.1126/science.1237973
- Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8), 817-825. doi:10.1038/nbt.1662
- Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3), 215-216. doi:10.1038/nmeth.1906
- Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, 12(12), 2478-2492. doi:10.1038/nprot.2017.124
- Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shores, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis

- of chromatin state dynamics in nine human cell types. *Nature*, 473(7345), 43-49. doi:10.1038/nature09906
- Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B. R., Leonhardt, H., . . . Mirny, L. A. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, 570(7761), 395-399. doi:10.1038/s41586-019-1275-3
- Fan, Y., Nikitina, T., Zhao, J., Fleury, T. J., Bhattacharyya, R., Bouhassira, E. E., . . . Skoultschi, A. I. (2005). Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell*, 123(7), 1199-1212. doi:10.1016/j.cell.2005.10.028
- Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A. D., & Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res*, 26(12), 1345-1348. doi:10.1038/cr.2016.137
- Farabella, I., Di Stefano, M., Soler-Vila, P., Marti-Marimon, M., & Marti-Renom, M. A. (2021). Three-dimensional genome organization via triplex-forming RNAs. *Nat Struct Mol Biol*, 28(11), 945-954. doi:10.1038/s41594-021-00678-3
- Filion, G. J., van Bemmelen, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., . . . van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell*, 143(2), 212-224. doi:10.1016/j.cell.2010.09.009
- Finch, J. T., & Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A*, 73(6), 1897-1901. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/1064861>
- Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., . . . Bickmore, W. A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet*, 4(3), e1000039. doi:10.1371/journal.pgen.1000039
- Finn, E. H., Pegoraro, G., Brandao, H. B., Valton, A. L., Oomen, M. E., Dekker, J., . . . Misteli, T. (2019). Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell*, 176(6), 1502-1515 e1510. doi:10.1016/j.cell.2019.01.020

- Fitz-James, M. H., & Cavalli, G. (2022). Molecular mechanisms of transgenerational epigenetic inheritance. *Nat Rev Genet.* doi:10.1038/s41576-021-00438-5
- Flavahan, W. A., Drier, Y., Liao, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., . . . Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584), 110-114. doi:10.1038/nature16490
- Flyamer, I. M., Gassler, J., Imakaev, M., Brandao, H. B., Ulianov, S. V., Abdennur, N., . . . Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648), 110-114. doi:10.1038/nature21711
- Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., . . . Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624), 265-269. doi:10.1038/nature19800
- Franklin, R. E., & Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171(4356), 740-741. doi:10.1038/171740a0
- Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., & Mirny, L. A. (2017). Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb Symp Quant Biol*, 82, 45-55. doi:10.1101/sqb.2017.82.034710
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep*, 15(9), 2038-2049. doi:10.1016/j.celrep.2016.04.085
- Fudenberg, G., & Mirny, L. A. (2012). Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev*, 22(2), 115-124. doi:10.1016/j.gde.2012.01.006
- Fullwood, M. J., Wei, C. L., Liu, E. T., & Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*, 19(4), 521-532. doi:10.1101/gr.074906.107

- Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, 13(12), 840-852. doi:10.1038/nrg3306
- Fussner, E., Djuric, U., Strauss, M., Hotta, A., Perez-Iratxeta, C., Lanner, F., . . . Bazett-Jones, D. P. (2011). Constitutive heterochromatin reorganization during somatic cell reprogramming. *EMBO J*, 30(9), 1778-1789. doi:10.1038/emboj.2011.96
- Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., . . . Bazett-Jones, D. P. (2012). Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Rep*, 13(11), 992-996. doi:10.1038/embor.2012.139
- Fyodorov, D. V., Zhou, B. R., Skoultchi, A. I., & Bai, Y. (2018). Emerging roles of linker histones in regulating chromatin structure and function. *Nat Rev Mol Cell Biol*, 19(3), 192-206. doi:10.1038/nrm.2017.94
- Gagliardi, A., Mullin, N. P., Ying Tan, Z., Colby, D., Kousa, A. I., Halbritter, F., . . . Chambers, I. (2013). A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J*, 32(16), 2231-2247. doi:10.1038/emboj.2013.161
- Galupa, R., & Heard, E. (2017). Topologically Associating Domains in Chromosome Architecture and Gene Regulatory Landscapes during Development, Disease, and Evolution. *Cold Spring Harb Symp Quant Biol*, 82, 267-278. doi:10.1101/sqb.2017.82.035030
- Gangaraju, V. K., & Bartholomew, B. (2007). Mechanisms of ATP dependent chromatin remodeling. *Mutat Res*, 618(1-2), 3-17. doi:10.1016/j.mrfmmm.2006.08.015
- Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., & Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384), 102-105. doi:10.1126/science.aar7831
- Gavrilov, A., Razin, S. V., & Cavalli, G. (2015). In vivo formaldehyde cross-linking: it is time for black box analysis. *Brief Funct Genomics*, 14(2), 163-165. doi:10.1093/bfpg/elu037

- Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., & Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet*, 51(8), 1272-1282. doi:10.1038/s41588-019-0462-3
- Gibson, B. A., Doolittle, L. K., Schneider, M. W. G., Jensen, L. E., Gamarra, N., Henry, L., . . . Rosen, M. K. (2019). Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell*, 179(2), 470-484 e421. doi:10.1016/j.cell.2019.08.037
- Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., . . . Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4), 950-963. doi:10.1016/j.cell.2014.03.025
- Golfier, S., Quail, T., Kimura, H., & Bruges, J. (2020). Cohesin and condensin extrude DNA loops in a cell cycle-dependent manner. *Elife*, 9. doi:10.7554/eLife.53885
- Goloborodko, A., Marko, J. F., & Mirny, L. A. (2016). Chromosome Compaction by Active Loop Extrusion. *Biophys J*, 110(10), 2162-2168. doi:10.1016/j.bpj.2016.02.041
- Gomez-Diaz, E., & Corces, V. G. (2014). Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol*, 24(11), 703-711. doi:10.1016/j.tcb.2014.08.003
- Gonzalez-Sandoval, A., Towbin, B. D., Kalck, V., Cabianca, D. S., Gaidatzis, D., Hauer, M. H., . . . Gasser, S. M. (2015). Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in *C. elegans* Embryos. *Cell*, 163(6), 1333-1347. doi:10.1016/j.cell.2015.10.066
- Grosberg, A., Nechaev, S. K., & Shakhnovich, E. I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. *Journal De Physique* 49, 2095-2100.
- Gross, D. S., Chowdhary, S., Anandhakumar, J., & Kainth, A. S. (2015). Chromatin. *Curr Biol*, 25(24), R1158-1163. doi:10.1016/j.cub.2015.10.059
- Gu, B., Swigut, T., Spencley, A., Bauer, M. R., Chung, M., Meyer, T., & Wysocka, J. (2018). Transcription-coupled changes in nuclear

- mobility of mammalian cis-regulatory elements. *Science*, 359(6379), 1050-1055. doi:10.1126/science.aao3136
- Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., . . . van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197), 948-951. doi:10.1038/nature06947
- Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1), 77-88. doi:10.1016/j.cell.2007.05.042
- Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., . . . Wu, Q. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4), 900-910. doi:10.1016/j.cell.2015.07.038
- Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yanez-Cuna, J. O., Amendola, M., van Ruiten, M. S., . . . Rowland, B. D. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, 169(4), 693-707 e614. doi:10.1016/j.cell.2017.04.013
- Hacisuleyman, E., Goff, L. A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., . . . Rinn, J. L. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol*, 21(2), 198-206. doi:10.1038/nsmb.2764
- Haddad, N., Vaillant, C., & Jost, D. (2017). IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Res*, 45(10), e81. doi:10.1093/nar/gkx036
- Hansen, A. S. (2020). CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. *Nucleus*, 11(1), 132-148. doi:10.1080/19491034.2020.1782024
- Hansen, A. S., Cattoglio, C., Darzacq, X., & Tjian, R. (2018). Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*, 9(1), 20-32. doi:10.1080/19491034.2017.1389365

- Happel, N., & Doenecke, D. (2009). Histone H1 and its isoforms: contribution to chromatin structure and function. *Gene*, *431*(1-2), 1-12. doi:10.1016/j.gene.2008.11.003
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, *39*(3), 311-318. doi:10.1038/ng1966
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, *38*(4), 576-589. doi:10.1016/j.molcel.2010.05.004
- Heurteau, A., Perrois, C., Depierre, D., Fosseprez, O., Humbert, J., Schaak, S., & Cuvier, O. (2020). Insulator-based loops mediate the spreading of H3K27me3 over distant micro-domains repressing euchromatin genes. *Genome Biol*, *21*(1), 193. doi:10.1186/s13059-020-02106-z
- Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, *169*(1), 13-23. doi:10.1016/j.cell.2017.02.007
- Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., . . . Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), 1454-1458. doi:10.1126/science.aad9024
- Hon, G., Wang, W., & Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, *5*(11), e1000566. doi:10.1371/journal.pcbi.1000566
- Hou, C., Li, L., Qin, Z. S., & Corces, V. G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell*, *48*(3), 471-484. doi:10.1016/j.molcel.2012.08.031
- Hsieh, T. H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, *162*(1), 108-119. doi:10.1016/j.cell.2015.05.048

- Hsieh, T. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., & Darzacq, X. (2020). Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol Cell*. doi:10.1016/j.molcel.2020.03.002
- Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., & Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23), 3131-3133. doi:10.1093/bioinformatics/bts570
- Huang, C., Su, T., Xue, Y., Cheng, C., Lay, F. D., McKee, R. A., . . . Carey, M. (2017). Cbx3 maintains lineage specificity during neural differentiation. *Genes Dev*, 31(3), 241-246. doi:10.1101/gad.292169.116
- Huang, X., & Wang, J. (2014). The extended pluripotency protein interactome and its links to reprogramming. *Curr Opin Genet Dev*, 28, 16-24. doi:10.1016/j.gde.2014.08.003
- Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., . . . Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet*, 46(2), 205-212. doi:10.1038/ng.2871
- Hwang, W. W., Salinas, R. D., Siu, J. J., Kelley, K. W., Delgado, R. N., Paredes, M. F., . . . Lim, D. A. (2014). Distinct and separable roles for EZH2 in neurogenic astroglia. *Elife*, 3, e02439. doi:10.7554/eLife.02439
- Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., . . . Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9(10), 999-1003. doi:10.1038/nmeth.2148
- Izzo, A., Kamieniarz-Gdula, K., Ramirez, F., Noureen, N., Kind, J., Manke, T., . . . Schneider, R. (2013). The genomic landscape of the somatic linker histone subtypes H1.1 to H1.5 in human cells. *Cell Rep*, 3(6), 2142-2154. doi:10.1016/j.celrep.2013.05.003
- Jackson, D. A., Hassan, A. B., Errington, R. J., & Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. *EMBO J*, 12(3), 1059-1065. Retrieved from

