CHROMATIN ORGANIZATION: META-ANALYSIS FOR THE IDENTIFICATION AND CLASSIFICATION OF STRUCTURAL PATTERNS

Silvia Galan Martínez

TESI DOCTORAL UPF / any: 2020

Directors de la tesi:

Marc A. Marti-Renom i François Serra

STRUCTURAL GENOMICS,
GENE REGULATION, STEM CELLS AND CANCER
CENTRE NACIONAL D'ANÀLISI GENÒMICA
CENTRE FOR GENOMIC REGULATION

UNIVERSITAT POMPEU FABRA, DEPARTAMENT DE
CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra *Barcelona*

CRG Centre for Genomic Regulation

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

Acknowledgements

Primer de tot vull agrair a **Marc**, per donar-me aquesta oportunitat, per totes les discussions i per la teva transparència.

También quiero agradecerle a mi supervisor y amigo, François, por toda la ayuda, crítica y soporte durante todos estos años.

I would like to thank all the past and present members of the Marcius Lab, specially **Juan**, por saber tanto de todo, **Yasmina** per sempre superar-te, **David** por toda la ayuda, and **Francesca**, por enseñarme a ser un alma libre.

También quiero agradecer a mis dos chicas guapas por dentro y por fuera, **Mar** y **Alejandra**, las mejores para compartir paellas, vermuts y consejos.

I also met an amazing group of people the **Compactos,** especially **Tobias, Marcos** and **Artem,** the best supporters and debaters.

I would also want to thank **Monika**, my adopted Catalan, for her empathy, strength and fun.

Sé que això no hagués estat possible sense els entrenaments previs amb **Celdillas etmoidales. Patri, Ester** i **Mari** per ser-hi sempre, per ser tant bones. **Clàudia** per ser increïble, perquè ens entenem. **Miquel** per la teva màgia, felicitat i riures. **Javi** por ser el major compañero de cervezas y risas. **Laura** per ser tan dolça i clara. **Ricard**, perquè no entenc com, t'has convertit en família.

Gracias **Mónica**, por tu locura contagiosa, que te hace única y necesaria.

Gràcies, **Lucas,** per fer-me feliç en pensar que et trobaré a casa i compartirem sopes de peix, *bailoteos*, i confinaments, *'Cause somewhere in the crowd there's you…*

A la generació cor, **Mireia**, **Nuria, Ari** i **Estefania**, per tots aquests anys d'amistat, i els que falten; pels bingos, viatges i riures.

Gràcies **Robert**, per fer-me riure tant i per totes les converses interminables.

Vull agrair a la meva família. Pel vostre suport infinit, per ser els meus herois, **Mama** i **Papa**, sempre m'heu ajudat a ser la persona més feliç, per ara i sempre.

A tu, **Gemma**, perquè formes part de mi, i sense tu no hauria tingut la valentia de prendre aquest camí; *You are my person.*

Finalment, vull agrair a les meves àvies, **Iaia Maruja** i **Iaia Montse**, ja no esteu aquí, però sé que us tinc al meu costat, us estimo.

Finally, in this thesis I wanted to highlight brilliant women, as a representation of women in STEM, who were/are fully dedicated and determined to science, and with their commitment helped to fill the gap for future generations.

Abstract

High-throughput Chromosome Conformation Capture (3C) experiments, provide detailed three-dimensional (3D) information about genome organization. Specially, Hi-C, a 3C derivative, has become the standard technique to investigate the 3D chromatin structure, and its functional implication into cell fate determination. However, the correct bioinformatic analysis and interpretation of this data is still an active field of development.

In this thesis, we explore the ability of CTCF to form chromatin loops and their epigenetic signature, by developing Meta-Waffle, an artificial neural network to classify structural patterns without any prior information. This classification, was used to generate a convolutional neural network to *de novo* detect chromatin loops from Hi-C contact matrices, called LOOPbit.

We also present CHESS, a bioinformatic tool for the comparison of chromatin contact maps and differential 3D feature extraction, such as Topologically Associating Domains, stripes or loops.

Resum

L'avenç de mètodes experimentals basats en la captura de la conformació genòmica (3C), estant aportant informació valuosa sobre l'estructura tri-dimensional (3D) del genoma. En especial, Hi-C, un derivat del 3C, s'ha convertit en la tècnica estàndard per estudiar l'estructura 3D de la cromatina i la seva implicació funcional en la determinació de la identitat cel·lular. De fet, el anàlisi i la interpretació correcta d'aquesta informació és encara un camp de desenvolupament bioinformàtic actiu.

En aquesta tesi, explorem la capacitat del CTCF de formar bucles de cromatina i la seva signatura epigenètica, desenvolupant Meta-Waffle, una xarxa neuronal artificial per la classificació de patrons estructurals sense informació prèvia. Aquesta classificació, permet la generació d'una xarxa neuronal convolutiva per la detecció *de novo* de bucles de cromatina en matrius de contacte Hi-C, anomenada LOOPbit.

També presentem CHESS, una eina bioinformàtica per la comparació de mapes de contactes i l'extracció d'estructures diferencials, tals com dominis (anomenats TADs), ratlles o bucles.

Preface

All living organisms are made of cells, the smallest unit of life. Interestingly, human cells contain a subcellular compartment of few micrometres in size, the nuclei. This compartment contains the genomic information, the DNA, which folding is not arbitrary.

Microscopy and Chromosome Conformation Capture (3C) technologies, helped to unveil the complex hierarchical organization of chromatin. The chromatin fiber has to be sufficiently accessible for DNA-binding proteins, which will be crucial for cell maintenance and fate, such as transcription factors, polymerases and chromatin modifiers, but maximizing its compaction. Recent advances in 3C-based technologies allowed the inspection and acquisition of increasing evidences indicating how the genome architecture influence the regulation of gene transcription. The interplay of genome architecture and function is a paramount to understand multiple biological scenarios, such as cell identity, development and disease.

This thesis consists of multiple chapters. First of all, in the Introduction I review each chromatin organization layer and its impact on the transcription regulation. Moreover, I provide information of the tools available to identify structural patterns, which can be key to regulate cell expression in different scenario.

The results obtained in the two main publications of the candidate, are presented in the core chapters I and II. In chapter I, first I present Meta-Waffle, an algorithm to deconvolve and classify the structural pattern of DNA-binding proteins, specifically CTCF, the master regulator of chromatin loops. Then, I present LOOPbit, a convolutional neural network, trained with CTCF loops, which is able to retrieve chromatin loops probabilities genome-wide. In chapter II, I introduce CHESS, a new bioinformatics tool to systematically compare and identify differential structural features between contact matrices.

Finally, the conclusions for both chapters are added to highlight the main contributions of this thesis.

Objectives

The broad goal of this thesis was to provide an in-depth analysis of how mammalian genomes are organized in 3D. Specifically, we studied the role of the insulator protein CCCTC-Binding Factor (CTCF) in the formation of chromatin loops and its biological implications. To achieve this broad goal, were carried out two specific objectives:

1. Specific Aim 1 includes three tasks: (1) to develop a computational tool for deconvoluting the mean interaction signal between pairs of CTCF in an unsupervised manner. (2) To obtain structural subpopulations and their epigenetic signature. And (3) to develop a chromatin loop detection method.

2. Specific Aim 2 includes two tasks: (1) the design and development of an algorithm for the assessment of structural similarity between genomic regions. And (2) the identification and classification of the genome differential structures.

Table of contents

Table of figures

Introduction

"Identity is not an object; it is a process with addresses for all the different directions and dimensions in which it moves, and so it cannot so easily be fixed with a single number."

Lynn Margulis, 1991

The term *cell* (from Latin *cella*, meaning "small room") was coined by Robert Hooke (1665), which has been described to be the smallest unit of life. In 1839, Theodor Schwann and Matthias Jakob Schleiden proposed the cell theory, which says that all living organisms are made of cells, whose size, number and type, will ultimately define the structure and functions of the organism.

Currently, living organisms on Earth are classified in three large kingdoms: archaea, bacteria and eukaryotes. Bacteria and archaea are also called prokaryotes and are simpler, single-celled organisms. While eukaryotes cells contain compartments, called organelles surrounded by membranes, including the mitochondria, the chloroplasts, the smooth and rough endoplasmic reticula and the nucleus. Lynn Margulis proposed the endosymbiotic theory for the origin of eukaryotic cells (Sagan, 1967), however, it is still unclear and have been proposed more than 20 different versions (Martin, Garg, & Zimorski, 2015). This subcellular organization in eukaryotes gives the opportunity to separate metabolic processes, leaving the cell nucleus an unique structure.

The nucleus

> "If you know you are on the right track, if you have this inner knowledge, then nobody can turn you off... no matter what they say…"
>
> Barbara McClintock.

The nucleus represents a milestone in evolution transition. It is enclosed by the nuclear envelope, which separates transcription from translation. This separation still needs the bidirectional trafficking allowed by the nuclear pore complex. The nuclear envelope is also essential for the anchoring of lamins for mechanical support and chromosomal positioning and segregation (Devos, Graf, & Field, 2014). Many nuclear functions such as these complex interactions governing chromosome positioning with respect to the nuclear envelope and their dynamics, are conserved across eukaryotes.

The nucleus was the first organelle discovered by Antoine van Leeuwenhoek (1632-1723), who observed a lumen, the nucleus, in salmon red blood cells. The nucleus is the largest and most easily discernible organelle in eukaryotic cells. The nucleus is a fascinating structure to study, as it regulates in space and time, genomic and metabolic functions such as transcription and genome stability.

The genetic material

> "The results suggest a helical structure (which must be very closely packed) containing 2, 3 or 4 co-axial nucleic acid chains per helical unit, and having the phosphate groups near the outside."
>
> Rosalind Franklin, 1951.

It was not until 1869 that Fiedrich Miescher identified what he called "nuclein" inside the nuclei of human white blood cells (Dahm, 2005). By the twentieth century, Miescher's term fell into oblivion, and nowadays it is known as deoxyribonucleic acid or DNA. In 1909, Phoebus Levene discovered the deoxyribose (carbohydrate element of DNA) and in 1929 the ribose (carbohydrate element of ribonucleic acid or RNA). Moreover, in 1919, Levene proposed the polynucleotide model, which claims that nucleic acids were composed by a series of nucleotides, and each nucleotide was composed of just one of the four nitrogen-containing bases, a glucose molecule and a phosphate group (Levene, 1919). There are two main nitrogen-containing bases classes: purines (adenine (A) and guanine (G); with two fused rings each) and pyrimidines (cytosine (C), thymine (T) and uracil (U); each with a single ring). It is also known that RNA contains A, G, C and U, while DNA contains A, G, C, and T. In 1944, Erwin Chargaff studied the differences on DNA composition between species (Chargaff et al., 1950) and proposed the "Chargaff's rule", which claims that the amount of A was similar to T and the amount of G was similar to C. He shared his studies with James Watson and Francis Crick, who were benefited by it and a DNA X-ray image generated by Rosalind Franklin (Franklin & Gosling, 1953), to claim that the DNA is based by two polynucleotide chains twisted around each other to form a double helix (Watson & Crick, 1953).

The chromatin

Each human diploid cell contains a 2 meters length of DNA fiber, which needs to be efficiently accessible to DNA-binding proteins

(DBPs) and at the same time strengthen and compact. This is possible thanks to the wrapping of DNA fiber around octamers of histone proteins. Together this structure, the nucleosome, is composed by two of each of the four histones (H2A, H2B, H3 and H4), discovered in 1884 by Albrecht Kossel, with ≅ 147 bp of DNA wrapped around. In the 1970s was identified the histone H1 to be responsible to link adjacent nucleosomes (with 20-80 bp distance). Altogether, the nucleosomes and the linker DNA connecting them, adopt an open beads-on-a-string-like structure (Van Holde, 1989)(Figure 1). Each histone has highly disordered N-terminal and C-terminal tails that are susceptible to post-translational modifications (PTMs), which are relevant for the maintenance of the proper cell identity. PTMs are reversible and can regulate the binding of DBPs. There are described a large number of PTMs including methylation, acetylation or phosphorylation. The interaction between the DNA and the histone tails is proposed to drive the liquid-liquid phase separation, which is also facilitated by the linker histone H1 (Gibson et al., 2019). Using live-cell super-resolution imaging, it was observed that chromatin domains behave as "liquid drops", suggesting that are phase-separated through self-assembly providing plasticity to the chromatin to conduct many functions such as DNA repair, gene expression and cell-cycle (Nozaki et al., 2017). The macromolecular complex composed by nucleosomes, linker DNA and other DBPs is called chromatin. Under physiological salt conditions this structure is condensed *in vitro* into a fiber of around 30 nm of diameter, referred to as 30 nm chromatin fiber (Horowitz-Scherer & Woodcock, 2006)(Figure 1). However, this structure is not still clear *in vivo*. The chromatin fiber can be highly packed in a more condensed structure, the chromosomes, which were observed in 1842 by Karl Wilhelm von Nägeli. It was not until 1902, thanks to the work of Walter Sutton and Theodor Boveri, when it was described the chromosome theory of inheritance, in which chromosomes are identified as the carriers of the genetic information.

Considering all the mentioned discoveries, chromatin organization started to be considered of fundamental importance for basic biological processes, such as gene expression. In order to study chromatin structure is key to obtain the genomic sequence, being the first human genome sequence published twenty years ago (Venter et al., 2001). Its discovery contributed into the understanding of the human evolution, disease, and the interplay between environment and heredity.

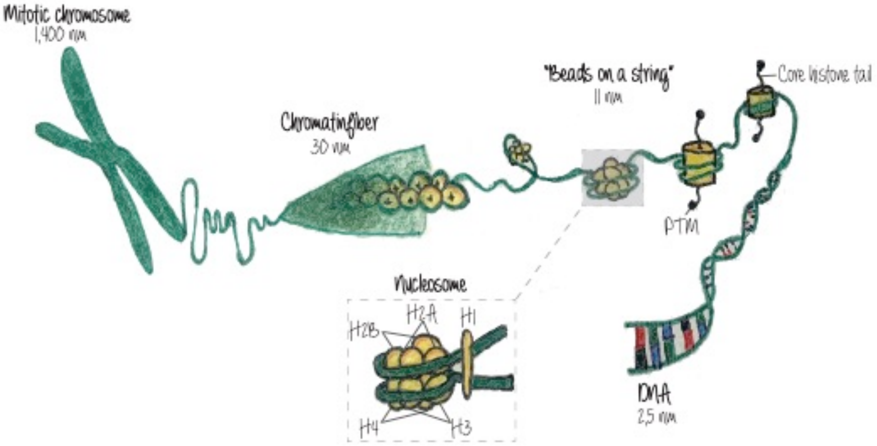


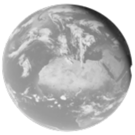
Figure 1. Illustration of the chromatin organization hierarchy in an interphase nucleus. First the double helix DNA is wrapped around the nucleosome, which is shown in detail in the dashed square. The histone tails can present PTMs, enabling chromatin plasticity. The nucleosomes and the linker DNA, form the "beads-on-a-string" like structure. This structure, can be packed to form the chromatin fiber. The chromatin will ultimately be packed conforming the chromosome.

Nuclear organization

The biggest obstacle for women to remain our best in science, I think it is really the combination of career and science. You could say this is not unique to science except that science is really a time eater.

Susan M. Gasser, 2013

The spatial organization of the human genome in the nucleus is known to play a key role in transcriptional regulation. In order to shed light into the assessment of the three-dimensional (3D) architecture of the nucleus, various methods are used, i.e. microscopy and 3C-based technology. Thanks to them, currently it is known that the genome is organized in multiple layers, and each layer has its own regulatory system. Here are discussed the main features in each layer, from larger to finest structures.



Chromosome territories
~100 Mb

In 1885 Carl Rabl suggested a territorial organization of chromosomes, but was in 1909, when Theodor Boveri introduced the term chromosome territories (CTs). He described the non-random distribution of chromosomes during interphase and the maintained structure in the daughter nuclei (Bovery, 1909). From 1970 to 1980 the scientific community believed that chromatin fibers were almost randomly intermingled, picturing the still today very spread image of the spaghetti dish. In 1988 it was possible to directly observe CTs in a microscope using chromosome painting (Lichter, Cremer, Borden, Manuelidis, & Ward, 1988).

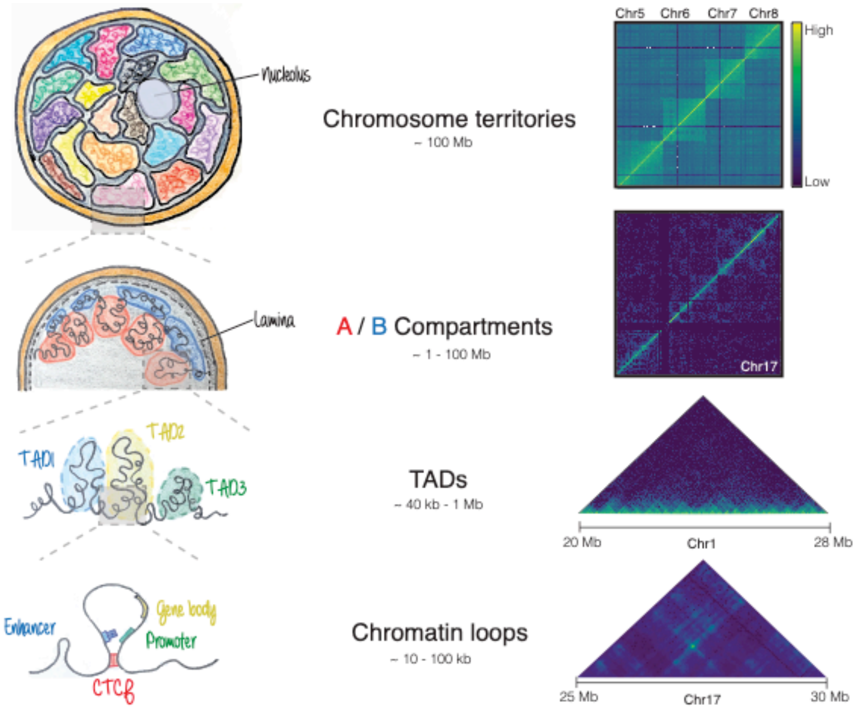


Figure 2. Schematic view of the chromatin organization inside the nucleus. In the left column are illustrated the different chromatin organization layers from chromosome territories to chromatin loops. In the right column a representation of Hi-C maps from GM12878 cells (S. S. Rao et al., 2014) at different genomic scales, reflecting the different layers of chromatin organization. First, in the illustration, chromosomes are represented by different colours, depicting the chromosome territories. This organization is also visible in the Hi-C map, showing higher interaction within chromosomes. Then, chromatin compartments are represented in the illustration using red for the A compartment, and blue for the B compartment, this organization is reflected in the Hi-C map by the checkboard pattern segmentation. At 40 kb – 1 Mb are represented the TAD structures, observed in the rotated Hi-C map as triangles, with high interaction frequencies. Finally, are illustrated chromatin loops, which are demarcated by CTCF and help to bring in close proximity genes and their regulatory units. This structure is found in the Hi-C map by a strong dark peak.

Since then, the details about the organization of chromosomes have been emerging. It has been described that larger chromosomes, that contain higher number of heterochromatic regions, are more present on the nuclear periphery. Whereas smaller chromosomes with higher euchromatic regions are localized at the centre of the nucleus (Tanabe et al., 2002). The boundaries between these CTs have also been precisely described using laser UV microbeam, with the discovery of "intermingling" or "kissing" between chromosomes (Cremer, Cremer, Baumann, et al., 1982). Later, fluorescence *in situ* hybridization (FISH) techniques indicated a probability of interchromosomal associations to occur (Bolzer et al., 2005). This event changed the way to understand how genes can be coregulated considering the third dimension. For example, the olfactory receptor genes, which are around 1,400 genes located across 18 different chromosomes, are gathered at the time of expression into the same interchromatin space, called "olfactosome" (Horta, Monahan, Bashkirova, & Lomvardas, 2019).

Chromosome compartments
~ 1 Mb

It was not until the emergence of a new molecular technique, the high-throughput conformation capture (Hi-C), to be able to discern finer structures inside CTs, which can be as large as several Mb (Lieberman-Aiden et al., 2009). Two major chromatin compartments were elucidated: the A compartment containing more active and open chromatin, and the B compartment containing more inactive and closed chromatin. These two major compartments match with euchromatin and heterochromatin regarding compaction, replication timing, repetitive elements distribution, gene location and expression (Croft et al., 1999). This organization has been maintained in eukaryotes over more than 500 million years, with only few exemptions. Chromosome compartments are cell-type specific, defining the identity of the cell. The 3D organization of the genome is partly driven by the high affinity of active regions for other active regions, which can explain the infrequent euchromatin-heterochromatin interactions. Euchromatin regions are enriched in housekeeping genes, and replicate early in S-phase, while heterochromatin is gene-poor, with tissue-specific genes and with late S-phase replication. Moreover, euchromatin contains most of the short interspersed repetitive sequences (SINEs), whereas retrotransposon-related long elements (LINEs) and long terminal repeats (LTRs) are located preferentially in heterochromatin. Mitosis disrupts this euchromatin-heterochromatin segregation, and then this separation is gradually restored in two phases during interphase. Which are the mechanisms that maintain this organization is not fully known. Various studies have proposed that sequences exhibit high affinity to each other, causing the separation of the two compartments. This can be caused by the attraction of homotypic chromatin marked by the same repetitive sequences and enforced by the binding of architectural and epigenetic factors (Gibcus & Dekker, 2013). Moreover, the level of segregation has been seen to correlate with the cell differentiation state. For instance, embryonic stem cells (ESCs) are devoid of compact heterochromatin domains, but possess hyperdynamic organization, are transcriptionally promiscuous and have an open chromatin configuration. This flexibility is thought to be required for the maintenance of the pluripotent state and for its progression. With cell differentiation there is a loss in potency and the genome is partitioned into larger euchromatin and heterochromatin domains with replication synchrony. For instance, in human ESCs around the 40 % of the genome switches towards B compartment during cell differentiation (Dixon et al., 2015).

Subdomains
~ 50 kb

The chromatin is organized in a subscale into multiple subdomains, which are important for the proper nuclear homeostasis. According to their nuclear localization, and their contact regions, it have been mainly described the topologically associating domains (TADs), the nucleolus associated domains (NAR) and the lamin associated domains (LADs).

Topologically Associating Domains (TADs)

Hi-C techniques revealed that within compartments, chromatin is organized in sub-domains called TADs. TADs are characterized to interact more within themselves, than between adjacent domains. TADs might serve as structural platforms for dynamic *cis*-interactions between regulatory elements. In fact, TADs have helped to improve our understanding of the relation between enhancers and their target genes. Nowadays, it has been observed that actively transcribed genes can form mini-domains that interact more frequently with other active genes. Then, clusters of active genes can form multi-gene domains, with all the belonging genes with a similar transcriptional activity. Within the TAD, chromatin presents a common epigenetic signature and replication timing. Moreover, TADs are highly conserved across cell types, substantiating their importance in cell homeostasis regulation (Dixon et al., 2012; Nora et al., 2012). The high conservation of TAD features in mammals (Dixon et al., 2012; Vietri Rudan et al., 2015), suggests that TADs might represent a basic and ancient structure of the chromatin organization in eukaryotes. The differential TAD sizes between organisms can be explained by the relative sizes of active and inactive chromatin segments, for example, in *Drosophila*, TADs have a mean size of 60 kb, while in mammals it is around 800 kb (Dekker & Heard, 2015). As TADs can differ in size, chromatin features and formation mechanisms, they can be classified into different classes or subtypes, each with a specific structural and functional properties. It is relevant to notice that the identification of TADs is highly dependent on the experiment resolution and the software used to annotate them. Recently, high sequencing depths and resolution experiments have revealed a finer TAD patterning (Forcato et al., 2017; S. S. Rao et al., 2014; Rowley et al., 2017; Zufferey, Tavernari, Oricchio, & Ciriello, 2018). TADs are demarcated by boundaries, which are enriched in multiple genomic features, such as transcription start sites (TSS) and binding sites of CCCTC-binding factor (CTCF). Multiple studies showed the importance of TAD borders to regulate gene expression, for instance their deletion can lead to TAD-fusion events and gene deregulation (Nora et al., 2012). Another studies showed that the disruption or relocation of TAD borders, lead to ectopic contacts between *cis*-regulatory elements, and finally contributing to developmental disorders or cancer (Flavahan et al., 2016; Franke et al., 2016; Hnisz et al., 2016; Kraft et al., 2019; Lupianez et al., 2015). Due to the hierarchical fashion of chromatin organization, the compartmentalization of the genome (see Chromosome compartments section) is responsible for both long-range (as for genomic compartments) and local domains (as for TADs). However, their regulatory cross-talk is not fully described. A study of CTCF in loop and TAD formation in mammals, showed that the loss of CTCF using an auxin-mediated system, was translated into a loss of TAD domains, while compartments were maintained (E. P. Nora et al., 2017). Furthermore, the transcriptional activity was slightly affected, suggesting TAD compartmentalization as a fundamental support for the regulation of transcription. In order to shed light into the interplay between transcription and TAD formation, it has been recently suggested that it can take place even after the inhibition of transcription in *Drosophila* embryos. However, it is not totally clear as under the triptolide treatment used, the RNA-Pol II (RNAPII) remains bound to promoter genes (Hug, Grimaldi, Kruse, & Vaquerizas, 2017). Nevertheless, differential gene expression in multiple cell types, can result in the formation of distinct compartmental domains. Supporting the idea that TADs, which are formed by compartments, may be different between cell types with different transcriptional patterns. The regulation of the chromatin organization by gene expression, can be explained in the organisms, like in *Drosophila*, in which CTCF is not found in TADs. However, as mentioned before, in mammals CTCF is an essential architectural protein, and super-resolution microscopy experiments suggested that CTCF together with cohesin are required to for TAD boundaries position rather than for TAD formation. FISH labelling, proved the presence of TAD-like structural units in single cells, in wild type and in cohesin-depleted conditions. While the position of the TAD boundaries in the wild-type lies more often at CTCF sites, it is random in cohesin-depleted samples (Bintu et al., 2018). Some studies showed that there is ~ 20 % of TAD boundaries independent of CTCF, suggesting their resilience after the loss of CTCF (E. P. Nora et al., 2017). These boundaries might be associated to transcription, as proposed before for *Drosophila*, (Bonev et al., 2017; Dixon et al., 2012) or can correspond to frontier between active and inactive chromatin regions, such as A and B compartment types (S. S. Rao et al., 2014; Rowley et al., 2017). Nevertheless, recent studies showed using CRISPR-dCas9-mediated transcriptional activation its inability to create new TAD borders (Bonev et al., 2017). This result suggests that transcription is not sufficient to demarcate CTCF-independent TAD boundaries. TADs have been observed to appear gradually in early mouse embryogenesis, and they can still be observed after inhibiting transcription with α-amanitin. However, the inhibition of DNA-replication with aphidicolin blocked the TAD establishment, indicating the potential of replication for the primary establishment of TADs (Du et al., 2017; Ke et al., 2017). Finally, it exists a correlation between the conserved TADs and the conservation of the CTCF binding sites and their motif orientation at their boundaries. Interestingly, it has been observed that changes within TADs are correlated with changes in the binding and orientation of CTCF. Thus, TADs are maintained as core structures during evolution, being each a platform imprinted with a specific functional regulatory scenario. Indeed genomic sequences at TAD boundaries are hotspots for genomic rearrangements as they appear to be locally open chromatin with higher frequency of double strand breaks (Guelen et al., 2008).

Lamina-Associated Domains (LADs)

The nuclear lamina is a meshwork of intermediate filament proteins called lamins, which is subjacent to the internal side of the nuclear membrane. These lamina consist of lamins A and C (also referred as lamins A/C or lamin A; both splice variants of the *LMNA* gene) and lamins B1 and B2 (products of *LMNB1* and *LMNB2* genes) (de Leeuw, Gruenbaum, & Medalia, 2018). Specifically, Lamin B-receptor (LBR) and Lamin A/C have been described as major LAD tethers (Solovei et al., 2013). LADs are enriched in LINEs, and in heterochromatin compartment and have a size between 100 kb and 10 Mb (Guelen et al., 2008). Various DNA motifs and proteins have been described to play a role driving LADs to the nuclear periphery. The use of the DNA adenine methyltransferase identification (DamID) method (van Steensel & Henikoff, 2000) was key to infer the LAD composition and function, disentangling the basic principles of genome organization. Lamin B1 is observed at the nuclear periphery, while lamin A is also present in the nuclear interior (Kind et al., 2013; E. Lund et al., 2013). While peripheral LADs are gene-poor, transcriptionally silent and heterochromatic, intranuclear LADs tend to be more gene-rich, transcriptionally active and euchromatic (Gesson et al., 2016; E. G. Lund, Duband-Goulet, Oldenburg, Buendia, & Collas, 2015). This spatial distribution of lamin A, may explain its impact on the radial positioning of chromatin and its dynamics (Solovei et al., 2013; Vivante, Brozgol, Bronshtein, Levi, & Garini, 2019). However, it is relevant to know that lamin A is not sufficient to anchor heterochromatin into the nuclear periphery, suggesting the need of lamin-associated protein complexes containing integral proteins of the inner nuclear membrane (Buchwalter, Kaneshiro, & Hetzer, 2019). Lamin B1, despite its exclusively peripheric location, also seems to be associated to a relatively complex regulation mechanism as it can be found in euchromatin, presenting a dynamic role in the execution of the Epithelial-Mesenchymal Transition (EMT) transcriptional program (Pascual-Reguant et al., 2018)(see Annex). Recent experiments using FISH have shown a radial repositioning of TADs dependent on the presence of lamin B1 LADs (Forsberg, Brunet, Ali, & Collas, 2019). The interplay between TADs and LADs to orchestrate spatial genome topology has been studied in a differentiation system showing its fundamental role to shape the 4-dimensional genome during differentiation (Paulsen et al., 2019).

Nucleolus-Associated Domains (NADs)

The largest substructure in the nucleus is the nucleolus, which is the responsible of the ribosome biogenesis. This process is initiated by the transcription of ribosomal RNA (rRNA) genes, the basic component of the nucleolus. The ribosome biogenesis is vital for the assembly of the ribosome to regulate the translational state of the cell. The chromatin regions in contact with the nucleolus are referred as NADs. These subdomains were firstly observed through a very specific sonication of the nuclei that allowed to obtain unbroken nucleoli (Sullivan et al., 2001). Some years later, NADs were mapped in HeLa, IMR90 and HT1080 human cell lines (Nemeth et al., 2010; van Koningsbruggen et al., 2010). NADs are composed by regions with low gene density and transcriptional levels and enriched in repressive histone modifications (H3K27me3, H3K9me3 and H4K20me3). Microscopy studies identified centromeric and pericentromeric satellite repetitive regions and subtelomeric regions as NADs. Moreover, NADs have been observed to cover around 40 % of the genome, and revealed a considerable overlap with LADs (van Koningsbruggen et al., 2010).

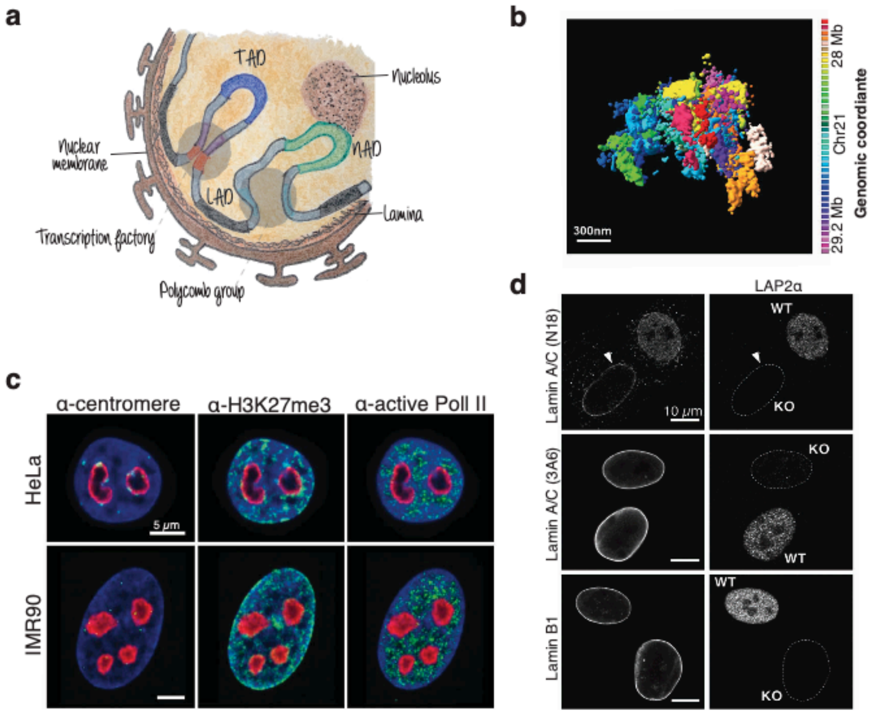
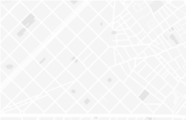


Figure 3. Subdomains organization. a, Illustration of a cell nuclei with the three subdomains explained in the Subdomains section. The transcription factory is highlighted in a grey circle next to the TAD represented as a blue rectangle, helping to gather regulatory units highlighted in coloured boxes, to enhance gene expression. LADs are represented as black rectangles, being in contact with the lamina. Polycomb group proteins are highlighted with a grey circle, being relevant to regulate cell identity. Finally, NADs are represented using green rectangles, being in contact with the nucleolus. **b,** 3D STORM images of 41 consecutive 30 kb chromatin segments from a 1.2 Mb of IMR90 chromosome 21 (Adapted from (Bintu et al., 2018)). **c,** Immunostaining in HeLA and IMR90 cells of α-centromere, α-H3K27me3 and α-active Pol II signals shown in green, nucleolar staining in red, and DAPI stain in blue (Adapted from (Nemeth et al., 2010)). **d,** Confocal immunofluorescence microscopic images of wild-type and knock-out mouse dermal fibroblast double-stained for LAP2α and lamins. The unstained nuclei are highlighted using a white dashed line (Adapted from: (Gesson et al., 2016)).

The unveiling of the role of the nucleolus in genome architecture has been impeded by its membrane-less substructure, which makes difficult the precise mapping of the chromatin domains that are contacting it.

This factor limits the application of the DamID technology as done for the genome-wide mapping of the LADs. However, it has been hypothesized that alterations at rRNA repeats would alter the nucleolus in its structure and protein composition, promoting the required concentration of rRNA regulatory factors to establish repressive states (Bersaglieri & Santoro, 2019). Finally, both NADs and LADs are suggested to represent the hub for organizing the inactive or heterochromatin genome regions.



DNA-looping

~ 10 – 100 kb

Genetic studies have demonstrated that translocations or deletions can affect the correct transcription regulation of genes located at distal regions (tens to hundreds of kilobases). It indicates the ability of regulatory elements to exert their function over large genomic distances. Enhancers have been described as key gene-regulatory elements that can control gene expression in a cell-specific and spatiotemporal manner through long-range chromosomal interactions. In mammal genomes, such as mouse and human, hundreds of thousands of putative enhancers have been mapped. The elucidation of enhancer-promoter pairs solely based on the linear distance leads to high number of false positive assignments. Thanks to major technological advances that allowed the genome-wide mapping of enhancer-promoter contacts at high resolution, elucidated the principles of enhancer function and enhancer-promoter communication. These type of events are known as DNA-looping or loops, which permit the regulation of genes by distant regulatory elements.

Loop extrusion model

Over the years the study of how genes can be regulated from distant elements has been of high interest, being the "loop model" the more prevalent (Ptashne, 1986). As mentioned in previous sections (see Topologically Associating Domains), CTCF has been observed to play a major role in chromatin organization. CTCF was first described as a repressor of the *C-MYC* oncogene in chicken, mouse and human (Filippova et al., 1996). Classically, CTCF has been observed as an insulator, blocking the communication between gene promoters and distal enhancers. The first study proving this CTCF capacity was in transgenic assays in the chicken β–globin locus, where the enhancer and the reporter gene are separated by 1.2 kb DNA segment. This study revealed that the insulator ability of CTCF was dependent on its relative position, as was only observed when placed between the enhancer and the promoter (Bell & Felsenfeld, 1999). CTCF is a zinc-finger protein highly conserved across bilaterians and absent in yeast, derived nematodes as *Caenorhabditis elegans*, fungi and plants (Heger, Marin, Bartkuhn, Schierenberg, & Wiehe, 2012). The central region of CTCF consists of 11 zinc fingers, with around 20 bp well-conserved and non-palindromic as a core region, also referred as M1 motif (Schmidt et al., 2012)(Figure 4). What makes CTCF binding motif unique, in comparison to the majority of the known transcription factors (TFs), is that it is long and information-enriched, meaning that it can possess pleiotropic functions by diverse combination of its 11 zinc fingers (Nakahashi et al., 2013). Moreover, in mammals have been described shorter motifs of around 10 bp, up- and downstream to M1, which can help to stabilize or destabilize the CTCF binding, respectively. Despite some CTCF sites lack any sequence motif, they present a markedly lower affinity than those with the core motif (Nakahashi et al., 2013)(Figure 4).

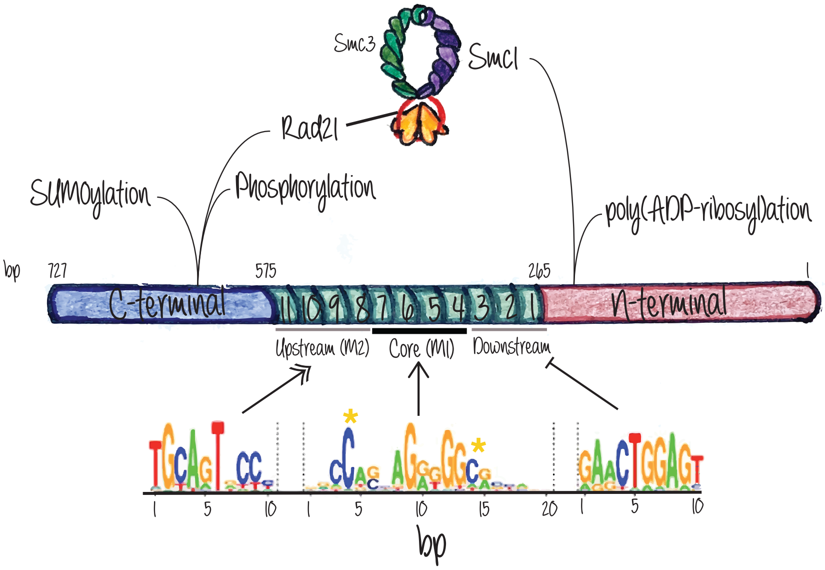


Figure 4. Illustration of CTCF regulation. CTCF is depicted in a reverse orientation to reflect its binding. CTCF can be regulated by diverse post-translation modifications (PTMs), such as SUMOylation, phosphorylation (by casein kinase 2) or Poly(ADP-ribos)ylation by PARP1. While the C-terminal of CTCF can interact with the cohesin subunit, Rad21, the N-terminal interacts with the cohesin subunit, Smc1. CTCF can regulate its binding to chromatin by the central 11 zinc finger domains. The core motif (M1) consists of the 4 central zinc fingers; whereas the up and downstream motifs can stabilize or destabilize its binding, respectively. The methylation of two cytosines located in the core motif, depicted with yellow asterisks, can impede the CTCF binding.

The binding of CTCF into chromatin correlates to gene density, especially in intergenic regions, gene bodies and near TSSs (Holwerda & de Laat, 2013). Moreover, its binding capacity can be regulated at multiple levels. First, the methylation of cytosines in the M1 motif will impair the CTCF binding (Wang et al., 2012). Second, the binding has been observed normally to happen in nucleosome-free regions (Fu, Sinha, Peterson, & Weng, 2008). Finally, PTMs in the CTCF protein such as SUMOylation, poly(ADP-ribosyl)ation and phosphorylation by casein-kinase 2 (CK2), can interfere CTCF function without affecting its chromatin binding (El-Kady & Klenova, 2005; Kitchen & Schoenherr, 2010; Pavlaki et al., 2018)(Figure 4).

Mammalian genomes contain around 50,000 CTCF binding sites, with around 10-20 % located at TAD boundaries and 60-70 % in intra-domain regions (Cuddapah et al., 2009; S. S. Rao et al., 2014; Tang et al., 2015). CTCF has been described to be an architectural protein, as well as, cohesin and Mediator, all three are involved in chromatin-looping (Parelho et al., 2008; Rubio et al., 2008), but are not fully required for the maintenance of TADs. Cohesin is a highly conserved protein complex assembled in a ring-like structure by two proteins from the "Structural Maintenance of Chromosomes" family, Smc1 and Smc3, and a kleisin family subunit, Rad21. The cohesin complex is mainly known to be responsible for tethering sister chromatids during mitosis by encircling the DNA fiber. However, cohesin has been found in non-dividing cells, and mutations on its subunits can lead to developmental impairments in humans, suggesting its role in gene regulation (Brooker & Berkowitz, 2014). In vertebrates, it was observed that around 50 to 80 % of CTCF sites were co-occupied by cohesin, which can interact directly with the C-terminal of CTCF through the Rad21 subunit (Xiao, Wallace, & Felsenfeld, 2011)(Figure 4). While cohesin is not needed for the binding and positioning of CTCF, CTCF is necessary for the positioning of cohesin. As expected and according to all these observations, CTCF and cohesin have been found to participate in the chromatin loop formation, having a direct effect on gene activation.

Recent studies have showed how the CTCF N-terminus interacts with Smc1 subunits of human cohesin, probing that this interaction is required for the positioning of cohesin in CTCF-anchored loops (Li et al., 2020)(Figure 4). Nevertheless, CTCF is able to form loops via homodimerization, or by interacting with other proteins: such as, nucleophosmin (NPM) (Yusufzai, Tagami, Nakatani, & Felsenfeld, 2004), thyroid hormone receptor (TR) (Lutz et al., 2003), chromodomain helicase 8 (CHD8) (Ishihara, Oshimura, & Nakao, 2006) and transcription factor Yin Yang 1 (YY1)(C. Guo et al., 2011). Despite all the possible interactors of CTCF, none has been observed to co-localize genome-wide as extensively as with cohesin. This suggests that CTCF might have different interactors depending on the epigenetic scenario, participating in a wide range of protein complex modulations to adapt its function to specific loci in a given conditions. For example, various loops can appear during differentiation, without changes in CTCF binding, which have been observed to be enriched in other TFs binding motifs (Bonev et al., 2017; Phanstiel et al., 2017).

Hi-C experiments showed that CTCF chromatin loops occur preferentially between motifs in convergent orientation (S. S. Rao et al., 2014). This orientation preference was tested by CRISPR-mediated inversions of different CTCF motifs, which disrupted the loops and created new enhancer-promoter interactions (Y. Guo et al., 2015). Moreover, it is possible to accurately predict the changes in CTCF loops after inversions or deletions of CTCF motifs (Sanborn et al., 2015). This preference on motifs orientation suggests that CTCF pairs are formed in a dimension-restricted space, via an extrusion process mediated by the cohesin complex (Fudenberg et al., 2016; Nichols & Corces, 2015; Sanborn et al., 2015). Loop formation can be explained by the loop extrusion model, which suggests that SMC cohesin subunits progressively extrude chromatin until halted by convergent-oriented CTCFs (Fudenberg et al., 2016)(Figure 5a). The capacity of CTCF to stall the extrusion may be caused by its ability to induce chromatin conformational changes, such as nucleosome repositioning (Clarkson et al., 2019; Fu et al., 2008). Atomic force microscopy experiments showed that DNA wraps around bound CTCF, forming structures of around 67-80 nm in diameter (Mawhinney et al., 2018), which are larger than proteins able to block cohesin sliding *in vitro* (Davidson et al., 2016; Stigler, Camdere, Koshland, & Greene, 2016)(Figure 5c).

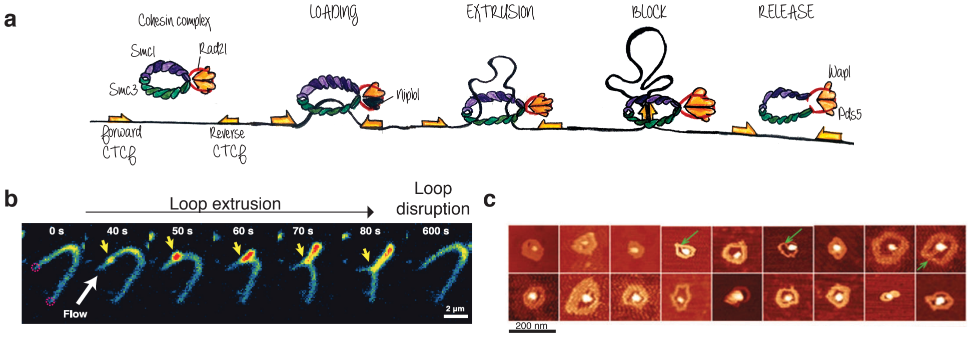


Figure 5. Loop extrusion illustration and supporting *in vitro* experiments. a, Schematic view of the loop extrusion model, from the loading of cohesin into chromatin until its release. Here we can observe CTCF pairs in a convergent oriented manner, as yellow arrows, which will halt the extrusion. Are also depicted Nipbl, the cohesin loader, and Wapl the release factor (Adapted from (Rowley & Corces, 2018)). **b,** Snapshots showing DNA loop extrusion by condensin on a SxO-stained double-tethered γ-DNA. The constant flow (white arrow) maintains the DNA in the imaging plane and extrudes the loop. The position of the loop base is pointed by a yellow arrow. At 40 s starts to appear a small loop that will grow over time until 80 s. The loop is disrupted after 600 s (Adapted from (Ganji et al., 2018)). **c,** Atomic force images of DNA in the presence of the 11 zinc fingers domains of CTCF. Green arrows indicate multiple DNA strands (Adapted from (Mawhinney et al., 2018)).

Several *in vitro* studies proved the diffusing capacity of cohesin along the DNA, until being blocked by CTCF (Davidson et al., 2016; Stigler et al., 2016). Importantly, it has been observed the condensin mediating loop extrusion *in vitro* (Ganji et al., 2018; Terakawa et al., 2017)(Figure 5b). Interestingly, a single cohesin ring can capture two DNA fragments, but only if one is single stranded (Murayama, Samora, Kurokawa, Iwasaki, & Uhlmann, 2018), suggesting that a single ring is able to lead this process. Polymer physics simulations, ChIP-exo and ChIP-nexus experiments have shed light on understanding how the cohesin ring is randomly loaded into chromatin, until blocked by CTCF when reaching the 3' end of the motif (i.e. within the loop) (Fudenberg et al., 2016; Nagy et al., 2016; Tang et al., 2015). Moreover, while cohesin depletion leads to CTCF loops loss (S. S. P. Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017), the depletion of WAPL, a cohesin release factor, results in the formation of longer loops due to an extended period of cohesin bound into chromatin, and a decreased long-range interactions between compartmental domains (Gassler et al., 2017; Haarhuis et al., 2017)(Figure 6e). The deletion of Nipbl, the cohesin loader, and Rad21 subunit, did not affected the CTCF binding, but resulted in a general loss of CTCF loops and a stronger compartmental segregation (S. S. P. Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017)(Figure 6c). Auxin-mediated degradation of Rad21 showed that 6 h were enough to lose CTCF loops, suggesting the need of being constantly extruded by cohesin (Fudenberg et al., 2016). Then, after 40 min of restoring cohesin, CTCF loops as large as 900 kb were established (S. S. P. Rao et al., 2017). For example, it has been proposed that cohesin can be relocated by transcription, directly pushed by RNAPII (Ocampo-Hafalla, Munoz, Samora, & Uhlmann, 2016) or indirectly by chromatin supercoiling (Racko, Benedetti, Dorier, & Stasiak, 2018).

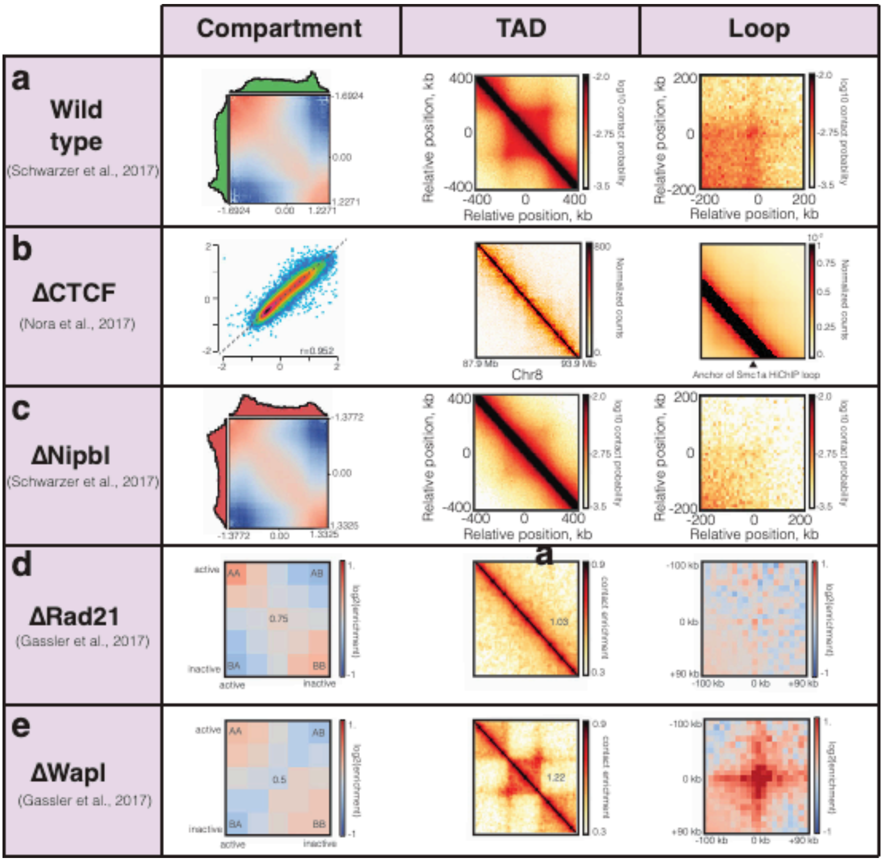


Figure 6. Summary of chromatin organization effects by cohesin and CTCF mutants. Each row represents a condition, and each column represent the chromatin organization layer being analysed. **a**, In the first row the results from a wild-type mouse are shown. The compartmentalization is represented by saddle plots in which the average interaction frequencies between pairs of loci (100 kb bins) arranged by their compartment signal (eigenvector). The histograms on the axis show the distributions of eigenvector values. TADs are shown as an average Hi-C map of all the TADs called in the dataset. The loop information is shown as the average Hi-C map around 102 peaks (Adapted from (Schwarzer et al., 2017)). **b**, Results of degrading CTCF in mESCs. The compartments are not affected, as can be observed by the correlation between the mutant and the wild-type, of the first eigenvector values. However, TADs disappeared, as shown in a 6 Mb segment of Hi-C data at 20 kb resolution. Loops were also affected, visualized by aggregating the Hi-C signal from Smc1

Hi-ChIP loops separated by 280-380 kb (Adapted from (E. P. Nora et al., 2017)). **c**, Results from a mouse mutant, with a deletion of Nipbl. Here the compartment information is retrieved like in the wild-type (panel a). Notice an enrichment of AA and a depletion of AB interactions. Can also be observed a TAD and loop lose (Adapted from (Schwarzer et al., 2017). **d**, Deletion of Rad21 results in mouse zygotes. The average loops, TADs and compartments are done by pooling together the maternal and paternal data. TADs and loops were entirely absent. However, the compartmentalization of active and inactive chromatin was increased (Adapted from (Gassler et al., 2017)). **e**, Study deleting WAPL in mouse zygotes. Here the same analysis as in Rad21 knock-out is applied (panel d). This mutant showed stronger signal of TADs and loops, while the compartments became weaker (Adapted from (Gassler et al., 2017)).

A study showed that transcription elongation resulted in a decrease in CTCF looping. Strikingly this decrease was weaker than upon ATP depletion, showing the emergence of hundreds of CTCF-independent loops (Vian et al., 2018). Moreover, loop extrusion via RNAPII fails to explain the formation of inactive domains and present a slower elongation rate compared to the estimated loop extrusion speed, ~ 9-90 bps/s versus ~ 374-850 bp/s (Jonkers & Lis, 2015; S. S. P. Rao et al., 2017). However, it can suggest that RNAPII will interfere cohesin extrusion over transcriptionally active regions. Overall, the proposed models are supported by some evidence, nevertheless each presents its limitations, suggesting that a combination between some may underlie the extrusion process.

Finally, CTCF can mediate cell-to-cell gene expression variability by regulating enhancer-promoter interactions (Ren et al., 2017). It has been shown that around the 41 % differential CTCF binding through 19 human cell types is due to methylation, as disruption of CTCF binding is associated with increased methylation at promoter sites (Wang et al., 2012). Genome-wide association studies (GWAS) have revealed various mutations in CTCF binding sites, which can affect TAD organization or enhancer-promoter interactions, leading to an increased variability of gene expression (Hnisz et al., 2016; Katainen et al., 2015). Thus, CTCF can contribute to cellular heterogeneity in mammals by mediating transcriptional pausing (Paredes, Melgar, & Sethupathy, 2013) and alternative mRNA splicing (Shukla et al., 2011). In the future, single-cell Hi-C will shed light into the dynamics of chromatin looping, allowing the comparison between pre-established versus *de novo* loops, and how are the loops maintained in non-expressing cells. These data together with epi-transcriptomics, will be needed to provide more evidences of the contribution of CTCF to cellular heterogeneity and its relevance in diverse biological scenario.

Chromatin loop detection methods

> "Impostor syndrome is the frequent feeling of not deserving one's success, and of being of a failure despite a sustained record of achievements. Highly successful people often experience it throughout their careers, especially when they are members of a group that is underrepresented in their profession—such as female scientists or engineers".
>
> Maria Klawe, 2014

The advent of genome-wide chromatin architecture (mainly thanks to Hi-C) has made necessary the development of loop detection methods. During the last two decades, a lot effort has been put to develop a loop detection software. There is a plethora of different software (Table 1). Among these, a selection has been benchmarked in detail in (Forcato et al., 2017). Here there is a brief description of their methodology:

1. *Hi-C Computational Unbiased Peak Search (HiCCUPS) (Durand et al., 2016; S. S. Rao et al., 2014)*

HiCCUPS is implemented as a part of Juicer suite of tools for the analysis and visualization of Hi-C experiments. This software by default needs as input the normalized contact matrix with Knight-Ruiz matrix balancing (Knight & Ruiz, 2013). Its algorithm looks for clusters of contact matrix entries in which the contact frequency is enriched compared to the local background. It scans each pixel and compares its number of contacts with four neighbouring areas, giving the possibility to cluster the significant peaks. It is implemented using GPUs, as it has to analyse trillions of pixels in kilobase-resolution experiments. It is programmed in JAVA language.

2. *Fit-Hi-C (Ay, Bailey, & Noble, 2014) and Fit-Hi-C 2 (Kaul, Bhattacharyya, & Ay, 2020)*

This software is designed to identify mid-range intra-chromosomal contacts, and inter-chromosomal with the new version. The algorithm relies in a model with two splines. The first spline models the observed counts according to the genomic distance between all the possible pairs. The second spline is used to fit to calculate a refined null model. The biases are computed using the Iterative Correction and Eigenvector decomposition normalization (ICE) (Imakaev et al., 2012) and are incorporated in the expected contact probability. Using a binomial distribution are calculated the p-values and corrected for multiple testing. Fit-Hi-C is programmed in Python language.

3. *GOTHiC (Mifsud et al., 2017)*

This software takes as input the aligned reads, it pairs them, and assigns the pairs to enzyme-specific restriction fragments and finally discards the ones separated by less than 10 kb. To normalize the counts it applies a similar Vanilla Coverage (Lieberman-Aiden et al., 2009), a binomial test to capture the significant interactions and a Benjamini-Hochberg multiple testing correction (Benjamini & Hochberg, 1995). The output contains both *cis* and *trans* interactions with the $\log_2$ of observed over expected counts, p-value, false discovery rate (FDR) and the number of read pairs that support the interaction. GOTHiC is programmed in R.

4. *HOMER (Heinz et al., 2010)*

HOMER is a command-line based software. As input it takes the aligned reads, which are paired and filtered. In order to identify the significant interaction bins, HOMER generates a background model that normalizes the genomic interactions for linear distance and coverage at a specific bin size. This model permits to estimate the expected read count and to apply a binomial test to get the significant *cis* and *trans* interactions. Finally, HOMER outputs a p-value, the FDR, the number of read pairs supporting the interaction, and the interaction distance. HOMER is programmed in Perl and R languages.

5. *High-throughput Identification Pipeline for Promoter Interacting Enhancer elements (HIPPIE) (Hwang et al., 2015)*

HIPPIE identifies the significant interactions at a restriction fragment-level. To be run it requires a computing cluster with Open Grid Scheduler or other Sun Grid Engine (SGE) compatible with job schedulers. HIPPIE consists in five steps: mapping, quality control, Hi-

C interactions identification, enhancer-target gene interactions and its prediction analysis. It classifies the read pairs as specific or non-specific if the sum of the distance of the reads from the closest restriction enzyme site is smaller or bigger than a given size selection parameter (Yaffe & Tanay, 2011). The biases are computed as in (Jin et al., 2013) which estimates the expected random contact frequencies considering mappability, GC content, fragment length and fragment distance (for intra-chromosomal read pairs). The significant interactions are identified by fitting a negative binomial distribution. HIPPIE retrieves the *cis* and *trans* restriction fragment-based interactions with the associated p-value. The software is programmed in Python, Perl and R languages.

6. *DiffHiC (Lun & Smyth, 2015)*

DiffHiC is intended to identify significant differential interactions between Hi-C experiments. However, it contains a method to call the interactions on individual samples based on the signal enrichment over a local background. The software consists on read mapping, filtering and genome partitioning to obtain the counts between the genomic bin pairs. It applies a "local enrichment" algorithm to identify the bin pairs with more reads than their neighbours. It computes the $\log_2$ fold change between the number of read pairs of the target-bin and the neighbour region with greatest abundance. DiffHiC does not applies statistical test, meaning that there is no significance value associated to the interaction. It is programmed in R and Python languages.

Software	Input data	Algorithm
HMRFBayesHiC (Xu, Zhang, Jin, et al., 2016)	O/E Hi-C matrices	HMRF model
FastHiC (Xu, Zhang, Wu, Li, & Hu, 2016)	O/E Hi-C matrices	HMRF based on simulated field approximation
Binless (Spill, Castillo, Vidal, & Marti-Renom, 2019)	Hi-C mapped reads (TSV)	Negative binomial likelihood
r3C-seq (Thongjuea, Stadhouders, Grosveld, Soler, & Lenhard, 2013)	3C mapped reads (BAM)	Background scaling method (Z-scores)
ChiaSig (Paulsen, Rodland, Holden, Holden, & Hovig, 2014)	ChIA-PET interaction file (BEDPE)	NCHG distribution
CHiCAGO (Cairns et al., 2016)	Capture Hi-C data mapped reads (BAM)	Convolution background model
Mustache (Ardakany, Gezer, Lonardi, & Ay, 2020)	Raw contact map (TSV)	2D Gaussian

Table *1.* Other available software for the identification of significant chromatin interactions. For each software the input file, and the used algorithm, are specified. All the software of the table run statistical test to retrieve the significant chromatin interactions. Bayesian Additive Regression Trees (BART); Hidden Markov random field model (HMRF); Non-central hypergeometric (NCHG).

Technology to assess nuclear organization

> "Although we don't know what is outside our universe, astronomers still wonder. Several pictures of what there might be have been dreamed up. An interesting one, called multiverse, has lots of universes. Picture it as a foam of bubbles. Our universe would be one bubble, and we'd be surrounded by lots of other bubbles."
>
> Jocelyn Bell, 2013

Understanding the three-dimensional spatial genome organization is paramount to fully characterize its function. Therefore, it is fundamental to use and develop microscopy technologies, both conventional and super-resolution, as well as chromosome conformation capture (3C) techniques. Each of these two types of technologies allow a given resolution and the inspection of specific chromatin topology levels (François Le Dily, Serra, & Marti-Renom, 2017).

Microscopy

As mentioned in the "Nuclear organization" section, many of the basic principles of genome architecture were discovered by microscopy techniques. For instance light and electron microscopy proved the existence of nuclear bodies such as the nucleolus, nuclear speckles, Cajal bodies and polycomb bodies (Denker & de Laat, 2016). Using this technology was possible to observe the nuclear localization of active euchromatin, which is less densely stained than heterochromatin (Heitz, 1928). Moreover, it has been shown that CTs occupy a specific position with a slight intermingling (Cremer, Cremer, Schneider, et al., 1982). It was also observed that some genes occupy specific nuclear positions according to their transcriptional status (Brown et al., 1997). Thanks to recent advances in super-resolution microscopy techniques, it is now possible to visualize from whole chromosomes to few kilobases interactions among nearby cis-regulatory elements. Thanks to advances in robotics and microfluidics, as well as accelerated super-resolution, it is also possible to inspect thousands to hundreds of thousands of individual cells, thereby increasing robustly the statistical power of imaging. The breakthrough of the Oligopaints technology has transformed the sequence-specific imaging of genome organization. Stochastic optical reconstruction microscopy (STORM) technology, captured images of active, inactive and Polycomb-repressed (PC-repressed) epigenetic states in *Drosophila.* This study revealed that each state obeys a different power-law scaling of domain size as function of length. Meaning that these three states have different packaging density, levels of self-interaction and levels of interaction between them. For instance, in *Drosophila* and mouse, PC domains are compacted globular structures, contrasting with active state structures, which are more extended. Due to the epigenetic role of PC in maintaining transcriptionally silent throughout cell divisions, it has been hypothesized that its characteristic densely packed structure and high degree of spatial exclusion of neighbouring active domains, contributes to this ability (Boettiger et al., 2016; Kundu et al., 2018). Recently, the optical reconstruction of chromatin architecture (ORCA) method was able to visualize very small TADs (< 20 kb) in several cells. This method allowed to study a strong cell-type specific physical partition of the posterior Hox complex in *Drosophila*, in which two transcriptionally active genes are separated. Surprisingly, the boundary of these two genes did not corresponded to a divergent epigenetic state, as both Ubx and abd-A gene regions were in an active epigenetic state, and a deletion of ~ 4 kb between these two domains disrupted their correct expression (Mateo et al., 2019)(Figure 7).

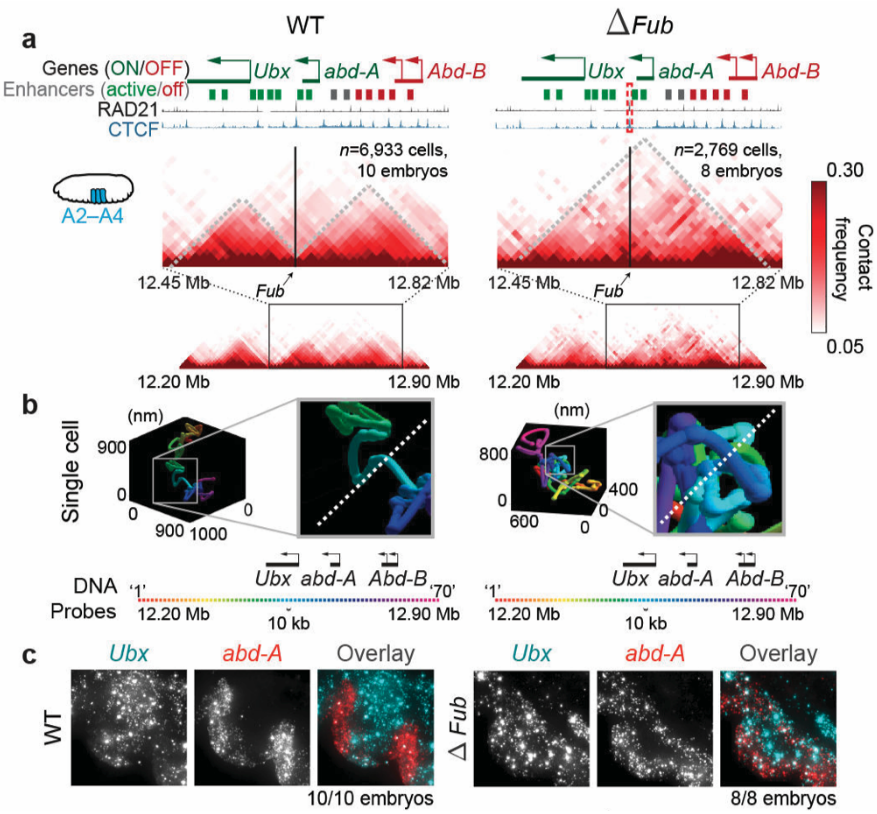


Figure 7. Deletion of TAD borders leading to TAD fusion and enhancer crosstalk. a, Contact frequency computed from ORCA for A2-A4 in wild-type (WT) and Fub mutants, with a deletion of a TAD boundary (mark with red dotted lines), at 10 kb resolution. Dashed grey lines mark the TADs**.** Below are marked the bithorax-complex (BX-C) genes, and the ChIP-seq of CTCF and Rad21 from WT embryos. **b,** ORCA snapshots from one experiment of a 700 kb domain containing the BX-C genes. Barcode position is marked by colour. The dashed line in the zoom-in images, indicate the 3D separation of up and downstream regions of the Fub locus. **c,** Comparison of the expression of two genes: abd-A and Ubx in WT and Fub mutant embryos in individual regions (Adapted from (Mateo et al., 2019))**.**

This proved that the epigenetic state is not the unique mechanism determining the chromatin organization, and showed the contribution of the boundary elements *in vivo*. Since the discovery of TADs thanks to 3C-based methods, which are mainly population-based run in millions of cells, their existence in single-cells has been very controversial. Microscopy techniques, showed that while CTs correspond to a physical structure, TADs only exist as statistical features in population of cells. However, only around 100 cells are needed to provide a proper delineation of TADs, sub-TADs and loops. The globular structures or TAD-like features are present in single cells, and persist even in the absence of cohesin, whereas TADs in identical compartment types (sharing an epigenetic state) disappear (Gassler et al., 2017; Haarhuis et al., 2017; S. S. P. Rao et al., 2017). Microscopy methods explained the emerging of this pattern as an uniform statistical position of the boundaries instead of biased to CTCF binding sites (Bintu et al., 2018)(Figure 8). This new data has promoted multiple models to explain the competing mechanistic of chromatin folding (Barbieri et al., 2012; Brackley et al., 2018; Fudenberg et al., 2016; Giorgetti et al., 2014; Jost, Carrivain, Cavalli, & Vaillant, 2014; Sanborn et al., 2015).

Finally, recent advances in live imaging techniques validated the dynamic and transient nature of chromatin interactions (Bintu et al., 2018; Finn et al., 2019; Mateo et al., 2019). It allowed to study the transcriptional dynamics, memory, and gene co-regulation (Alexander, Guan, Huang, Lomvardas, & Weiner, 2018; Chen et al., 2018; Ferraro et al., 2016). For the future, it will be highly important to combine the flexibility and ease use of microscopy methods, to enhance the resolution, image DNA and protein at the same time, and increase its throughput and robustness.

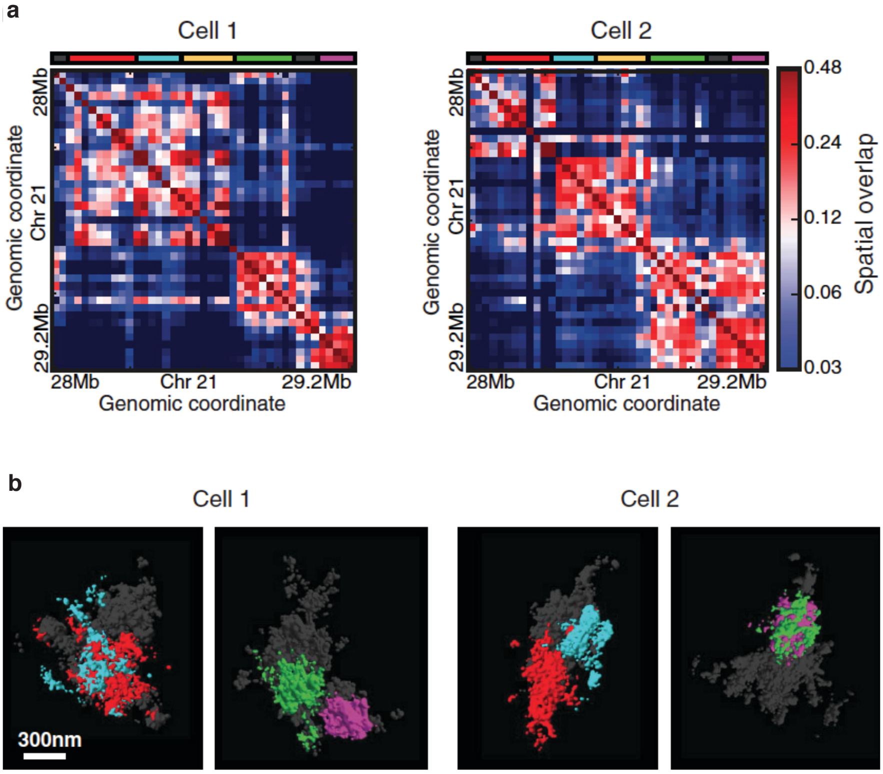


Figure 8. Visualization of TAD-like globular domains in single cells. a, Matrices from the spatial-overlap from a 1.2 Mb region (chr21: 28-29.2 Mb) from two individual IMR90 cells. The colour barcode below the contact matrix indicate the five sub-TADs identified at population level. **b,** Images from multiplexed 3D STORM corresponding to the two cells shown before. Per cell are shown two images, in each highlighting two sub-TADs and with a specific rotation to ease their visualization (Adapted from (Bintu et al., 2018)).

Chromosome conformation capture technology

In the last decade, the development of 3C-based technology and thereby their modelling and analysis, has revolutionized the understanding of 3D chromatin architecture. First, 3C techniques were applied to yeast chromosomes (Dekker, Rippe, Dekker, & Kleckner, 2002), then, as explained before, to observe the long distance chromatin loops between enhancer and the target genes in the β-globin locus (Tolhuis, Palstra, Splinter, Grosveld, & de Laat, 2002).

Technology	Type of approach	Advantage	Limitation
3C (Dekker et al., 2002)	One-to-one	Cheap and simple	Amplification efficiency
4C (Simonis et al., 2006)	One-to-all	No need a priori knowledge	Amplification efficiency
5C (Dostie et al., 2006)	Many-to-many	Unbiased readout and great coverage	Need a priori knowledge
ChIA-PET (Fullwood et al., 2009)	Many-to-many	Enrichment in rare interactions	Difficult quantification
Capture-C (Hughes et al., 2014)	Many-to-all	Detection of SNPs	Sequence enrichment efficiency
Micro-C (T. H. Hsieh et al., 2015)	All-to-all	High resolution	Increase noise in long distance
HiChIP (Mumbach et al., 2016)	Many-to-all	Low input requirement	RE site proximity / Antibody efficiency

Table 2. Existing 3C-derived methods. In the table are listed the main advantages and drawbacks of the most used technologies, as well as the information retrieved.

Finally, this technique was adapted to its high-throughput variants (Table 2). All these techniques are based on formaldehyde crosslinking of chromatin, which allows the capture of a snapshot of the interactions among any pair of genomic loci in the three-dimensional nuclear space. Then chromatin is fragmented by digestion, and re-ligated in order to convert the interacting loci into unique DNA ligation products to be finally detected by different methods. Originally, PCR with locus-specific primers was used to detect ligation products one at a time. The development of deep-sequencing techniques allowed the high-throughput detection of ligation products and permitted the interrogation of the chromatin architecture genome-wide, the most popular implementation of this method is called Hi-C (Lieberman-Aiden et al., 2009).

High-throughput conformation capture (Hi-C)

This technique allows the interrogation of all loci at once. The experiment requires around 5 million of cells as an input material in order to have enough library complexity for sequencing. Although this limitation has been addressed in recent versions of the Hi-C protocol, it is still normally conducted on a population of cells and the resulting 3D genome organization represents the ensemble of single cells of the population. The development of the Hi-C assay represents a breakthrough in the understanding of genome organization as it is unbiased and unsupervised. As in the others 3C-based methods, the chromatin is crosslinked with formaldehyde and digested by a restriction enzyme (RE), which leaves a 5' overhang to be filled by free nucleotides some of which biotinylated. Then the repaired blunt ends are ligated and the DNA is sheared and purified using a biotin pull-down with streptadivin beads. The final product of the experiment before sequencing is a library of ligation junctions that will be sequenced using a paired-end approach. On the bioinformatics side, both reads will be mapped to a reference genome, obtaining the *cis* and *trans* contacts (respectively intra- and inter-chromosomal contacts) to generate the contact matrices. It has to be considered that the resolution of the experiment is highly dependent on the RE sites frequency. One solution to increase x-fold the resolution is to sequence x^2 more pairs, keeping in mind that the resolution could never go deeper that the RE fragment sizes. The first version of Hi-C corroborated the existence of the A and B chromatin compartments. Since the first Hi-C assay it has been exponentially improved retrieving kilobase contact matrices allowing the detection of sub-compartmentalization of the chromatin (Imakaev et al., 2012; S. S. Rao et al., 2014). The bioinformatics analysis of Hi-C experiments consists mainly on 5 steps:

I. Hi-C quality check and mapping to a reference genome.

II. Read filtering: removal of PCR artifacts as well as products from non-standard ligation. Only the valid pairs will be kept.

III. Building of contact matrices: the genome is chunked into non-overlapping bins of fixed size.

IV. Bin filtering: removal of matrix columns with low counts. If the observed number of reads in a bin is much lower than average, it is expected that they belong only from mapping artifacts.

V. Matrix normalization: remove the inherent biases of the experiment. There are two types of normalization approaches: explicit methods, that suppose that all the biases are known, like GC content, mappability or number of RE sites per bin; such as OneD normalization (Vidal et al., 2018). And implicit or balancing methods, which assume an equal experimental visibility of each bin, including ICE (Imakaev et al., 2012), square root vanilla coverage, vanilla coverage normalization (S. S. Rao et al., 2014) and Knight-Ruiz Matrix Balancing (KR)(Knight & Ruiz, 2013).

However, Hi-C protocols presented caveats, including an overrepresentation of inter-chromosomal contacts (~ 60 %) due to random ligations between unrelated DNA fragments (i.e., uncross-linked). This high percentage will not affect the intra-chromosomal contacts measured from short and medium-range distance (< 2 Mb), but the long-distance (> 10 Mb) between and within chromosomes. The inter-chromosomal contacts were reduced to < 20 % by placing the ligation *in situ* inside the nuclei instead of *in solution* (Nagano et al., 2015). Another critical step and potential source of biases is the use of formaldehyde to fix DNA fragments. Formaldehyde links proteins or protein and DNA, however the fixation could differ depending on the proteins and potentially chromatin state, being a drawback to capture protein-mediated loops. Moreover, highly dynamics and fluctuating interactions might not be captured as formaldehyde takes at least 5 seconds to crosslink. Another important issue is to consider the meaning of the retrieved quantitative contact by 3C experiments, as these methods assess the ligation frequencies between cross-linked and fragmented DNA segments. For a DNA segments to participate in the 3C contact profiles it needs to i) be cross-linkable, ii) have DNA ends available to ligate, and iii) in the cross-linked DNA-protein aggregates compete with other fragments to ligate. These three requisites depends on size, chromatin composition, duration and strictness of fixation (Dekker, 2006). The ligation efficiency can work as a proxy for contact frequencies, which must be validated by microscopy or genetics (i.e., DNA deletions).

Despite its limitations, Hi-C is the chosen method to obtain the genome's interactome. During the past 10 years, in the majority of the labs the Hi-C technology became ordinary, thus there are Hi-C contact maps available for the most used cell types from human and mouse, human tissues and organoids. The 3D genome organization together with epigenomics and transcriptomics data would help to interpret the cell homeostasis and the genetic variations associated to disease. All the recent advances in single-cell Hi-C, will help to reveal the heterogeneity of specific chromatin contacts within cell populations and cell-to-cell during multiple biological processes (i.e., development).

Function-Structure dogma

> "I don't want to say epigenetics isn't exciting … [but] there's a gap between the fact and the fantasy. Now the facts are having to catch up."
>
> Edith Heard, 2013

The relationship between structure and function is a paramount in structural biology. First, was reported that changes in structure were directly related to sequence modifications in proteins (Chothia & Lesk, 1986). In the case of 3D genome organization, as described in previous sections (see Nuclear Organization), TADs may represent regulatory domains due to their high conservation in mammals and the co-regulation of their contained genes (Dixon et al., 2015; F. Le Dily et al., 2014). Moreover, the functional relevance of TADs has been validated by studies in diseases caused by structural variations (SVs), which the disruption of TADs caused congenital limb malformations, gene misexpression and cancer (Dixon et al., 2018; Franke et al., 2016; Lupianez et al., 2015; Spielmann, Lupianez, & Mundlos, 2018). However, neither CTCF nor cohesin depletion resulted in large repercussions on gene expression (Elphege P Nora et al., 2017; S. S. P. Rao et al., 2017; Wutz et al., 2017). These studies challenge the crucial implication of 3D genome organization for gene regulation, suggesting a more complex and multi-layer effect of chromatin architecture on gene regulation. Moreover, high resolution experiments showed that not all the regions of the genome follow the same rules. For example, gene dense regions do not present a canonical 3D structure, as they can form multiple smaller interacting domains, suggesting an alternative regulation as large TADs and loops. Recently, thanks to higher resolution and more detailed analysis, 3D architectural stripes have been described, which are asymmetrical patterns of contacts that can span several 100 kb and are reflecting an unidirectional loop extrusion process. Architectural stripes are associated with strong and active enhancers, which are scanning the genome for a target gene, in close proximity to a CTCF boundary, defined as stripe anchor. (Barrington et al., 2019; Kraft et al., 2019; Vian et al., 2018). Similarly, the development of the micro-C method, which provides 3D chromatin structure at nucleosomal resolution, it revealed the existence of micro-TADs, with a median size from 5.4 to 40 kb, which are reflecting individual transcriptional units (T. S. Hsieh et al., 2020). It makes essential to understand the relationship between 3D chromatin structure and gene regulation to decipher the possible regulatory scenarios.

Software for the identification of differential chromatin interactions

Humans are allergic to change. They love to say, "We've always done it this way". I try to fight that. That's why I have a clock on my wall that runs counter-clockwise.

Grace Murray Hopper, 1987

Nowadays there are a large number of 3C experiments coming from different cell types, tissues and/or conditions or diseases. This has made necessary the development of software tools to compare them to shed light on the biological implications of genome structural changes.

The vast majority of available software to identify differential interactions are bin or pixel based, limiting the identification of chromatin loops (Table 2). Moreover, some of the tools presented in chromatin loop detection methods section present a version to assess differential interactions between experiments. Here are summarized the most used:

1. *HOMER (Heinz et al., 2010)*

First, HOMER identifies the interactions in the first experiment as described in the "Chromatin loop detection methods" section. Then, quantifies the number of reads per interaction in the second experiment. In order to retrieve significant interactions independent statistics are computed for the second experiment on its background model and compared to the first experiment. The output contains the Z-score, logP and number of reads supporting the interaction between the experiment and the background.

2. *DiffHiC (Lun & Smyth, 2015)*

As previously explained, the main purpose of DiffHiC is to capture the differential significant interactions between experiments. In this case applies a quasi-likelihood F-test (QLF-test) that yields a p-value per bin. It also corrects for multiple testing to control the FDR with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). These two values can be used to threshold the differential interactions.

	Datatype	Normalization	Algorithm
HOMER	Hi-C	✓	Background model
DiffHiC	Hi-C, Capture Hi-C*	✓	QLF-test
HiBrowse (Paulsen, Sandve, et al., 2014)	Hi-C, ChIA-PET	✓	Monte Carlo
HiCdat (Schmid, Grob, & Grossniklaus, 2015)	Hi-C, ChIP-Seq, RNA-seq, BS-seq, genome annotation	✓	Signed difference matrices
FIND (Djekidel, Chen, & Zhang, 2018)	Hi-C	✓	Spatial Poisson
Selfish (Ardakany, Ay, & Lonardi, 2019)	Hi-C		Self-similarity metric
Chicdiff (Cairns, Ochard, Malysheva, & Spivakov, 2019)	Promoter Capture Hi-C	✓	IHW
HiCcompare (Stansfield, Cresswell, Vladimirov, & Dozmorov, 2018)	Hi-C	✓	Z-score from loess
MultiHiCcompare (Stansfield, Cresswell, & Dozmorov, 2019)	Hi-C	✓	GLM
ACCOST (Cook, Hristov, Le Roch, Vert, & Noble, 2020)	Hi-C		Extended model used by DEseq**

Table 3. Comparison between features of some currently available software for differential chromatin contact data analysis. In this table are included the methods explained before as well as other available methods. All the software result in statistical information of the structural changes. * might be implemented in future versions. ** (Anders & Huber, 2010). Independent Hypothesis Weighting (IHW); locally weighted linear regression (loess); general linear model (GLM).

Chapter I

Identification of chromatin loops from Hi-C interaction matrices by CTCF-CTCF topology classification

Galan S, Serra F, Marti-Renom MA.

Identification of chromatin loops from Hi-C interactions matrices by CTCF-CTCF topology classification.

doi: https://doi.org/10.1101/2020.07.21.214585

Identification of chromatin loops from Hi-C interaction matrices by CTCF-CTCF topology classification

Silvia Galan[1,3], François Serra[#,*] and Marc A. Marti-Renom[1,2,3,4,*]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri 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.

#Current address: Computational Biology Group - Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain

*To whom correspondence should be addressed: M.A.M-R. martirenom@cnag.crg.eu, F.S. francois.serra@bsc.es

ABSTRACT

Genome-wide profiling of long-range interactions has revealed that the CCCTC-Binding factor (CTCF) often anchors chromatin loops and is enriched at boundaries of the so-called Topologically Associating Domains or TADs, which suggests that CTCF is essential in the 3D organization of chromatin. However, the systematic topological classification of pairwise CTCF-CTCF interactions has not been yet explored.

Here, we developed a computational pipeline able to classify all CTCF-CTCF pairs according to their chromatin interactions from Hi-C experiments. The interaction profiles of all CTCF-CTCF pairs were further structurally clustered using Self-Organizing Feature Maps (SOFM) and their functionality characterized by their epigenetic states. The resulting cluster were then input to a convolutional neural network aiming at the *de novo* detecting chromatin loops from Hi-C interaction matrices.