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8458323

- Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., . . . Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5), 1369-1384 e1319. doi:10.1016/j.cell.2016.09.037
- Jost, D., Carrivain, P., Cavalli, G., & Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res*, 42(15), 9553-9561. doi:10.1093/nar/gku698
- Jost, D., & Vaillant, C. (2018). Epigenomics in 3D: importance of long-range spreading and specific interactions in epigenomic maintenance. *Nucleic Acids Res*, 46(5), 2252-2264. doi:10.1093/nar/gky009
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., . . . Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314), 430-435. doi:10.1038/nature09380
- Kaneko, S., Son, J., Shen, S. S., Reinberg, D., & Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol*, 20(11), 1258-1264. doi:10.1038/nsmb.2700
- Kang, J., Xu, B., Yao, Y., Lin, W., Hennessy, C., Fraser, P., & Feng, J. (2011). A dynamical model reveals gene co-localizations in nucleus. *PLoS Comput Biol*, 7(7), e1002094. doi:10.1371/journal.pcbi.1002094
- Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., . . . Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A*, 111(17), 6131-6138. doi:10.1073/pnas.1318948111
- Kempfer, R., & Pombo, A. (2020). Methods for mapping 3D chromosome architecture. *Nat Rev Genet*, 21(4), 207-226. doi:10.1038/s41576-019-0195-2
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome

- browser at UCSC. *Genome Res*, 12(6), 996-1006. doi:10.1101/gr.229102
- Kieffer-Kwon, K. R., Nimura, K., Rao, S. S. P., Xu, J., Jung, S., Pekowska, A., . . . Casellas, R. (2017). Myc Regulates Chromatin Decompaction and Nuclear Architecture during B Cell Activation. *Mol Cell*, 67(4), 566-578 e510. doi:10.1016/j.molcel.2017.07.013
- Kieffer-Kwon, K. R., Tang, Z., Mathe, E., Qian, J., Sung, M. H., Li, G., . . . Casellas, R. (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 155(7), 1507-1520. doi:10.1016/j.cell.2013.11.039
- Kim, S., & Shendure, J. (2019). Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol Cell*, 76(2), 306-319. doi:10.1016/j.molcel.2019.08.010
- Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., . . . Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6), 1231-1245. doi:10.1016/j.cell.2006.12.048
- Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J., & Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. *Science*, 366(6471), 1345-1349. doi:10.1126/science.aaz4475
- Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*, 20(4), 207-220. doi:10.1038/s41576-018-0089-8
- Kloet, S. L., Makowski, M. M., Baymaz, H. I., van Voorthuijsen, L., Karemaker, I. D., Santanach, A., . . . Vermeulen, M. (2016). The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat Struct Mol Biol*, 23(7), 682-690. doi:10.1038/nsmb.3248
- Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schopflin, R., Markowski, J., . . . Mundlos, S. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat Cell Biol*, 21(3), 305-310. doi:10.1038/s41556-019-0273-x

- Krietenstein, N., Abraham, S., Veney, S. V., Abdennur, N., Gibcus, J., Hsieh, T. S., . . . Rando, O. J. (2020). Ultrastructural Details of Mammalian Chromosome Architecture. *Mol Cell*, 78(3), 554-565 e557. doi:10.1016/j.molcel.2020.03.003
- Kubben, N., Adriaens, M., Meuleman, W., Voncken, J. W., van Steensel, B., & Misteli, T. (2012). Mapping of lamin A- and progerin-interacting genome regions. *Chromosoma*, 121(5), 447-464. doi:10.1007/s00412-012-0376-7
- Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J. T., Sadreyev, R. I., . . . Kingston, R. E. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell*, 65(3), 432-446 e435. doi:10.1016/j.molcel.2017.01.009
- Lancot, C., Cheutin, T., Cremer, M., Cavalli, G., & Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet*, 8(2), 104-115. doi:nrg2041 [pii]
10.1038/nrg2041
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921. Retrieved from
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11237011
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357-359. doi:10.1038/nmeth.1923
- Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F., & Orlando, V. (2007). Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat Cell Biol*, 9(10), 1167-1174. doi:ncb1637 [pii]
10.1038/ncb1637
- Laugsch, M., Bartusel, M., Rehim, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., . . . Rada-Iglesias, A. (2019). Modeling the Pathological Long-Range Regulatory Effects of Human

- Structural Variation with Patient-Specific hiPSCs. *Cell Stem Cell*, 24(5), 736-752 e712. doi:10.1016/j.stem.2019.03.004
- Le Dily, F., Baù, D., Pohl, A., Vicent, G. P., Serra, F., Soronellas, D., . . . Beato, M. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev*, 28(19), 2151-2162. doi:10.1101/gad.241422.114
- Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., . . . Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2), 84-98. doi:10.1016/j.cell.2011.12.014
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289-293. doi:10.1126/science.1181369
- Lin, D., Bonora, G., Yardimci, G. G., & Noble, W. S. (2019). Computational methods for analyzing and modeling genome structure and organization. *Wiley Interdiscip Rev Syst Biol Med*, 11(1), e1435. doi:10.1002/wsbm.1435
- Liu, X. S., Wu, H., Krzisch, M., Wu, X., Graef, J., Muffat, J., . . . Jaenisch, R. (2018). Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell*, 172(5), 979-992 e976. doi:10.1016/j.cell.2018.01.012
- Liu, Y. R., Laghari, Z. A., Novoa, C. A., Hughes, J., Webster, J. R., Goodwin, P. E., . . . Scotting, P. J. (2014). Sox2 acts as a transcriptional repressor in neural stem cells. *BMC Neurosci*, 15, 95. doi:10.1186/1471-2202-15-95
- Lodato, M. A., Ng, C. W., Wamstad, J. A., Cheng, A. W., Thai, K. K., Fraenkel, E., . . . Boyer, L. A. (2013). SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and