Our new method, called LOOPbit, is able to automatically detect higher number of pairwise interactions with functional significance compared to other loop callers highlighting the link between chromatin structure and function.

INTRODUCTION

The human genome consists of 2 meters of DNA and its folding in the cell nucleus is not random (Bonev and Cavalli, 2016). During the last decade, the three-dimensional (3D) organization of the genome have been associated to the regulation of multiple nuclear functions like DNA replication, repair, rearrangement and recombination, RNA processing, and transcription (Bonev and Cavalli, 2016; Dekker and Mirny, 2016; Sexton and Cavalli, 2015; Stadhouders et al., 2019). Indeed, the 3D genome is organized in a hierarchical fashion, with each layer regulating different functions. For instance, chromosomes are found in preferential areas of the nucleus, called chromosome territories (Cremer and Cremer, 2001). The molecular evidence of this observation was first brought by high-throughput Chromosome Conformation Capture (3C) experiments (Hi-C) (Lieberman-Aiden et al., 2009), which also revealed a multi-Megabase (Mb) scale structure, called chromatin compartments. In mammal genomes, there are mainly two compartments, A and B, which correlate with active, open and gene-rich and inactive, closed and gene-poor chromatin, respectively (Lieberman-Aiden et al., 2009). At a sub-Megabase level later 3C-based experiments unveiled that the genome can be further partitioned into self-interacting regions with sizes between 40 kb and 3 Mb, called Topologically Associating Domains (TADs) (Dixon et al., 2012; Nora et al., 2012). TADs modulate interactions between regulatory elements, such as promoters and enhancers, playing a role in transcriptional regulation. At their borders, TADs often harbor regions enriched in CCCTC-binding factor (CTCF) protein (Rao et al., 2014). CTCF is a transcription factor formed by 11 DNA-binding Zinc finger (ZF) that plays a major role in genome architecture (Nichols and Corces, 2015). Its functionality is dependent on the location and the relative orientation of its binding sites. Interacting CTCF pairs tend to be organized in a convergent orientation, with the binding motifs facing each other (Rao et al., 2014). The distribution and organization of CTCF leading to TAD formation has been explained by the loop extrusion model that involves CTCF but also other protein complexes like cohesin (Fudenberg et al., 2016; Sanborn et al., 2015). Briefly, the cohesin complex, which forms a ring-shaped structure upon loading onto chromatin, extrudes chromatin resulting in a growing loop of DNA. The extrusion is blocked when in both sides cohesin encounters two CTCF in a convergent orientation (Fudenberg et al., 2016; Sanborn et al., 2015). Various experiments have described the relevance of the binding motif orientation for loop formation by inverting or deleting CTCF binding sites using CRISPR/Cas9 experiments (Sanborn et al., 2015; Wutz et al., 2017). Moreover, a number of studies revealed the effect of CTCF depletion, resulting in a fainting of TADs and loop structures, while maintaining compartment organization. It suggests that CTCF and cohesin play a role in TAD and loop formation, but not necessarily in compartment maintenance (Guo et al., 2015; Hyle et al., 2019; Nora et al., 2017).

Recently, several studies have applied Aggregate Peak Analysis (APA) (Rao et al., 2014), or pile-up methods (Lekschas et al., 2018) to analyze Hi-C datasets using the mean signal from selected regions of interest (Bonev et al., 2017; de Wit et al., 2013; McLaughlin et al., 2019; Pekowska et al., 2018; Ruiz et al., 2018; Schwarzer et al., 2017), such as CTCF loops. At the level of single chromatin loops, their biological relevance in gene regulation, has led to the development of several loop callers. However, most of the methods present low reproducibility between biological replicates and a low correlation with different biological markers (Forcato et al., 2017). Here, we propose to use Self-Organizing Feature Maps (SOFM) (Kohonen, 1987), an artificial neural network, to classify the signal from pairs of CTCF without any prior information about their topological organization. This approach allows us to obtain sub-populations of CTCF-CTCF interacting structures and to identify their epigenetic signature in an unsupervised manner. As a result, we next trained a convolutional neural network (CNN), and used the generated model in our tool, called LOOPbit, to identify chromatin loops in a robust, fast and genome-wide manner.

RESULTS

Classifying CTCF-CTCF interactions using Self-Organizing Feature Maps (SOFMs) and Uniform Manifold Approximation and Projection (UMAP).

SOFMs are dependent on several parameters, such as the grid size (number of neurons or clusters), the learning radius, the step size, the standard deviation and the number of iterations or epochs (Kohonen, 1987). To unveil the best combination of SOFM parameters in the context of CTCF-CTCF submatrices classification, we first extracted the Hi-C submatrices of all possible *cis* pairs of CTCF peaks linearly separated between 45 kb and 1.5 Mb in the genome (**Fig. 1a**). These submatrices, here referred to as "CTCF-CTCF submatrices", spanned over 45 kb (20 kb of each side of the 5 kb bin with a given CTCF peak). Next, all extracted CTCF submatrices were input to a SOFM that resulted in a total of 1,152 SOFMs each with different combinations of parameters (**Methods**). To select the optimal set of parameters, we assessed three measures: (i) the percentage of classified submatrices where singletons (SOFM neurons with only one submatrix) were considered as non-classified; (ii) variability between neurons; and (iii) average compartment-type segregation across neurons. These three quality measures aimed at identifying the "best" classification that maximized the number of CTCF-CTCF submatrices classified, the separation between neurons (i.e., increased variability inter-neuron), and the homogenous genome compartment type within neurons. Of all varied parameters, the SOFM grid and the step sizes were the most sensitive to the final classification (**Fig. 1b** and **Supplementary Fig. 1**). The three measures mentioned above were optimal for grid size of 30, learning radius of 5, standard deviation of 0.5, step of 0.01 and, 30 epochs. Such optimal SOFM map can be represented as a grid map where each cell (neuron) is composed by a set of similar CTCF-CTCF submatrices represented by its medoid submatrix (**Fig. 1c**). The SOFM map clearly reflects a variety of CTCF-CTCF submatrices from loop forming pairs (lower-left corner) to non-interacting pairs (upper-right corner). Next, the resulting 900 SOFM neurons were further classified into CTCF-CTCF clusters by computing the Euclidean distance between the medoid submatrices and projecting it into a two-dimensional UMAP (McInnes, Healy, & Melville, 2018), a non-linear manifold dimension reduction followed by a density-based clustering algorithm (HDBSCAN) (Campello et al., 2013). The two-dimensional UMAP resulted in 10 clusters of unique CTCF-CTCF pairing patterns, from the most structured canonical cross pattern (CTCF-CTCF cluster 1 including 11 SOFM neurons with a total of 2,685 CTCF-CTCF submatrices) to a completely flat pattern with no interaction (CTCF-

CTCF cluster 10 with 7 SOFM neurons and 1,565 CTCF-CTCF submatrices) (**Fig. 1d**). Our results suggest that CTCF-CTCF pairs can adopt a variety of well-defined topological signatures observed in a Hi-C matrix. Next, we set ourselves to functionally characterize each of the detected CTCF-CTCF clusters.

Functional characterization of the CTCF-CTCF clusters.

As described before, segregation of A/B compartment types between neurons was used as a metric to select the optimal set of parameters for the SOFM classifier. However, such measure was agnostic to the medoid submatrix topology representing each neuron. Interestingly, clusters with clear loop topology (that is, CTCF-CTCF clusters 1 to 5) are enriched in A compartment at the anchor points of the loops. Conversely, clusters with non-interacting CTCF-CTCF (that is, clusters 6 to 10) are enriched in B compartment (**Fig. 2a**). This separation between clusters 1 to 5 and 6 to 10 is also observed when revealing enrichments in CTCF pairs with different directionalities**Error! Bookmark not defined.** (Rao et al., 2014), CTCF-CTCF clusters with clear interaction patterns are enriched in convergent-oriented CTCF-CTCF (clusters 1 to 4, **Fig. 2b**), while non-interacting patterns are enriched in divergent-oriented CTCF-CTCF (clusters 7 to 10, **Fig. 2c)**. Parallel-oriented CTCF-CTCF are enriched in mid-interacting signal clusters (clusters 5 and 6, **Fig. 2d**). CTCF-CTCF clusters with loop topology (that is, clusters 1 to 3) have a mean genomic distance between CTCF peaks of between ~634 kb and ~680 kb (**Fig. 2e**). CTCF-CTCF clusters 7 to 10, which correspond to enrichment of B compartment and divergent orientation, present shorter mean genomic distances spanning from ~510 kb to ~629 kb (**Fig. 2e**). To assess whether chromatin state correlate with the CTCF-CTCF clusters, we next assessed the chromatin state enrichment in the anchors of CTCF sites for each cluster (**Methods**). Interestingly, pairs of CTCFs that form loop-like structures are enriched with enhancer-promoter interactions (with one anchor points labelled as "enhancer" chromatin state and the other as "promoter", **Methods**) (**Fig. 2f**). Anchors falling in "heterochromatin" state have an almost opposite distribution among clusters. CTCF-CTCF clusters forming loops (clusters 1 to 5) are depleted of heterochromatin types while non-loop clusters (cluster 6 to 9) are enriched in heterochromatic marks (**Fig. 2g**). Interestingly, CTCF-CTCF cluster number 10, which corresponds to the most B compartment cluster, has no heterochromatic anchors but is enriched in polycomb-promoter pairs (**Fig. 2h**). CTCF-CTCF pairs with polycomb in both anchors are enriched in mid-interacting clusters (that is, clusters 4 to 6) and depleted in non-interacting pairs (that is, clusters 7 to 10) (**Fig. 2i**). Finally, enhancer-enhancer pairs, as expected, are more present in interacting clusters (clusters 1 to 4) and less abundant in non-interacting clusters (clusters 7 to 10) (**Fig. 2j**). Our analysis indicates the existence of clusters or types of CTCF-CTCF pairs that gradually expand from enhancer-promoter, convergent, mid-range interacting pairs in A compartment, to a polycomb-heterochromatin, divergent, short-range, non-interacting pairs in B compartment. Taken together, our results based solely in the interaction pattern of CTCF-CTCF proteins reveal two major types of CTCF-CTCF pairing that can further characterize well defined functional states.

Loop calling using LOOPbit, a CNN trained with looping and non-looping CTCF-CTCF pairs.

Next, we randomly subset a total of 6,000 CTCF-CTCF submatrices from clusters 1 to 5 as loop forming pairs and another 6,000 CTCF-CTCF submatrices from clusters 7 to 10 as non-loop forming pairs. The two datasets, loop and non-loop, where next used to train a convolutional neural network (CNN), that we called LOOPbit, that aims at automatically assign a probability of a CTCF-CTCF pair forming a loop in the central bin of a 9x9 submatrix extracted from genome-wide Hi-C experiments (**Fig. 3a** and **Methods**). LOOPbit, which was trained by a 20% leave-out of the used data, was then assess for accuracy and compared to other loop-calling methods using a recently published benchmark (Forcato et al., 2017). LOOPbit, similarly to other published methods including HICCUPS (Rao et al., 2014), GOTHiC (Mifsud et al., 2017), HOMER (Heinz et al., 2010), diffHiC (Lun and Smyth, 2015), HIPPIE (Hwang et al., 2015), and Fit-Hi-C (Ay et al., 2014), detects more interacting loops with an increase of the number of valid interactions in the Hi-C experiment (**Fig. 3b**), which indicates the dependency of such loop-detection methods on the depth of the Hi-C experiment. A solution to minimize this effect is to reduce data resolution (*i.e.*, 5 kb to 40 kb) (**Supplementary Fig. 2a**). Loops detected by LOOPbit were of similar average size as of HICCUPS, diffHiC and HIPPIE (~200 kb) and larger than those by GOTHiC (~80 kb) and HOMER (~100 kb) but shorter than those by Fit-Hi-C (~10 Mb) (**Fig. 3c**). This loop size measure is again dependent on the resolution of the data as loops called at 40 kb resolution increased to ~1 Mb of size for all methods (**Supplementary Fig. 2b**). Next, we assessed the ability of LOOPbit to reproduce loop detection by measuring the Jaccard Index

(JI) using replicates of previously published Hi-C experiments (**Methods** and **Supplementary Table 2**). Despite that LOOPbit yields continuous probability clouds instead of pinpointing single cells in the matrix, we used the exact same benchmark measures previously published (Forcato et al., 2017) to calculate JI between replicates (that is to consider two loops identical if both their anchoring bins are the same). With such benchmark, LOOPbit results are slightly more reproducible in average than compared to HOMER and Fit-Hi-C, similar to GOTHiC, diffHiC and HIPPIE and lower reproducibility than HICCUPS (**Fig. 3d**). Next, we compared the accuracy of the loop callers in terms of the biological relevance of the predicted loops. LOOPbit was able to detect higher percentage of enhancer-promoter loops than any other caller (**Fig. 3e**, top panel). Concomitantly, LOOPbit also detects less loops between heterochromatic anchors (**Fig. 3e**, middle panel) and has similar levels of non-expected loops (that is, between enhancer and heterochromatin, **Fig. 3e**, bottom panel). Interestingly, and likely particular to LOOPbit as it was trained using CTCF-CTCF selected loops, ~56% of all detected loops in the training had an annotated CTCF site in both anchor points (**Methods**). Altogether, our results indicate that LOOPbit has similar tolerance to sparse input Hi-C data as other loop callers, a characteristic related to the intrinsic difficulty of detecting loops (Forcato et al., 2017). Nonetheless, chromatin loops detected by LOOPbit show an enrichment in functional signatures when compared to the other methods. Therefore, we next focused on assessing LOOPbit applicability in a real case biological scenario.

Usability of LOOPbit to capture chromatin loops in a CTCF-targeted degradation system.

Next, we assessed whether LOOPbit was able to detect differential loops between three previously published Hi-C datasets where CTCF was targeted for degradation (Nora et al., 2017). The three datasets included non-treated cells, auxin treated cells with degraded CTCF, and auxin wash-off cells where CTCF recovered non-treated levels. These three experiments, which had around 450 million valid Hi-C pairs each, allowed to test whether LOOPbit was sensitive to the depletion of CTCF for detecting loops. First, we assessed the JI value between the three experiments (**Fig. 4a**). As expected, the JI between non-treated and wash-off experiments is higher (~0.1 JI and **Fig. 4a**). Importantly, LOOPbit results confirmed that the auxin-treated loops detected by LOOPbit were different than those for non-treated or wash-off replicas (0.0 JI and **Fig. 4a**). Next, we assessed whether the detected loops have enrichment of CTCF ChIP-Seq signal around the anchor points. Indeed, the majority of the loops detected by LOOPbit are enriched in CTCF at anchoring bins. This enrichment drops for the auxin-treated cells and is recovered for the wash-off cells (**Fig. 4b**). Then we identified that the distance between loop anchors in auxin-treated cells was significantly higher than in non-treated and wash-off datasets (**Fig. 4c**). Finally, we show an example from the region chr3:47,600,000-49,000,000 where LOOPbit detects two loops for non-treated and wash-off, that vanish for auxin-treated cells (**Fig. 4d**). In summary, LOOPbit detects chromatin loops from Hi-C interaction matrices that are enriched for functional interactions and, as expected, for those mediated by the CTCF transcription factor.

DISCUSSION AND CONCLUSION

In this work we introduce the use of the structure signal deconvolution in the context of colocalizing DNA-binding proteins. This methodology aims at identifying clusters of different structural patterns. Until now, most methods detecting structural patterns associated to DNA-binding proteins use aggregate peak analysis (APA) to show an average interaction pattern for different targets of interest. Unfortunately, APA is blind to small subsets with specific interaction patterns, different from the average structure that could potentially be related to specific functions.

Here, we deconvolved the genomic average CTCF-CTCF interaction pattern, and, based solely on structural features, we were able to classify each CTCF-CTCF interaction into distinct subpopulations within the range of distances between CTCF pairs of 45 kb to 1.5 Mb. According to their structural patterns, ten CTCF clusters were obtained (**Fig 1d**). Each CTCF-CTCF cluster has a specific pattern in terms of genome compartment location and epigenetic state. The first observation is that the genomic distance between the CTCF pairs as well as their orientation are relevant features to classify CTCF-CTCF interactions between those that structurally form and do not form loops (Bonev and Cavalli, 2016; Rao et al., 2014). Also expected, but less evident is that loop forming clusters (that is, those from cluster 1 to cluster 5) are enriched in A compartment and drive principally enhancer-enhancer and enhancer-promoter interactions. Then CTCF-CTCF clusters with sparse or blur signal of interaction are often appearing in heterochromatin and polycomb chromatin states spanning shorter genomic distance, which may indicate their participation in silent chromatin (Ogiyama et al., 2018) or simply highlight the overlapping of the signal of polycomb-polycomb driven interactions with CTCF-CTCF driven interactions (Narendra et al., 2015; Van Bortle and Corces, 2012). Interestingly, we could capture the polycomb interacting network, which has been observed to be essential for cell differentiation and identity. Thus, cluster number 10, which is mostly associated to compartment B, no loop structure, few interactions, divergent CTCF binding sites and short genomic distances between anchor points, contained a surprising enrichment of promoter-polycomb states which could be explained by polycomb driving interactions here instead of CTCF. While clusters 4 to 6, presented a loop pattern and were mostly associated to A compartment, were enriched in polycomb at both loop anchors, suggesting the formation of chromatin loops for a proper cell identity regulation. Together these findings define subcategories of CTCF-CTCF interactions within which few (CTCF clusters 1 to 5 representing ~30% of the CTCF-CTCF interactions studied) are consistent with the most accepted model of CTCF loops bringing together promoters and enhancers to initiate transcription (Guo et al., 2015).

The deconvolution of the average signal between CTCF-CTCF pairs allowed us thus to identify two major types of interactions, one forming a canonical loop (clusters 1 to 5) and a second not forming the canonical loop (clusters 7 to 10). This allowed us to generate a bona fide set of CTCF-CTCF submatrices that generate loops versus those that do not. Next, we used those two sets to train a CNN to develop the loop-caller LOOPbit, which was technically and biologically tested against several previously benchmarked methods (Forcato et al., 2017). LOOPbit, which was applied to 33 Hi-C experiments at 5 kb resolution and 5 experiments at 40 kb resolution, resulted in a number and length of loops similar to all other benchmarked methods and suffered from the same limitations. Indeed, loop-callers cannot easily replicate findings when comparing replicates of low sequencing depths or binned at very high resolutions. Fortunately, when high sequencing depth data is available, LOOPbit increases its reproducibility.

Importantly, loops called by LOOPbit were found to be particularly relevant in terms of biological function. We found a clear enrichment of promoter-enhancer loops and depletion of loops between anchor points in heterochromatin state (Guo et al., 2015; Rao et al., 2014). When compared to other callers, LOOPbit detected almost twice as many promoter-enhancer loops. This feature is likely to be a direct consequence of the SOFM classification and its ability to capture the functional information embedded in CTCF-pairings structures. LOOPbit was also able to identify differentially called loops between experiments where the CTCF protein was degraded by auxin-induced degron system (Nora et al., 2017). Interestingly, loops detected by LOOPbit in auxin-treated samples were significantly enriched in long-range distances, this is in agreement with the loss of local organization of TADs and the preserved regulation of long-range enhancer-promoter interactions by the remaining ~20% of CTCF (Hyle et al., 2019; Elphege P Nora et al., 2017; Splinter et al., 2006).

In summary, we have shown that signal deconvolution is able to define sub-classes of CTCF-driven chromatin loops with specific structural features, which allowed us to train with an increased specificity an Artificial Intelligence algorithm to finally call functionally relevant loops. Beyond the scope of this study, our methodology opens the possibility to train our CNN on loops driven by other DNA binding proteins.

METHODS

CTCF ChIP-Seq and Hi-C data.

CTCF ChIP-Seq experiments of Human B-lymphocyte cell line (GM12878) were downloaded from the ENCODE database (https://www.encodeproject.org/; dataset ids: ENCSR000DRZ, ENCSR000AKB and ENCSR000DZN) (Davis et al., 2018). After processing the peaks following the ENCODE pipelines available at https://github.com/ENCODE-DCC, only common peaks from all three experiments were kept resulting in a total of 52,844 CTCF peaks. Next, the orientation of the CTCF binding motifs was assessed by means of the MEME and FIMO motif-based sequence analysis tools (Grant et al., 2011; Ma et al., 2014). Only peaks with a statistically significant CTCF motif were kept (p-value < 0.05), which resulted in a total of 41,816 *ChIP-Seq+motif* peaks. This filter discarded ~21% of the original detected CTCF ChIP-Seq peaks.

In situ Hi-C datasets for GM12878 cell line (Rao et al., 2014) were downloaded from the GEO database (**Supplementary Table 1**) and its replicates merged and parsed using TADbit as previously described (Serra et al., 2017). Two resolution matrices (that is, at 100 kb and at 5 kb) were obtained and normalized using OneD (Vidal et al., 2018) with default parameters. The 100 kb Hi-C matrices were next used to calculate chromosome compartmentalization (Imakaev et al., 2012; Lieberman-Aiden et al., 2009) using TADbit (Serra et al., 2017). The

5 kb resolution matrices were further parsed to subtract 45 kb squared submatrices centered in each axis of the matrices to any two of the 15,597 isolated CTCF *ChIP-Seq+motif* peaks (that is, ~38% of all selected peaks above). Isolated peaks were defined as those *ChIP-Seq+motif* peaks with no other peak within a 45 kb window span from the center of the peak. This additional filter ensured that the observed signal in a Hi-C submatrix was due to a particular pair of CTCF-CTCF peak and not multiple pairs. Finally, we subtracted a total of 130,655 submatrices between any pair of isolated CTCF-CTCF peaks spanning any distance between 45 kb and 1.5 Mb, which ensured to select most pairs within the size of a typical TAD in the human genome (~900 kb).

Submatrix analysis, deconvolution, classification and clustering. All 130,655 submatrices between any pair of isolated CTCF-CTCF peaks were next analyzed with our Python based package called Meta-Waffle and available in GitHub (https://github.com/3DGenomes/metawaffle). Meta-Waffle takes as input a list of coordinates pairs corresponding to the selected pairs of selected peaks. Meta-Waffle then will extract, analyze, deconvolve and classify the submatrices using a Self-Organizing Feature Map (SOFM) approach (Kohonen, 1987) available in http://neupy.com/pages/home.html. Briefly, Meta-Waffle first extracts the submatrices from an input Hi-C map, which in this application were 45 kb per 45 kb (9x9 bins) of size. Second, the submatrices values are next re-scaled between 0 and 1 using a sigmoid function. Third, the re-scaled submatrices are input to the SOFM based on a series of empirically optimized parameters including the SOFM grid size (*i.e.,* 10x10, 20x20, 30x30, 40x40, 45x45, and 50x50), learning radius (*i.e.,* 1,2, and 5), standard deviation (*i.e.,* 0.5, 0.1, and 0.01), steps (*i.e.,* 1, 0.5, 0.1, and 0.01), and Epochs (*i.e.,* 30, 90, 130, and 200). Combination of the tested parameters resulted in a total of 1,152 generated SOFMs. Assessing which set of parameters results in the best sub-matrices classification is not trivial as there is no a "standard of gold" for classifying CTCF-CTCF Hi-C sub-matrices. Therefore, we devised three measures to be maximized by the selected SOFM optimal parameters. First, a percentage of classified sub-matrices, which allowed to select the optimal parameters that result in minimal singleton cells (that is, cells with only one sub-matrix) after the SOFM classification. Second the variability between neurons. And third, a compartment segregation score that maximizes the segregation of SOFM cells in the two main genome compartments. Using the mean compartment signal per SOFM cell, the segregation compartment score was calculated as the subtraction between A and B compartment frequencies in each SOFM map. The SOFM optimal parameters that maximized both of the devised scores were: Grid size: 30 x 30; Learning radius: 5; Standard deviation: 0.5; Step: 0.01; and Epochs: 30. Finally, the optimal SOFM cells were used to generate a hierarchical clustering by means of the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), which can be used as an effective pre-processing step to enhance the performance of density-based clustering. The UMAP was computed using a local neighboring size of 6 for the manifold approximation, an effective minimum distance between embedded points of 0.3 and 100 epochs to increase its accuracy. The final 10 clusters of Hi-C CTCF-CTCF cells were obtained by using HDBSCAN (Campello and D.M.J., 2013) with default parameters, considering a minimum clusters size of 6 and a minimum of 6 samples in a neighborhood for a point to be considered a core point.

Chromatin states integration.

The chromatin 15-state model for the GM12878 cell line was downloaded from the Roadmap Epigenomics Consortium (Roadmap Epigenomics et al., 2015). Following a similar previously published protocol (Forcato et al., 2017), all 15 states were merged into 4 major classes: promoter (Active TSS, Flanking Active TSS, Bivalent/poised TSS Flanking bivalent TSS), enhancer (Enhancers, Genic enhancers, Bivalent enhancers), repressed polycomb (Repressed Polycomb, Weak repressed Polycomb) and heterochromatin (Heterochromatin, Quiescent/Low). Next, the genome was segmented into 5 kb bins, and each bin was classified based on the overlap (> 50 bp) with any of the four major chromatin states. A bin could be assigned to one or more categories. Next, interactions between bins were classified as "Not expected" (interacting promoter/enhancer bin with heterochromatin bin), "Promoter-Enhancer" (interacting promoter bin with an enhancer bin), and "Heterochromatin - Heterochromatin", (interacting heteroch romatin bins).

The LOOPbit CNN.

LOOPbit is a trained Convolutional Neural Network (CNN) to predict the localization of loops from Hi-C interaction matrices. LOOPbit was trained using the TensorFlow platform (https://www.tensorflow.org/) using a common pattern: input matrix flattening - dense layer (ReLu) - dropout - dense layer (Softmax). As an input, the CNN takes tensors of shape (9, 9) that will be flattened in the first layer. The dense layers were used to perform the classification of input Hi-C cells into two different classes: loop and no-loop. The dropout layer is needed to avoid overfitting of the model (**Fig 3a**). The LOOPbit CNN was trained by a 20% leave-out of the data used as test and 80% of the data for training. The training dataset obtained by sub-sampling a total of 6,000 randomly selected CTCF-CTCF submatrices from the SOFM clusters with clear signal of looping (loops) and another 6,000 randomly selected CTCF-CTCF submatrices from the SOFM clusters with no signal of looping (no-loop) (**Fig 3a**). In particular, loop submatrices were extracted from clusters 1, 2, 3, 4 and 5, and no-loop submatrices from clusters 7, 8, 9 and 10 (**Fig 1d**). The trained model used a 0.2 drop of the data and 1,024 neurons resulting in a classification accuracy of 85.9% and 89.6% for the loop and no-loop class, respectively. As LOOPbit retrieves a loop probability per matrix cell, the chromatin loops are defined according to the probability values similarity within the neighbors in a delimited area, which was set to 3 bins for the benchmark and 5 bins for the CTCF depleted experiment. Thus, multiple neighboring matrix cells with high loop probability would be considered to be the same chromatin loop.

LOOPbit benchmark.

The trained model was then use to predict loop localization in a different Hi-C matrix with different goals:

1. *Large-scale benchmark for loop detection:* To benchmark the accuracy of LOOPbit detecting loops from a Hi-C input matrix, we used 38 previously published Hi-C datasets (**Supplementary**

Table 2). To avoid biases derived from the data processing pipelines, all 38 datasets were all processed using the same protocol as the training datasets from GM12878 cell line (see above). Then, the Hi-C experiments were analyzed with LOOPbit, scanning chromosome-wise using a windows of 9x9 bins and a step of 1 bin, to predict loops at two different resolutions (5 kb and 40 kb). The predicted localization of loops was next assessed using several measures. First, a Jaccard Index (JI) was computed to assess reproducibility between replicates of the same experiment. Two predicted loops were considered to be identical when they shared exactly the same anchoring bins in both replicates. We also computed the reproducibility allowing the 70% of overlapping between the chromatin loops (**Supplementary Fig. 3**). Second, to characterize the possible biological relevance of the predictions, the enrichment of diverse chromatin marks at loop anchors was calculated. For the benchmarking, the 15-states chromatin models for GM12878, IMR90 and h1-ESC cell lines were downloaded from the Roadmap Epigenomics Consortium (Roadmap Epigenomics et al., 2015) and analyzed as described above. For the fly late embryos dataset, the 16-chromatin states model was download from modENCODE (Ho et al., 2014). As previously, the states were also merged into 4 major classes: promoter (Promoter), enhancer (Enhancer 1, Enhancer 2), repressed polycomb (PC repressed 1, PC repressed 2), and heterochromatin/Low (Heterochromatin1, Heterochromatin 2, Low signal 1, Low signal 2, Low signal 3). Similarly, the enrichment analysis of the chromatin states at the loop anchors was done described above.

Additionally, the assessment presence of CTCF sites and their orientation in the base of the predicted loops was performed only for *cis* interactions identified in the Hi-C maps at 5 kb resolution. Briefly, CTCF ChIP-seq experiments were downloaded (**Supplementary Table 3**) and peaks processed using HOMER motif analysis with default parameters. The *cis* interactions conserved in at least 2 replicates within each dataset (with exception of Jin H1-hESC with one replicate) were intersected with the list of motifs of the CTCF peaks. Then, an interaction was considered convergent if the interacting bin closer to the p-terminus of the chromosome contained one CTCF motif on the forward strand (+ orientation), and the interacting bin closer to the q-terminus of the chromosome contained one CTCF motif on the reverse strand (– orientation) (Forcato et al., 2017).

2. *Control of a CTCF depleted experiment.* To assess whether LOOPbit still detects loops in a CTCF depleted cell, we downloaded the Hi-C and CTCF ChIP-Sequencing data from the GSE98671 GEO entry (Nora et al., 2017). The dataset include three samples: (i) untreated cells, (ii) cells treated for 2 hours with a degron system targeting CTCF with about 20% remaining CTCF after treatment, and (iii) wash-off cells with recovered CTCF levels. The three biological replicates were merged and parsed as indicated above. As in the other benchmark, we assess the accuracy of LOOPbit detecting loops by means of the Jaccard Index between replicates as well as chromatin marks enrichment at loop anchors (see above).

CODE AVAILABILITY

Meta-Waffle as well as LOOPbit are available on GitHub: (https://github.com/3DGenomes/metawaffle and https://github.com/3DGenomes/loopbit, respectively).

ACKNOWLEDGEMENTS

We thank the authors of (Forcato et al., 2017) for providing the dataset used to make possible the comparison of LOOPbit with other loop callers. We thank the authors of (E. P. Nora et al., 2017) for helping with the interpretation of CTCF targeted-degraded Hi-C datasets. We thank David Castillo, Luciano Di Croce, Sandra Peiró and François Le Dily for their continuous discussions and support to the development of Meta-Waffle and LOOPbit.

FUNDING

This work was partially supported by the European Research Council under the 7th Framework Program FP7/2007-2013 (ERC grant agreement 609989), the European Union's Horizon 2020 research and innovation programme (grant agreement 676556) and the Spanish Ministerio de Ciencia e Innovación (BFU2017-85926-P). We also knowledge support from "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208, the Spanish ministry of Science and Innovation to the EMBL partnership and the CERCA Programme/Generalitat de Catalunya to the CRG. Finally, we acknowledge support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement and the Co-financing by the Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014-2020 Smart Growth Operating Program to CNAG.

CONFLICT OF INTEREST

None declared.

REFERENCES

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

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat Rev Genet *17*, 661-678.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A.*, et al.* (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557-572 e524.

Campello, R.J.G.B., and D.M.J., S. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Advances in Knowledge Discovery and Data Mining *7819*, 160-172.

Campello, R.J.G.B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates (Berlin, Heidelberg: Springer Berlin Heidelberg).

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet *2*, 292-301.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K.*, et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794-D801.

de Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., Welling, M.*, et al.* (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227-231.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110-1121.

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.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat Methods *14*, 679-685.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell reports *15*, 2038-2049.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y*., et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900-910.

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.

Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A*., et al.* (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449-452.

Hwang, Y.C., Lin, C.F., Valladares, O., Malamon, J., Kuksa, P.P., Zheng, Q., Gregory, B.D., and Wang, L.S. (2015). HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics *31*, 1290-1292.

Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., Tian, L., Feng, R., Xu, P., and Li, C. (2019). Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. Nucleic Acids Res *47*, 6699-6713.

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.

Kohonen, T. (1987). Adaptive, associative, and self-organizing functions in neural computing. . Appl Opt, *26(23)*, 4910-4918.

Lekschas, F., Bach, B., Kerpedjiev, P., Gehlenborg, N., and Pfister, H. (2018). HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples. IEEE transactions on visualization and computer graphics *24*, 522-531.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O*., et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Lun, A.T., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics *16*, 258.

Ma, W., Noble, W.S., and Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc *9*, 1428-1450.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv *1802.03426*.

McLaughlin, K., Flyamer, I.M., Thomson, J.P., Mjoseng, H.K., Shukla, R., Williamson, I., Grimes, G.R., Illingworth, R.S., Adams, I.R., Pennings, S*., et al.* (2019). DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. Cell reports *29*, 1974-1985 e1976.

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., and Luscombe, N.M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. PLoS One *12*, e0174744.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017-1021.

Nichols, M.H., and Corces, V.G. (2015). A CTCF Code for 3D Genome Architecture. Cell *162*, 703-705.

Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930-944 e922.

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.

Ogiyama, Y., Schuettengruber, B., Papadopoulos, G.L., Chang, J.M., and Cavalli, G. (2018). Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. Mol Cell *71*, 73-88 e75.

Pekowska, A., Klaus, B., Xiang, W., Severino, J., Daigle, N., Klein, F.A., Oles, M., Casellas, R., Ellenberg, J., Steinmetz, L.M*., et al.* (2018). Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency. Cell Syst *7*, 482-495 e410.

Rao, S.S., 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.

Roadmap Epigenomics, C., 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.

Ruiz, J.L., Tena, J.J., Bancells, C., Cortes, A., Gomez-Skarmeta, J.L., and Gomez-Diaz, E. (2018). Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum. Nucleic Acids Res *46*, 9414-9431.

Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J.*, et al.* (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*, E6456-6465.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., C, H.H., Mirny, L.*, et al.* (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51-56.

Serra, F., Bau, D., Goodstadt, M., Castillo, D., Filion, G.J., and 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*, e1005665.

Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. Cell *160*, 1049-1059.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345-354.

Van Bortle, K., and Corces, V.G. (2012). Nuclear organization and genome function. Annu Rev Cell Dev Biol *28*, 163-187.

Vidal, E., le Dily, F., Quilez, J., Stadhouders, R., Cuartero, Y., Graf, T., Marti-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.

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J.*, et al.* (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. EMBO J *36*, 3573-3599.

Yang, T., Zhang, F., Yardimci, 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.

FIGURES

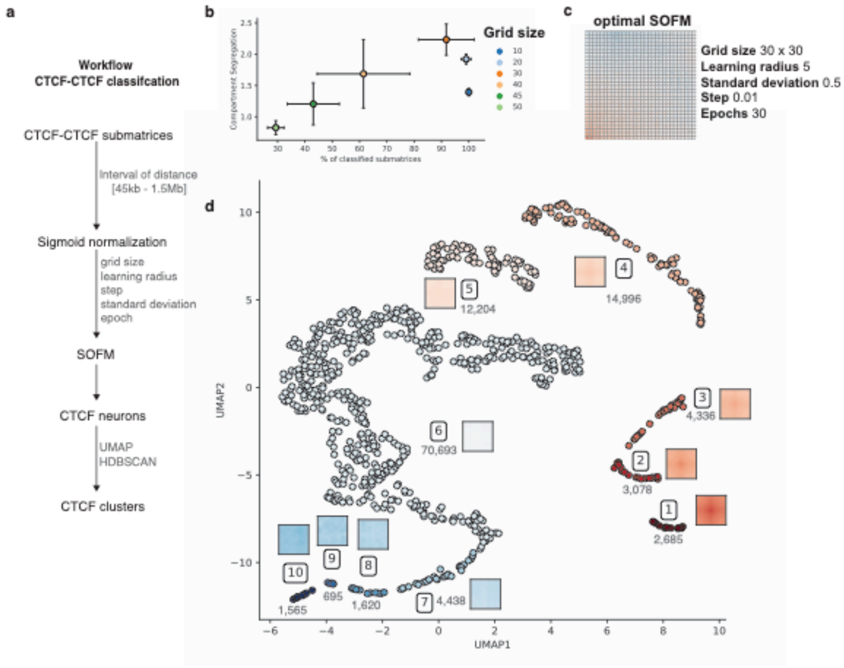


Figure 1. General overview of signal structure deconvolution. a. Schematic workflow to obtain CTCF clusters. **b.** SOFM parameters selected based on the percentage of classified matrices and the compartment segregation value (see also **Supplementary Fig. 1**). **c.** SOFM map showcasing the medoid of each neuron for optimal SOFM with the highest compartment segregation as well as highest percentage of classified CTCF-CTCF submatrices. **d.** Low dimensional representation of the SOFM neurons using a UMAP algorithm followed by a clustering method using HBDSCAN. A total of 10 clusters were obtained. Each point corresponds to a neuron in the SOFM map in panel c. Clusters are represented by the medoid signal and the number of CTCF-CTCF submatrices per cluster.

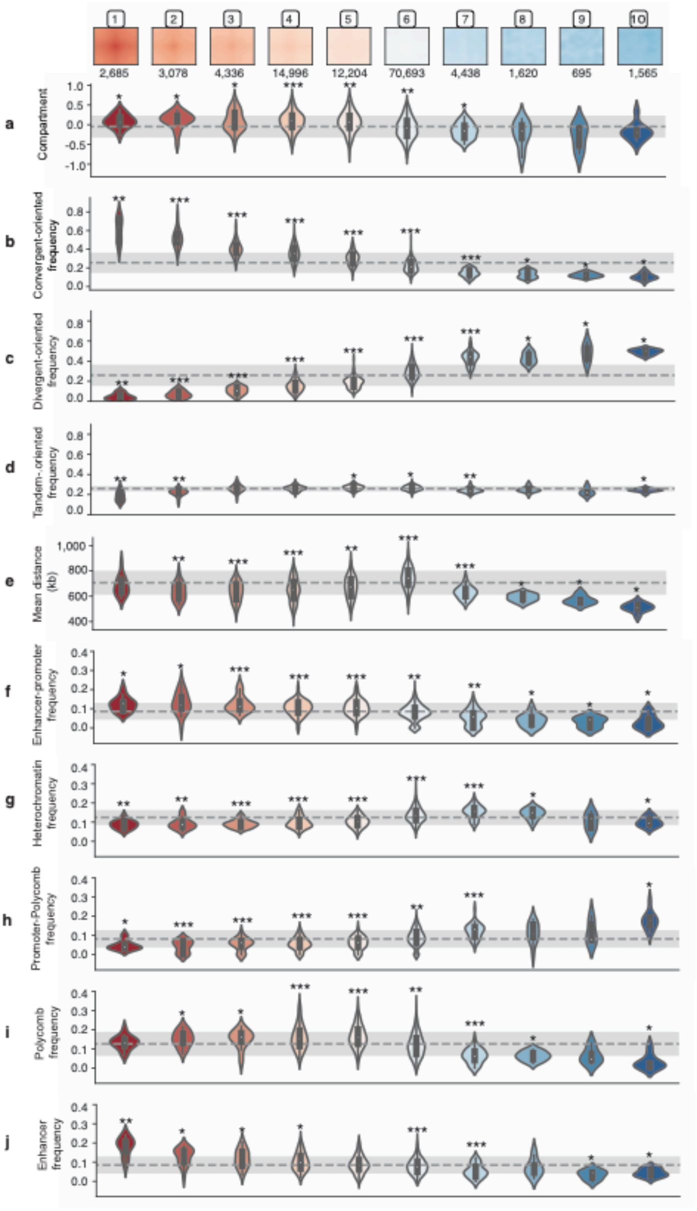


Figure 2. Distribution of multiple genomic features throughout CTCF clusters. In the first row the representative medoid of each CTCF cluster is represented. Below the different genomic features per cluster. In the first row the compartment type distribution, then the binding CTCF motif orientation, the distance between the two CTCFs and finally the chromatin state enrichment of the CTCF pairs. The dashed gray lines mark the mean, whereas the standard deviation is marked with lighter grey. Statistical significance

against the mean is calculated using Wilcoxon test, $p < 0.05$ (*), $p < 0.001$ (**), $p < 0.0001$ (***).

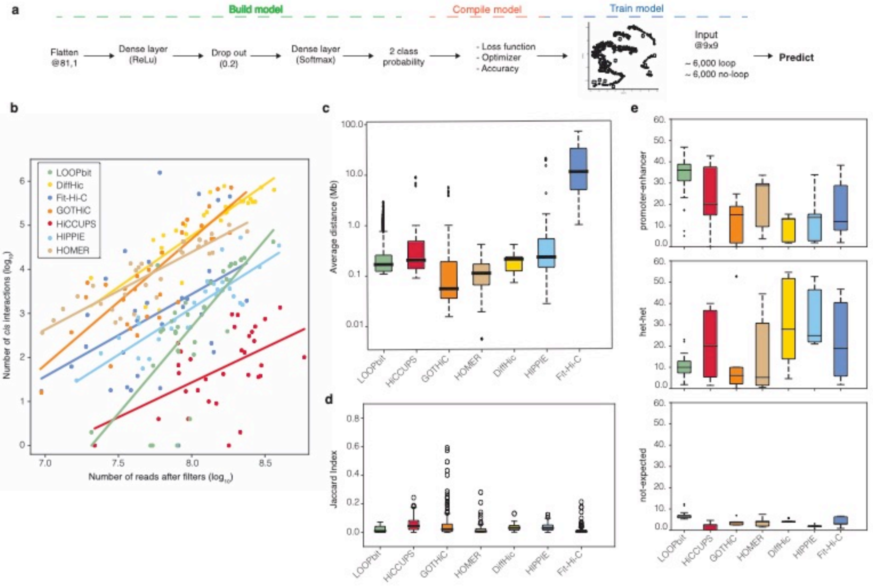


Figure 3. LOOPbit technical and biological benchmarking. a. CNN workflow, model building with the multiple layers to transform the input data. Model compilation to assess the accuracy and optimization of the model. Finally, training of the model using the data coming from the CTCF-CTCF SOFM deconvolution. **b.** Representation of the number of reads after filters and the number of identified *cis*-interactions by LOOPbit in all experiments at a 5 kb resolution (n=32). **c.** Average distance between the identified loop-anchors of all the Hi-C experiments at 5 kb resolution (n=32). **d.** Boxplot representing the Jaccard Index, in here the overlapping between exact loop-anchors were considered to be the same loop between the same replicates (n=39). **e.** Proportion of the identified cis interactions based on the chromatin states at their anchoring points of the datasets at 5 kb resolution (n=32).

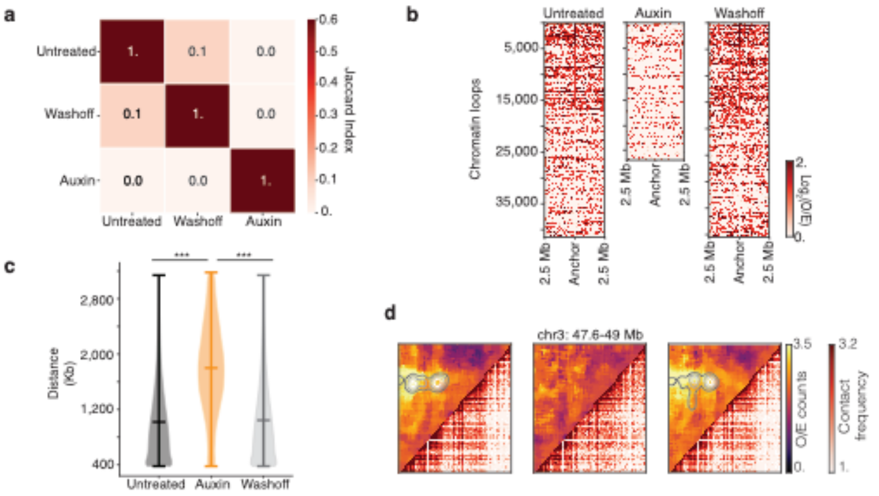


Figure 4. Applicability of LOOPbit to AID-CTCF dataset (E. P. Nora et al., 2017). a. Jaccard Index (JI) between the three samples. **b.** CTCF enrichment in the loop anchors in the untreated, CTCF-degraded and wash-off datasets. **c.** Distance between loop-anchors in the three conditions. Statistical significance against the mean is calculated using Wilcoxon test, p < 0.05 (*), p < 0.001 (**), p < 0.0001 (***). **d.** Example of an output of LOOPbit from a genomic region (chr3:47,600,000-49,000,000), showing half of the normalized Hi-C matrix, and half with the contact frequencies. The density of probabilities provided by LOOPbit define the position of the chromatin loops (white and grey lines).

SUPPLEMENTARY

ID	Reestriction enzyme	Filtered reads
GSM1551552	MboI	361,207,349
GSM1551553	MboI	136,507,541
GSM1551554	MboI	255,191,180
GSM1551555	MboI	129,354491
GSM1551556	MboI	148,151,521
GSM1551557	MboI	172,219,884
GSM1551558	MboI	99,044,057
GSM1551559	MboI	48,503,486
GSM1551560	MboI	48,052,662
GSM1551561	MboI	115,277,508
GSM1551562	MboI	58,351,032
GSM1551563	MboI	209,244,601
GSM1551564	MboI	94,795,755
GSM1551565	MboI	94,272,224
GSM1551566	MboI	147,025,950
GSM1551567	MboI	119,160,683
GSM1551571	MboI	228,444,864
GSM1551572	MboI	218,414,758
GSM1551573	MboI	77,453,007
GSM1551577	MboI	54,523,851
GSM1551578	MboI	121,751,924
GSM1551588	DpnII	58,470,721
GSM1551589	DpnII	76,770,682
GSM1551590	DpnII	61,529,714
GSM1551591	DpnII	93,923,134
GSM1551598	MboI	82,406,062

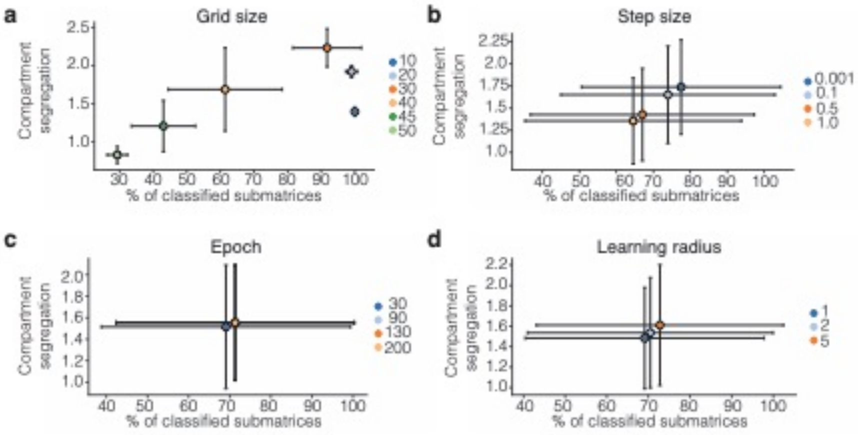
Supplementary Table 1. GM12878 Hi-C experiments used to deconvolve the CTCF-CTCF topology (Rao et al., 2014).

Dataset	ID	Reestriction enzyme	Cell type	Filtered reads	Resolution (kb)
JIN	GSM1055805	HindIII	H1-hESC	134,678,276	5
JIN	GSM1055800	HindIII	IMR90	104,929,328	5
JIN	GSM1055801	HindIII	IMR90	175,071,658	5
JIN	GSM1154021	HindIII	IMR90	97,395,328	5
JIN	GSM1154022	HindIII	IMR90	79,559,232	5
JIN	GSM1154023	HindIII	IMR90	52,425,794	5
JIN	GSM1154024	HindIII	IMR90	54,082,516	5
RAO	GSM1551552	MboI	GM12878	361,207,349	5
RAO	GSM1551569	MboI	GM12878	72,934,660	5
RAO	GSM1551570	MboI	GM12878	77,651,974	5
RAO	GSM1551571	MboI	GM12878	228,444,864	5
RAO	GSM1551572	MboI	GM12878	218,414,758	5
RAO	GSM1551573	MboI	GM12878	77,453,007	5
RAO	GSM1551574	MboI	GM12878	81,318,602	5
RAO	GSM1551575	MboI	GM12878	80,613,339	5
RAO	GSM1551576	MboI	GM12878	80,438,152	5
RAO	GSM1551577	MboI	GM12878	54,523,851	5
RAO	GSM1551578	MboI	GM12878	121,751,924	5
RAO	GSM1551587	DpnII	GM12878	63,854,975	5
RAO	GSM1551588	DpnII	GM12878	58,470,721	5
RAO	GSM1551589	DpnII	GM12878	76,770,682	5
RAO	GSM1551590	DpnII	GM12878	61,529,714	5
RAO	GSM1551591	DpnII	GM12878	93,923,134	5
RAO	GSM1551599	MboI	IMR90	164,365,813	5
RAO	GSM1551600	MboI	IMR90	181,640,359	5
RAO	GSM1551601	MboI	IMR90	20,676,015	5
RAO	GSM1551602	MboI	IMR90	90,970,187	5
RAO	GSM1551603	MboI	IMR90	186,707,523	5
RAO	GSM1551604	MboI	IMR90	198,492,270	5
RAO	GSM1551605	MboI	IMR90	216,948,852	5
Dixon 2015	GSM1267196	HindIII	H1-hESC	172,971,685	5
Dixon 2015	GSM1267197	HindIII	H1-hESC	103,476,074	5
Sexton	GSM849422	DpnII	Fly embryo	31,357,023	40
Dixon 2012	GSM862723	HindIII	H1-hESC	21,292,727	40
Dixon 2012	GSM892306	HindIII	H1-hESC	134,678,276	40
Dixon 2012	GSM862724	HindIII	IMR90	102,906,483	40
Dixon 2012	GSM892307	HindIII	IMR90	104,974,904	40

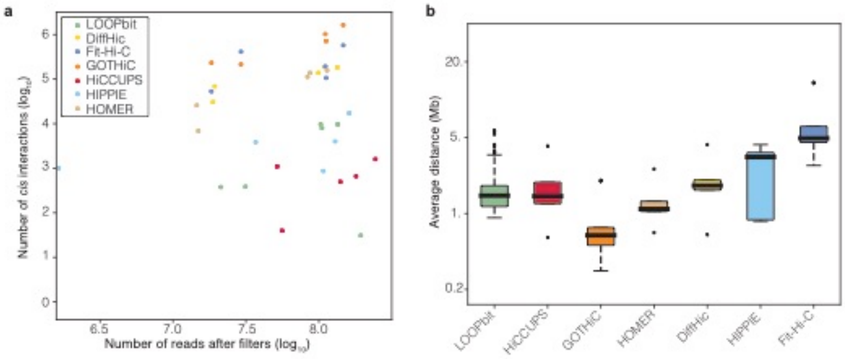
Supplementary Table 2. Hi-C experiments used for LOOPbit benchmarking. All the experiments were preprocessed and filtered using TADbit (Serra et al., 2017) and OneD normalized (Vidal et al., 2018).

Experiment	Cell type	Accession number
CTCF ChIP-seq	H1-hESC, GM12878	GSE29611
CTCF ChIP-seq	IMR90	GSE31477
CTCF ChIP-seq	Embryo 14-16hr Oregon-R	GSE47264

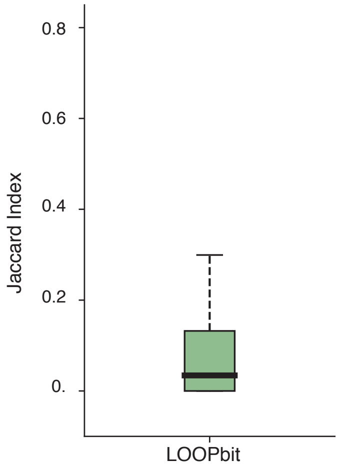
Supplementary Table 3. CTCF ChIP-seq experiments used in the analysis.



Supplementary Figure 1. SOFM parameters. SOFM parameters, grid size, step size, epoch and learning radius, based on the percentage of classified matrices and the compartment segregation value (see also **Fig. 1**), **a**, **b**, **c** and **d,** respectively.



Supplementary Figure 2. Results benchmark at 40 kb resolution. a. Number of reads after filters and the number of identified *cis*-interactions by LOOPbit in all experiments at a 40 kb resolution (n=5). **b.** Average distance between the identified loop-anchors of all the Hi-C experiments at 40 kb resolution (n=5).



Supplementary Figure 3. Jaccard Index allowing 70% of overlapping between chromatin loops. Boxplot representing the Jaccard Index, in here an overlap of 70% between chromatin loops were considered to be the same loop between the same replicates (n=39).

Chapter II

Quantitative comparison and automatic feature extraction for chromatin contact data

Galan S., Machnik N., Kruse K., Díaz N., Marti-Renom MA., Vaquerizas JM.
Quantitative comparison and automatic feature extraction for chromatin contact data.
Accepted in Nature Genetics, to be published.

Quantitative comparison and automatic feature extraction for chromatin contact data

Silvia Galan[1,2‡], Nick Machnik[1,3‡], Kai Kruse[1], Noelia Díaz[1], Marc A. Marti-Renom[2,4,5,6] and Juan M. Vaquerizas[1,7*¶]

Affiliations:

1. Max Planck Institute for Molecular Biomedicine, Roentgenstrasse 20, 48149 Muenster, Germany.
2. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
3. Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria.
4. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Carrer del Doctor Aiguader, 88, 08003 Barcelona, Spain
5. Universitat Pompeu Fabra (UPF), Carrer del Doctor Aiguader 88, Barcelona 08003, Spain.
6. ICREA, Pg Lluís Companys 23, 08010 Barcelona, Spain.
7. MRC London Institute of Medical Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, UK.

‡ These authors contributed equally.

* Corresponding author.

Correspondence to:

Juan M. Vaquerizas (jmv@mpi-muenster.mpg.de; @vaquerizasjm).

Keywords:

Hi-C, Capture-C, structural similarity index, chromosome conformation capture, comparative genomics, algorithm.

Abstract

Dynamic changes in the three-dimensional organisation of chromatin are associated with central biological processes such as transcription, replication, and development. The comprehensive identification and quantification of these changes is therefore fundamental to our understanding of evolutionary and regulatory mechanisms. Here, we present CHESS (Comparison of Hi-C Experiments using Structural Similarity), an algorithm for the comparison of chromatin contact maps and automatic differential feature extraction. We demonstrate the robustness of CHESS to experimental variability and showcase its biological applications on: i) inter-species comparisons of syntenic regions in human and mouse; ii) intra-species identification of conformational changes in Zelda depleted *Drosophila* embryos; iii) patient-specific aberrant chromatin conformation in a diffuse large B-cell lymphoma sample, and, iv) the systematic identification chromatin contact differences in high resolution Capture-C data. In summary, CHESS is a computationally efficient method for the comparison and classification of changes in chromatin contact data.

Introduction

Eukaryotic genomes follow similar global organisational principles: a multi-layer, hierarchical organisation into domains, with specific three-dimensional (3D) interactions between individual genomic regions[1]. Local chromatin conformation, however, can be variable across species[2–6], developmental stages[7–9], cell types[10,11], and can change dynamically with transcription[12], during replication[13], and cell division[14], among others. Mutations affecting nuclear architecture have been shown to cause misregulation of gene expression leading to developmental disorders and disease (reviewed in [15];[16]). It is therefore of paramount importance to elucidate the relationship between nuclear architecture, evolution, and fundamental biological processes.

Existing approaches to discover changes in the 3D conformation of genomic regions have relied partially on visual identification of differences, such as a side-by-side evaluation of Hi-C matrices[17,18] or fold-change maps[19]. While visual comparisons can highlight specific changes in Hi-C matrices, results are often difficult to quantify and, by nature, cannot be automated to compare large numbers of matrices. More quantitative approaches have been developed. One class of tools focuses on the assessment of the degree of similarity or reproducibility between full chromatin contact matrices / datasets and does not allow for the identification of regions with particularly strong similarities or differences[20–24]. Another focuses on the comparison of specific features, such as topologically associating domains (TADs)[25,26], or loops[10,27], which limits the discovery of differences to the specific feature analysed. A third class of tools aims to find single pairs of bins with significantly differential interactions[28–32] without providing any information about the specific type of structural feature that changes. Until now, there is no method that allows a systematic comparison of the 3D conformation of genomic regions, that is at the same time quantitative, able to identify and classify a range of structural variations, and corresponds well to the visual perception of differences.

Here, we describe CHESS, an algorithm to robustly identify and classify specific similarities or differences and features in chromatin contact data using a feature-free approach. CHESS applies the concept of the structural similarity index widely used in image analysis[33,34] to chromatin contact matrices, assigning a structural similarity score and an associated p-value to pairs of genomic regions. Next, CHESS uses image processing approaches to automatically extract three-dimensional chromatin conformation features, such as TADs, stripes or loops. We first demonstrate the robustness of CHESS scores by evaluating the method on artificially generated and real, Hi-C matrices of different sizes, sequencing depths, and varying levels of noise. We then highlight the utility of CHESS in different real-world applications: i) genome-wide structural comparisons of syntenic regions between human and mouse; ii) the detection of conformational changes in *Drosophila melanogaster* upon knockdown of the transcription factor Zelda during early embryonic development; iii) the systematic detection of 3D chromatin conformation changes in B-cells of a diffuse large B-cell lymphoma (DLBCL) patient; and, iv) the automatic detection and classification of subtle changes in chromatin conformation from genome editing experiments. Overall, our results demonstrate that CHESS can be successfully applied to diverse chromatin contact datasets to quantitatively determine structural differences between them.

Results

Overview of the CHESS algorithm

The main aim of CHESS is to assess the degree of similarity between any pair of normalised chromatin contact matrices, and thereby provide a measure that allows to identify particularly similar or dissimilar matrices. To calculate the comparison, we use the structural similarity index (SSIM), a widely used metric for similarity of matrices. While it was initially developed for the evaluation of image quality[33,34], it does not make any assumptions about its input data other than that it comes in the form of two matrices of same dimensionality, with numerical entries, irrespective of how the data in these matrices have been generated. Outside of the computer vision field, it is for example also used in transportation research to compare matrix representations of origin-destination graphs, which differ from chromatin contact graphs conceptually only in that they are directed graphs[35–37], and as a similarity metric for acoustic pressure signals[38–40].

CHESS accepts as input any pair of normalised chromatin contact matrices[41,42], such as those produced from Hi-C or tiled Capture-C experiments (Fig. 1a). Matrices can originate from different regions in the same genome, the same genomic region across two experimental conditions, developmental time points, or even regions from different species. To calculate the similarity between the two matrices, which we refer to as reference (R) and query (Q), matrix entries (i.e., contact pairs or pixels in the maps) are first divided by the expected contact intensity at the respective distance. This transformation is necessary since Hi-C matrices display a characteristically high signal at the diagonal that corresponds to an increased contact probability of nearby regions in linear distance[43]. CHESS comparisons on matrices that are not corrected for this distance-dependency of contact probabilities are sensitive to varying experimental noise and relative region sizes (Supplementary Fig. 1) (Methods). Therefore, the distance-dependency is removed by dividing the observed contact values by the average contact value at the respective genomic distance. CHESS then scales the matrices to equal size and calculates the SSIM between R and Q. The SSIM score consists of comparisons of the brightness, contrast and structure between two matrices. Brightness is calculated as the mean of the signal intensity. Contrast is calculated as the variance in signal. The structure term is calculated as the correlation between signal values of two matrices. Then, SSIM is defined as a weighted product of these three components, which are scaled such that SSIM ranges between -1 and 1. A SSIM of 0 indicates no similarity, 1 reflects identity and -1 a perfect inverse relationship. Typically, SSIM is computed for many small windows within the R and Q matrix comparison. The full SSIM score (in the following referred to as S) for the particular R vs Q pair is then the average of all small comparisons. We performed an in-depth evaluation of the three components of SSIM (Supplementary Fig. 2) which showed that the main contributor to similarity when applied to Hi-C matrices is the product of contrast and structure terms.

S can be used directly to quantify changes in the Hi-C matrix of the same region: identical matrices have a "baseline" score of $S = 1$, and the lower the score the larger the change. This makes it easy to quantify how much the structure of any given region is changing, for example under different experimental conditions. However, when comparing different regions, a baseline is necessary to determine the statistical significance of S. It is therefore essential to put the obtained scores into the context of an appropriate background model. For example, a region R containing just a single TAD might obtain a high score when compared to a particular query Q, but might be equally similar to other regions in the genome. The score for the comparison of R vs Q should then be assigned a low significance. To compute a suitable background model, CHESS compares the reference matrix R to all other regions of the same size across the genome (referred to as ${Q^B}_i$ in Fig. 1b). The distribution of scores from the background model is then used to calculate a z-score, corresponding to a normalised effect size, and a p-value, denoting the frequency of scores equal to or higher than S in the background model (Fig. 1c) (Methods). Therefore, CHESS enables a quantitative comparison and assessment of statistical significance of contact matrix similarities.

Automatic feature selection and classification of structural changes

In addition to the identification of changes in chromosome conformation data, a major analytical task consists in the recognition of changes in specific 3D chromatin organisation features, such as TADs, among others. To specifically determine which 3D genome features are gained or lost between a pair of samples, CHESS implements a simple and fast workflow of image filters that allow the automatic identification and classification of such features in these samples (Fig. 1d). In brief, first, a differential contact matrix is computed between all sets of R vs Q matrices identified by CHESS, and gained and lost contacts in each matrix are separated for further analysis. Then, the matrices are denoised, smoothed and binarised to apply a close morphology filter to extract the individual areas that change in each comparison. Subsequently, the 2D cross-correlation between all the extracted areas in a dataset are computed. Finally, K-means clustering is used to detect the main structural features identified in these areas, such as TADs, loops, stripes and borders (Fig. 1d). Overall, this strategy allows us to automatically identify the precise 3D structural features, such as TADs, loops and stripes, which are present in a particular differential region between samples captured by CHESS.

CHESS requires only low sequencing depth and tolerates a high level of noise

To robustly estimate the performance of CHESS with regards to different experimental conditions (e.g., noise, sequencing depth) and matrix parameters (e.g., size, size difference), we generated a set of synthetic Hi-C data designed to reflect features commonly observed in real-world Hi-C matrices, including an exponential decay of contact frequency with genomic distance, TADs, and loops between genomic regions (Methods). Typical numbers of valid read pairs in current Hi-C studies range from 0.1 million per megabase (M/Mb)[10,18] to 1 M/Mb[12]. To test the performance of CHESS on datasets with different sequencing depths, we generated synthetic matrices within a range of numbers of simulated read pairs. As an example of a deeply sequenced dataset, we generated matrices with an equivalent depth of 1.5 M/Mb, corresponding to ~4.5 billion mapped reads across the whole genome in a Hi-C experiment on human cells. Datasets at lower sequencing depths were then generated by downsampling the number of read pairs in the original matrix by randomly removing pairs of contacts (Supplementary Fig. 3) (Methods). This ensures that the overall structure of the dataset is maintained for the evaluation of sequencing depth-related effects. In addition, experimental noise was also simulated in the synthetic datasets by removing a number of contacts and adding them at random locations (Supplementary Fig. 3). This allows us to model the effect of random ligations, a main contributor to noise in chromatin contact maps.

To test the sensitivity of CHESS to noise and sequencing depth, we used the algorithm to compare an artificial matrix R to a copy Q of itself, while adjusting the sequencing depth and adding noise to both of the matrices independently. As a background to calculate p-values and z-scores, similarity scores were additionally calculated in comparisons of R to 1,000 randomly generated artificial matrices at the same sequencing depth and noise level (Fig. 2a). These simulations show that Q is correctly assigned the best CHESS score for "deeply sequenced" matrices up to a noise level of 90 % (Fig. 2b). Beyond that, ranking quickly becomes random, which is reflected in a uniform distribution of p-values (Supplementary Fig. 4). For artificial matrices with less contacts, CHESS still tolerates noise levels of 60 - 80 %. Correspondingly, z-scores are consistently high for the same levels of noise depending on sequencing depth (Fig. 2c). Interestingly, z-scores do not peak at 0 % noise. This can be explained by the changing standard deviations of the background scores: with increasing noise, the similarity of random matrices increases, leading to progressively narrower distributions of CHESS scores. This leads to a slight increase in z-scores up until the point that CHESS is no longer able to identify Q as the top-ranking hit in these comparisons (Supplementary Fig. 4).

To verify these results in a real-world setting, we repeated the above analyses on a deeply sequenced mouse embryonic stem cell Hi-C dataset[12]. Interestingly, the results on real data indicate an even higher robustness of CHESS to high noise and low sequencing depth when compared with the synthetic datasets; while high robustness requires sufficiently large region sizes (> 2.5 Mb, which is likely due to the increased amount of distinctive features in larger regions), CHESS tolerates sequencing depths as low as 0.06 M/Mb and 80 % noise (Supplementary Fig. 5), demonstrating the applicability of this approach in shallow-sequenced datasets. The increased robustness to noise is likely due to a stronger signal enrichment in structural features, compared to the artificial data, since structural features remain visible by eye even at 95 % noise (Supplementary Fig. 5). Overall these results demonstrate the ability of CHESS to reliably detect similarities between Hi-C matrices even at very low sequencing depths and with high amounts of noise.

Then we tested the reproducibility of CHESS on real datasets[12,44] by performing a parameter sweep analysis, testing the effect of different values for window span (250 kb - 3 Mb), step size (25 kb - 1 Mb), resolution (10 kb and 25 kb) and sequencing depth (percentage of original reads: 20 - 80) (Supplementary Fig. 6). The Jaccard Index (JI) was used to assess the overlap between differential regions obtained by CHESS runs with combinations of the parameter values listed above. The results showed high JI across all tested ranges of parameters values (Supplementary Fig. 6). This analysis shows the high robustness of CHESS across different parameter values, in capturing structural changes between genomic regions.

Finally, to benchmark CHESS, we compared it to three widely used differential interaction detection packages: HOMER[32], diffHiC[30] and ACCOST[31]. All methods were run using default parameters on Hi-C interaction matrices at 5 kb resolution for chromosome 19 from mESC (mouse embryonic stem cells) and NPC (neural progenitor cells)[12]. Since a gold-standard for assessing accuracy of differential chromatin interactions does not exist, we performed the analysis by examining the degree of overlap in differential interacting regions identified by the three methods. Overall, we find a high level of overlap between the three methods (Supplementary Fig. 7). However, it is important to note that CHESS identifies entire regions with differences while diffHiC, HOMER and ACCOST identify specific pairs of bins with significant differences in contact counts. A small proportion of differences were reported only by HOMER, diffHiC and ACCOST. It is important to note that many of these differential interactions were filtered out by CHESS due to low signal to noise ratios.

Together, these results demonstrate that CHESS is able to robustly identify changes in chromatin conformation features over a wide range of experimental conditions.

CHESS similarity scores are consistent for matrices of different sizes

An immediate advantage of the analytical strategy behind CHESS is that it allows the calculation of S for matrices of different sizes. This is needed to measure, for example, the similarity between Hi-C maps of different species, or for paralog-containing regions within the same genome. To do so, we implemented an upscaling transformation of the smaller of the two matrices (in case these are of different sizes) using nearest-neighbour interpolation (Methods). To test the performance of CHESS on matrices of different sizes, we calculated S using an artificial matrix R and a matrix Q that maintains the relative positions, sizes and intensities of all features in R (i.e., TADs and loops), but differs in size by a certain scaling factor (Fig. 2d, Supplementary Fig. 8, Methods). Randomly generated matrices of the same size as Q serve as background to calculate statistical significance. In a "deeply sequenced" Hi-C matrix of 1.5 M/Mb, divided into equally sized regions of at least 60 bins, CHESS consistently ranks Q as the matrix most similar to R even if Q is less than half the size of R (Fig. 2e, f). Small matrices Q (smaller than 30 bins) do not rank higher than random matrices, since they do not provide enough space to fit the features (i.e., TADs and loops) of the reference matrix. A test with a simulated sequencing depth of 100 k/Mb and 25 % noise led to similar results, demonstrating that the method's ability to detect similarities between matrices of different sizes is robust to experimental noise and different levels of sequencing depth (Supplementary Fig. 9).

Comprehensive ranking of syntenic regions by structural similarity

Having validated the ability of CHESS to reliably and robustly detect similarities and differences between synthetic and real Hi-C matrices, we next showcase its use in a real research scenario. Previous studies have examined the level of chromatin conformation similarity for regions of synteny (genomic blocks with high degrees of sequence conservation between species) finding a high degree of structural conservation between them[2,10,18]. These comparisons have mainly focused on visual examination of individual examples[10,18], correlation analyses of specific 3D genome features, such as the binding of architectural proteins[2], measures of insulation[45], or the contact strength correlation within syntenic regions[2]. However, a genome-wide quantification of the degree of similarity at the contact matrix level has not yet been performed.

We used CHESS to determine the level of chromatin conformation conservation for 175 regions of synteny between human and mouse obtained from Synteny Portal[46]. To calculate statistical significance for the degree of conservation, we computed S scores for 100 random permutations of syntenic region pairs. Similarity scores for true syntenic region pairs were strongly and consistently higher than those of random pairs (Fig. 3a, $p < 0.01$), permutation test). Therefore, in agreement with previous observations[2,10,18], these results demonstrate genome-wide that overall, regions of synteny between human and mouse share a similar three-dimensional chromatin organisation. However, our results highlight that not all regions of synteny have the same degree of structural similarity (Fig. 3b), suggesting that the evolutionary constraints on three-dimensional chromatin structure are not uniform across the genome, resulting in some regions evolving at different rates than others. In summary, our results demonstrate that CHESS can be used to automatically quantify and rank 3D structural similarity genome-wide between species.

Detection of structural variation upon genetic perturbation

A common problem in comparative genomics is the detection of emerging changes in a system with different experimental conditions, including targeted introduction of disturbances into the system. When applied to chromatin conformation, this approach has been fundamental to determine the contribution to 3D chromatin organisation of different factors, such as CTCF, cohesin or Wapl, among others[47–52]. However, these studies have mostly relied on the detection of visual differences between Hi-C maps or the comparison of measurements derived from these maps, such as the directionality or insulation indices.

Using the insulation score[19], a metric that is low at TAD borders and high within TADs, we have previously shown that depletion of the pioneer transcription factor Zelda during early embryonic development in *Drosophila* leads to a weakening of insulation at TAD boundaries in loci strongly bound by Zelda in wild type embryos[9]. Therefore, we sought to evaluate the sensitivity of CHESS in detecting these changes, as well as its ability to detect further modifications in chromatin conformation that would have escaped detection by a simple comparison of insulation scores.

Running CHESS in a comparison between wild type nuclear cycle 14 and Zelda-depleted embryos resulted in the detection of 65 regions in the genomes with changes in chromatin conformation. Out of the 62 differential boundaries identified before[9], 29 were contained in regions marked as changing by CHESS. Visual inspection of the structural changes at the remaining 33 differential boundaries revealed that the differences in contact intensities were typically small and primarily caused by a decline in short distance contacts around the boundary (Supplementary Fig. 10). We reasoned that a smaller matrix size should increase the sensitivity of CHESS with regards to these types of changes, since they would correspond to a larger fraction of the input matrix pixels. Indeed, after reducing the size of the input matrices from

250 kb to 125 kb, we detected 51 out of 63 differential boundaries as changing, along with 163 additional regions. It is important to note that this approach is likely to miss changes occurring far away from the diagonal, such as differences in the contact probability decay, long-range loops, or large TADs. Therefore, we conclude that, by altering the size of compared matrices, it is possible to fine-tune CHESS to the scale of changes it can detect.

Visual examination of the regions captured in the first run, as well as of control regions, confirmed the detected differences. Notably, besides the already reported loss of insulation at a subset of Zelda-bound TAD boundaries (Fig. 4a)[9], the newly identified regions highlighted a range of structural changes, such as differing signal intensity away from the main diagonal of the Hi-C matrix, suggesting changing levels of chromatin compaction, and varying contact intensities inside TADs and at long distances (Fig. 4b-e). These results demonstrate that CHESS is able to systematically identify regions that undergo structural changes upon genetic perturbation, and that the identified differences cover a broad spectrum of structural features which correspond well to visual perception of differences.

CHESS identifies structural abnormalities in diffuse large B-cell lymphoma

We next sought to determine whether CHESS is able to detect structural differences in clinically relevant samples. The characterisation of these differences is of prime importance since changes in the 3D structure of chromatin in mammals can have a strong impact on genomic regulation and thereby give rise to disease phenotypes[53,54] and the activation of oncogenes[55,56]. We reasoned that by comprehensively scanning a genome for structural abnormalities compared to a healthy control, CHESS can greatly aid our understanding of the relationship between nuclear architecture and disease. To test this, we performed a CHESS comparison using a recent Hi-C dataset from primary diffuse large B-cell lymphoma (DLBCL) and healthy B-cells[44]. Across the whole genome, CHESS identified 810 regions of 2 Mb size with prominent structural variations in DLBCL (Fig. 5a, b). After filtering these for regions with high experimental noise (Methods), we obtained a high-confidence set of 112 regions exhibiting clear changes between healthy and diseased B-cells (Fig. 5c-e). One of the most striking examples displayed the emergence of well-defined TAD structures in a region that seemed devoid of structural features in healthy cells (Fig. 5e). Despite these differences, our analysis also revealed that the majority of structures remained unchanged (Fig. 5f). To gain further insight into the nature of the changes, we applied the feature extraction component of CHESS to the 112 selected regions. This resulted in the identification of 144 gained features (104 stripes and 40 TADs) and 53 lost loops in the DLBCL sample compared to the control (Fig. 5g). This illustrates the application of CHESS to examine disease-related processes by systematically identifying genomic regions and characterizing the specific features whose 3D structure differs between healthy and diseased cells.

Identification and automatic classification of structural features in Capture-C data

Finally, we investigated whether CHESS is also applicable to additional types of chromatin conformation capture datasets, such as tiled Capture-C experiments[57]. To do so, we analysed previously published Capture-C experiments for CRISPR-Cas9 mediated genome edits of architectural features, such as deletion of CTCF binding sites, modifications of TAD boundaries, and a TAD inversion at the *Sox9/Kcnj2* locus[58]. CHESS identified all previously described 3D rearrangements in the different mutants compared to wild-type mice (Fig. 6 and Supplementary Fig. 11). In addition, CHESS identified marked differences that had not been reported in the original study.

For example, besides the previously reported TAD fusion resulting from the *Sox9* regulatory domain inversion (*InvC*), CHESS captured the loss of chromatin loops between the two TADs (Fig. 6a). A similar inversion not including the TAD boundary (*Inv-Intra*) did not result in a TAD fusion. However, CHESS captured an increase in contact frequencies across the boundary in the form of a stripe (Fig. 6b). Applying CHESS to all generated mutants systematically characterised the set of very subtle differences across these samples (Supplementary Fig. 11). This demonstrates that CHESS is able to automatically identify and classify chromatin contact differences.

Discussion

The increasing wealth of available Hi-C datasets calls for fast, quantitative algorithms that enable a systematic comparison of local chromatin structure. However, currently there are no algorithmic approaches for Hi-C data analysis that allow automated comparisons and classification of the identified 3D genome changes directly on the matrix level. This results in a lack of identification and characterisation of a broad spectrum of differences in chromatin conformation maps that can be visually recognised, but that may be missed by more specialised approaches relying on pre-processed features. We have developed CHESS to fill this gap by providing automated, systematic Hi-C matrix comparisons, and feature classification that correspond well to the visual perception of structural differences (Fig. 1). A major feature of CHESS is that it is not limited to comparing regions within a single dataset, but comparisons can be made between samples, cell types, developmental stages, and even across different species, which makes it widely applicable. We demonstrate that CHESS is robust to experimental noise and usable on shallow sequenced datasets (Fig. 2). Furthermore, we show that CHESS can be used to perform cross species comparisons (Fig. 3) and that it is able to detect 3D genome changes in genomes of different sizes (Fig. 4-5). Finally, we demonstrate that CHESS can be used to analyse chromatin conformation capture datasets generated using different experimental approaches, such as Capture-C[57] (Fig. 6). Therefore, we expect CHESS to be immediately applicable to other datasets, including tethered chromatin conformation capture (TCC)[59], digestion-ligation-only Hi-C (DLO Hi-C)[60], genome architecture mapping (GAM)[61], and microscopy based methods, such as Hi-M[62].

An additional advantage of CHESS is the fast and highly efficient implementation of the structural similarity algorithm that has a very small memory footprint, as only the two matrices that are being compared need to be loaded. As a comparison, when scanning a whole chromosome for structural differences between conditions, CHESS achieves a 4-320 times speedup at 3 times lower memory consumption compared to HOMER, diffHiC and ACCOST [30,32,31] (Supplementary Fig. 7). This makes the approach usable without requiring an advanced computational infrastructure. In addition, the nature of CHESS comparisons makes them trivially parallelizable, so that the algorithm can be efficiently sped up by dedicating more computational resources to it. This allows CHESS to make the myriad of comparisons necessary for more complex biological questions, including the background computation for comparing regions of different origin.

Within this context, a promising outlook for CHESS applications is the *de novo* discovery of structurally similar regions between two genomes using an all-against-all comparison approach. These "structurally syntenic" region pairs could provide fundamental insights on the evolution of nuclear architecture and its 3D constraints, including the effects of processes affecting 3D chromatin organisation such as rearrangements or changes in the binding of architectural proteins. Despite the efficiency of CHESS, further work and heuristics would be necessary to make this computationally tractable. Highlighting the importance of considering the 3D genome in evolutionary analyses, we find different degrees of structural conservation across mammalian evolution (Fig. 4), suggesting different rates of evolutionary change in these regions. This demonstrates how CHESS can already facilitate the study of evolutionary genomics in the context of 3D structure.

Similarly, the identification of structural variation and its association with abnormalities in three-dimensional genome organisation and gene expression misregulation is central to evaluate the contribution of chromatin organisation to disease-generating processes. As a proof of principle, here we demonstrate how CHESS can be used to detect a number of chromatin conformation alterations genome-wide in B-cells from a DLBCL patient without the need of previous knowledge regarding the nature of the aberrations. Interestingly, our analysis identified regions in the genome gaining structural features, such as TADs and loops, despite the lack of protein coding genes in these regions. Instead, these regions frequently contained long non-coding RNAs and pseudogenes. Future work integrating other patient-matched genome-wide datasets, such as chromatin opening or RNA-seq, will be necessary to determine the cause and consequence of these changes in relation to disease.

Future improvements of CHESS might benefit from a further dissection of the structural similarity index, which may allow us to pinpoint the contributions of individual regions to overall matrix similarity. In general, structural similarity of images is an active field of research. Modifications improving the robustness of SSIM to small shifts in position [63] or its power to identify similar sub-images [64] are promising, but it remains to be determined which of the algorithms developed for assessing image similarity are compatible with the specific requirements of Hi-C matrix comparisons.

In conclusion, CHESS is an algorithm to quantitatively assess and classify the structural similarity of two genomic regions from chromosome conformation capture data - without the need for feature selection prior to comparison. CHESS is highly tolerant of differences in chromatin conformation capture library size and the noise level of datasets. Its applications include the ranking of known region pairs by similarity, such as syntenic regions in different species, and the discovery of structural changes, such as chromatin conformational changes of the same genomic region in two different conditions. CHESS has great utility in the field of chromatin conformation and can simplify the identification of disease-associated structural variation in clinical applications.

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 661–678 (2016).
2. Vietri Rudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* **10**, 1297–1309 (2015).
3. Acemel, R. D., Maeso, I. & Gómez-Skarmeta, J. L. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *WIREs Developmental Biology* **6**, e265 (2017).
4. Lazar, N. H. *et al.* Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* (2018) doi:10.1101/gr.233874.117.
5. Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E. & Gilad, Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet* **15**, (2019).
6. Yang, Y., Zhang, Y., Ren, B., Dixon, J. R. & Ma, J. Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF. *Cell Systems* **8**, 494-505.e14 (2019).
7. Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381.e20 (2017).
8. Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232–235 (2017).
9. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228.e19 (2017).
10. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
11. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
12. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).
13. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
14. Gibcus, J. H. *et al.* A pathway for mitotic chromosome formation. *Science* **359**, (2018).
15. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).

16. Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771–782 (2016).

17. Darrow, E. M. *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *PNAS* **113**, E4504–E4512 (2016).

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

19. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).

20. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* **27**, 1939–1949 (2017).

21. Sauria, M. E. & Taylor, J. QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data. *bioRxiv* 204438 (2017) doi:10.1101/204438.

22. Ursu, O. *et al.* GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).

23. Yan, K.-K., Yardımcı, G. G., Yan, C., Noble, W. S. & Gerstein, M. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199–2201 (2017).

24. Shavit, Y. & Lio', P. Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol Biosyst* **10**, 1576–1585 (2014).

25. Huynh, L. & Hormozdiari, F. Contribution of structural variation to genome structure: TAD fusion discovery and ranking. *bioRxiv* 279356 (2018) doi:10.1101/279356.

26. Paulsen, J. *et al.* HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics* **30**, 1620–1622 (2014).

27. Lareau, C. A. & Aryee, M. J. diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics* **34**, 672–674 (2018).

28. Djekidel, M. N., Chen, Y. & Zhang, M. Q. FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Res* **28**, 412–422 (2018).

29. Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I. & Dozmorov, M. G. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics* **19**, (2018).

30. Lun, A. T. L. & Smyth, G. K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, 258 (2015).

31. Cook, K. B., Hristov, B. H., Le Roch, K. G., Vert, J. P. & Noble, W. S. Measuring significant changes in chromatin conformation with ACCOST. *Nucleic Acids Res* **48**, 2303–2311 (2020).

32. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).

33. Wang, Zhou & Bovik, A. C. A universal image quality index. *IEEE Signal Processing Letters* **9**, 81–84 (2002).

34. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* **13**, 600–612 (2004).

35. Behara, K. N. S., Bhaskar, A. & Chung, E. Classification of typical Bluetooth OD matrices based on structural similarity of travel patterns- Case study on Brisbane city. in (2018).

36. Behara, K. N. S., Bhaskar, A. & Chung, E. *Geographical window based structural similarity index for OD matrices comparison*. https://eprints.qut.edu.au/133466/ (2020).