- NPCs to specify cell state. *PLoS Genet*, 9(2), e1003288. doi:10.1371/journal.pgen.1003288
- Loubiere, V., Martinez, A. M., & Cavalli, G. (2019). Cell Fate and Developmental Regulation Dynamics by Polycomb Proteins and 3D Genome Architecture. *Bioessays*, 41(3), e1800222. doi:10.1002/bies.201800222
- Loubiere, V., Papadopoulos, G. L., Szabo, Q., Martinez, A. M., & Cavalli, G. (2020). Widespread activation of developmental gene expression characterized by PRC1-dependent chromatin looping. *Sci Adv*, 6(2), eaax4001. doi:10.1126/sciadv.aax4001
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12), 550. doi:10.1186/s13059-014-0550-8
- Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat Rev Mol Cell Biol*, 13(7), 436-447. doi:10.1038/nrm3382
- Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648), 251-260. doi:10.1038/38444
- Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., . . . Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5), 1012-1025. doi:10.1016/j.cell.2015.04.004
- Lupianez, D. G., Spielmann, M., & Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet*, 32(4), 225-237. doi:10.1016/j.tig.2016.01.003
- Lutz, M., Burke, L. J., Barreto, G., Goeman, F., Greb, H., Arnold, R., . . . Renkawitz, R. (2000). Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Res*, 28(8), 1707-1713. doi:10.1093/nar/28.8.1707
- Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., . . . Duan, Z. (2014). Fine-scale chromatin interaction maps reveal the cis-

- regulatory landscape of human lincRNA genes. *Nat Methods*. doi:10.1038/nmeth.3205
- Malik, L., & Patro, R. (2018). Rich Chromatin Structure Prediction from Hi-C Data. *IEEE/ACM Trans Comput Biol Bioinform*. doi:10.1109/TCBB.2018.2851200
- Marco-Sola, S., Sammeth, M., Guigo, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, 9(12), 1185-1188. doi:10.1038/nmeth.2221
- Margueron, R., & Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet*, 11(4), 285-296. doi:10.1038/nrg2752
- Margueron, R., Trojer, P., & Reinberg, D. (2005). The key to development: interpreting the histone code? *Curr Opin Genet Dev*, 15(2), 163-176. doi:10.1016/j.gde.2005.01.005
- Marti-Renom, M. A., & Mirny, L. A. (2011). Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput Biol*, 7(7), e1002125. doi:10.1371/journal.pcbi.1002125
- Mas, G., Blanco, E., Ballare, C., Sanso, M., Spill, Y. G., Hu, D., . . . Di Croce, L. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet*, 50(10), 1452-1462. doi:10.1038/s41588-018-0218-5
- Mateo, L. J., Murphy, S. E., Hafner, A., Cinquini, I. S., Walker, C. A., & Boettiger, A. N. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, 568(7750), 49-54. doi:10.1038/s41586-019-1035-4
- McAninch, D., & Thomas, P. (2014). Identification of highly conserved putative developmental enhancers bound by SOX3 in neural progenitors using ChIP-Seq. *PLoS ONE*, 9(11), e113361. doi:10.1371/journal.pone.0113361
- McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. *Mol Cell*, 77(4), 688-708. doi:10.1016/j.molcel.2019.12.021
- Mendieta-Esteban, J., Di Stefano, M., Castillo, D., Farabella, I., & Marti-Renom, M. A. (2021). 3D reconstruction of genomic regions

- from sparse interaction data. *NAR Genom Bioinform*, 3(1), lqab017. doi:10.1093/nargab/lqab017
- Meshorer, E., Yellajoshula, D., George, E., Scambler, P. J., Brown, D. T., & Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell*, 10(1), 105-116. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16399082
- Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., . . . Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*, 47(6), 598-606. doi:10.1038/ng.3286
- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., . . . Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), 553-560. doi:nature06008 [pii] 10.1038/nature06008
- Millan-Arino, L., Islam, A. B., Izquierdo-Bouldstridge, A., Mayor, R., Terme, J. M., Luque, N., . . . Jordan, A. (2014). Mapping of six somatic linker histone H1 variants in human breast cancer cells uncovers specific features of H1.2. *Nucleic Acids Res*, 42(7), 4474-4493. doi:10.1093/nar/gku079
- Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res*, 19(1), 37-51. doi:10.1007/s10577-010-9177-0
- Mirny, L. A., Imakaev, M., & Abdennur, N. (2019). Two major mechanisms of chromosome organization. *Curr Opin Cell Biol*, 58, 142-152. doi:10.1016/j.ceb.2019.05.001
- Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell*, 128(4), 787-800. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17320514
- mod, E. C., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., . . . Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012), 1787-1797. doi:10.1126/science.1198374

- Morey, L., Aloia, L., Cozzuto, L., Benitah, S. A., & Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. *Cell Rep*, 3(1), 60-69. doi:10.1016/j.celrep.2012.11.026
- Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods*, 13(11), 919-922. doi:10.1038/nmeth.3999
- Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., . . . Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469), 59-64. doi:10.1038/nature12593
- Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., & Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, 347(6225), 1017-1021. doi:10.1126/science.1262088
- Nasmyth, K., & Haering, C. H. (2009). Cohesin: its roles and mechanisms. *Annu Rev Genet*, 43, 525-558. doi:10.1146/annurev-genet-102108-134233
- Nguyen, H. Q., Chattoraj, S., Castillo, D., Nguyen, S. C., Nir, G., Lioutas, A., . . . Wu, C. T. (2020). 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat Methods*, 17(8), 822-832. doi:10.1038/s41592-020-0890-0
- Nichols, M. H., & Corces, V. G. (2015). A CTCF Code for 3D Genome Architecture. *Cell*, 162(4), 703-705. doi:10.1016/j.cell.2015.07.053
- Nir, G., Farabella, I., Perez Estrada, C., Ebeling, C. G., Beliveau, B. J., Sasaki, H. M., . . . Wu, C. T. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet*, 14(12), e1007872. doi:10.1371/journal.pgen.1007872
- Nishi, Y., Zhang, X., Jeong, J., Peterson, K. A., Vedenko, A., Bulyk, M. L., . . . McMahon, A. P. (2015). A direct fate exclusion mechanism by Sonic hedgehog-regulated transcriptional