37. Djukic, T., Hoogendoorn, S. & Van Lint, H. Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index. in (2013).

38. Hines, A., Skoglund, J., Kokaram, A. & Harte, N. ViSQOL: The Virtual Speech Quality Objective Listener. in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement* 1–4 (2012).

39. Breakey, D. & Meskell, C. Comparison of metrics for the evaluation of similarity in acoustic pressure signals. *Journal of Sound and Vibration* **332**, 3605–3609 (2013).

40. Hines, A. & Harte, N. Speech intelligibility prediction using a Neurogram Similarity Index Measure. *Speech Communication* **54**, 306–320 (2012).

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

42. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J Numer Anal* **33**, 1029–1047 (2013).

43. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

44. Díaz, N. *et al.* Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat Commun* **9**, 1–13 (2018).

45. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8**, 1–13 (2017).

46. Lee, J. *et al.* Synteny Portal: a web-based application portal for synteny block analysis. *Nucleic Acids Res.* **44**, W35-40 (2016).

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

48. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).

49. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707.e14 (2017).

50. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).

51. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).

52. Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).

53. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

54. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

55. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).

56. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).

57. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).

58. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).

59. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).

60. Lin, D. *et al.* Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat Genet* **50**, 754–763 (2018).

61. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

62. Cardozo Gizzi, A. M. *et al.* Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol. Cell* **74**, 212-222.e5 (2019).

63. Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C. & Markey, M. K. Complex Wavelet Structural Similarity: A New Image Similarity Index. *IEEE Transactions on Image Processing* **18**, 2385–2401 (2009).

64. Homola, T., Dohnal, V. & Zezula, P. Searching for Sub-images Using Sequence Alignment. in *Proceedings of the 2011 IEEE International Symposium on Multimedia* 61–68 (IEEE Computer Society, 2011). doi:10.1109/ISM.2011.19.

Acknowledgements

Work in the Vaquerizas lab is supported by the Max Planck Society, the Deutsche Forschungsgemeinschaft (DFG) Priority Programme SPP2202 Spatial Genome Architecture in Development and Disease (Project Ref VA 1456/1 to J.M.V.), the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 643062 - ZENCODE-ITN, and the Medical Research Council, UK. This research was partially funded by the European Union's H2020 Framework Programme through the ERC (grant agreement 609989 to M.A.M-R.). We also acknowledge the support of Spanish Ministerio de Ciencia, Innovación y Universidades through BFU2017-85926-P to M.A.M-R. CRG thanks the support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208, the CERCA Programme/Generalitat de Catalunya, Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement and the Co-financing by the Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014-2020 Smart Growth Operating Program.

Author Contributions

Conceptualisation: N.M. and J.M.V.; Methodology: S.G., N.M. and K.K.; Investigation: N.M. and J.M.V.; Resources: S.G., K.K. and N.D.;

Writing - original draft preparation: S.G. N.M., K.K., M.A.M-R., and J.M.V.; Writing - reviewing & editing: S.G., N.M, K.K., N.D., M.A.M-R., and J.M.V.; Supervision: J.M.V. Funding acquisition: M.A.M-R. and J.M.V.

Data Availability

The datasets analysed in this study have been obtained from Gene Expression Omnibus (GEO; Rao, et al., 2014: GSE63525[10]; Bonev, et al., 2017: GSE96107[12]; Despang, et al., 2019: GSE125294[58]) and ArrayExpress (Hug, et al., 2017: E-MTAB-4918[9]; Díaz, et al., 2018: E-MTAB-5875[44]).

Code Availability

The CHESS source code, as well as code for generating synthetic Hi-C matrices and running tests on them is available on GitHub: (https://github.com/vaquerizaslab/CHESS).

Competing interests

The authors declare no competing interests.

Figures

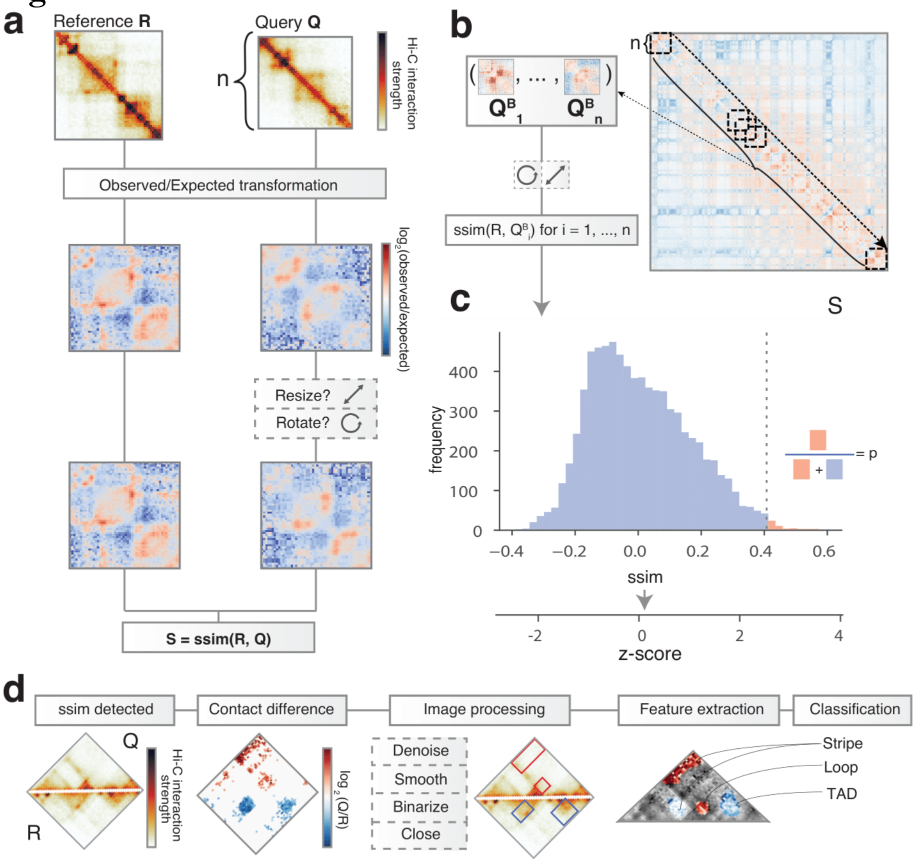


Figure 1. CHESS overview and examples. a, CHESS workflow, showing the observed/expected transformation, size/orientation adjustments, and structural similarity score S calculation on two example matrices R and Q. **b**, Example of a background model for z-score and p-value calculation. Specifically, similarity scores are calculated for each $n\ x\ n$ matrix $Q^{B}{}_{i}$ at every position i along the diagonal of the whole chromosome matrix. **c**, Distribution of similarity scores for $Q^{B}{}_{i}$ can be used to calculate p-values and z-scores for S. **d**, Feature extraction workflow, showing the contact difference map between R and Q matrices identified by CHESS, and the list of image filters applied. The specific differential gained and lost structures are highlighted in red and blue boxes, respectively. Finally, all the features are classified according to their structural pattern, such as TADs, loops and stripes.

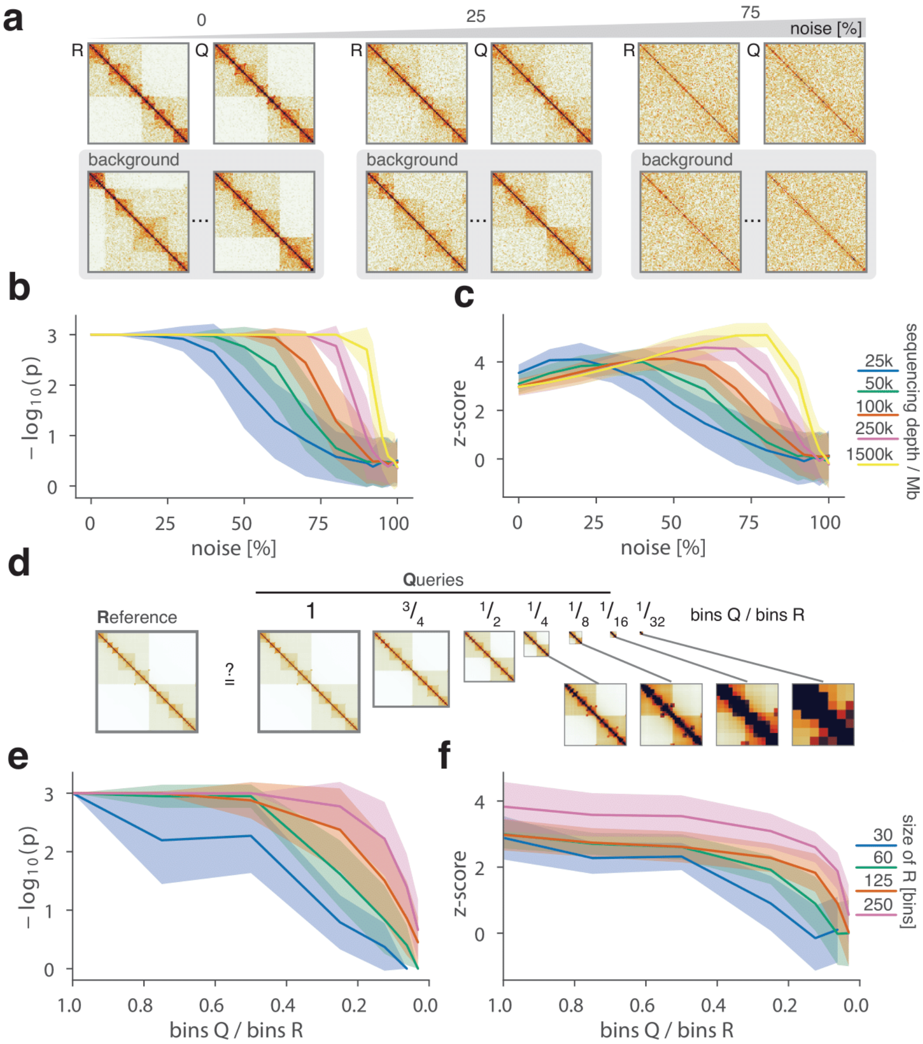


Figure 2. CHESS evaluation on synthetic Hi-C matrices. a-c, Tests for the tolerance of CHESS to increasing levels of simulated experimental noise. **a**, Schematic representation of the tests for the sensitivity of CHESS to noise. **b**, CHESS p-values on synthetic matrices with increasing levels of noise. **c**, CHESS z-scores on synthetic matrices with increasing levels of noise. **d-e**, Tests for the performance of CHESS when comparing matrices of different sizes. **d**, Schematic representation of different scaling factors used to generate a query matrix Q from a reference R. **e**, Dependence of CHESS p-values on the scaling factor of synthetic matrices Q and R. **f**, Dependence of CHESS z-scores on the scaling factor. **b, c** and **e, f**, Solid lines indicate the mean, shaded areas the standard deviation. Lowest possible p-value in all tests: 0.001.

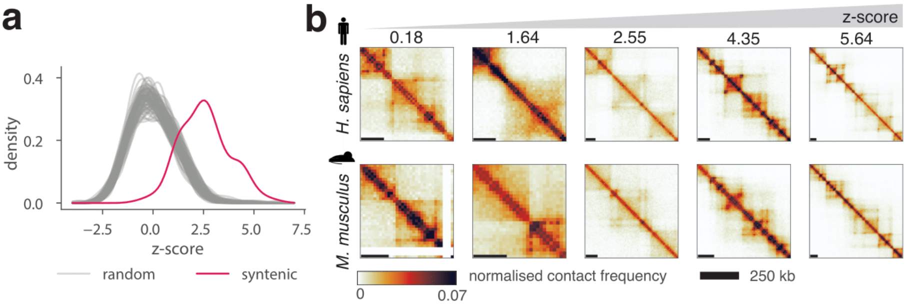


Figure 3. Global comparison of syntenic region similarity between *H. sapiens* and *M. musculus* using CHESS. a, Distributions of CHESS z-scores for 175 syntenic region pairs in *H. sapiens* and *M. musculus* (red) and 100 random permutations of region pairs (grey). Permutation test p-value < 0.01, comparing the mean scores of randomly permuted pairs to the mean score of real syntenic regions. **b**, Examples of syntenic regions with increasing CHESS z-scores from left to right.

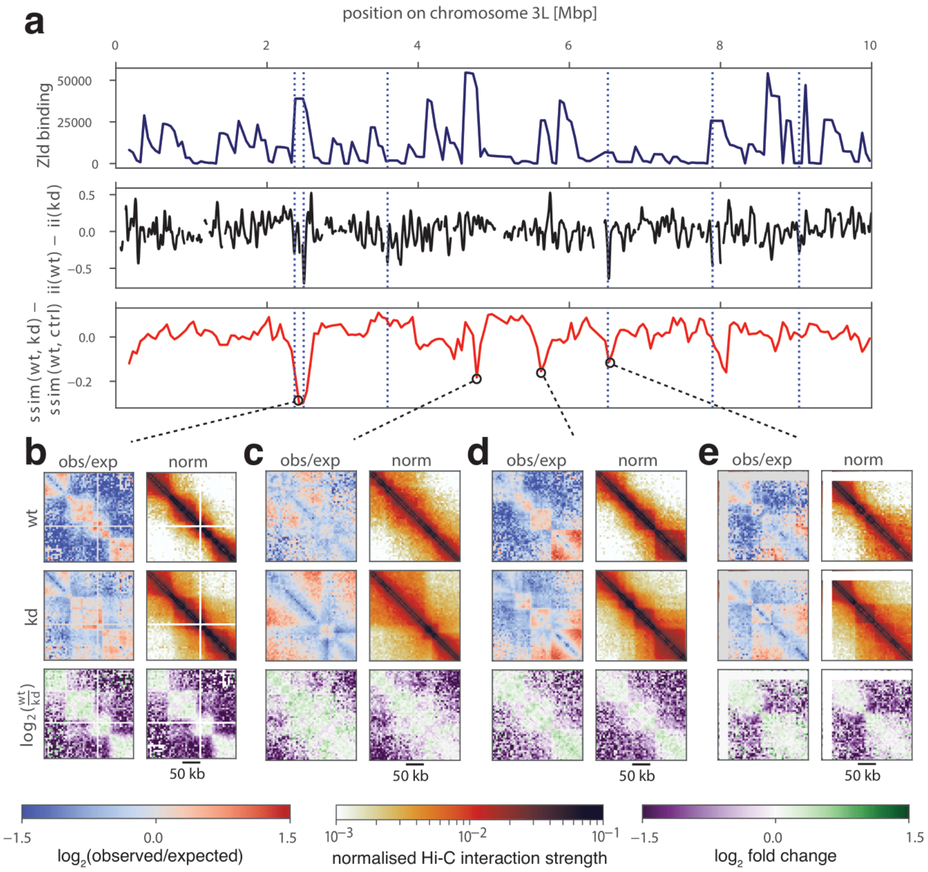


Figure 4. Identification of chromatin conformational changes in *D. melanogaster* embryos after Zelda (*zld*) knockdown. a, Zelda binding signal in wild type (wt) (top), insulation score difference between wt and *zld* knockdown (kd) (middle, smoothed), and difference between similarity scores calculated on wt to kd and wt to water injection control (ctrl) (bottom) for regions on a subset of chromosome 3L. Dotted blue lines indicate differential boundaries as identified by [9]. **b-e**, Examples of regions with the strongest conformational changes between wt and kd, showing observed/expected (obs/exp) and normalised Hi-C matrices, and $\log_2$-fold-change matrices for wt/kd. White lines on the Hi-C plots correspond to regions of the genome masked from the analysis due to low mappability.

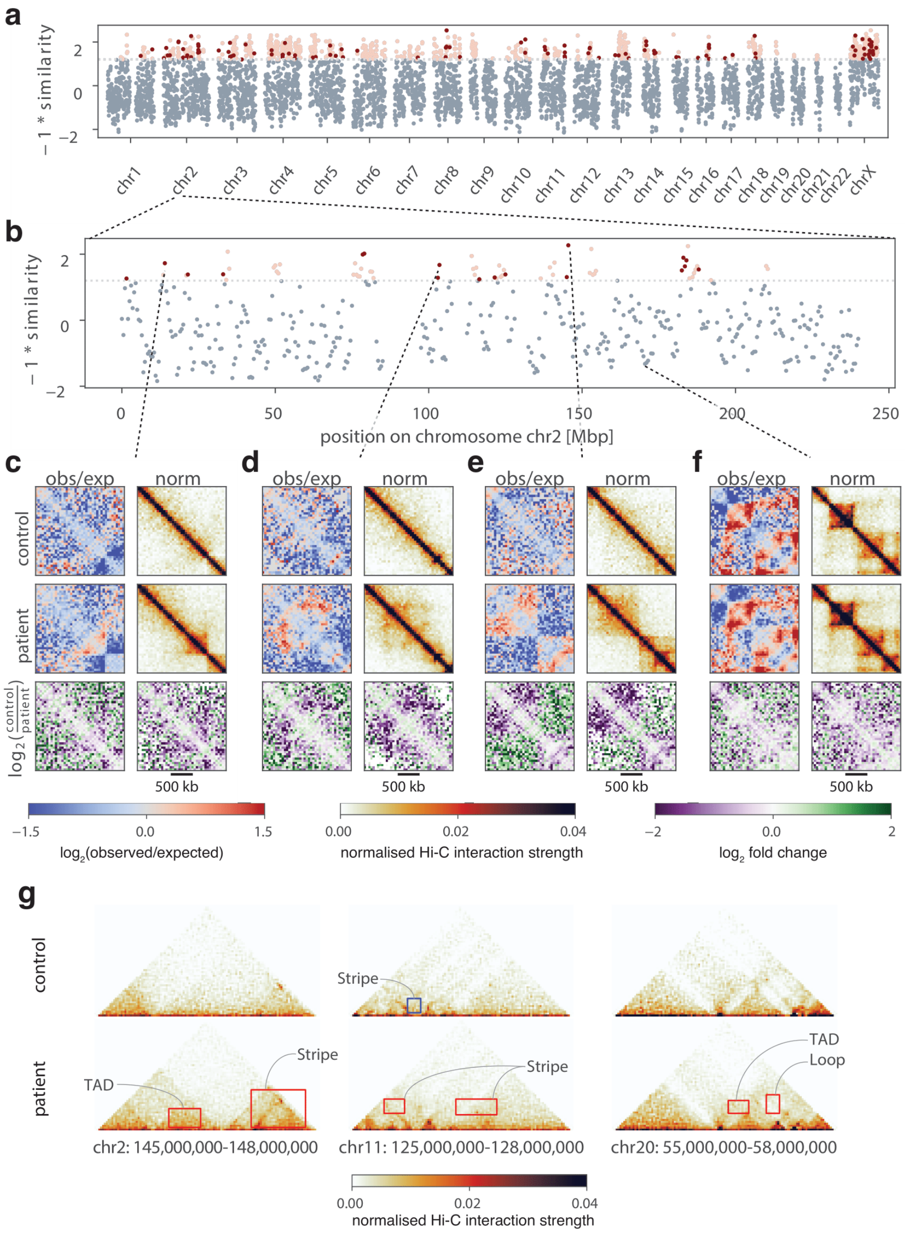


Figure 5. Identification of structural changes in a diffuse large B-cell lymphoma. a, b Similarity of Hi-C data generated from healthy B-cells (control) and a diffuse large B-cell lymphoma (patient), as assessed by CHESS for 2 Mb regions. Highly dissimilar regions (z-normalised similarity score <= -1.2) are coloured in red, where noisy regions (signal to noise ratio < 0.6) are in light red. **c-f**, Examples of regions with conformational changes (**c-e**) and

conservation (**f**) between healthy and diseased B-cells, showing observed/expected (obs/exp) and normalised Hi-C matrices, and $\log_2$-fold-change matrices for control/patient. **g**, Three examples of highly dissimilar regions identified by CHESS, with the gained and lost features highlighted in red and blue, respectively. The features are annotated according to their structural category.

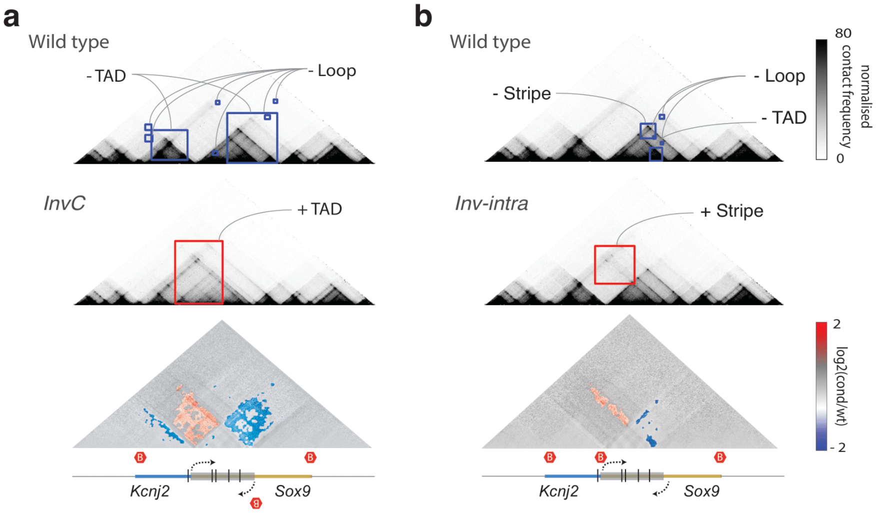
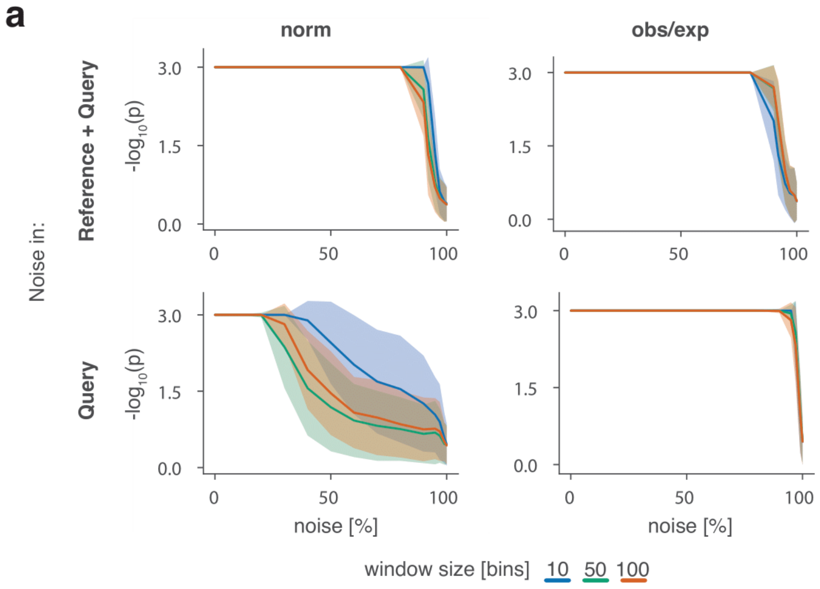
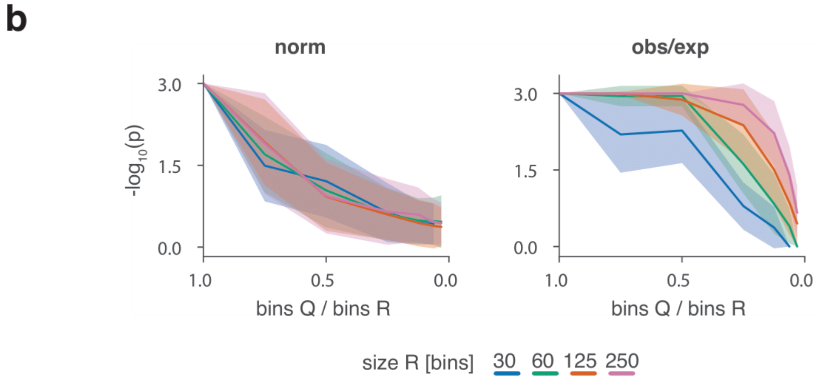


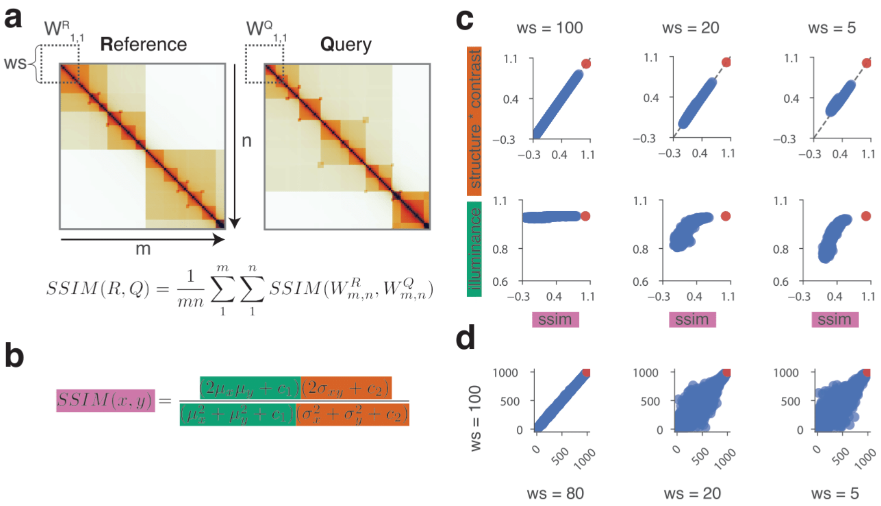
Figure 6. Feature extraction from Capture-C data from Despang et al.[58]. Lost and gained structures in the mutant are highlighted in blue and red squares, respectively. Red hexagons demarcate the positions of the TAD boundaries. **a,** Feature extraction of the wt against the *InvC* mutant, which presents an inversion in the *Sox9* sequence represented in a grey box. **b,** Feature extraction of the wt versus the *Inv-intra* mutant, which has the same sequence inverted, but now not including the border between the two TADs.

Supplementary Figures

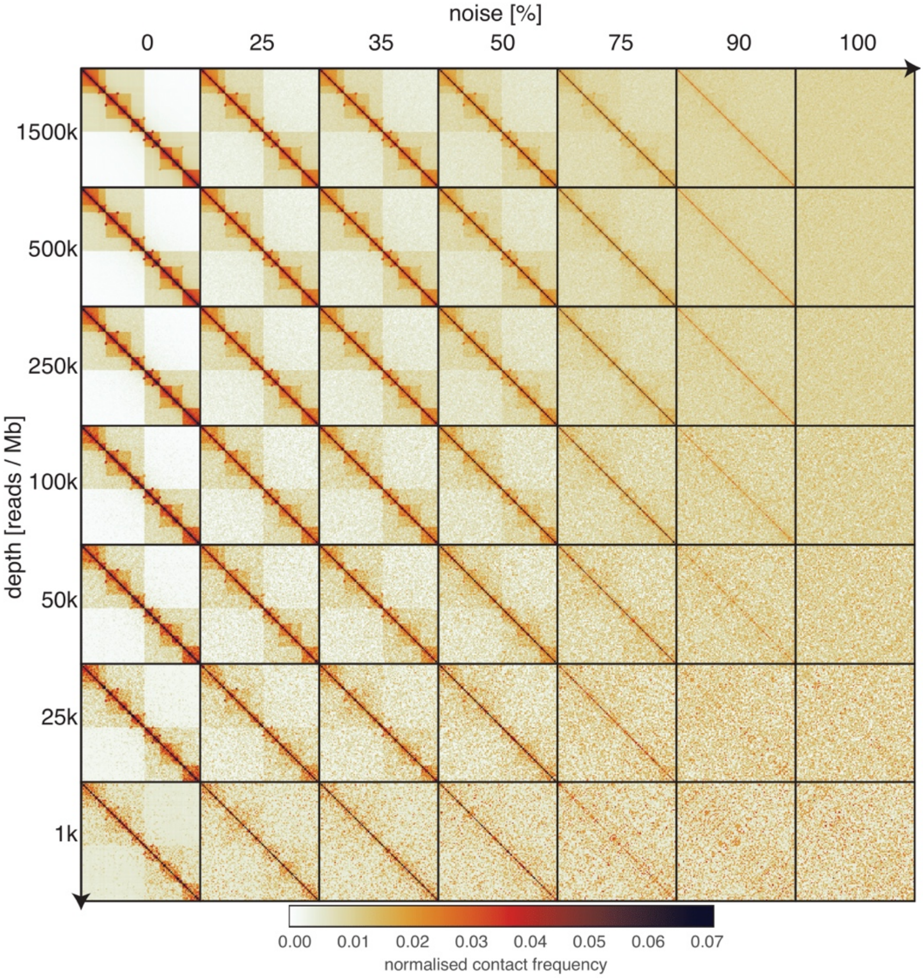




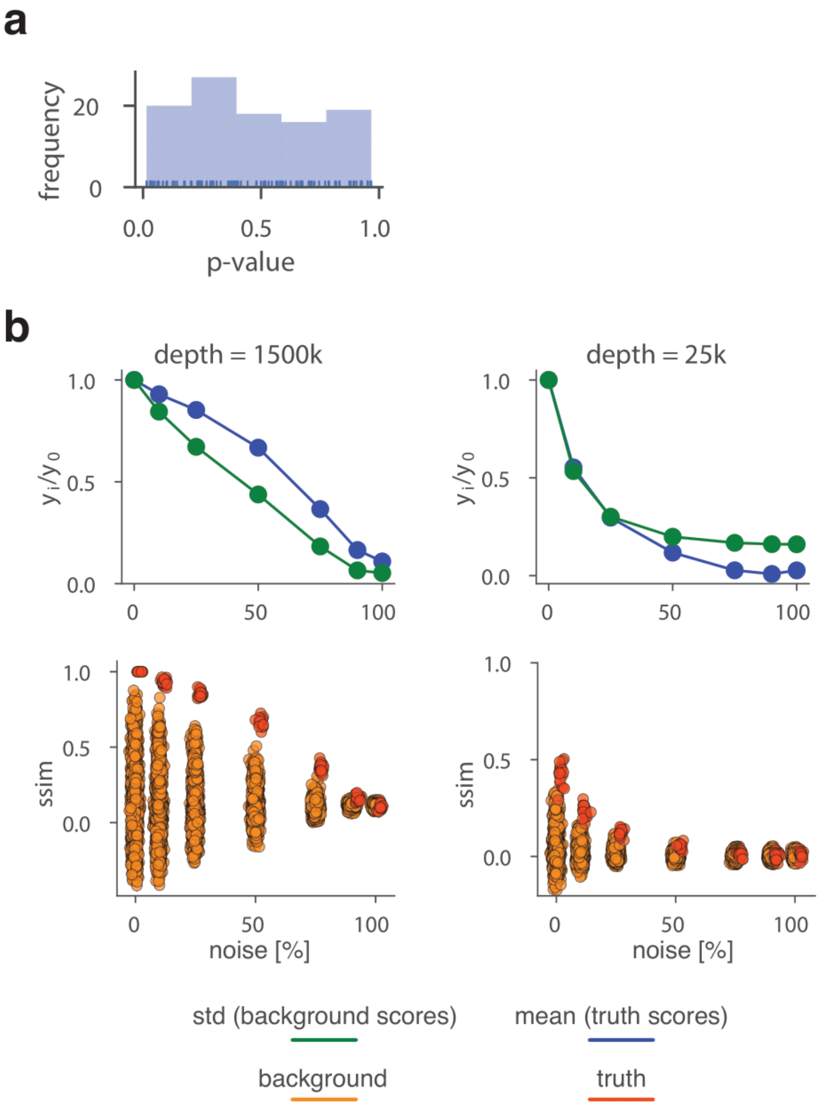
Supplementary Figure 1. Performance analysis of the CHESS algorithm. a, CHESS p-values in dependence of the relative noise level in synthetic matrices. Shown are the cases of equal amounts of noise in reference R and query Q (top) and different amounts of noise (bottom, noise only added to Q). Each case is examined for normalised and observed/expected (obs/exp) matrices, and different window sizes in the SSIM algorithm. **b**, CHESS p-values in dependence of the size factor between R and Q for normalised (left) and observed/expected (obs/exp) matrices (right). **a, b**, Solid lines indicate the mean, shaded areas the standard deviation.



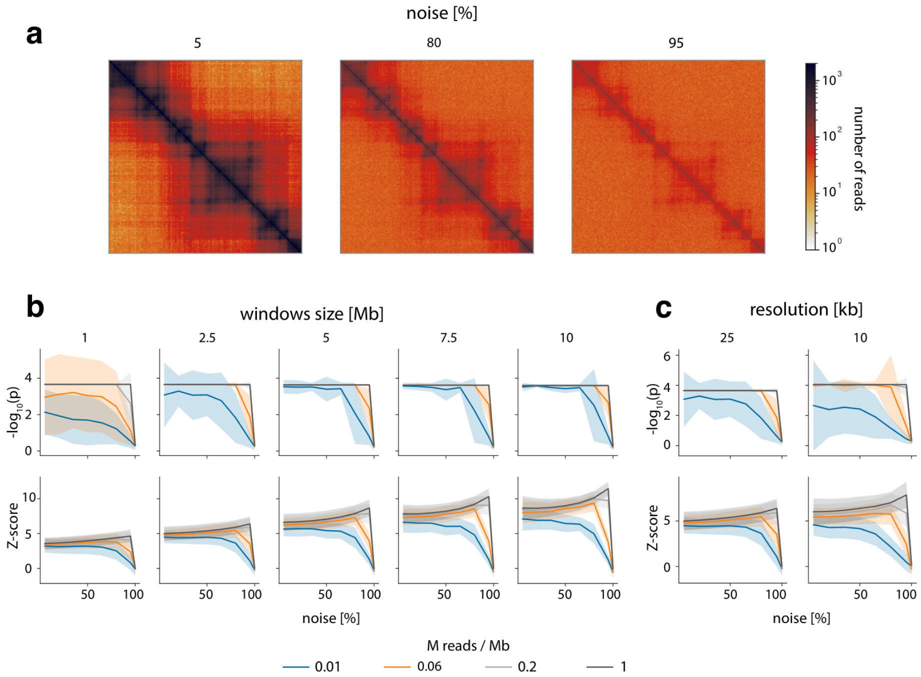
Supplementary Figure 2. Technical details of the SSIM algorithm applied to Hi-C matrices. a, Schematic overview of the structural similarity algorithm (SSIM). SSIM scores are calculated on all submatrices of R / Q at a given window size (WS). The final SSIM score is the mean of all SSIM submatrix scores. **b**, SSIM submatrix formula. Different components are coloured: illuminance (green), structure * contrast (red). x, y refer to submatrices (at the same positions) of the two full matrices for which the SSIM average is computed (see panel a). µ indicates the mean, σ the standard deviation, c1 and c2 are small constants that are introduced only for numerical reasons. **c** and **d**, SSIM comparisons of a matrix to itself (red dots) and 1,000 random matrices of the same size (blue dots). **c**, SSIM component values in dependence of SSIM score for different SSIM window sizes. **d**, Scatterplots of ranked SSIM scores at window size 100 vs. ranked scores at smaller window sizes.



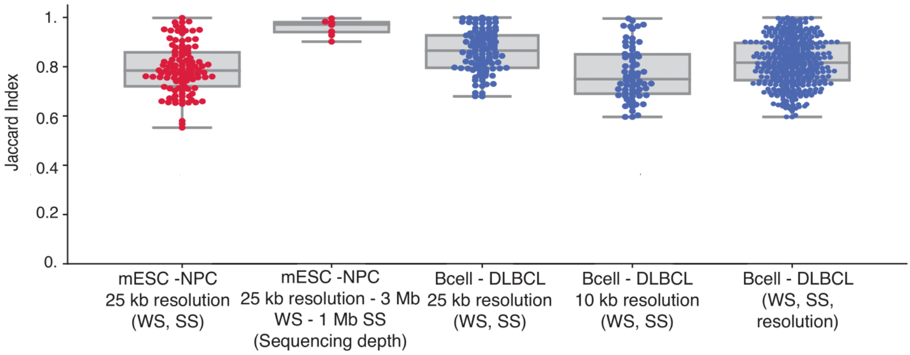
Supplementary Figure 3. Examples of synthetic matrices with varying levels of noise and simulated number of read pairs (depth). In all panels, depth was adjusted before adding noise.



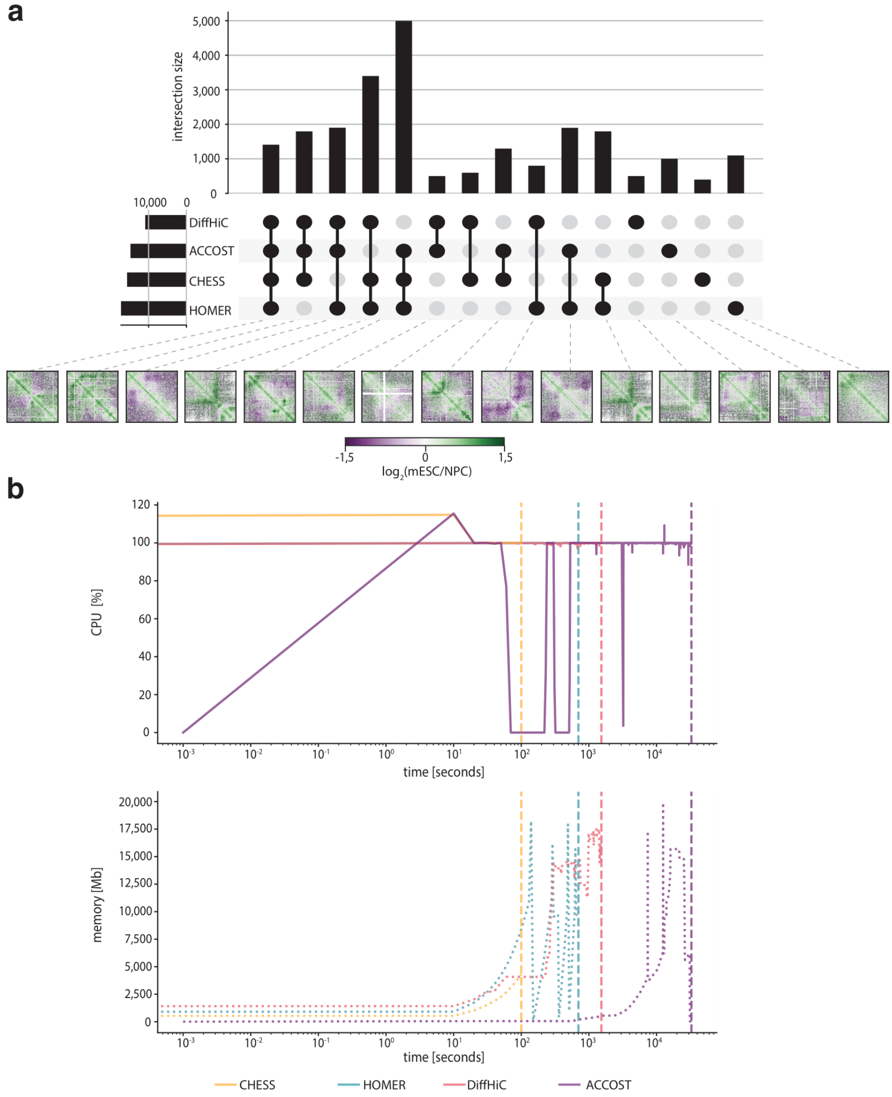
Supplementary Figure 4. Additional analysis of the CHESS algorithm. **a**, Uniform distribution of p-values for comparisons of matrices with 100 % noise added. **b**, Distribution of structural similarity scores (ssim) for background and truth comparisons at 25 k/Mb and 1.5 M/Mb simulated sequencing depth. Above each: Fractional change (value at x % noise/value at 0 % noise) of the standard deviation (std) of background scores and mean of truth scores.



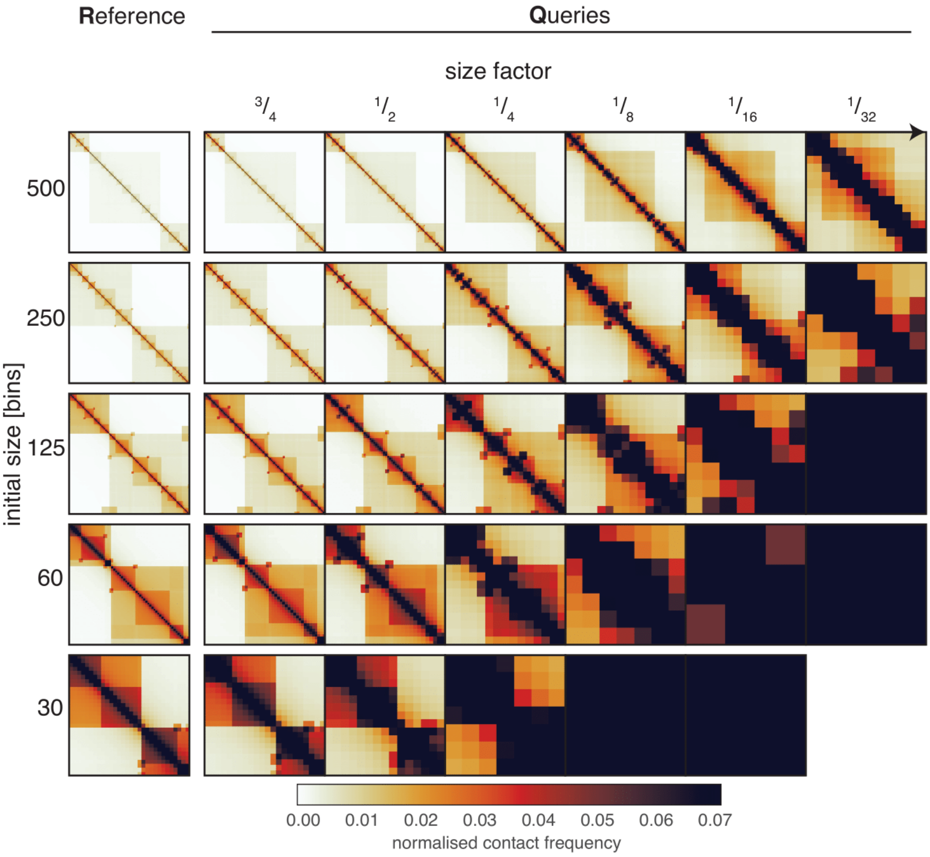
Supplementary Figure 5. CHESS is robust to changes in noise due to random ligations and sequencing depth in real Hi-C data. **a,** Examples of matrices used in this analysis including a 5, 80 and 95 % of added noise (random ligations between pairs of loci). We tested to what extent CHESS is able to identify two matrices as being identical, after noise and sequencing depth were adjusted independently in them. Matrices are based on chromosome 19 data from Bonev et al. 2017[12]. **a,** examples of the data with different amounts of noise. **b,** p-values and z-scores of CHESS runs with different window sizes, noise levels and simulated sequencing depths. Step size and matrix resolution were both 25 kb. Lines for 2 x 10^5 and 1 x 10^6 overlap for runs with window sizes > 1 Mb. **c,** As in panel a, but comparing CHESS runs with 2.5 Mb window size on matrices binned at 25 kb and 10 kb. **b, and c,** solid lines indicate the mean, shaded areas the standard deviation.