- repressors. *Development*, 142(19), 3286-3293. doi:10.1242/dev.124636
- Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., & Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *Elife*, 3, e02557. doi:10.7554/eLife.02557
- Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., & Duboule, D. (2011). The dynamic architecture of Hox gene clusters. *Science*, 334(6053), 222-225. doi:10.1126/science.1207194
- Nora, E. P., Goloborodko, A., Valton, A. L., Gibcus, J. H., Uebersohn, A., Abdennur, N., . . . Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169(5), 930-944 e922. doi:10.1016/j.cell.2017.05.004
- Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., . . . Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398), 381-385. doi:10.1038/nature11049
- Norton, H. K., Emerson, D. J., Huang, H., Kim, J., Titus, K. R., Gu, S., . . . Phillips-Cremins, J. E. (2018). Detecting hierarchical genome folding with network modularity. *Nat Methods*, 15(2), 119-122. doi:10.1038/nmeth.4560
- Norton, H. K., & Phillips-Cremins, J. E. (2017). Crossed wires: 3D genome misfolding in human disease. *J Cell Biol*, 216(11), 3441-3452. doi:10.1083/jcb.201611001
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A*, 115(29), E6697-E6706. doi:10.1073/pnas.1717730115
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1), D733-745. doi:10.1093/nar/gkv1189

- Ogiyama, Y., Schuettengruber, B., Papadopoulos, G. L., Chang, J. M., & Cavalli, G. (2018). Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Mol Cell*, 71(1), 73-88 e75. doi:10.1016/j.molcel.2018.05.032
- Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N. O., Lubling, Y., . . . Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature*, 540(7632), 296-300. doi:10.1038/nature20158
- Oluwadare, O., Highsmith, M., & Cheng, J. (2019). An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol Proced Online*, 21, 7. doi:10.1186/s12575-019-0094-0
- Ong, C. T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*, 15(4), 234-246. doi:10.1038/nrg3663
- Ou, H. D., Phan, S., Deerinck, T. J., Thor, A., Ellisman, M. H., & O'Shea, C. C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349). doi:10.1126/science.aag0025
- Oudelaar, A. M., Davies, J. O. J., Hanssen, L. L. P., Telenius, J. M., Schwessinger, R., Liu, Y., . . . Hughes, J. R. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat Genet*, 50(12), 1744-1751. doi:10.1038/s41588-018-0253-2
- Oudelaar, A. M., & Higgs, D. R. (2021). The relationship between genome structure and function. *Nat Rev Genet*, 22(3), 154-168. doi:10.1038/s41576-020-00303-x
- Palstra, R. J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., & de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet*, 35(2), 190-194. doi:10.1038/ng1244
- Papantonis, A., Kohro, T., Baboo, S., Larkin, J. D., Deng, B., Short, P., . . . Cook, P. R. (2012). TNFalpha signals through specialized factories where responsive coding and miRNA genes are

- transcribed. *EMBO J*, 31(23), 4404-4414. doi:10.1038/emboj.2012.288
- Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., . . . van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*, 38(4), 603-613. doi:S1097-2765(10)00321-7 [pii] 10.1016/j.molcel.2010.03.016
- Phair, R. D., & Misteli, T. (2000). High mobility of proteins in the mammalian cell nucleus. *Nature*, 404(6778), 604-609. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10766243
- Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., . . . Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6), 1281-1295. doi:10.1016/j.cell.2013.04.053
- Pierce, B. A. (2012). Genetics: A conceptual approach.
- Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol*, 16(4), 245-257. doi:10.1038/nrm3965
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842. doi:10.1093/bioinformatics/btq033
- Quinodoz, S. A., Jachowicz, J. W., Bhat, P., Ollikainen, N., Banerjee, A. K., Goronzy, I. N., . . . Guttman, M. (2021). RNA promotes the formation of spatial compartments in the nucleus. *Cell*, 184(23), 5775-5790 e5730. doi:10.1016/j.cell.2021.10.014
- Quinodoz, S. A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J. M., Detmar, E., . . . Guttman, M. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, 174(3), 744-757 e724. doi:10.1016/j.cell.2018.05.024
- Rajapakse, I., & Groudine, M. (2011). On emerging nuclear order. *J Cell Biol*, 192(5), 711-721. doi:10.1083/jcb.201010129

- Rajapakse, I., Perlman, M. D., Scalzo, D., Kooperberg, C., Groudine, M., & Kosak, S. T. (2009). The emergence of lineage-specific chromosomal topologies from coordinate gene regulation. *Proc Natl Acad Sci U S A*, 106(16), 6679-6684. doi:10.1073/pnas.0900986106
- Ramachandrareddy, H., Bouska, A., Shen, Y., Ji, M., Rizzino, A., Chan, W. C., & McKeithan, T. W. (2010). BCL6 promoter interacts with far upstream sequences with greatly enhanced activating histone modifications in germinal center B cells. *Proc Natl Acad Sci U S A*, 107(26), 11930-11935. doi:10.1073/pnas.1004962107
- Ramani, V., Cusanovich, D. A., Hause, R. J., Ma, W., Qiu, R., Deng, X., . . . Duan, Z. (2016). Mapping 3D genome architecture through in situ DNase Hi-C. *Nat Protoc*, 11(11), 2104-2121. doi:10.1038/nprot.2016.126
- Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., . . . Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nat Methods*, 14(3), 263-266. doi:10.1038/nmeth.4155
- Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Gruning, B. A., Villaveces, J., . . . Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*, 9(1), 189. doi:10.1038/s41467-017-02525-w
- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665-1680. doi:10.1016/j.cell.2014.11.021
- Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., . . . Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2), 305-320 e324. doi:10.1016/j.cell.2017.09.026
- Reddy, K. L., Zullo, J. M., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, 452(7184), 243-247. doi:10.1038/nature06727

- Ricci, M. A., Manzo, C., Garcia-Parajo, M. F., Lakadamyali, M., & Cosma, M. P. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, 160(6), 1145-1158. doi:10.1016/j.cell.2015.01.054
- Rosa, A., & Everaers, R. (2008). Structure and dynamics of interphase chromosomes. *PLoS Comput Biol*, 4(8), e1000153. doi:10.1371/journal.pcbi.1000153
- Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nat Rev Genet*, 19(12), 789-800. doi:10.1038/s41576-018-0060-8
- Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., . . . Corces, V. G. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell*, 67(5), 837-852 e837. doi:10.1016/j.molcel.2017.07.022
- Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., . . . Sali, A. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol*, 10(1), e1001244. doi:10.1371/journal.pbio.1001244
- PBIOLOGY-D-11-02156 [pii]
- Rust, M. J., Bates, M., & Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat Methods*, 3(10), 793-795. doi:nmeth929 [pii] 10.1038/nmeth929
- Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., . . . Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400). doi:10.1126/science.aar3958
- Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., . . . Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 112(47), E6456-6465. doi:10.1073/pnas.1518552112
- Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414), 109-113. doi:10.1038/nature11279

- Sati, S., Bonev, B., Szabo, Q., Jost, D., Bensadoun, P., Serra, F., . . . Cavalli, G. (2020). 4D Genome Rewiring during Oncogene-Induced and Replicative Senescence. *Mol Cell*, 78(3), 522-538 e529. doi:10.1016/j.molcel.2020.03.007
- Schalch, T., Duda, S., Sargent, D. F., & Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047), 138-141. doi:10.1038/nature03686
- Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., . . . Fraser, P. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, 42(1), 53-61. doi:10.1038/ng.496
- Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B. M., Armstrong, H., Mifsud, B., . . . Elderkin, S. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet*, 47(10), 1179-1186. doi:10.1038/ng.3393
- Schuettengruber, B., Bourbon, H. M., Di Croce, L., & Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell*, 171(1), 34-57. doi:10.1016/j.cell.2017.08.002
- Schuettengruber, B., Oded Elkayam, N., Sexton, T., Entrevan, M., Stern, S., Thomas, A., . . . Cavalli, G. (2014). Cooperativity, specificity, and evolutionary stability of Polycomb targeting in Drosophila. *Cell Rep*, 9(1), 219-233. doi:10.1016/j.celrep.2014.08.072
- Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., . . . Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678), 51-56. doi:10.1038/nature24281
- Segal, E., & Widom, J. (2009). What controls nucleosome positions? *Trends Genet*, 25(8), 335-343. doi:10.1016/j.tig.2009.06.002
- Serra, F., Bau, D., Goodstadt, M., Castillo, D., Filion, G. J., & Marti-Renom, M. A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol*, 13(7), e1005665. doi:10.1371/journal.pcbi.1005665

- Serra, F., Di Stefano, M., Spill, Y. G., Cuartero, Y., Goodstadt, M., Bau, D., & Marti-Renom, M. A. (2015). Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett*, 589(20 Pt A), 2987-2995. doi:10.1016/j.febslet.2015.05.012
- Sexton, T., & Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, 160(6), 1049-1059. doi:10.1016/j.cell.2015.02.040
- Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., . . . Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3), 458-472. doi:10.1016/j.cell.2012.01.010
- Shachar, S., & Misteli, T. (2017). Causes and consequences of nuclear gene positioning. *J Cell Sci*, 130(9), 1501-1508. doi:10.1242/jcs.199786
- Shachar, S., Voss, T. C., Pegoraro, G., Sciascia, N., & Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. *Cell*, 162(4), 911-923. doi:10.1016/j.cell.2015.07.035
- Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., . . . Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*. doi:10.1038/nature11243
nature11243 [pii]
- Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., & Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*, 44(7), e70. doi:10.1093/nar/gkv1505
- Shin, Y., & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science*, 357(6357). doi:10.1126/science.aaf4382
- Shin, Y., Chang, Y. C., Lee, D. S. W., Berry, J., Sanders, D. W., Ronceray, P., . . . Brangwynne, C. P. (2018). Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. *Cell*, 175(6), 1481-1491 e1413. doi:10.1016/j.cell.2018.10.057