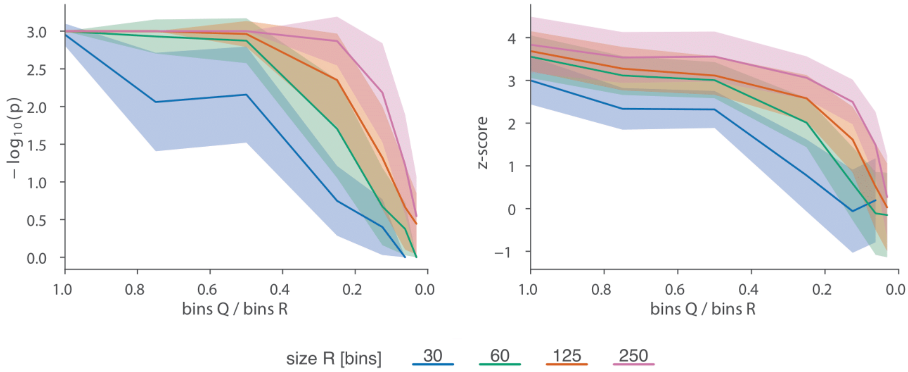
Supplementary Figure 6. Reproducibility of CHESS using different window (WS) and step sizes (SS), sequencing depths and resolutions. For this analysis were tested the WS (250 kb - 3 Mb), SS (25 kb - 1 Mb), sequencing depths (percentage of reads between 20 and 80) and resolutions (10 kb and 25 kb) (details in Methods). The first two boxplots with red dots represent the Jaccard indices (JI) between CHESS results in Bonev et al. 2017[12] using different WS, SS and sequencing depths. The boxplots with blue dots correspond to the Díaz et al.[44] dataset; in this case using different WS, SS, and then between different WS, SS and resolutions. mESC mouse embryonic stem cells, NPC neural progenitor cells. Boxplot elements: centre line: median, whiskers: 1.5x interquartile range, box limits: upper-lower quartile.



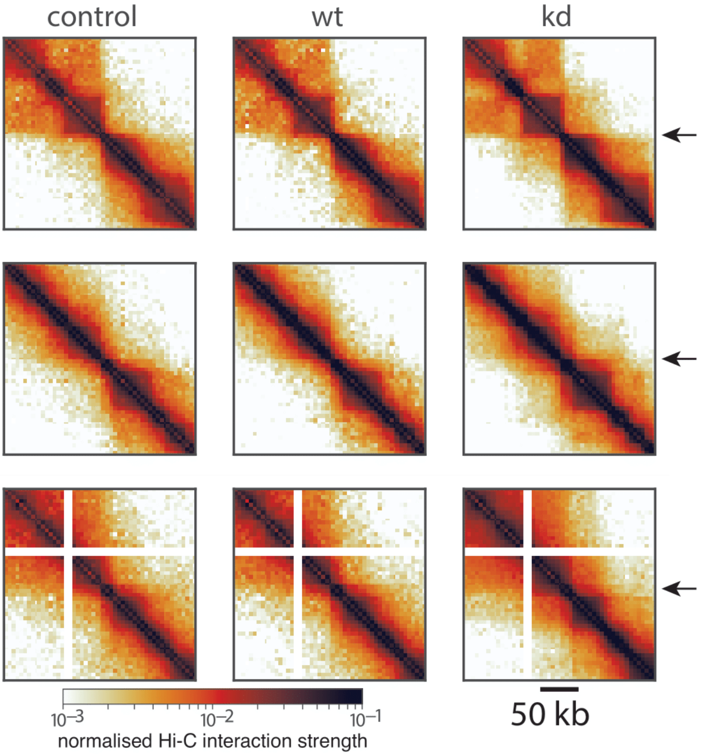
Supplementary Figure 7. CHESS benchmark against HOMER, diffHiC and ACCOST. a, Upset plot representing the intersection size between differential interactions of CHESS, HOMER, diffHiC and ACCOST**.** Below, an example is shown for each intersected group. **b**, Computational requirements of CHESS, HOMER, diffHiC and ACCOST**.** The first line plot shows the CPU usage, the second the memory consumption. The vertical dashed line represents the end of the run.



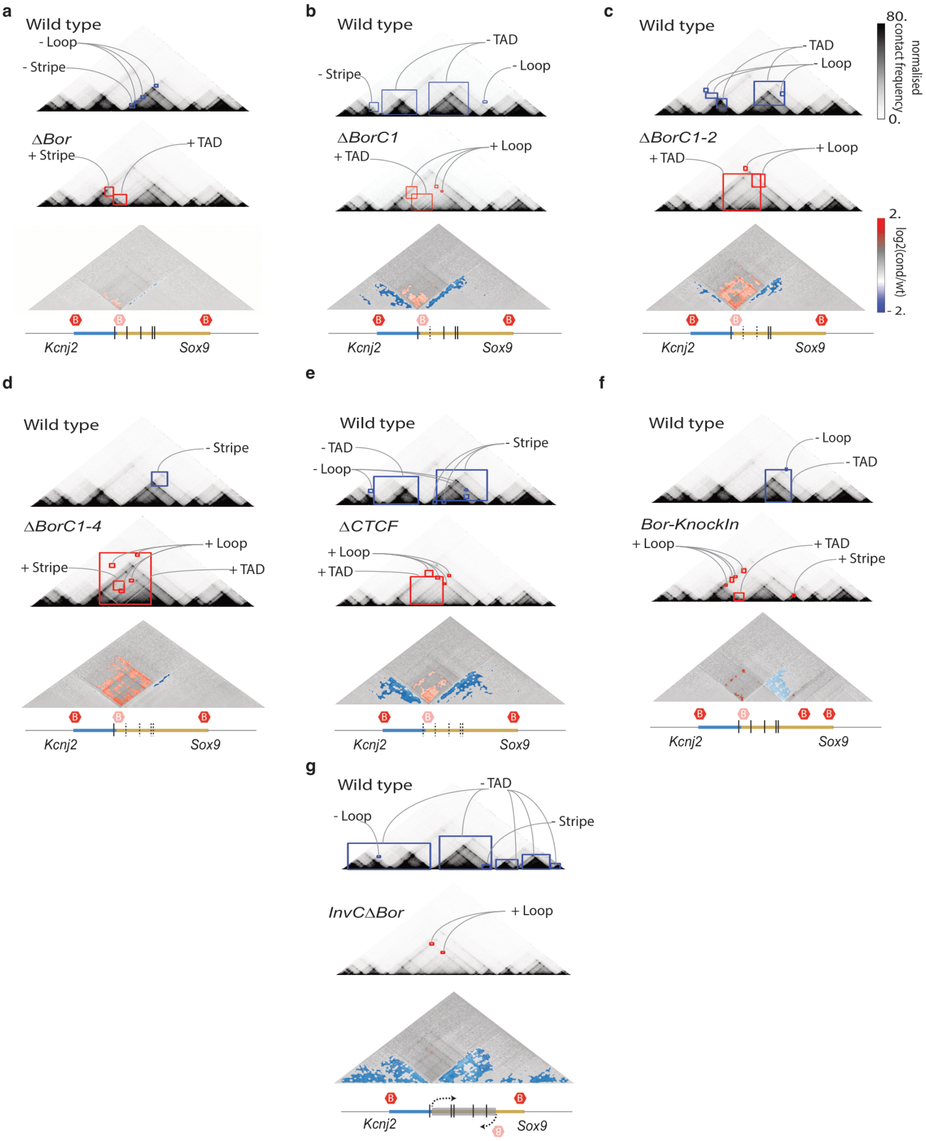
Supplementary Figure 8. Examples of synthetic matrices scaled to different sizes. Matrix bins were rounded up to the nearest integer after scaling. Only matrices with at least two bins were plotted.



Supplementary Figure 9. CHESS performance on differently sized simulated matrices with realistic noise and sequencing depth. Shown are CHESS p- and z-scores for comparisons of R with a read depth of 100 read pairs / 100 bins and a resized copy Q. Scaling factor is indicated on the X axis. A noise level of 25 % was added to both matrices independently. Sequencing depth was adjusted to 100 k/Mb. Solid lines indicate the mean, shaded areas the standard deviation. Colours correspond to the different sizes of R.

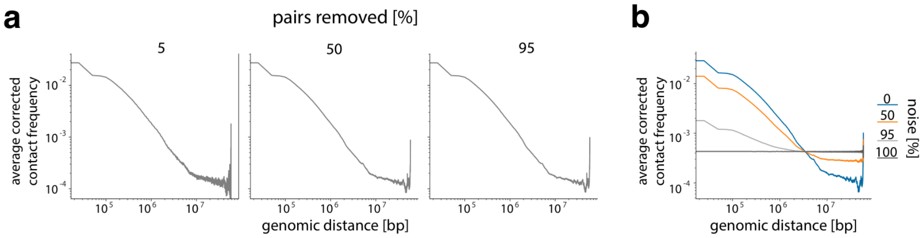


Supplementary Figure 10. Examples of minor changes in boundary strength after *zld* knockdown in *D. melanogaster*. Hi-C matrices are shown for a water injected control, wild_type (wt) cells and cells after *zld* knockdown (kd). Arrows indicate positions of differential boundaries between wt and kd as identified previously[9]. White lines correspond to regions of the genome masked from the analysis due to low mappability.

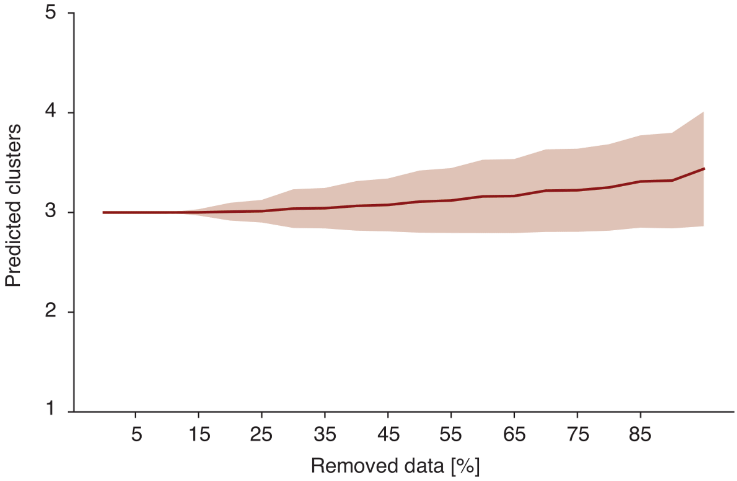


Supplementary Figure 11. Feature extraction from Capture-C data. Examples of differential feature extraction with CHESS between the wt and different mutants in the Despang et al.[58] dataset. The lost and gained structures in the mutants are highlighted in blue and red squares, respectively. Below each comparison, the genomic annotation is represented, highlighting the modification of each mutant. The vertical lines define the CTCF binding motifs, dashed when deleted. Red hexagons demarcate TAD boundaries. Feature extraction between wt and **a,** Δ*Bor*, in which the border was deleted.

b, Δ*BorC1*, in which the border and the first CTCF binding motif were deleted. **C,** Δ*BorC1-2*, in which the border and the two first CTCF binding motifs were deleted. **d,** Δ*BorC1-4*, in which the border and four CTCF binding motifs were deleted. **e,** Δ*CTCF,* in which the border and all the CTCF binding motifs were removed. **f,** *Bor-KnockIn*, in which the border was moved to a new location within the *Sox9* locus. **g,** *InvC*Δ*Bor*, in which the *Sox9* sequence was inverted and the border was removed.



Supplementary Figure 12. Effects of downsampling and noise on decay of average contact frequency with genomic distance. **a,** after downsampling to lower sequencing depths, **b,** with increased percentage of random ligations (= noise). Decay functions were computed on data from Bonev et al. 2017[12], on chromosome 19.



Supplementary Figure 13. Effect of downsampling on number of optimal clusters. The optimal number of clusters was computed after downsampling all extracted features from the Hi-C data generated from healthy B-cells (control) and a diffuse large B-cell lymphoma (patient. This process was repeated 1,000 times. The solid line indicates the mean, shaded area corresponds to the standard deviation.

Methods

The CHESS pipeline

The CHESS pipeline is illustrated in Fig. 1. CHESS takes two normalised, whole-genome Hi-C matrices as input. We recommend to use matrices 20×20 bins in size with no more than 10 % of all bins unmappable (without signal). In a first step, these are transformed to observed/expected (obs/exp) matrices[43] by dividing each matrix entry by the average of all entries at the same distance (see below). From those, the submatrices corresponding to the specified regions of interest are extracted, forming a comparison pair R (reference) and Q (query). Subsequently, R or Q are resized to the dimensions of the larger matrix using nearest neighbour interpolation (skimage.transform.resize in the scipy package[65]). If necessary, the matrices are rotated by 180 degrees (for example in case a syntenic region is annotated on the reverse strand). All bins marked as unmappable in either of the matrices are removed from both matrices. The resulting processed versions of R and Q are handed to the structural similarity function, yielding a raw similarity score. We use an implementation of the original structural similarity algorithm[33,34] available for the Python programming language in the scikit-image module[65].

In some applications, R may be compared to a pool of matrices P forming the background model. The process described above for the pair R, Q is repeated for each pair $R, Q^B \in P$. We use the similarity scores B obtained from these background comparisons to calculate a p-value and a z-score for the raw score $s = ssim(R, Q)$:

$$p = \frac{|\{x \in B \mid x \geq s\}|}{|B|},$$

$$z = \frac{s - \mu_B}{\sigma_B}$$

where μ_B denotes the mean, and σ_B the standard deviation of scores in B. We used two kinds of background models for this manuscript. (1) all submatrices of Q's size located on the same chromosome as Q for the comparison of chromatin structures between syntenic regions, and (2) a pool of synthetic matrices built with the same parameters as Q but randomly generated features for the tests of CHESS on synthetic Hi-C data. CHESS p-values are not automatically corrected for multiple testing, as this is not necessary for all use cases. If CHESS is used to identify significantly similar or different regions across the genome with a fixed acceptance threshold, the CHESS p-values need to be corrected for multiple testing.

Next, CHESS extracts individual features that are different between two genomic regions (Fig. 1d). First, gained and lost contacts in the R matrix are computed and separated as increased/decreased interactions with respect to Q. Then, a set of image filters, with the parameters automatically adjusted according to the matrix size or user-defined, are applied to these two matrices that are from now on considered as images:

1. Denoise the image using a bilateral filter[66]: this is an edge-preserving filter that averages pixels based on their spatial closeness and their radiometric similarity, by default they are computed using a window size of 3. The Gaussian function of the Euclidean distance between two pixels and its standard deviation is used to obtain the spatial closeness. The Euclidean

distance between two colour values is used for the radiometric similarity, CHESS by default uses the mean value of the matrix. It has to be noted that higher values of spatial closeness and radiometric similarity will average the bins with larger differences.

2. Smooth the image using a median filter: this scans the image using a square shaped array with an area computed automatically depending on the picture size. This array scans the image using a windows size, and computes the median of the pixels, smoothing the signal. Higher values will smooth larger structures, while smaller values will consider more subtle signals.
3. Image binarization using Otsu's method[67]: it returns a threshold value that separates the pixels in two classes. This algorithm searches for the threshold that minimises the intra-class variance. This threshold by default is calculated using whole matrix values to be more refined.
4. Morphological closing of the image: this filter is used to remove small dark spots and connect small bright cracks. This helps to remove the remaining noise and to enclose the individual structures. By default CHESS uses a square of 8 bins, higher values will enclose larger structures while lower will consider smaller or more punctuated signals.

With the four filters, CHESS extracts individual structures, which can be used to get the main structural clusters according to their pattern of interactions. First, the 2D-cross correlation between all the individual features is computed. Finally, the K-means clustering algorithm is applied to obtain the main structural clusters. The optimal number of clusters is computed according to the Elbow method by fitting the model within a range of 1 to 15 clusters, which may vary depending on the number of identified differential features. The robustness of the clustering was assessed by downsampling the identified structural features from the Hi-C data generated from healthy B-cells (control) and a diffuse large B-cell lymphoma (patient) and computing the optimal number of clusters. This process was repeated 1,000 times. The clustering step proved to be highly robust to data sparsity (Supplementary Fig. 13).

Calculation of observed/expected matrices

We calculated the observe /expected form $M^{obs/exp}$ of a balanced matrix M by first computing the expected matrix M^{exp} by determining the average value of each diagonal in M:

$$M^{exp}{}_{i,j} = \frac{\sum_{n=1}^{N-D} M_{D+n,n}}{N-D} \text{ with } D = i - j, i \geq j$$

As M is symmetric around $M_{i=j}$, we computed this only for $i \geq j$ and then set $M^{exp}{}_{i,j} = M^{exp}{}_{j,i}$.

We then calculated $M^{obs/exp}$:

$$M^{obs/exp}{}_{i,j} = \frac{M_{i,j}}{M^{exp}{}_{i,j}}.$$

As for the matrix balancing, the observed/expected calculation was performed on a per-chromosome basis for real Hi-C data.

Generation of synthetic Hi-C matrices

To generate a synthetic matrix, we performed the following steps: first, we produced an empty matrix M of dimensions n^2. We then filled M with simulated pairs of reads, modelling the power-law decay of signal away from the main diagonal[43,68] by:

$$x_i = (10^{-4} + \frac{1 - 10^{-4}}{1000} i)^{-0.85}, i = 0,1\ldots, n$$

where x_i denotes the read counts at the ith diagonal, counted as moving away from the main diagonal at $i = 0$. At this point, the number of reads is uniform in each diagonal and the mean number of reads per bin is inversely proportional to the matrix size. We then added structural features resembling TADs and loops to M.

First, three layers of TADs were added. TAD size was randomly determined by drawing a size s from a truncated normal distribution, while TAD intensity was modelled by adding a constant read count to a square of area s^2. For each consecutive layer the TAD size decreased while the TAD intensity increased (Supplementary Table 1). To start with, we placed a first TAD at a randomly chosen position on the main diagonal at least $0.1n$ away from each end of the diagonal. We then filled the main diagonal to both sides of this initial TAD with adjacent TADs. In cases where a space smaller than the lower bound of the truncated normal occurred at the ends of a diagonal, we covered it by adding a small TAD that can be thought of as being part of a bigger TAD reaching into the field of view from an adjacent genomic region. In the second and third round, smaller TADs were placed inside the TADs generated in the previous round.

Second, corner loops were added to TADs with a chance of $1/3$. Loops were modelled as squares with an additional, randomly selected intensity of either 90 %, 140 % or 220 % of the intensity and side lengths of either 5 %, 7 %, or 10 % of the side length of the corresponding TAD.

Simulation of different sequencing depths and experimental noise

Hi-C matrix quality and resolution are primarily affected by two properties: (1) random ligations, which are distinct from proximity ligations; and (2) sequencing depth. These properties have distinct effects: random ligations are an indicator of poor library quality and introduces "noise" into the Hi-C matrix, while sequencing depth determines the achievable matrix resolution. However, they are not entirely independent. Increased sequencing depth can mitigate the effects of random ligations by enriching contacts in regions with "true" proximity signal.

In all our tests of CHESS, lower sequencing depths were simulated by subsampling, i.e. the random removal of "ligation fragment" pairs in a high-resolution matrix. Random ligations, on the other hand, were simulated by the random replacement of pairs in a matrix. In particular, to model different sequencing depths, we lowered the density of read pairs in M by removing random read pairs until we reached the desired number of read pairs d. Subsequently, we simulated an experimental noise level ε by reassigning $r = \varepsilon \times d$ read pairs to randomly selected pairs of loci.

As stated above, our noise models random ligations in the Hi-C experiment that can occur after the genomic material has been digested. These random ligations are intramolecular ligations (not within a crosslinked pair). The probability of a random ligation of two fragments is therefore not related to the linear genomic distance between them. In consequence, the distance decay graph is expected to approach a flat line as the fraction of random ligations in the Hi-C library increases. Our noise procedure moves a fraction of the reads in each bin to randomly chosen bins, where each bin in the map has the same chance of receiving a read. This procedure results in the expected behaviour of the distance decay graph, as shown in Supplementary Fig. 12.

The downsampling procedure on the other hand removes randomly chosen reads, without adding them anywhere. As the fraction of reads removed is on average the same at all genomic distances, the distance decay graph does not change significantly due to sampling. Slight deviations occur only close to the maximum distance, where the numbers of reads and bins are small enough to allow for random fluctuations of the mean after sampling. We show the largely unchanged distance decay in Supplementary Fig. 12.

We simulated the size difference of regions with similar structural features by first generating a reference matrix of a certain size n_r and saving the relative positions and intensities of structural features in it. We then generated query matrices of smaller sizes n_q and placed the structural features at the same relative positions (rounded to the next full bin) with the same intensities. To ensure equal sequencing depth relative to the matrix size between the scaled matrices, we subsequently adjusted the depth of the scaled matrices to a scaled depth d_q in relation to the depth of the reference matrix d_r:

$$d_q = d_r \frac{n_q}{n_r}.$$

The structural similarity algorithm in chromatin contact map comparisons

The SSIM score for whole matrices is calculated from the average of multiple “sub-scores” obtained on smaller subsets of a matrix, the size of which can be controlled with a dedicated window size parameter (Supplementary Fig. 2). Each sub-score consists of three components: corrections for illuminance (differences in brightness), corrections for contrast, and the correlation coefficient between the two matrices (Supplementary Fig. 2)[33,34].

We quantified the contribution of each component to the final score in comparisons of a random synthetic Hi-C reference matrix to an identical copy of itself and to a pool of 1,000 randomly generated matrices of the same size. We assessed the dependence of the final score on each component using multiple window sizes (Supplementary Fig. 2). For sufficiently large window sizes, SSIM sub-scores are perfectly reflected by the combination of the contrast and correlation components. Only for very small window sizes does the illuminance play a minor role. While different window sizes affect the scores and relative rank of random matrices (Supplementary Fig. 2), the comparison of the reference matrix to its identical copy yields a perfect score in all comparisons, independently of the window size.

Tests on synthetic Hi-C matrices

We tested CHESS on synthetic matrices in two main test scenarios: noise/sequencing depth tests and size/size difference tests. The test setup was similar in both scenarios. For each test run, we first defined the test conditions by setting the following parameters: the size of reference matrices, query matrices, the reference noise level, query noise level and the sequencing depth (always the same for reference and query), the type of the input matrices (normalised or observed/expected), and the window size parameter of the structural similarity function. Using these parameters, we then generated a reference set of 100 synthetic Hi-C matrices. Corresponding to these references we then produced 100 query matrices, differing in a certain parameter (noise level or size) by a certain factor, but with the same structural features in the same positions (see 'Generation of synthetic Hi-C matrices'). We then generated a decoy pool of 1,000 synthetic matrices with all parameters equal to the query matrices, but with randomly generated features. Each reference matrix was compared to its corresponding query using the structural similarity algorithm with the specified parameters, and also to each matrix in the decoy pool, which we used as a simulation of the genomic background. The p and z-scores were then calculated as described in 'The CHESS pipeline'. The best possible p-value in this test ($p \leq \frac{1}{1000}$) was achieved when the comparison score of R vs Q was greater than all scores from comparisons to the 1,000 random matrices.

Processing of Hi-C matrices

We obtained Hi-C sequencing reads for human IMR90[10], mouse CH12.LX[10] (GSE63525), mouse ESC and NPC[12], and fly embryos at nuclear cycle 14, in wild-type (wt), Zelda knockdown cells (*zld* kd) and injected water control (wc)[9]: ArrayExpress: E-MTAB-4918), as well as B-cell and diffuse large B-cell lymphoma[44].The B-cell and DLBCL data was processed as described previously [44]. Fly data was processed as described in previously[9].

All human and the and CH12.LX mouse paired-end FASTQ files were mapped independently to the reference genome (hg19 and mm10, respectively) in an iterative fashion using Bowtie 2.2.4 with the

“--very-sensitive” preset. Briefly, unmapped reads were truncated by 15 bp and realigned iteratively, until a valid alignment could be found or the truncated read was shorter than 25 bp. Only uniquely mapping reads with a mapping quality (MAPQ) >= 30 were retained for downstream analysis. mESC and NPC FASTQ files were mapped using BWA mem version 0.7.17-r1188 in a non-iterative fashion with default parameters.

Restriction fragments were computationally predicted using the Biopython[69] “Restriction” module. Reads are assigned to fragments, and fragments pairs are formed according to read pairs. Pairs are then filtered for self-ligated fragments, PCR duplicates (both read pairs mapping within 1 bp of each other), read pairs mapping further that 5 kb from the nearest restriction sites, and ligation products indicating uninformative ligation products[70]. The Hi-C matrix is built by binning each genome at a given resolution of 10 kb and 25 kb and counting valid fragment pairs falling into each respective pair of bins. Finally, bins that have less than 25 % (human) or 10 % (mouse) of the median number of fragments per bin are masked, and the matrix is normalised using KR matrix balancing[42] on each chromosome independently.

Tests on real Hi-C matrices

The robustness of CHESS was also tested using real Hi-C data from [12] and from [44] experiments. We show the results in the Supplementary Fig 5. and Supplementary Fig. 6.

Firstly, the data from Bonev et al. 2017[12], binned at 25 kb, was used to repeat the analysis performed on synthetic matrices (see ‘Tests on synthetic Hi-C matrices’). First, different levels of noise (5 %, 20 %, 35 %, 50 %, 65 %, 80 %, 95 %) were added to the raw Hi-C matrix of chromosome 19. This was done twice independently to obtain versions A and B of the matrix, in order to model matrices coming from independent experiments. Each of these was then downsampled (to 1 %, 5 %, 50 %, 95 % of the original number of reads), corrected and transformed to observed/expected values. For each combination noise and sampling depth, CHESS was run in default mode (using comparisons to the rest of the chromosome as background model) to compare the same regions in the A and B matrices. These region pairs were obtained from a sliding window of sizes 1 Mb, 2.5 Mb, 5 Mb, 7.5 Mb and 10 Mb, with a step size of 25 kb. The resulting mean p-values and z-scores, as well as their variances were plotted as shown in Supplementary Fig. 5. The best possible p-value, or perfect performance, was achieved when no other region got an equal or higher similarity score than the region with identical positions in A and B. We found the dependence on window size to be the main parameter governing the robustness of CHESS; smaller bin size, i.e. higher resolution of the maps, did not qualitatively change the results (Supplementary Fig. 5).

Secondly, Hi-C data from mouse embryonic stem cells (mESC) and neural progenitor cells (NPC) at a resolution of 25 kb from the same dataset[12] was used for the reproducibility analysis of CHESS. CHESS was run using different combinations of window span (250 kb, 500 kb, 1 Mb, 2 Mb and 3 Mb) and step sizes (25 kb, 250 kb, 500 kb and 1 Mb). The Jaccard Index (JI) was calculated to obtain the overlap between the identified genomic regions. To check the reproducibility of CHESS using different sequencing depths (percentage of reads: 80, 60, 40 and 20), we applied CHESS using 25 kb resolution, 3 Mb windows span and 500 kb step size. The data from Díaz et al.[44], was used to check how consistent was CHESS when using different values of windows span (250 kb, 500 kb, 1 Mb, 2 Mb and 3 Mb), step sizes (25 kb, 250 kb, 500 kb and 1 Mb) and resolutions (25 kb and 10 kb). One point in the plot (Supplementary Fig. 6) corresponds to the Jaccard Index computed for a pair of CHESS runs with different combinations of parameter values. All possible pairs were compared.

Benchmark analysis

Hi-C interaction matrices at 5 kb resolution for the chromosome 19 from mESC and NPC from Bonev et al. 2017[12] were scanned for differences using different tools. HOMER[32], diffHiC[30] and ACCOST[31] were run using default parameters. CHESS was run using a windows span of 1 Mb and a step size of 500 kb. All tools were run using a single CPU computational machine of the following characteristics: Intel Xeon W @ 3GHZ with 128 Gb of RAM.

In particular, CHESS was run using a windows span of 1 Mb and a step size of 500 kb. CHESS ran ~7 times faster than HOMER, ~15 times faster than diffHiC and ~320 times faster than ACCOST and had a ~4 times lower peak memory consumption than the two other tools (Supplementary Fig. 7).

To assess the similarities and differences between the three methods, we selected for each method, all bins that were involved in a significant difference between mESC and NPC contact maps. Then, the selected bins were intersected to identify those bins common to the three methods, any common bin by at least two of the three methods and, finally, any bin identified only by one of the methods (Supplementary Fig. 7). CHESS and HOMER identified about 9,000 bins with differential interactions between mESC and NPC while diffHiC identified about 4,000 and ACCOST about 6,000. Of the total identified differences, ~12 % were identified by CHESS, HOMER, diffHiC and ACCOST. About 50 % of differences were identified by CHESS and HOMER alone.

Comparison of syntenic regions between *H. sapiens* and *M. musculus*

We retrieved the annotations for syntenic blocks between hg19 (selected as reference) and mm10 with a resolution of 300 kb using SynBuilder[46]. We used CHESS to compare syntenic region pairs in Hi-C matrices at 25 kb resolution for the human fibroblasts and the mouse lymphoblasts. As control, we also did comparisons between region pairs with shuffled syntenic region IDs. This was repeated 100 times. To reduce the runtime of our method, we used a randomly chosen subset of 175 syntenic regions. For the same reason, we restricted the background calculation to the query chromosome the syntenic region was located on.

Detecting structural changes between wild type and zld knockdown in *Drosophila melanogaster*

We obtained the locations of differential boundaries in *Drosophila melanogaster* and Hi-C data at 5 kb resolution for the wild type (wt), *zld* knockdown (kd) and water control (wc) for nuclear cycle 14 from Hug et al.[9]. From these Hi-C data we computed insulation scores as described in the same publication. We smoothed the resulting index track with a Savitzky-Golay (implemented in Scipy[71]) filter (window = 29, polyorder = 2, derivative = 0). We obtained data for Zld binding ChIP-seq experiments from Blythe et al.[72].

We partitioned the *D. melanogaster* genome into 250 kb / 125 kb regions with a step size of 50 kb / 25 kb. We ran CHESS on the observed/expected transformed Hi-C matrices corresponding to these regions, always comparing a region in wt to the same region in wc and kd. Inside the same windows we summed the $log_{10}(q - values)$ for all Zld-peaks with a $log_{10}(q - value) > 10$ to generate the Zld binding tracks.

Using the CHESS comparisons between wt–wc and wt–kd, we defined regions with structural changes as regions located at local minima of the track with values smaller or equal to -0.1.

Differential boundaries were defined as boundaries present in wt c14 cells (calls available at https://github.com/vaquerizaslab/Hug-et-al-Cell-2017-Supp-Site) at which the difference in the $\log_2$(insulation index) between the wt c14 and the *zld* knockdown was greater or equal to 0.3.

We defined differential boundaries that were closer than 125 kb / 62.5 kb to the centre of a structurally changing region as captured by CHESS.

Detecting structural changes between healthy B-cells and a diffuse large B-cell lymphoma

We obtained Hi-C data from Díaz et al.[44], and processed it as described in the original publication. We partitioned the human hg19 genome into 2 Mb regions with a step size of 500 kb. We used CHESS to compare the corresponding regions in the observed/expected transformed Hi-C data from the healthy B-cells (control) and a diffuse large B-cell lymphoma (patient). To distinguish between actual structural differences and such attributable to noise we calculated a signal to noise ratio r for the differential signal of each matrix pair:

$$r = \frac{\mu_M}{{\sigma^2}_M}, M = M_{control} - M_{patient}.$$

This was done for a sliding window of 7×7 pixels on the matrix. The total signal to noise ratio was taken as the mean of all windows. Regions with a z-normalised similarity score ≤ -1.2 and a signal to noise ratio $r \geq 0.6$ were labelled and accepted as changing.

Feature extraction from Capture-C data

CHESS feature extraction was applied to Capture-C experiments from Despang et al.[58] (GSE125294). Interaction matrices normalised by the Knight-Ruiz (KR) balancing method were downloaded. All the mutants were compared to the wild type. All the differential features were extracted and clustered according to their interaction pattern (see '*The CHESS pipeline*'). Three structural clusters were obtained: TAD, loop and stripe. Some examples are shown in Fig. 6 and Supplementary Fig.11.

Availability of data and materials

The CHESS source code, as well as code for generating synthetic Hi-C matrices and running tests on them is available on GitHub: (https://github.com/vaquerizaslab/CHESS). This includes tools for the generation of synthetic Hi-C matrices as outlined in this paper.

CHESS was implemented in Python 3.6.1.

Internally CHESS uses the following packages: FAN-C[73], Cython[74], SciPy[71], Scikit-image[65], NumPy[75,76], Pandas[77], Pathos[78], Pybedtools[79], Kneed[80], intervaltree (https://github.com/chaimleib/intervaltree) and tqdm (https://github.com/tqdm/tqdm).

References

65. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

66. Tomasi, C. & Manduchi, R. Bilateral filtering for gray and color images. in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* 839–846 (1998). doi:10.1109/ICCV.1998.710815.

67. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).

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

69. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

70. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).

71. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

72. Blythe, S. A. & Wieschaus, E. F. Zygotic Genome Activation Triggers the DNA Replication Checkpoint at the Midblastula Transition. *Cell* **160**, 1169–1181 (2015).

73. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data. *bioRxiv* 2020.02.03.932517 (2020) doi:10.1101/2020.02.03.932517.

74. Behnel, S. *et al.* Cython: The Best of Both Worlds. (2011).

75. Oliphant, T. E. *A guite to NumPy.* vol. 1 (Trelgol Publishing USA, 2006).

76. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13**, 22–30 (2011).

77. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (2010). doi:10.25080/Majora-92bf1922-00a.

78. McKerns, M. M., Strand, L., Sullivan, T., Fang, A. & Aivazis, M. A. G. Building a Framework for Predictive Science. *arXiv:1202.1056 [cs]* (2012)

Supplementary Tables

Supplementary Table 1. Parameters of truncated normal distributions of TAD sizes used for the generation of random features in synthetic Hi-C matrices. Mean, standard deviation (μ, σ), upper and lower bounds indicate number of pixels, intensity is in number of pairs.

round	μ	σ	upper bound	lower bound	intensity
1	150	100	300	100	75
2	20	15	50	10	100
3	10	5	20	5	150

Discussion

Identification of chromatin loops from Hi-C interaction matrices by CTCF-CTCF topology classification

During the last years an outstanding number of studies proved the high relevance of the 3D organization of chromatin and its tight relationship with functional and regulatory processes. Specifically, chromatin loops are able to bring close into space regulatory units, forming transcription factories or hubs. The vast majority of these loops are flanked by two CTCFs in a convergent oriented manner. However, not all the chromatin loops present a canonical CTCF binding and the same functional signature. What is more, not all convergent CTCFs are observed to form chromatin loops and same-oriented CTCFs can also be part of 3D structural rearrangements. Until now, the studies to unveil the role of features, which may play a role in chromatin organization, have been based on their average interaction frequency. The main limitation of pilling up pairs of genomic coordinates, is the possibility to lose small clusters with a similar, and non-average, structure and function.

In Chapter I, we present an algorithm to deconvolve the structural signal of Hi-C experiments in the context of colocalizing DNA-binding proteins, called *Meta-Waffle*. Specifically, we deconvolved the genomic average CTCF-CTCF interaction pattern within 45 kb and 1.5 Mb, due to its relevant role in formation and maintenance of chromatin loops. A total of 10 CTCFs subpopulations were identified after applying *Meta-Waffle* and clustering. Thanks to this classification, we were able to check the distribution of various features, such as binding motif orientation, compartment type and genomic distance. First, we observed that the genomic distance between pairs of CTCF is relevant to form a chromatin loop. It can be explained by the chromatin loop extrusion speed and the exchange dynamics of the SMC complex, and its requirement of ATP to translocate the DNA (Ganji et al., 2018; Terakawa et al., 2017). Moreover, polymer simulation modelling allowed the study of TAD borders, probing the interval of distance to insulate neighbouring TADs, being comparable to our results (Fudenberg et al., 2016). As expected, the convergent oriented CTCFs, highly correlated with the loop structure pattern, whereas divergent oriented CTCFs presented the opposite trend. It has been observed that the directionality imposed by the DNA bending-initiated loop extrusion model produces a higher interaction frequency with the DNA on one side of it, which agrees with the low interacting frequency between divergent CTCF sites (Y. Guo et al., 2015; S. S. Rao et al., 2014). As for same-oriented CTCF pairs, which can be encountered during loop extrusion, nonetheless their anti-parallel orientation would be unfavourable for dimerization, and the extrusion will continue until finding a convergent site. Interestingly, this approach allowed us to separate the A/B compartment types, with an enrichment of A compartment in the clusters presenting a loop structure pattern. We wanted to study in detail the chromatin states distribution through the CTCF-CTCF structural clusters. The promoter-enhancer state correlated with the A compartment type, being enriched in the canonical loop structure clusters, in line with the loop function, which is to bring together regulatory units for the proper gene expression. The heterochromatin state was found in between the two cluster extremes, which had the expected structural pattern according to the genomic distance between CTCF pairs. It may be caused by the role of active and bivalent chromatin into 3D chromatin organization, consequently organizing the heterochromatin domains. The two clusters with depleted interactions were enriched by promoter-polycomb state, which may represent cell-specific "forbidden" loops. Surprisingly, the polycomb-polycomb state was enriched in those clusters with more interactions than expected. We hypothesize that they can contribute to gene silencing of the already described polycomb-dependent loops (Eagen, Aiden, & Kornberg, 2017), or simply appear by the overlapping of polycomb-polycomb and CTCF-CTCF driven interactions. Here has been inspected the interval of distance from 45 kb to 1.5 Mb in which CTCF it is known to play a relevant role. However, it can be applied to other distances, which would shed light on the complex regulation of the hierarchical structural layers.

The signal deconvolution of CTCFs pairs, also allowed us to obtain those specifically forming chromatin loops, which were used to train a CNN as a loop caller, called here *LOOPbit*. The lack of ground-truth-positive and ground-truth-negative controls, hinders the robust quantification of the specificity and the sensitivity of its performance. To overcome this limitation, *LOOPbit* was compared to a previously published benchmark in which a large set of experiments by 6 different loop callers were analysed (Forcato et al., 2017). *LOOPbit* was applied to 33 Hi-C experiments at 5 kb resolution and to 4 experiments at 40 kb resolution. It presented the same trend to identify more *cis* interactions when having greater number of filtered reads. However, this trend was more pronounced in *LOOPbit*, as the algorithm relies on the immediate surrounding of the loop for its identification. Nevertheless, this higher tendency was not observed in the Hi-C datasets at 40 kb resolution. The loops identified by *LOOPbit* presented a similar average distance between their anchors when compared to other methods. All the loop callers presented poor reproducibility between biological replicates, which normally are pooled together before the analysis to generate a unique sample with higher number of reads. Then, it was not surprising to obtain a similar reproducibility for our method. However, *LOOPbit* was more affected by low coverage experiments, as expected by the higher tendency of *cis*-interaction identification by greater number of filtered reads. Nonetheless, it is expected to perform better with higher coverage samples, as observed in the analysis of Hi-C experiments with targeted degraded CTCF (E. P. Nora et al., 2017), which presented a high reproducibility between untreated and recovered CTCF samples, and low reproducibility of these two samples with the CTCF degraded replicates. This overall poor reproducibility between replicates, may be explained by the fact that Hi-C experiments are an ensemble of cells in different cell states and cell cycle phases, thus not having identical chromatin contacts. Interestingly, anchors of the detected chromatin loops by *LOOPbit*, were highly enriched in promoter-enhancer state, and depleted in heterochromatin and a biologically less plausible state, being consistent with their biological description (Y. Guo et al., 2015; S. S. Rao et al., 2014). This can be a result of the training set used to build *LOOPbit*, which consisted in CTCF-CTCF interactions that were classified by *Meta-Waffle* according to their structural pattern. Moreover, the chromatin loop anchors identified in (E. P. Nora et al., 2017), were more enriched of CTCF ChIP-Seq peaks in the untreated and CTCF-recovered samples, than in the degraded CTCF experiments. Our results indicate that *LOOPbit* is more specific than other tools available as it captured higher proportion of functional loops.