- Siggens, L., & Ekwall, K. (2014). Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. *J Intern Med*, 276(3), 201-214. doi:10.1111/joim.12231
- Simon, M. D., Pinter, S. F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S. K., . . . Lee, J. T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, 504(7480), 465-469. doi:10.1038/nature12719
- Soler-Vila, P., Cusco, P., Farabella, I., Di Stefano, M., & Marti-Renom, M. A. (2020). Hierarchical chromatin organization detected by TADpole. *Nucleic Acids Res*, 48(7), e39. doi:10.1093/nar/gkaa087
- Soshnikova, N., Montavon, T., Leleu, M., Galjart, N., & Duboule, D. (2010). Functional analysis of CTCF during mammalian limb development. *Dev Cell*, 19(6), 819-830. doi:10.1016/j.devcel.2010.11.009
- Spitz, F. (2016). Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin Cell Dev Biol*, 57, 57-67. doi:10.1016/j.semcdb.2016.06.017
- Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., . . . Graf, T. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet*, 50(2), 238-249. doi:10.1038/s41588-017-0030-7
- Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., . . . Laue, E. D. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648), 59-64. doi:10.1038/nature21429
- Strom, A. R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., & Karpen, G. H. (2017). Phase separation drives heterochromatin domain formation. *Nature*, 547(7662), 241-245. doi:10.1038/nature22989
- Sutherland, H., & Bickmore, W. A. (2009). Transcription factories: gene expression in unions? *Nat Rev Genet*, 10(7), 457-466. doi:nrg2592 [pii]
- 10.1038/nrg2592

- Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., & Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev Cell*, 39(5), 529-543. doi:10.1016/j.devcel.2016.10.015
- Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., . . . Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res*, 24(3), 390-400. doi:10.1101/gr.163519.113
- Szabo, Q., Donjon, A., Jerkovic, I., Papadopoulos, G. L., Cheutin, T., Bonev, B., . . . Cavalli, G. (2020). Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat Genet*, 52(11), 1151-1157. doi:10.1038/s41588-020-00716-8
- Szabo, Q., Jost, D., Chang, J. M., Cattoni, D. I., Papadopoulos, G. L., Bonev, B., . . . Cavalli, G. (2018). TADs are 3D structural units of higher-order chromosome organization in *Drosophila*. *Sci Adv*, 4(2), eaar8082. doi:10.1126/sciadv.aar8082
- Taddei, A., Van Houwe, G., Nagai, S., Erb, I., van Nimwegen, E., & Gasser, S. M. (2009). The functional importance of telomere clustering: global changes in gene expression result from SIR factor dispersion. *Genome Res*, 19(4), 611-625. doi:10.1101/gr.083881.108
- Tan, L., Xing, D., Chang, C. H., Li, H., & Xie, X. S. (2018). Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405), 924-928. doi:10.1126/science.aat5641
- Tan-Wong, S. M., Wijayatilake, H. D., & Proudfoot, N. J. (2009). Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. *Genes Dev*, 23(22), 2610-2624. doi:10.1101/gad.1823209
- Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., . . . Proudfoot, N. J. (2012). Gene loops enhance transcriptional directionality. *Science*, 338(6107), 671-675. doi:10.1126/science.1224350
- Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J., & Bickmore, W. A. (2014). Chromatin decondensation is

- sufficient to alter nuclear organization in embryonic stem cells. *Science*, 346(6214), 1238-1242. doi:10.1126/science.1259587
- Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6), 1453-1465. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12504019
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111. doi:10.1093/bioinformatics/btp120
- Travers, A., & Muskhelishvili, G. (2015). DNA structure and function. *Febs J*, 282(12), 2279-2295. doi:10.1111/febs.13307
- Tremethick, D. J. (2007). Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, 128(4), 651-654. doi:10.1016/j.cell.2007.02.008
- Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., . . . Razin, S. V. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res*, 26(1), 70-84. doi:10.1101/gr.196006.115
- Valton, A. L., & Dekker, J. (2016). TAD disruption as oncogenic driver. *Curr Opin Genet Dev*, 36, 34-40. doi:10.1016/j.gde.2016.03.008
- van Bemmell, J. G., Galupa, R., Gard, C., Servant, N., Picard, C., Davies, J., . . . Heard, E. (2019). The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist. *Nat Genet*, 51(6), 1024-1034. doi:10.1038/s41588-019-0412-0
- van de Werken, H. J., Landan, G., Holwerda, S. J., Hoichman, M., Klous, P., Chachik, R., . . . de Laat, W. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods*, 9(10), 969-972. doi:10.1038/nmeth.2173
- van Steensel, B., & Furlong, E. E. M. (2019). The role of transcription in shaping the spatial organization of the genome. *Nat Rev Mol Cell Biol*, 20(6), 327-337. doi:10.1038/s41580-019-0114-6

- van Steensel, B., & Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*, 18(4), 424-428. doi:10.1038/74487
- Vangala, P., Murphy, R., Quinodoz, S. A., Gellatly, K., McDonel, P., Guttman, M., & Garber, M. (2020). High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. *Mol Cell*, 80(2), 359-373 e358. doi:10.1016/j.molcel.2020.09.005
- Vian, L., Pekowska, A., Rao, S. S. P., Kieffer-Kwon, K. R., Jung, S., Baranello, L., . . . Casellas, R. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell*, 173(5), 1165-1178 e1120. doi:10.1016/j.cell.2018.03.072
- Vidal, E., le Dily, F., Quilez, J., Stadhouders, R., Cuartero, Y., Graf, T., . . . Filion, G. J. (2018). OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res*, 46(8), e49. doi:10.1093/nar/gky064
- Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D., & Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proc Natl Acad Sci U S A*, 112(15), 4672-4677. doi:10.1073/pnas.1504783112
- Voigt, P., Tee, W. W., & Reinberg, D. (2013). A double take on bivalent promoters. *Genes Dev*, 27(12), 1318-1338. doi:10.1101/gad.219626.113
- Wang, H., Nakamura, M., Abbott, T. R., Zhao, D., Luo, K., Yu, C., . . . Qi, L. S. (2019). CRISPR-mediated live imaging of genome editing and transcription. *Science*, 365(6459), 1301-1305. doi:10.1126/science.aax7852
- Wang, L., Gao, Y., Zheng, X., Liu, C., Dong, S., Li, R., . . . Li, P. (2019). Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism. *Mol Cell*, 76(4), 646-659 e646. doi:10.1016/j.molcel.2019.08.019
- Wang, S., Su, J. H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T., & Zhuang, X. (2016). Spatial organization of chromatin domains