Altogether, *LOOPbit* is comparable to other previously published methods for loop detection in terms of reproducibility and sensitivity, but results in a better performance on detecting functional loops, as it detected twice as many promoter-enhancer loops than most of the other callers. We can conclude, that there is no gold standard algorithm to identify chromatin loops. However, it is important to consider the usability, interoperability, stability of the implementation, and the computing resources required for each algorithm. Considering the fast pace on data production, it is key to provide computational tools as *LOOPbit* and *Meta-Waffle* able to deal with high resolution datasets with reasonable amounts of computational resources with easily exchangeable data formats, following the FAIR principles (findable, accessible, interoperable and re-usable) (Marti-Renom et al., 2018).

Quantitative comparison and automatic feature extraction for chromatin contact data

The development of C-based experiments and their technical improvements, allowed their implementation as a lab routine technique, highly increasing the number of publicly available datasets. It opened the need of new algorithms able to compare chromatin structure between conditions and/or species. Nowadays, there are no available methods able to automatically compare and classify the identified 3D structural changes between contact matrices. To overcome this limitation, we developed *CHESS*, an automated, systematic and feature extraction tool, which correspond to the visual perception of structural differences. It is important to notice that *CHESS* is not limited to compare specific regions within a dataset, but genome-wide comparisons between samples, cell types, developmental stages and species. To note that other available methods for Hi-C comparison are bin-based, meaning that they rely on the comparison between individual bins from the contact matrices, or that will biased towards detecting local structural patterns such as loops or TADs. What makes *CHESS* unique is that it is structure-free and region-based. It will scan full genomic regions, and will compute their degree of similarity, from which the differential structural features would be extracted and classified. Furthermore, *CHESS* performance was faster and highly efficient with a small memory footprint, compared to *diffHiC*, *HOMER* and *ACCOST*. *CHESS* achieved 4, 7 and 320 times speed up, respectively, at 3 times lower memory consumption than the three other methods, making *CHESS* more usable without the need of an advanced computational infrastructure. Moreover, *CHESS* can be parallelizable, allowing to do a countless number of comparisons if wanted, to address more complex biological questions.

CHESS also proved to be highly robust to experimental noise and usable even on shallow sequenced datasets. This makes *CHESS* widely applicable to a large amount of additional types of chromatin conformation capture datasets. For instance, *CHESS* was applied to mouse high-resolution Capture-C data (Despang et al., 2019) from the *Sox9-Kcnj2* locus, where they studied how genome-editing affected TAD function and gene expression. *CHESS* proved to identify the same structural rearrangements as in the experimental study with manually curated detection of structural differences. Nonetheless, it was able to spot subtle changes, such as the gain of a stripe in the *Inv-Intra* mutant, which had the *Kcnj2* TAD inverted without affecting the border between the two genes. This mutant showed a structural rearrangement without drastic effects on gene expression. Loop extrusion models have suggested that stripes/flames at CTCF sites are formed when an extruder finds a single barrier in one orientation but can continue tracking along the chromatin fiber in the other direction. Stripes have been identified in population-based methods, which accumulate snapshots of this dynamic process, giving rise to an increase of contact frequency coming out from the barrier site (Mirny, Imakaev, & Abdennur, 2019).

As stated before, *CHESS* is a structure-free method, which is very promising to *de novo* discovery of structurally similar regions between two experiments. For instance, *CHESS* was applied to mice and human "structurally syntenic" region pairs providing fundamental insights on the evolution of nuclear architecture and its 3D constraints. *CHESS* also identified different degrees of structural conservation across mammalian evolution, indicating variable evolutionary 3D genomic rearrangements paces. The results, thus, suggest that evolutionary conserved regions will require a specific regulatory requisites, which will need the formation of large structured domains, while less conserved regions will not have the same demands and will depend less on such structures.

Additionally, several studies identified the association between structural variation and 3D rearrangements, being translated to gene expression mis-regulation. The identification of structural changes will help to shed light to evaluate the contribution of chromatin organization and disease-generating processes. To demonstrate how *CHESS* can contribute, it was applied to detect chromatin conformation alterations genome-wide in human B-cells from a diffuse large B-cell lymphoma (DLBCL) patient in comparison with normal B-cells, without the need of previous knowledge of the aberrations. Interestingly, the identified differential regions contained long non-coding RNAs (lncRNAs) and pseudogenes instead of protein coding genes. For instance, lncRNAs have been observed to participate in normal B-cell differentiation, and its deregulation can lead to B-cell malignancies such as DLBCL, being their discovery of prognostic value (Dahl, Kristensen, & Gronbaek, 2018). Moreover, from those differential regions, the gain of structural features, classified as TADs and loops, were systematically identified.

As mentioned before, the vast majority of available algorithms to detect structural differences between experiments are structure-biased, such as TAD and loop callers. It limits the study between structure and function relationship as they are focused on the role of mammalian TADs in spatially regulating *cis*-regulatory elements, which it is not clear to apply to all scenarios and species. Nonetheless, as *CHESS* is a feature-free algorithm, it will not be biased and will provide all the set of structural changes regardless their structure. An integrative analysis of other patient-matched genome-wide datasets, such as RNA-seq, ATAC-seq, will be highly useful to determine the cause and consequence of the *CHESS* detected 3D structural rearrangements and its relation to disease. Furthermore, apart from the mentioned scenario in which *CHESS* can be applied, it might also be used to assess the reproducibility between biological replicates. However, this utility has not been inspected in detail in Chapter II.

In conclusion, *CHESS* is an unsupervised algorithm that can automatically retrieve a quantification of the structural similarity and the differential structural feature classification between two genomic regions from chromosome conformation data. Moreover, *CHESS* proved to be highly robust and tolerant to library size and noise level of datasets. Finally, *CHESS* can be applied to a wide range of biological scenarios, including the ranking of known region pairs by structural similarity, such as syntenic regions, and the discovery of structural changes between conditions, such as disease-associated structural variations for clinical applications.

Conclusions

From **Chapter I**, we can specifically infer:

I. We developed *Meta-Waffle*, an algorithm to deconvolve the structural patterns of DNA binding proteins within specific intervals of distances by combining Hi-C and ChIP-seq data.

II. We revealed the presence of 10 CTCF-CTCF subpopulations within 45 kb and 1.5 Mb interval of distance, with different structural patterns and functional signatures.

III. Each CTCF subpopulation presented an epigenetic enrichment, for instance the canonical loop was enriched in enhancer-promoter and enhancer mark, whereas polycomb-promoter mark was found in those regions with less interactions than expected.

IV. The classification of *Meta-Waffle* of the CTCF loops, was used to train a CNN, *LOOPbit*, which can be used to scan genome-wide and identify *cis*-interactions.

V. *LOOPbit* proved useful to capture functional chromatin loops, which were enriched in enhancer-promoter and enhancer marks.

VI. *Meta-Waffle* and *LOOPbit* are publicly-available and open-source python packages.

From **Chapter II**, we can specifically infer:

I. We developed *CHESS*, an algorithm to systematically compare and classify differential structural features between contact matrices.

II. *CHESS* is highly robust and consistent over different levels of experimental noise, resolutions and sequencing depths.

III. *CHESS* is highly efficient and fast, with a very small memory footprint in comparison to other similar methods.

IV. *CHESS* can be used to study multiple biological scenarios, by comparing samples, cell types, developmental stages and even species.

V. *CHESS* can be used to analyse chromatin conformation capture datasets, as Capture-C, by extracting and classifying differential structural features, such as TADs and loops.

VI. *CHESS* is a publicly-available and open-source python package.

References

Alexander, J. M., Guan, J., Huang, B., Lomvardas, S., & Weiner, O. D. (2018). Live-Cell Imaging Reveals Enhancer-dependent Sox2 Transcription in the Absence of Enhancer Proximity. *bioRxiv*, 409672. doi:10.1101/409672

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol, 11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Ardakany, A. R., Ay, F., & Lonardi, S. (2019). Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics, 35*(14), I145-I153. doi:https://doi.org/10.1093/bioinformatics/btz362

Ardakany, A. R., Gezer, H. T., Lonardi, S., & Ay, F. (2020). Mustache: Multi-scale Detection of Chromatin Loops from Hi-C and Micro-C Maps using Scale-Space Representation. *bioRxiv*. doi:https://doi.org/10.1101/2020.02.24.963579

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

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L. M., Dostie, J., Pombo, A., & Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A, 109*(40), 16173-16178. doi:10.1073/pnas.1204799109

Barrington, C., Georgopoulou, D., Pezic, D., Varsally, W., Herrero, J., & Hadjur, S. (2019). Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat Commun, 10*(1), 2908. doi:10.1038/s41467-019-10725-9

Bell, A. C., & Felsenfeld, G. (1999). Stopped at the border: boundaries and insulators. *Curr Opin Genet Dev, 9*(2), 191-198. doi:S0959-437X(99)80029-X [pii]

10.1016/S0959-437X(99)80029-X

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B, 57*, 289-300.

Bersaglieri, C., & Santoro, R. (2019). Genome Organization in and around the Nucleolus. *Cells, 8*(6). doi:10.3390/cells8060579

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

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

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., . . . Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol, 3*(5), e157. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15839726

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

Bovery, T. (1909). Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität. *Arch Zellforsch, 3*, 181-268.

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

Brooker, A. S., & Berkowitz, K. M. (2014). The roles of cohesins in mitosis, meiosis, and human health and disease. *Methods Mol Biol, 1170*, 229-266. doi:10.1007/978-1-4939-0888-2_11

Brown, K. E., Guest, S. S., Smale, S. T., Hahm, K., Merkenschlager, M., & Fisher, A. G. (1997). Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell, 91*(6), 845-854. doi:10.1016/s0092-8674(00)80472-9

Buchwalter, A., Kaneshiro, J. M., & Hetzer, M. W. (2019). Coaching from the sidelines: the nuclear periphery in genome regulation. *Nat Rev Genet, 20*(1), 39-50. doi:10.1038/s41576-018-0063-5

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Varnai, C., Dimond, A., Plagnol, V., . . . Spivakov, M. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol, 17*(1), 127. doi:10.1186/s13059-016-0992-2

Cairns, J., Ochard, W. R., Malysheva, V., & Spivakov, M. (2019). Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data. *Bioinformatics, 35*(22), 4764-4766. doi:https://doi.org/10.1093/bioinformatics/btz450

Chargaff, E., Magasanik, B., Vischer, E., Green, C., Doniger, R., & Elson, D. (1950). Nucleotide composition of pentose nucleic acids from yeast and mammalian tissues. *J Biol Chem, 186*(1), 51-67. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14778803

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J. B., & Gregor, T. (2018). Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet, 50*(9), 1296-1303. doi:10.1038/s41588-018-0175-z

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J, 5*(4), 823-826. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3709526

Clarkson, C. T., Deeks, E. A., Samarista, R., Mamayusupova, H., Zhurkin, V. B., & Teif, V. B. (2019). CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Res, 47*(21), 11181-11196. doi:10.1093/nar/gkz908

Cook, K. B., Hristov, B. H., Le Roch, K. G., Vert, J. P., & Noble, W. S. (2020). Measuring significant changes in chromatin conformation with ACCOST. *Nucleic Acids Res, 48*(5), 2303-2311. doi:10.1093/nar/gkaa069

Cremer, T., Cremer, C., Baumann, H., Luedtke, E. K., Sperling, K., Teuber, V., & Zorn, C. (1982). Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum Genet, 60*(1), 46-56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7076247

Cremer, T., Cremer, C., Schneider, T., Baumann, H., Hens, L., & Kirsch-Volders, M. (1982). Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum Genet, 62*(3), 201-209. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7169211

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P., & Bickmore, W. A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol, 145*(6), 1119-1131. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10366586

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T. Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res, 19*(1), 24-32. doi:gr.082800.108 [pii]

10.1101/gr.082800.108

Dahl, M., Kristensen, L. S., & Gronbaek, K. (2018). Long Non-Coding RNAs Guide the Fine-Tuning of Gene Regulation in B-Cell Development and Malignancy. *Int J Mol Sci, 19*(9). doi:10.3390/ijms19092475

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Dev Biol, 278*(2), 274-288. doi:10.1016/j.ydbio.2004.11.028

Davidson, I. F., Goetz, D., Zaczek, M. P., Molodtsov, M. I., Huis In 't Veld, P. J., Weissmann, F., . . . Peters, J. M. (2016). Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J, 35*(24), 2671-2685. doi:10.15252/embj.201695402

de Leeuw, R., Gruenbaum, Y., & Medalia, O. (2018). Nuclear Lamins: Thin Filaments with Major Functions. *Trends Cell Biol, 28*(1), 34-45. doi:10.1016/j.tcb.2017.08.004

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods, 3*(1), 17-21. doi:nmeth823 [pii]

10.1038/nmeth823

Dekker, J., & Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. *FEBS Lett, 589*(20 Pt A), 2877-2884. doi:10.1016/j.febslet.2015.08.044

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]

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

Devos, D. P., Graf, R., & Field, M. C. (2014). Evolution of the nucleus. *Curr Opin Cell Biol, 28*, 8-15. doi:10.1016/j.ceb.2014.01.004

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]

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., . . . Yue, F. (2018). Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet, 50*(10), 1388-1398. doi:10.1038/s41588-018-0195-8

Djekidel, M. N., Chen, Y., & Zhang, M. Q. (2018). FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Res, 28*(3), 412-422. doi:10.1101/gr.212241.116

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., . . . Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res, 16*(10), 1299-1309. doi:gr.5571506 [pii]

10.1101/gr.5571506

Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., . . . Xie, W. (2017). Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature, 547*(7662), 232-235. doi:10.1038/nature23263

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst, 3*(1), 95-98. doi:10.1016/j.cels.2016.07.002

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

El-Kady, A., & Klenova, E. (2005). Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Lett, 579*(6), 1424-1434. doi:10.1016/j.febslet.2005.01.044

Ferraro, T., Esposito, E., Mancini, L., Ng, S., Lucas, T., Coppey, M., . . . Lagha, M. (2016). Transcriptional Memory in the Drosophila Embryo. *Curr Biol, 26*(2), 212-218. doi:10.1016/j.cub.2015.11.058

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., . . . Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol, 16*(6), 2802-2813. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8649389

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

Flavahan, W. A., Drier, Y., Liau, 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

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., & Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. *Nat Methods, 14*(7), 679-685. doi:10.1038/nmeth.4325

Forsberg, F., Brunet, A., Ali, T. M. L., & Collas, P. (2019). Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. *Nucleus, 10*(1), 7-20. doi:10.1080/19491034.2019.1570810

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). Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature, 172*(4369), 156-157. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=13072614

Fu, Y., Sinha, M., Peterson, C. L., & Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet, 4*(7), e1000138. doi:10.1371/journal.pgen.1000138

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

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., . . . Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature, 462*(7269), 58-64. doi:nature08497 [pii]

10.1038/nature08497

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

Gassler, J., Brandao, H. B., Imakaev, M., Flyamer, I. M., Ladstatter, S., Bickmore, W. A., . . . Tachibana, K. (2017). A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J, 36*(24), 3600-3618. doi:10.15252/embj.201798083

Gesson, K., Rescheneder, P., Skoruppa, M. P., von Haeseler, A., Dechat, T., & Foisner, R. (2016). A-type lamins bind both hetero- and euchromatin, the latter being regulated by lamina-associated polypeptide 2 alpha. *Genome Res, 26*(4), 462-473. doi:10.1101/gr.196220.115

Gibcus, J. H., & Dekker, J. (2013). The hierarchy of the 3D genome. *Mol Cell, 49*(5), 773-782. doi:10.1016/j.molcel.2013.02.011

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

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:nature06947 [pii]

10.1038/nature06947

Guo, C., Yoon, H. S., Franklin, A., Jain, S., Ebert, A., Cheng, H. L., . . . Alt, F. W. (2011). CTCF-binding elements mediate control of V(D)J recombination. *Nature, 477*(7365), 424-430. doi:10.1038/nature10495

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

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., & Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci U S A, 109*(43), 17507-17512. doi:10.1073/pnas.1111941109

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

Heitz, E. (1928). Das Heterochromatin der Moose. *Jahrb Wiss Botanik, 69*, 762–818.

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

Holwerda, S. J., & de Laat, W. (2013). CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci, 368*(1620), 20120369. doi:10.1098/rstb.2012.0369

rstb.2012.0369 [pii]

Horowitz-Scherer, R. A., & Woodcock, C. L. (2006). Organization of interphase chromatin. *Chromosoma, 115*(1), 1-14. doi:10.1007/s00412-005-0035-3

Horta, A., Monahan, K., Bashkirova, E., & Lomvardas, S. (2019). Cell type-specific interchromosomal interactions as a mechanism for transcriptional diversity. *bioRxiv.*

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*, 108-119.

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

Hug, C. B., Grimaldi, A. G., Kruse, K., & Vaquerizas, J. M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell, 169*(2), 216-228 e219. doi:10.1016/j.cell.2017.03.024

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, Y. C., Lin, C. F., Valladares, O., Malamon, J., Kuksa, P. P., Zheng, Q., . . . Wang, L. S. (2015). HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics, 31*(8), 1290-1292. doi:10.1093/bioinformatics/btu801

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

Ishihara, K., Oshimura, M., & Nakao, M. (2006). CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell, 23*(5), 733-742. doi:10.1016/j.molcel.2006.08.008

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., . . . Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature, 503*(7475), 290-294. doi:10.1038/nature12644

Jonkers, I., & Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol, 16*(3), 167-177. doi:10.1038/nrm3953

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

Katainen, R., Dave, K., Pitkanen, E., Palin, K., Kivioja, T., Valimaki, N., . . . Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet, 47*(7), 818-821. doi:10.1038/ng.3335

Kaul, A., Bhattacharyya, S., & Ay, F. (2020). Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc.* doi:10.1038/s41596-019-0273-0

Ke, Y., Xu, Y., Chen, X., Feng, S., Liu, Z., Sun, Y., . . . Liu, J. (2017). 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell, 170*(2), 367-381 e320. doi:10.1016/j.cell.2017.06.029

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S. S., Janssen, H., . . . van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. *Cell, 153*(1), 178-192. doi:10.1016/j.cell.2013.02.028

S0092-8674(13)00217-1 [pii]

Kitchen, N. S., & Schoenherr, C. J. (2010). Sumoylation modulates a domain in CTCF that activates transcription and decondenses chromatin. *J Cell Biochem, 111*(3), 665-675. doi:10.1002/jcb.22751

Knight, P. A., & Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA journal of Numerical Analysis, 33*(3), 1029-1047.

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

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J. T., Sadreyev, R. I., . . . Kingston, R. E. (2018). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell, 71*(1), 191. doi:10.1016/j.molcel.2018.06.022

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

Le Dily, F., Serra, F., & Marti-Renom, M. A. (2017). 3D modeling of chromatin structure: is there a way to integrate and reconcile single cell and population experimental data? *Wiley Interdisciplinary Reviews: Computational Molecular Science*, e1308-n/a. doi:10.1002/wcms.1308

Levene, P. A. (1919). The structure of yeast nucleic acid: IV. Ammonia hydrolysis. *J. Biol. Chem., 40*, 415-424.

Li, Y., Haarhuis, J. H. I., Sedeno Cacciatore, A., Oldenkamp, R., van Ruiten, M. S., Willems, L., . . . Panne, D. (2020). The structural basis for cohesin-CTCF-anchored loops. *Nature*. doi:10.1038/s41586-019-1910-z

Lichter, P., Cremer, T., Borden, J., Manuelidis, L., & Ward, D. C. (1988). Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum Genet, 80*(3), 224-234. doi:10.1007/bf01790090

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

Lun, A. T., & Smyth, G. K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics, 16*, 258. doi:10.1186/s12859-015-0683-0

Lund, E., Oldenburg, A. R., Delbarre, E., Freberg, C. T., Duband-Goulet, I., Eskeland, R., . . . Collas, P. (2013). Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes. *Genome Res, 23*(10), 1580-1589. doi:10.1101/gr.159400.113

Lund, E. G., Duband-Goulet, I., Oldenburg, A., Buendia, B., & Collas, P. (2015). Distinct features of lamin A-interacting chromatin domains mapped by ChIP-sequencing from sonicated or micrococcal nuclease-digested chromatin. *Nucleus, 6*(1), 30-39. doi:10.4161/19491034.2014.990855

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

Lutz, M., Burke, L. J., LeFevre, P., Myers, F. A., Thorne, A. W., Crane-Robinson, C., . . . Renkawitz, R. (2003). Thyroid hormone-regulated enhancer blocking: cooperation of CTCF and thyroid hormone receptor. *EMBO J, 22*(7), 1579-1587. doi:10.1093/emboj/cdg147

Marti-Renom, M. A., Almouzni, G., Bickmore, W. A., Bystricky, K., Cavalli, G., Fraser, P., . . . Torres-Padilla, M. E. (2018). Challenges and guidelines toward 4D nucleome data and model standards. *Nat Genet, 50*(10), 1352-1358. doi:10.1038/s41588-018-0236-3

Martin, W. F., Garg, S., & Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci, 370*(1678), 20140330. doi:10.1098/rstb.2014.0330

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.* doi:10.1038/s41586-019-1035-4

Mawhinney, M. T., Liu, R., Lu, F., Maksimoska, J., Damico, K., Marmorstein, R., . . . Urbanc, B. (2018). CTCF-Induced Circular DNA Complexes Observed by Atomic Force Microscopy. *J Mol Biol, 430*(6), 759-776. doi:10.1016/j.jmb.2018.01.012

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., & Luscombe, N. M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS ONE, 12*(4), e0174744. doi:10.1371/journal.pone.0174744

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

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

Murayama, Y., Samora, C. P., Kurokawa, Y., Iwasaki, H., & Uhlmann, F. (2018). Establishment of DNA-DNA Interactions by the Cohesin Ring. *Cell, 172*(3), 465-477 e415. doi:10.1016/j.cell.2017.12.021

Nagano, T., Varnai, C., Schoenfelder, S., Javierre, B. M., Wingett, S. W., & Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol, 16*, 175. doi:10.1186/s13059-015-0753-7

Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., & Barta, E. (2016). Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics, 17*(1), 637. doi:10.1186/s12864-016-2940-7

Nakahashi, H., Kieffer Kwon, K. R., Resch, W., Vian, L., Dose, M., Stavreva, D., . . . Casellas, R. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep, 3*(5), 1678-1689. doi:10.1016/j.celrep.2013.04.024

Nemeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Peterfia, B., . . . Langst, G. (2010). Initial genomics of the human nucleolus. *PLoS Genet, 6*(3), e1000889. doi:10.1371/journal.pgen.1000889

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

Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., . . . Bruneau, B. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from higher-order genomic compartmentalization. *bioRxiv*. doi:10.1101/095802

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

Nozaki, T., Imai, R., Tanbo, M., Nagashima, R., Tamura, S., Tani, T., . . . Maeshima, K. (2017). Dynamic Organization of Chromatin Domains Revealed by Super-Resolution Live-Cell Imaging. *Mol Cell, 67*(2), 282-293 e287. doi:10.1016/j.molcel.2017.06.018

Ocampo-Hafalla, M., Munoz, S., Samora, C. P., & Uhlmann, F. (2016). Evidence for cohesin sliding along budding yeast chromosomes. *Open Biol, 6*(6). doi:10.1098/rsob.150178

Paredes, S. H., Melgar, M. F., & Sethupathy, P. (2013). Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index. *Bioinformatics, 29*(12), 1485-1487. doi:10.1093/bioinformatics/bts596

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., . . . Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell, 132*(3), 422-433. doi:10.1016/j.cell.2008.01.011

Pascual-Reguant, L., Blanco, E., Galan, S., Le Dily, F., Cuartero, Y., Serra-Bardenys, G., . . . Peiro, S. (2018). Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. *Nat Commun, 9*(1), 3420. doi:10.1038/s41467-018-05912-z

Paulsen, J., Liyakat Ali, T. M., Nekrasov, M., Delbarre, E., Baudement, M. O., Kurscheid, S., . . . Collas, P. (2019). Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat Genet, 51*(5), 835-843. doi:10.1038/s41588-019-0392-0

Paulsen, J., Rodland, E. A., Holden, L., Holden, M., & Hovig, E. (2014). A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res, 42*(18), e143. doi:10.1093/nar/gku738

Paulsen, J., Sandve, G. K., Gundersen, S., Lien, T. G., Trengereid, K., & Hovig, E. (2014). HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics, 30*(11), 1620-1622. doi:10.1093/bioinformatics/btu082

Pavlaki, I., Docquier, F., Chernukhin, I., Kita, G., Gretton, S., Clarkson, C. T., . . . Klenova, E. (2018). Poly(ADP-ribosyl)ation associated changes in CTCF-chromatin binding and gene expression in breast cells. *Biochim Biophys Acta Gene Regul Mech, 1861*(8), 718-730. doi:10.1016/j.bbagrm.2018.06.010

Phanstiel, D. H., Van Bortle, K., Spacek, D., Hess, G. T., Shamim, M. S., Machol, I., . . . Snyder, M. P. (2017). Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell, 67*(6), 1037-1048 e1036. doi:10.1016/j.molcel.2017.08.006

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature, 322*(6081), 697-701. doi:10.1038/322697a0

Racko, D., Benedetti, F., Dorier, J., & Stasiak, A. (2018). Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Res, 46*(4), 1648-1660. doi:10.1093/nar/gkx1123

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

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., . . . Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol Cell, 67*(6), 1049-1058 e1046. doi:10.1016/j.molcel.2017.08.026

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

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., . . . Krumm, A. (2008). CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A, 105*(24), 8309-8314. doi:0801273105 [pii]

10.1073/pnas.0801273105

Sagan, L. (1967). On the origin of mitosing cells. *J Theor Biol, 14*(3), 255-274. doi:10.1016/0022-5193(67)90079-3

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

Schmid, M. W., Grob, S., & Grossniklaus, U. (2015). HiCdat: a fast and easy-to-use Hi-C data analysis tool. *BMC Bioinformatics, 16*, 277. doi:10.1186/s12859-015-0678-x

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., . . . Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell, 148*(1-2), 335-348. doi:10.1016/j.cell.2011.11.058

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

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., . . . Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature, 479*(7371), 74-79. doi:10.1038/nature10442

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., . . . de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet, 38*(11), 1348-1354. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17033623

Solovei, I., Wang, A. S., Thanisch, K., Schmidt, C. S., Krebs, S., Zwerger, M., . . . Joffe, B. (2013). LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell, 152*(3), 584-598. doi:10.1016/j.cell.2013.01.009

Spielmann, M., Lupianez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nat Rev Genet, 19*(7), 453-467. doi:10.1038/s41576-018-0007-0

Spill, Y. G., Castillo, D., Vidal, E., & Marti-Renom, M. A. (2019). Binless normalization of Hi-C data provides significant interaction and difference detection independent of resolution. *Nat Commun, 10*(1), 1938. doi:10.1038/s41467-019-09907-2

Stansfield, J. C., Cresswell, K. G., & Dozmorov, M. G. (2019). multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics, 35*. doi:10.1093/bioinformatics/btz048

Stansfield, J. C., Cresswell, K. G., Vladimirov, I. V., & Dozmorov, M. G. (2018). HiCcompare: An R-package for Joint Normalization and Comparison of HI-C Datasets. *BMC Bioinformatics, 19*. doi:10.1186/s12859-018-2288-x

Stigler, J., Camdere, G. O., Koshland, D. E., & Greene, E. C. (2016). Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Rep, 15*(5), 988-998. doi:10.1016/j.celrep.2016.04.003

Sullivan, G. J., Bridger, J. M., Cuthbert, A. P., Newbold, R. F., Bickmore, W. A., & McStay, B. (2001). Human acrocentric chromosomes with transcriptionally silent nucleolar organizer regions associate with nucleoli. *EMBO J, 20*(11), 2867-2874. doi:10.1093/emboj/20.11.2867

Tanabe, H., Muller, S., Neusser, M., von Hase, J., Calcagno, E., Cremer, M., . . . Cremer, T. (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci U S A, 99*(7), 4424-4429. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11930003

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., . . . Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell, 163*(7), 1611-1627. doi:10.1016/j.cell.2015.11.024

Terakawa, T., Bisht, S., Eeftens, J. M., Dekker, C., Haering, C. H., & Greene, E. C. (2017). The condensin complex is a mechanochemical motor that translocates along DNA. *Science, 358*(6363), 672-676. doi:10.1126/science.aan6516

Thongjuea, S., Stadhouders, R., Grosveld, F., Soler, E., & Lenhard, B. (2013). r3Cseq: An R/Bioconductor Package for the Discovery of Long-Range Genomic Interactions From Chromosome Conformation Capture and Next-Generation Sequencing Data. *Nucleic Acids Res, 41*, e132. doi:10.1093/nar/gkt373

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

Van Holde, K. E. (1989). *Chromatin*. (C. Springer Ed.). Heidelberg.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G. J., Ariyurek, Y., . . . Lamond, A. I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell, 21*(21), 3735-3748. doi:10.1091/mbc.E10-06-0508

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

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Nodell, M. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. Retrieved from PM:11181995

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

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep, 10*(8), 1297-1309. doi:10.1016/j.celrep.2015.02.004

Vivante, A., Brozgol, E., Bronshtein, I., Levi, V., & Garini, Y. (2019). Chromatin dynamics governed by a set of nuclear structural proteins. *Genes Chromosomes Cancer, 58*(7), 437-451. doi:10.1002/gcc.22719

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., . . . Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res, 22*(9), 1680-1688. doi:10.1101/gr.136101.111

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

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., . . . Peters, J. M. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J, 36*(24), 3573-3599. doi:10.15252/embj.201798004

Xiao, T., Wallace, J., & Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol Cell Biol, 31*(11), 2174-2183. doi:10.1128/MCB.05093-11

Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T. S., Sullivan, P. F., . . . Li, Y. (2016). A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics, 32*(5), 650-656. doi:10.1093/bioinformatics/btv650

Xu, Z., Zhang, G., Wu, C., Li, Y., & Hu, M. (2016). FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics, 32*(17), 2692-2695. doi:10.1093/bioinformatics/btw240

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

Yusufzai, T. M., Tagami, H., Nakatani, Y., & Felsenfeld, G. (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell, 13*(2), 291-298. doi:10.1016/s1097-2765(04)00029-2

Zufferey, M., Tavernari, D., Oricchio, E., & Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biol, 19*(1), 217. doi:10.1186/s13059-018-1596-9

Annex

Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains

Pascual-Reguant L., […] Galan S., et al. Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. Nat Commun 9, 3420, doi:10.1038/s41467-018-05912-z (2018).

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05912-z OPEN

Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains

Laura Pascual-Reguant[1], Enrique Blanco[2], Silvia Galan[2,3], François Le Dily[2,4], Yasmina Cuartero[2,3], Gemma Serra-Bardenys[1,4], Valerio Di Carlo[2], Ane Iturbide[5], Joan Pau Cebrià-Costa[1], Lara Nonell[6], Antonio García de Herreros[4,7], Luciano Di Croce[2,8], Marc A. Marti-Renom[2,3,4,8] & Sandra Peiró[1]

Lamins (A/C and B) are major constituents of the nuclear lamina (NL). Structurally conserved lamina-associated domains (LADs) are formed by genomic regions that contact the NL. Lamins are also found in the nucleoplasm, with a yet unknown function. Here we map the genome-wide localization of lamin B1 in an euchromatin-enriched fraction of the mouse genome and follow its dynamics during the epithelial-to-mesenchymal transition (EMT). Lamin B1 associates with actively expressed and open euchromatin regions, forming dynamic euchromatin lamin B1-associated domains (eLADs) of about 0.3 Mb. Hi-C data link eLADs to the 3D organization of the mouse genome during EMT and correlate lamin B1 enrichment at topologically associating domain (TAD) borders with increased border strength. Having reduced levels of lamin B1 alters the EMT transcriptional signature and compromises the acquisition of mesenchymal traits. Thus, during EMT, the process of genome reorganization in mouse involves dynamic changes in eLADs.

[1] Vall d'Hebron Institute of Oncology, 08035 Barcelona, Spain. [2] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona, Spain. [3] Structural Genomic Group, CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona, Spain. [4] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [5] Institute of Epigenetics and Stem Cells, D-81377 München, Germany. [6] Servei d'Anàlisi de Microarrays Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain. [7] Programa de Recerca en Càncer, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain. [8] ICREA, Pg. Lluis Companys 23, Barcelona, Spain. Correspondence and requests for materials should be addressed to S.Pó. (email: speiro@vhio.net)

Nuclear genome folding occurs at multiple levels, and the dynamic folding of chromatin is known to be elemental in regulating gene expression. Alterations in these folding units are associated with multiple diseases and cancer[1]. One key level of organization involves the interaction between chromatin and the nuclear lamina (NL)[2,3]. Lamins (A/C and B) are type V intermediate filaments and are the major components of the NL. Chromatin regions that are in close contact with NL are called lamina-associated domains (LADs)[4–6]. These domains were initially identified using the DamID method, in which bacterial DNA adenine methyltransferase fused with lamin B1 methylate DNA regions that are in contact with NL[7]. LADs can be also detected by chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq)[8–10] and by fluorescent in situ hybridization. LADs are formed by heterochromatin defined as chromatin regions with low gene frequency, transcriptionally silent, and enriched in the repressive histone marks, H3K9me2/3[11]. Importantly, LADs are extremely conserved between species, although some show a certain degree of dynamism[11]. Despite the extensive data published about NLs, little is known regarding its structural organization. High-resolution confocal microscopy and three-dimensional (3D) structured illumination microscopy showed that A- and B-type lamins form separated but interconnected meshworks with distinct roles[12,13]. Recently, it has been demonstrated that A- and B-type lamins assemble into tetrameric filaments of 3.5 nm, a structure surprisingly different than that of other cytoskeletal elements[14]. Moreover, these filaments are variable in length and are found to form both sparsely and densely packed regions, which are both detected around dense nuclear material that could be chromatin[14]. Unlike A-type, B-type lamins remain permanent farnesylated and carboxymethylated,and thus remain tightly associated with the membrane[15]. There is also evidence of the existence of a nucleoplasmic pool of lamins (A/C and B) that are assembled into stable structures with characteristics different from the A- and B-type lamins located in the NL[16]. This finding suggests that nucleoplasmic lamins may have a role distinct from that of perinuclear lamins[6,17]. In fact, recent ChIP-seq genome-wide studies have shown that lamin A/C contact euchromatin[18,19] and have suggested a functional role for lamin A/C in creating a permissive environment for gene regulation[18]. These findings are of high interest for two main reasons: (1) they demonstrate interactions between large euchromatin regions and nucleoplasmic lamin A; and (2) methodologically, they show how enrichment of different chromatin fractions can reveal distinct lamin A-associated domains[20]. Importantly, although DamID maps of lamin A and B are similar, a fraction of lamin A is found throughout the nucleus that is not detected by DamID, for yet unknown reasons[11,21]. This fact, together with evidence that lamins form separate but interconnected networks[12,13] and interact with nuclear structures distinct from the NL[6,16,17], led us to hypothesize that lamin B1 filaments could also interact with euchromatin.

Here we used euchromatin enrichment and ChIP-seq to map the localization of lamin B1, and we then analyzed its dynamism using the epithelial-to-mesenchymal transition (EMT) model[22]. The EMT program describes a series of events by which epithelial cells lose many of their epithelial characteristics and take on properties that are typical of mesenchymal cells. These cells undergo complex changes in both cell architecture and behavior[23]. Developmental biologists have long recognized that EMT is a crucial process for the generation of tissues and organs during embryogenesis of both vertebrates and invertebrates, and it also has an important role in pathological processes, such as fibrosis and cancer[24]. During progression to metastatic competence, carcinoma cells acquire mesenchymal gene expression patterns and properties, resulting in modified adhesive characteristics and in activation of proteolysis capacity and motility; these changes allow tumor cells to metastasize and establish secondary tumors at distant sites[22]. Recent studies suggest that chromatin reorganization during EMT is an essential step for the conversion of an epithelial cell into a mesenchymal cell[25,26]. However, local and 3D chromatin roles in EMT and tumorigenesis are incompletely understood.

Here we define the existence of LADs formed when lamin B1 contacts euchromatin regions, which we term euchromatin LADs (eLADs) to differentiate them from conventional LADs. Lamin B1 ChIP-seq from an enriched euchromatin fraction, RNA sequencing (RNA-seq), and assay for transposase-accessible chromatin using sequencing (ATAC-seq) allowed us to detect these domains and to analyze their dynamism in the context of EMT. In combination with whole-genome chromatin conformation capture (Hi-C), we demonstrated that eLADs are located in the active (A) compartment and that, at the onset of EMT, the amount of lamin B1 increased at TAD borders concomitantly with increased border strength. Moreover, over the time course, additional eLADs were formed involving genes that belong to the EMT pathway. Finally, depletion of lamin B1 from the euchromatin fraction massively affected the gene expression profile (as determined by RNA-seq) at a key time point of this cellular transformation, resulting in impaired EMT.

Results

Lamin B1 associates with euchromatin regions that are dynamic during EMT. To induce EMT cellular transformation, normal mouse mammary epithelial cells (NMuMG) were treated with the transforming growth factor transforming growth factor (TGF)-β[27]. We first confirmed by confocal microscopy the presence of lamin B1 in the nuclear envelope as well as the nuclear interior of cells that were either untreated or treated with TGF-β for 8 or 24 h (Supplementary Fig. 1a). Colocalization with emerin, an integral protein of the NL[28], was only observed in the nuclear envelope (Supplementary Fig. 1b). We then performed ChIP-seq for lamin B1 on an enriched euchromatin fraction. As chromatin preparation is key to the identification of genomic-associated regions[18,29], chromatin was sheared by low sonication to obtain DNA fragments of 300–600 bp. This sonication condition favors shearing of accessible chromatin (e.g., euchromatin) but leaves heterochromatin regions intact[18,20,30]. In this way, heterochromatin regions can be excluded from Ilumina sequencing due to their size[18]; indeed, the bioanalyzer intensity profile showed a substantial fraction of fragments larger than 1 kb that were excluded from sequencing (Supplementary Fig. 1c). We identified significant lamin B1-positive sites (lamin B1 +) in chromatin in all three conditions (Fig. 1a), with a total of 4645 peaks in untreated cells, 10,484 peaks in 8 h TGF-β-treated cells, and 7083 peaks in 24 h TGF-β-treated cells. The distribution of ChIP-seq peaks across the genome showed a significant enrichment around the transcription start site (TSS) of specific genes (Fig. 1b) (of 4454 genes in untreated cells, 8755 genes in 8 h TGF-β-treated cells, and 6595 genes in 24 h TGF-β-treated cells) that did not overlap with canonical LADs (cLADs; Fig. 1a). Importantly, as all the immunoprecipitated chromatin could be analyzed by ChIP-quantitative PCR (qPCR) without size selection exclusion, we were also able to detect lamin B1 in a cLAD region (Fig. 1c). Moreover, the specificity of the antibody used for the genome-wide mapping of lamin B1 was confirmed by lamin B1 immunoprecipitation from NMuMG cells infected with lentivirus carrying either an irrelevant short hairpin RNA as a control (C), or specific for lamin B1 (knockdown, KD), in low-sonicated chromatin (Fig. 1d).