- and compartments in single chromosomes. *Science*, 353(6299), 598-602. doi:10.1126/science.aaf8084
- Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., . . . Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40(7), 897-903. doi:ng.154 [pii] 10.1038/ng.154
- Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737-738. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=13054692
- Wei, H., Dong, X., You, Y., Hai, B., Duran, R. C., Wu, X., . . . Wu, J. Q. (2021). OLIG2 regulates lncRNAs and its own expression during oligodendrocyte lineage formation. *BMC Biol*, 19(1), 132. doi:10.1186/s12915-021-01057-6
- Wei, Z., Gao, F., Kim, S., Yang, H., Lyu, J., An, W., . . . Lu, W. (2013). Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell Stem Cell*, 13(1), 36-47. doi:10.1016/j.stem.2013.05.010
- Weischenfeldt, J., Dubash, T., Drainas, A. P., Mardin, B. R., Chen, Y., Stutz, A. M., . . . Korbel, J. O. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat Genet*, 49(1), 65-74. doi:10.1038/ng.3722
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., . . . Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2), 307-319. doi:10.1016/j.cell.2013.03.035
- Wijchers, P. J., Krijger, P. H., Geeven, G., Zhu, Y., Denker, A., Verstegen, M. J., . . . de Laat, W. (2016). Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Mol Cell*, 61(3), 461-473. doi:10.1016/j.molcel.2016.01.001

- Wilkins, M. H., Stokes, A. R., & Wilson, H. R. (1953). Molecular structure of deoxypentose nucleic acids. *Nature*, *171*(4356), 738-740. doi:10.1038/171738a0
- Willcockson, M. A., Heaton, S. E., Weiss, C. N., Bartholdy, B. A., Botbol, Y., Mishra, L. N., . . . Skoultschi, A. I. (2021). H1 histones control the epigenetic landscape by local chromatin compaction. *Nature*, *589*(7841), 293-298. doi:10.1038/s41586-020-3032-z
- Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R. S., Paquette, D., . . . Bickmore, W. A. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev*, *28*(24), 2778-2791. doi:10.1101/gad.251694.114
- Wood, H. B., & Episkopou, V. (1999). Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mech Dev*, *86*(1-2), 197-201. doi:10.1016/s0925-4773(99)00116-1
- Woodcock, C. L. (2005). A milestone in the odyssey of higher-order chromatin structure. *Nat Struct Mol Biol*, *12*(8), 639-640. doi:10.1038/nsmb0805-639
- Wright, A. V., Nunez, J. K., & Doudna, J. A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell*, *164*(1-2), 29-44. doi:10.1016/j.cell.2015.12.035
- Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, *43*(11), 1059-1065. doi:10.1038/ng.947
- Yeo, J. C., & Ng, H. H. (2013). The transcriptional regulation of pluripotency. *Cell Res*, *23*(1), 20-32. doi:10.1038/cr.2012.172
- Yesudhas D, Anwar MA, & S, C. (2019). Structural mechanism of DNA-mediated Nanog–Sox2 cooperative interaction. *The Royal Society of Chemistry*, *9*, 8121-8130.
- Ying, Q. L., Stavridis, M., Griffiths, D., Li, M., & Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal

- precursors in adherent monoculture. *Nat Biotechnol*, 21(2), 183-186. doi:10.1038/nbt780
- You, Q., Cheng, A. Y., Gu, X., Harada, B. T., Yu, M., Wu, T., . . . He, C. (2021). Direct DNA crosslinking with CAP-C uncovers transcription-dependent chromatin organization at high resolution. *Nat Biotechnol*, 39(2), 225-235. doi:10.1038/s41587-020-0643-8
- Zabidi, M. A., & Stark, A. (2016). Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends Genet*, 32(12), 801-814. doi:10.1016/j.tig.2016.10.003
- Zaurin, R., Ferrari, R., Nacht, A. S., Carbonell, J., Le Dily, F., Font-Mateu, J., . . . Vicent, G. P. (2021). A set of accessible enhancers enables the initial response of breast cancer cells to physiological progesterone concentrations. *Nucleic Acids Res*, 49(22), 12716-12731. doi:10.1093/nar/gkab1125
- Zhang, L., Zhang, Y., Chen, Y., Gholamalamdari, O., Wang, Y., Ma, J., & Belmont, A. S. (2020). TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome Res*. doi:10.1101/gr.266239.120
- Zhang, P., Wu, W., Chen, Q., & Chen, M. (2019). Non-Coding RNAs and their Integrated Networks. *J Integr Bioinform*, 16(3). doi:10.1515/jib-2019-0027
- Zhang, T., Zhang, Z., Dong, Q., Xiong, J., & Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol*, 21(1), 45. doi:10.1186/s13059-020-01957-w
- Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., . . . Yue, F. (2018). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun*, 9(1), 750. doi:10.1038/s41467-018-03113-2
- Zhang, Y., Liu, T., Meyer, C. A., Eickhout, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9), R137. doi:10.1186/gb-2008-9-9-r137

Zheng, M., Tian, S. Z., Capurso, D., Kim, M., Maurya, R., Lee, B., . . . Ruan, Y. (2019). Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745), 558-562. doi:10.1038/s41586-019-0949-1