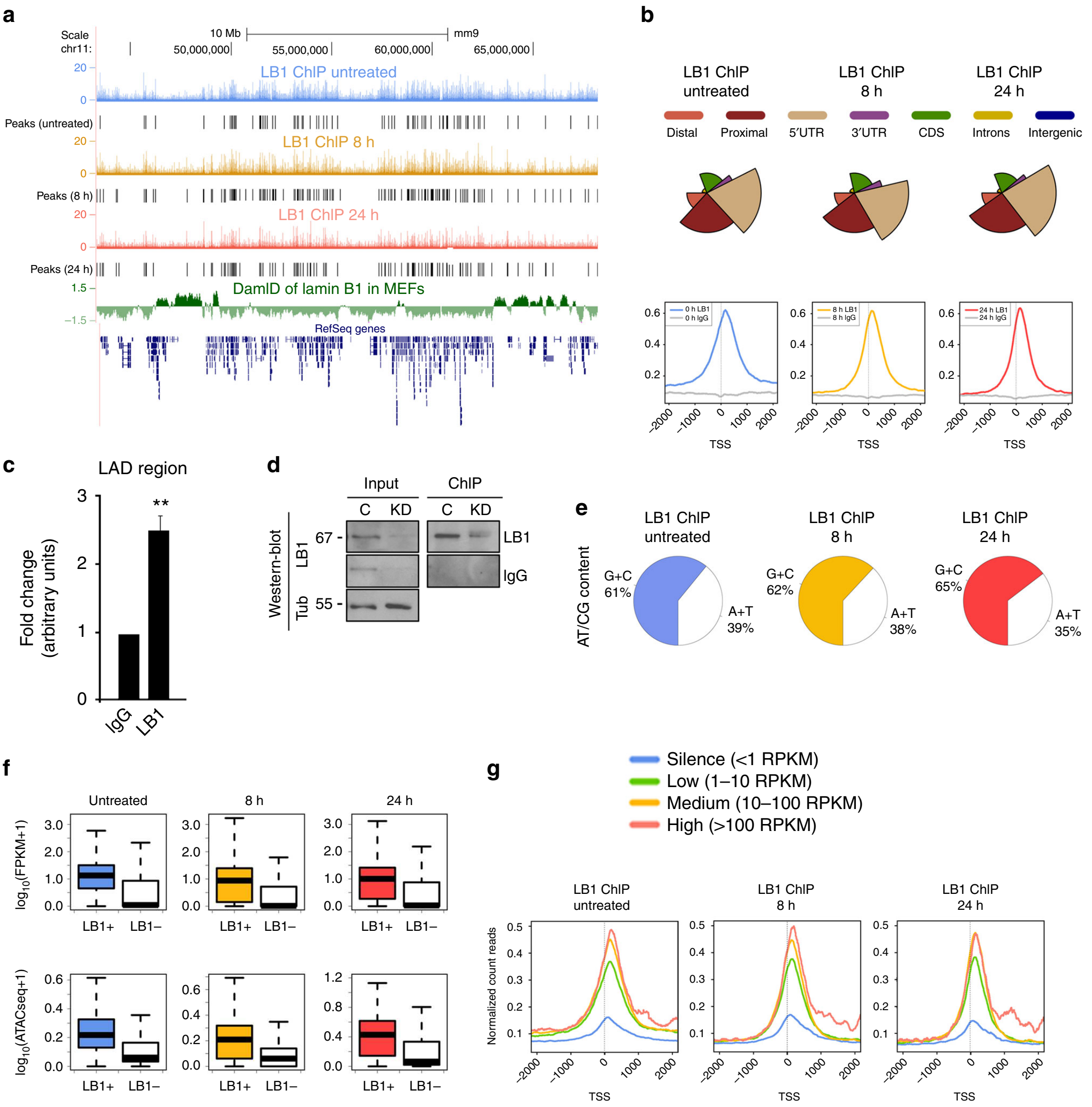


Fig. 1 Lamin B1 associates with euchromatin regions. **a** UCSC Genome Browser overview of one region across chromosome 11 (mm9) of the lamin B1 ChIP-seq profiles in NMuMG cells that were untreated (blue) or treated with TGF-β for 8 h (orange) or 24 h (red). Lamin B1 + sites identified at each condition (black), previously published lamina-associated domains (LADs) (green), and the RefSeq gene track (dark blue) are shown. **b** Genome distribution of lamin B1 ChIP-seq peaks. The distal region is that within 2.5 and 0.5 Kb upstream of a gene's TSS, and the proximal region, within 0.5 Kb of a gene's TSS (top). The average distribution of lamin B1 ChIP-seq reads is shown, as well as the corresponding IgG control experiments, at 2 Kb around the TSS of lamin B1 + genes (bottom). **c** ChIP-qPCR of a canonical LAD (cLAD)-selected region. Data are expressed as the fold-change relative to data obtained from the IgG experiments. Data are shown as mean ± SD, $n = 4$. **d** Western blotting of chromatin immunoprecipitated with lamin B1 or IgG antibodies from normal or KD cells. **e** AT/CG content of the sequences of lamin B1 + sites. **f** Expression of lamin B1 + genes computed in FPKM (fragment per kilobase of transcript per million mapped reads) in untreated cells (blue) or those treated with TGF-β for 8 h (orange) or 24 h (red), as compared with the rest of genes in the genome (white) (top). Promoter ATAC-seq enrichment of lamin B1 + genes measured in number of normalized reads in untreated cells (blue) or those treated with TGF-β for 8 h (orange) or 24 h (red), as compared with the rest of genes in the genome (white) (bottom). The bottom and top fractions in the boxes represent the first and third quartiles, and the line, the median. Whiskers denote the interval between 1.5 times the interquartile range (IQR) and the median. **g** Average distribution of lamin B1 ChIP-seq reads at 2 Kb around the TSS of mouse genes that were classified into four categories (silent, low, medium, and high) according to their expression levels

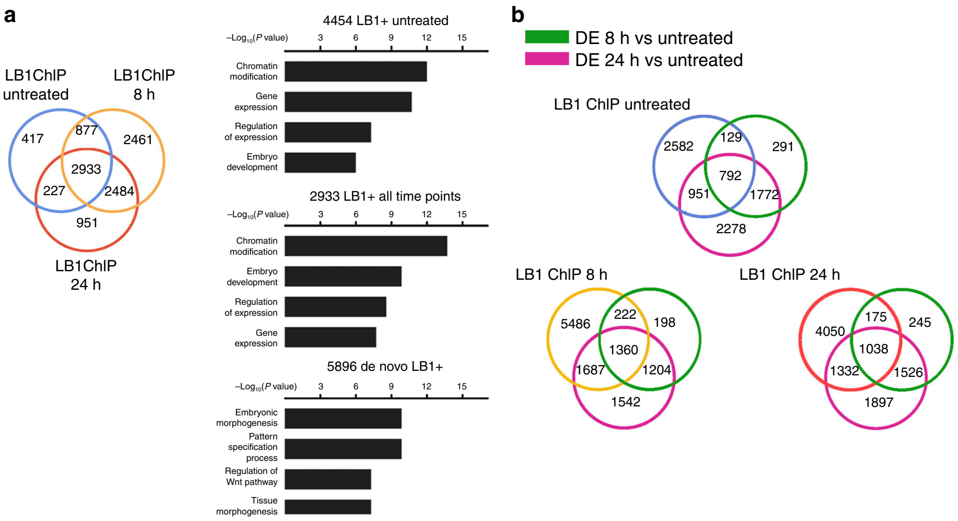


Fig. 2 Lamin-positive sites are dynamic during EMT. **a** Venn diagram showing the overlap between lamin B1 + genes in untreated cells (blue) and cells treated with TGF-β for 8 h (orange) or 24 h (red) (left). Gene ontology (GO) terms of lamin B1 + genes only present in untreated conditions, present in all three time points, or newly formed at the onset of the EMT are depicted (right). **b** Venn diagram showing the overlap between lamin B1 + genes in untreated conditions (blue) or after treatment with TGF-β for 8 h (orange) or 24 h (red); genes differentially expressed (DE) at 8 h (green) and 24 h (pink) after TGF-β treatment are shown

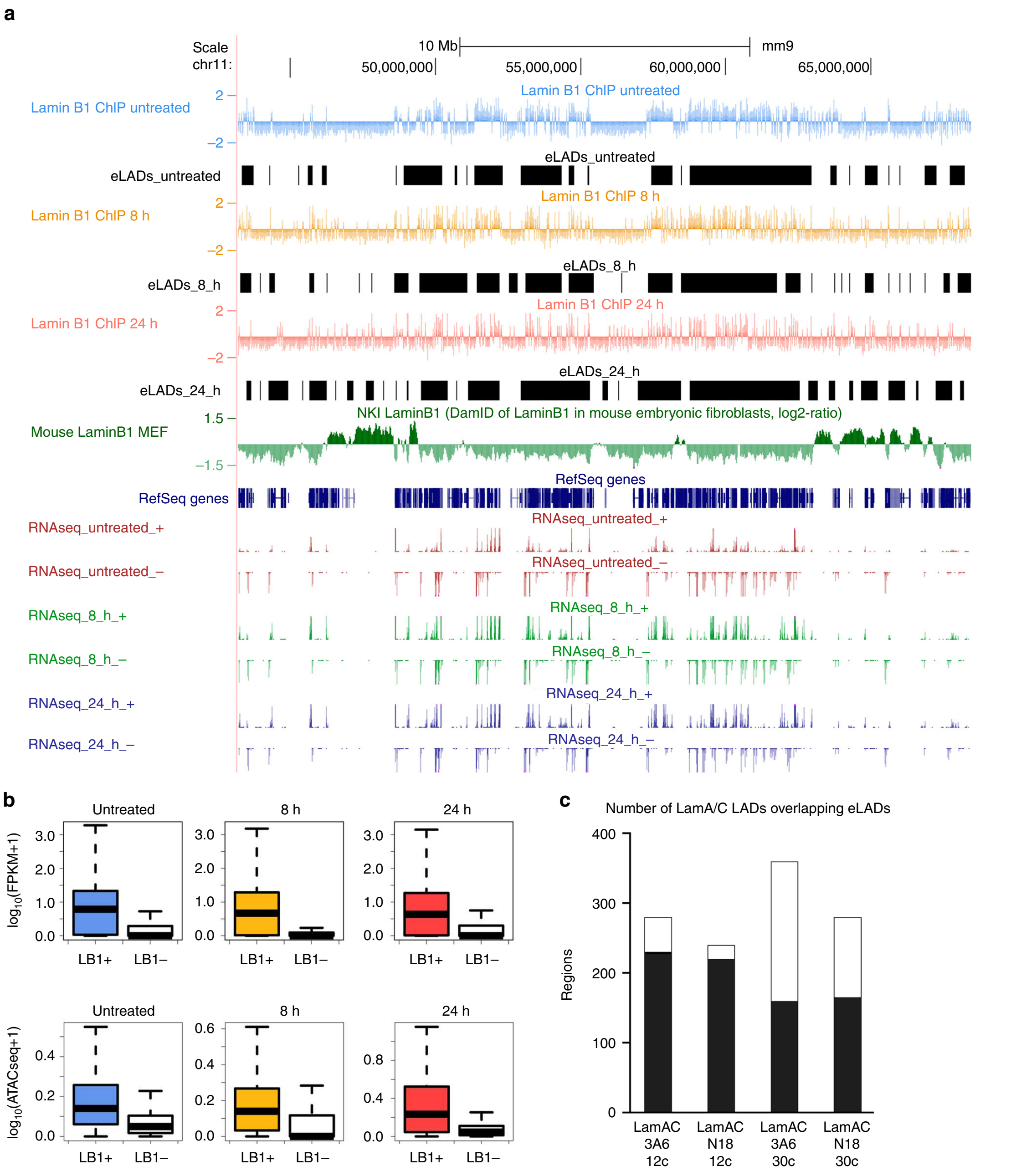
Further characterization showed that, in contrast to A/T-rich constitutive LADs[31], lamin B1 + sites were strongly associated with C/G-rich sequences (Fig. 1e) and contained a substantially higher number of genes. To determine whether these genes were decorated with canonical euchromatin or were instead associated to heterochromatin histone marks, bioinformatics Enrichr analysis[32] was performed. We found how histone marks associated to these genes were characteristic for euchromatin (Supplementary Fig. 2a). We repeated the same analysis on genes that were not enriched in lamin B1 at each time point and H3K9me2/3, the classical heterochromatin histone mark, was enriched (Supplementary Fig. 2b). In addition, RNA-seq and ATAC-seq data from these samples showed that lamin B1-enriched genes were actively transcribed, with their promoters located in accessible chromatin regions (Fig. 1f). On the other hand, non-lamin B1-enriched genes presented low expression rates and less chromatin accessibility (white plots; Fig.1f). To analyze whether a correlation between lamin B1 binding and gene expression exists during EMT, we stratified the full set of mouse genes on each time point into four groups based on RNA-seq data (of silent, low, medium, and high expression). We then plotted the lamin B1 levels obtained by ChIP-seq for each gene set and observed a strong correlation between expression and lamin B1 around the TSS (the higher the levels of expression, the higher the lamin B1 enrichment) (Fig. 1g).

Comparing lamin B1 + sites during EMT revealed that, of the total of 10,350 target genes identified in the three time points, only 2933 genes (28%) maintained lamin B1 throughout the EMT process, whereas a vast majority (7417; 72%) changed at the onset of EMT (Fig. 2a). Taken together, these data suggest that lamin B1 can be found in expressed euchromatin regions associated with C/G regions that are gene-rich, accessible, decorated with euchromatin histone marks, and change dynamically during EMT transformation.

Gene ontology (GO) showed that genes enriched in lamin B1 only in untreated cells or that maintain their level of lamin B1 during the entire EMT process belong to chromatin modification and general gene transcription (Fig. 2a; for the full list of GO categories in each time point, see Supplementary Fig. 3a). We next focused on the set of genes that emerged as new lamin B1 targets upon EMT induction by TGF-β treatment (8 and 24 h). We found 5896 genes positive for lamin B1, which encoded proteins with functions belonging to embryonic morphogenesis and pattern specification, both of which strongly related to the EMT process (Fig. 2a). In order to identify potential drivers responsible for the increase in lamin B1 occupancy at these regions, we determined the enrichment of transcription factor (TF) motifs at these new lamin B1 sites. The enriched factors identified were involved in developmental and differentiation process (Supplementary Fig. 3b, c; Supplementary Table 1), and some of them are members of the TGF-β signaling pathway[33,34]. These data suggest that TGF-β-activated TFs may have a role in recruiting lamin B1 to these specific sites. Finally, the percentage of lamin B1 + genes for which we observed a change in expression at the onset of EMT were 42% (1872/4454) in the epithelial state, 37% (3269/8755) at 8 h after TGF-β treatment, and 38% (2545/6595) at 24 h after TGF-β treatment (Fig. 2b).

Definition of eLADs and the putative role of lamin B1 as an architectural protein. cLADs have been traditionally highlighted as genomic regions enriched in lamins that emerged from comparison with the corresponding control experiments[4,18]. We similarly defined eLADs as clusters of neighboring lamin B1 + sites that occur within a delimited region of the genome (see Supplementary Methods). We identified 2051 eLADs in untreated cells, 2429 eLADs in 8 h-treated cells, and 2949 eLADs in 24 h-treated cells (Fig. 3a), with an average size of 0.34 Mb (and therefore smaller than LADs, which are about 1 Mb[4]). The coverage of the mouse genome for each set of eLADs was 25.9%, 31.7%, and 40.1%, respectively. As expected, genes within eLADs were active and located in accessible chromatin (Fig. 3b).

Given that lamin A/C also occupies euchromatin regions[18], we analyzed the degree of overlap between both sets of ChIP-seq



data. As expected, eLADs had an extended overlap with euchromatin regions enriched in lamin A/C (Fig. 3c).

Intermediate filaments are major contributors to cell architecture[35]. As genome function relies on the genomic spatial architecture[36], we hypothesized that lamin B1 could help to shape the new mesenchymal genome architecture. Genome organization and architecture was thus analyzed by Hi-C experiments[37] during EMT transformation. Sequencing Hi-C libraries from untreated cells or those treated with TGF-β (for 8 or 24 h) resulted in 89–106 M valid interactions per time point (Supplementary Table 2), which allowed us to generate chromosome-wide interaction maps at 100 and 40 Kb resolution for A/B (activated/repressed) compartments and TAD analyses, respectively. Hi-C maps showed overall compartment changes between untreated and 8 h-treated cells, with about 2400 bins of 100 Kb that changed compartments (Fig. 3a). These structural changes were further reinforced after a 24 h TGF-β treatment, with 2228 bins of EMT-related genes changing from B to A and 942

Fig. 3 Definition of eLADs. **a** UCSC Genome Browser overview of one region across the chromosome 11 (mm9) containing the following information (from top to bottom): lamin B1 ChIP-seq profiles subtracting the IgG control in untreated NMuMG cells (blue) or in cells treated with TGF-β for 8 h (orange) or 24 h (red); eLADs identified in each condition (black); previously published LADs from mouse embryonic fibroblast cells (green); and RNA-seq strand-specific expression profiles at each time point (red for untreated NMuMG cells, green for cells treated with TGF-β for 8 h, and blue for cells treated for 24 h). **b** Gene expression within eLADs (given in FPKM) in NMuMG cells that were untreated (blue) or treated with TGF-β for 8 h (orange) or 24 h (red), as compared with the rest of genes (white) (left). Promoter ATAC-seq enrichment of lamin B1 + genes (measured in number of normalized reads) in NMuMG cells that were untreated (blue) or treated with TGF-β for 8 h (orange) or 24 h (red), as compared with the rest of genes (white) (right). Genes from eLADs were more likely to be expressed and located in accessible chromatin than other genes. The bottom and top fractions in the boxes represent the first and third quartiles, and the line, the median. Whiskers denote the interval between 1.5 times the interquartile range (IQR) and the median. **c** Bar plot showing the overlap between eLADs and lamin A/C LADs obtained after low or high sonication (12 cycles or 30 cycles, respectively) and with two different antibodies (3A6 and N18)

changing from A to B (Fig. 4a, Supplementary Fig. 4). We next sought to analyze to which extent the dynamic changes of lamin B1 + sites and eLADs were related to genome architecture (i.e., in A/B compartments and at TAD borders). As expected, transcription, ATAC signal, lamin B1 + sites, and eLADs were enriched in A compartment (Fig. 3b), whereas cLADs were located in B compartments. In addition, we observed that lamin B1 + sites and eLADs decreased in A compartments and increased in B compartments during EMT (Fig. 4b); these corresponded to newly formed eLADs (Supplementary Table 3).

As borders between TADs are enriched in TSS, located in transcriptional active genomic regions, and enriched in architectural proteins[38–40], we analyzed TAD border behavior at the onset of the EMT and their relation to lamin B1. We observed that, although TAD borders were conserved during EMT transformation (with 50% of all borders conserved over the three time points), they increased in overall strength with longer treatment, indicative of more stable borders and increased intra-TAD interactions (Fig. 4c, left panel); this may reflect novel and biologically relevant chromatin interactions within TADs[36,41]. As changes in border strength could be due to changes in gene transcription[41], we checked the transcription rates after treatment at the TAD borders that had increased levels of lamin B1. Intriguingly, transcription was maintained in these TAD borders during EMT, suggesting that the increase in border strength is not due to changes in transcription (Fig. 4c, right panel). The increased number of lamin B1 + sites after TGF-β treatment (Fig. 2a) correlated with an increase in the percentage of lamin B1-containing borders, enhanced TAD border strength, and an enrichment of lamin B1 at TAD borders (with p-values of 10^{-28}, 10^{-68}, and 10^{-32} in untreated, 8 h-treated, and in 24 h-treated cells, respectively) (Fig. 4d). Overall, these results suggested that lamin B1 could contribute to 3D genome organization during EMT.

Altered lamin B1 levels impair EMT. Finally, we assessed the functional relevance of lamin B1 in a KD model, using NMuMG cells. It is noteworthy that the KD was kept at about 50% efficiency to avoid indirect effects (Fig. 5a). Given the low protein turnover of nuclear lamins once these are stably integrated in the NL[42], we reasoned that knocking down lamin B1 would mainly affect non- or less-NL-integrated lamin B1. Nuclei from control and KD NMuMG cells were extracted, and soluble and loosely bound chromatin proteins were separated from the insoluble chromatin fraction. Importantly, the soluble fraction of lamin B1 was affected to a greater extent in the KD cells as compared with control cells (Fig. 5b). To further confirm that the KD affected mainly the nucleoplasmic (or less-NL-integrated) lamin B1 fraction rather than the high-NL fraction, we used fluorescence recovery after photobleaching (FRAP). NMuMG cells transfected with human mCerulean-lamin B1 were subjected to short hairpin RNA (shRNA) lentivirus infection (with a control or human lamin B1-specific shRNA). Although recovery from bleaching was similar for the NL mCerulean-lamin B1 in both control and KD conditions, the nucleoplasmic fraction of mCerulean-lamin B1 recovered faster in the control cells than in the KD cells (Fig. 5c, left panel). Quantification data showed the mean recoveries for ten cells (Fig. 5c, right panel). Indeed, lamin B1 occupancy was reduced by 50% in lamin B1 + sites in KD cells, as shown by ChIP-qPCR (Fig. 4d), whereas cLADs were not affected (Fig. 5e). Under these conditions, KD cells had the same growth rate as control cells (Supplementary Fig. 5a) did not have a disrupted NL (Supplementary Fig. 5b) and did not enter into senescence (Supplementary Fig. 5c), all traits related with a massive loss of lamin B1[43].

Next, we analyzed the transcription profile of lamin B1 KD NMuMG cells treated with TGF-β during EMT by RNA-seq at each time point (Fig. 6). Principle component analysis revealed that KD cells at 8 h TGF-β were the most highly divergent from the control cells (Fig. 6a). Importantly, this is the time point in where we observed the highest number of newly formed lamin B1 + regions. Further, differentially expressed genes showed again a maximum dependence of lamin B1 at the 8 h TGF-β time point, with more than 2000 genes differentially expressed as compared with control cells (Fig. 6b). Strikingly, around 50% of these genes were direct lamin B1 targets (bar colors, Fig. 6b). Moreover, the differentially expressed genes that were upregulated in KD conditions did not belong to the EMT pathway, which completely changed the EMT transcriptional program (Fig. 6c, Supplementary Fig. 6). Indeed, western blot data showed alterations in classical EMT markers, a lack of fibronectin, N-cadherin upregulation, E-cadherin downregulation, and a loss of migrative and invasive capacities normally acquired during EMT (Fig. 6d, e). Finally, ChIP-qPCR in LB1 + EMT genes at 8 h TGF-β (such as fibronectin, vimentin, and twist) showed loss of lamin B1 enrichment in KD conditions as compared with control cells (Fig. 6f). Importantly, overexpressing human GFP-lamin B1 in KD cells restored the migratory capacity of mesenchymal cells (Fig. 6g).

Discussion

The presence of lamins in the nuclear interior has been long known but considered to be a transient pool on the way to assembly into the NL. Although B-type lamins remain permanently farnesylated and carboxymethylated (and therefore tightly associated to membranes[15]), they have also been reported to localize in the nuclear interior[44]. Polymerized lamins A and B exists within the nucleoplasm[6,16,17,45], but their specific organization, state of polymerization, solubility, and degree of integrity within the NL remains to be determined. Lamins are known to bind DNA, histones, and histone-binding proteins[46,47], and recent reports have also shown that lamin A/C bind to euchromatin regions, with an important role in gene regulation[18]. It is thus conceivable that all type of lamins in the

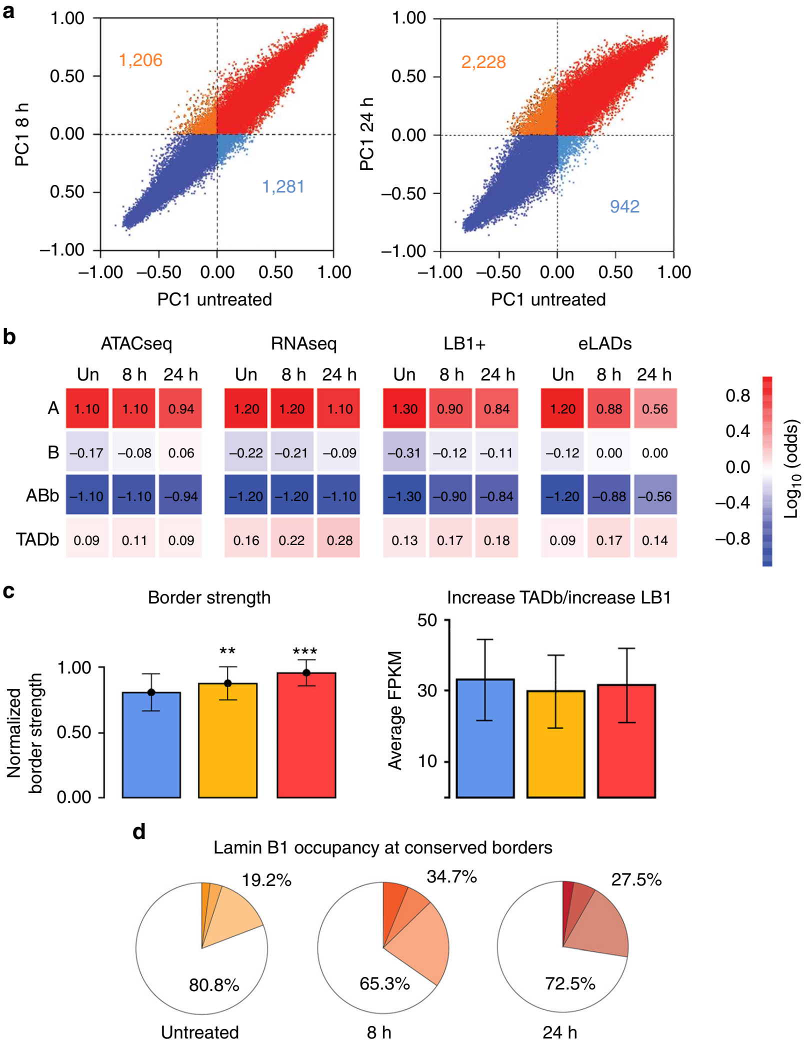


Fig. 4 Putative role of lamin B1 as an architectural protein. **a** A/B compartment dynamics. Scatterplots of PC1 values indicate compartment changes between untreated cells and cells treated with TGF-β for 8 or 24 h. Dots indicate 100 Kb bins that changed compartment with respect to untreated cells (from B to A, in orange; from A to B, in cyan). Numbers indicate the total bins that changed compartment. **b** Contingency table analysis for enriched and depleted features according to distinct genome architectural landmarks. Red to blue color indicate positive to negative log odds of the feature in genomic bins assigned to A or B compartment, or TAD borders. All log odds are statistically different than zero ($p < 0.05$, Fisher's exact *t*-test). **c** TAD border dynamics. Bar plot of normalized border strength for conserved borders in untreated cells (blue) or those treated with TGF-β for 8 h (orange) or 24 h (red) (left). Expression of genes within TAD borders in which lamin B1 increased (computed in FPKM) in untreated cells or those treated with TGF-β for 8 or 24 h (right). **d** Border occupancy by lamin B1 in conserved TADs. Pie charts of 1, 2, or ≥ 3 called peaks (dark to light color) of lamin B1 ChIP-seq in a conserved border

nuclear interior contribute to chromatin organization, although little is known about their potential roles in gene regulation. Lamins A/C have a functional role accommodating chromatin environment for gene regulation[18]. We now present evidence that the presence of lamin B1 in euchromatin is dynamic and has a role in the execution of the EMT transcriptional program, suggesting that lamin B1 also has a crucial role in gene regulation. We believe that this role is not exclusive to EMT. In fact, the GO categories found in genes enriched in lamin B1 in the epithelial state belong to general transcription, which suggests a general role of lamin B1 in gene regulation in response to a given stimulus.

Given that intermediate filaments are the major contributors of cell architecture[35] and, specifically, that nuclear lamins provide structural stability to the nucleus[48,49] and participate in multiple nuclear activities, we propose a role for lamin B1 in helping the 3D chromatin rearrangements occurring during EMT (Fig. 7). Hi-C data in combination with lamin B1 ChIP-seq showed that eLADs are present in the A compartment during the entire process, and that there is a significant increase of eLADs in the B

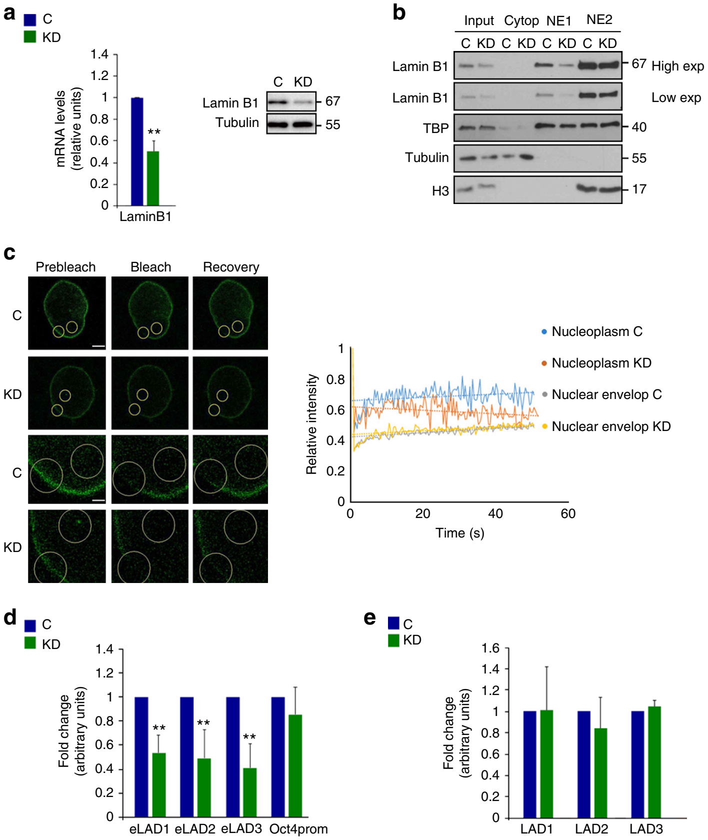


Fig. 5 Knockdown of lamin B1 mainly affects non-NL integrated lamin B1. **a** qRT-PCR showing mRNA levels of lamin B1 in KD NMuMG cells. Expression levels were normalized to an endogenous control and expressed relative to the control-infected cells (C), which was set as 1. Data are presented as mean ± SD, $n = 5$. Left panel, representative western blotting for LB1 and tubulin (as a loading control) from C and KD NMuMG cells. Images are representative of three independent replicates (right). **b** Fractionation of nuclei from C and KD NMuMG cells to obtain cytoplasmic proteins, soluble proteins (NE1), and insoluble proteins (NE2), as monitored by western blotting using TBP as a NE1 control, tubulin as a cytoplasmic control, and H3 as a NE2 control. Images are representative of two independent replicates and five technical replicates. **c** Confocal fluorescence recovery after bleaching (FRAP) in C and KD cells expressing lamin B1-mCerulean. Circles denote the areas of bleaching. Each column displays prebleach, bleach, and recovery images. Rows 3 and 4 are magnifications of rows 1 and 2, respectively. Scale bar: 5 μm (inset 1 μm) (left). Images are representative of ten cells for each condition. Graphic representation of the relative intensity after 60 s of nuclear envelope and nucleoplasm recovery from bleaching in C and KD cells (right). Data are shown as mean, $n = 6$. **d** ChIP-qPCR of three selected lamin B1 + regions and a negative control (Oct4prom) selected from a negative lamin B1 + region, in C and KD NMuMG cells. Data from real-time PCR (qPCR) amplifications were normalized to the input and expressed as the fold-change relative to data obtained in C condition, which was set as 1. Data are shown as mean ± SD, $n = 5$. **e** ChIP-qPCR of three cLAD-selected regions in C and KD NMuMG cells. Data from qPCR amplifications were normalized to the input and expressed as the fold-change relative to data obtained in C condition, which was set as 1. Data presented as mean ± SD, $n = 3$

compartment over time during EMT. Data analyses showed that this increase in the B compartment corresponds to newly formed eLADs, rather than simply reflecting a movement of eLADs from A to B. Interestingly, genes present in these new eLADs have to be in a repressive state when cells become mesenchymal, further supporting this function as a chromatin organizer. Moreover, the correlation of lamin B1 enrichment and border strength suggests a possible role of lamin B1 as an architectural protein that has a critical role in the establishment of a new genomic architecture during EMT (Fig. 7). The increase in border strength might be a consequence of transcriptional changes that occur during EMT. However, we did not observe dramatic transcriptional changes at TAD borders enriched in lamin B1, suggesting that transcription is not the main cause of these changes. Our results are in accordance with other models that have shown that redistribution of architectural proteins is responsible for establishing new genome interactions[41]. We cannot rule out that other proteins are also involved in establishing long-range interactions. Notably, RNA polymerase II and/or its associated—but not the process of transcription per se—have been identified as mediators of interactions throughout the A compartment[50,51].

Hundreds of proteins are able to interact with lamins[52]. As different tissues express different sets of lamina-associated proteins[53], the organization of lamina filaments can potentially vary significantly among tissues. Our results show that, once the EMT program is activated, new lamin B1 sites are formed that are enriched in TF-binding motifs belonging to developmental programs, including the TGF-β signaling pathway, suggesting a

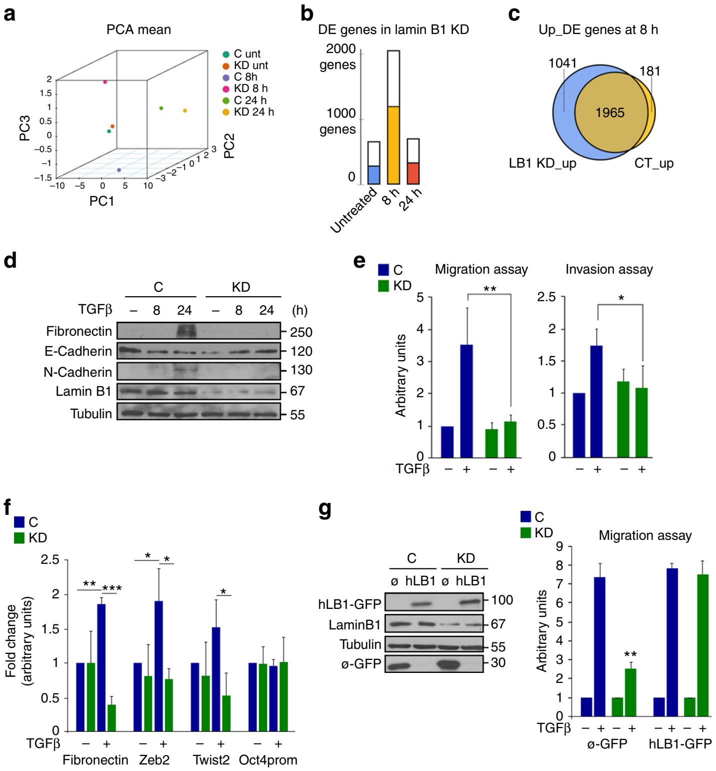


Fig. 6 Lamin B1 is essential for EMT. **a** Principle component analysis (PCA) plot of RNA-seq based expression profiling in KD conditions. Each time point was normalized to its respective control. **b** Number of genes that were differentially expressed (DE) upon lamin B1 knockdown (KD) during the EMT process. Inside each bar, the proportion of genes that are direct targets of lamin B1 in the ChIP-seq experiment of NMuMG cells untreated (blue) or treated with TGF-β for 8 h (orange) or 24 h (red) are shown. **c** Venn diagram representing the overlap between upregulated genes after 8 h of TGF-β treatment in KD and C conditions. **d** Western blotting showing protein levels of different EMT markers and lamin B1 in C and KD NMuMG cells that were untreated or treated with TGF-β for 8 or 24 h. Tubulin was used as a loading control. Images are representative of three independent replicates. **e** Migration (left panel) and invasion (right panel) assays performed with C and KD NMuMG cells after 24 h of TGF-β treatment. Data are presented as mean ± SD, $n = 3$. **f** Chip-qPCR of three selected EMT-related lamin B1+ regions at 8 h after TGF-β treatment in C and KD conditions. One of the selected regions from Fig. 4d was used as a negative control (prOCT4). Data from qPCR were normalized to the input and expressed as the fold-change relative to the untreated C condition, which was set as 1. Data are shown as mean ± SD; $n = 3$. **g** Western blotting of overexpressed GFP-control or GFP-hLB (human LB1) from C or KD NMuMG cells. Tubulin was used as a loading control (left panel). Images are representative of three independent replicates. Migration assay of C and KD NMuMG cells transfected with either GFP-control or GFP-hLB (right panel). Data are shown as mean ± SD; $n = 2$

putative role for these TFs in recruiting lamin B1 (although further experiments will be required to demonstrate this). The apparently distinct configuration between epithelial and mesenchymal cell types, together with the fact that different antibodies were used, might explain why Gesson et al.[18] did not detect lamin B1 in euchromatin regions, although they were also working with low sonication conditions.

Finally, it is still unclear why the presence of lamins (A/C and B1) in contact with euchromatin regions cannot be detected by the DamID method. Notably, DamID was recently proposed to favor the identification of stable interactions[21]. DamID methylation is modulated by the local chromatin structure, which means that open chromatin regions (euchromatin) may reach similar or even higher levels than the sites of specific methylation. Correction of the methylation levels by Dam unfused proteins could be responsible for a loss of lamin-euchromatin signal, and we now know that there is a more stable lamin pool in the NL that will always provide stronger signal.

To summarize, although lamin B1 filaments in contact with repressed chromatin are mainly located in the B compartment, where they form cLADs[4], our present work also shows that transcriptionally active chromatin in the A compartment can interact with lamin B1, to form eLADs. This is consistent with previous models showing lamin A/C also contact euchromatin[18], and with the fact that these two filaments form an interconnected meshwork[12,13]. Importantly, these two types of lamin B1 domains —cLADs and eLADs—behave differently: while cLADs are static, eLADs are dynamic and change during EMT. Moreover, the correlation of lamin B1 enrichment and border strength suggests a possible participation of these domains in establishing new genomic architecture during EMT (Fig. 7). Many questions still remain to be addressed: is the molecular structure of lamin B filament networks distinct in cLADs and eLADs? Is this lamin B1 pool soluble, or is it still membrane-bound but less stably integrated (and therefore more dynamic)? Are eLADs universal? Do they have a role in all types of cellular transformation processes?

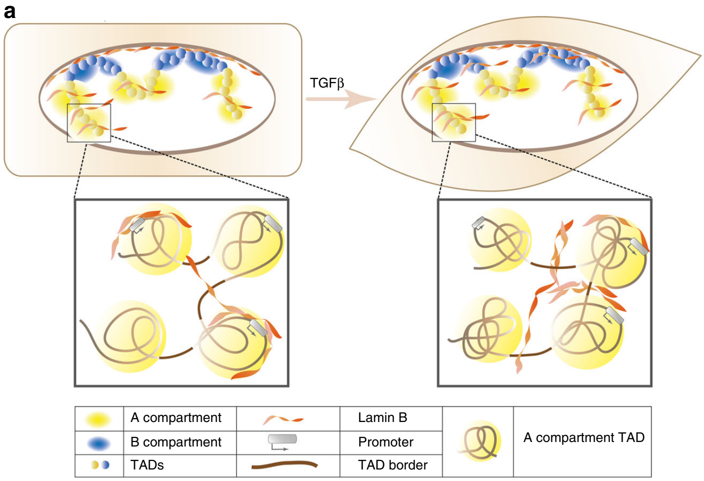


Fig. 7 Schematic representation of the dynamism of eLADs during the EMT process

Which are the signals and/or TFs responsible for recruiting lamin B1 to different loci? Are the changes in interactions the cause or consequence of lamin B1 enrichment at TAD borders? Answers to these questions will contribute to our understanding of how genome is reorganized during cellular transformation, and the role of euchromatin-lamin contacts in this re-organization and, therefore, in gene regulation.

Methods

Nuclear fractionation. Nuclear fractionation of infected NMuMG cells with either irrelevant shRNA or LB1 shRNA was performed using a Nuclear Extract Kit (ab219177, Abcam).

NMuMG cells were seeded in p150 plates and the assay was performed at 4 °C following manufacturer's instructions.

Cytochemical staining for *SA-β-gal*. Cytochemical staining for senescence-associated galactosidase (SA-β-gal) was performed using a Senescence β-Galactosidase Staining Kit (Cell Signaling Technology) at pH 6.0.

NMuMG cells infected and selected with puromycin were seeded in six-well plates, and 48 h after the assay was performed following manufacturer's instructions. In brief, culture medium was removed from the plates and two washes with phosphate-buffered saline (PBS) were performed. Cells were then fixed with Fixative solution provided with the Kit during 15 min at room temperature. Two more washes with PBS were performed before the addition of β-Galactosidase staining solution. Plates were sealed with parafilm and incubated overnight at 37 °C in a dry incubator. To avoid evaporation, corners of the plates were filled with PBS.

ChIP experiments. ChIP experiments were performed as described[54]. In brief, NMuMG cells were crosslinked in 1% formaldehyde for 5–10 min at 37 °C. Crosslinking was stopped by adding glycine to a final concentration of 0.125 M for 2 min at room temperature. For nuclear fractions, cells were scraped with cold soft-lysis buffer (50 mM Tris-HCl, 10 mM EDTA, 0.1% NP-40, and 10% glycerol) supplemented with protease inhibitors. Samples were then centrifuged at 800 × *g* for 15 min and the nuclei pellets were lysed with SDS-lysis buffer (1% SDS, 10 mM EDTA, and 50 mM Tris pH 8) supplemented with protease inhibitors. Extracts were sonicated to generate 200–600 bp DNA fragments, incubated on ice for 20 min, centrifuged at 16,000 × *g* for 10 min, and then diluted 1:10 with dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8 and 167 mM NaCl). For ChIP analysis, the primary antibody or an irrelevant antibody (IgG) was added to the sample, and the mixture was incubated overnight with rotation at 4 °C. Chromatin bound to the antibody was then immunoprecipitated using unblocked protein A beads (Diagenode) for 3 h with rotation at 4 °C. Precipitated samples were then washed three times with low-salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl) and with high-salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), and twice with LiCl buffer (250 mM LiCl, 1% Nonidet P-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0) using columns. To verify antibody specificity, samples were eluted with 2 × western blotting loading buffer for 5 min at 95 °C. Proteins were separated by SDS–polyacrylamide gel electrophoresis and probed with the anti-lamin B1 antibody. For qPCR detection of genomic regions of ChIP-sequencing analysis, washed samples were treated with elution buffer (100 mM Na_2CO_3 and 1% SDS) for 1 h at 37 °C and then incubated at 65 °C overnight with the addition of a final concentration of 200 mM NaCl to reverse the formaldehyde crosslinking. After proteinase K solution (0.4 mg/ml proteinase K (Roche), 50 mM EDTA, 200 mM Tris-HCl pH 6.5) treatment for 1 h at 55 °C, DNA was purified with MinElute PCR purification kit (Qiagen) and eluted in nuclease-free water. Genomic regions were detected by qPCR. Primers used are listed in Supplementary Table 4 (Supplementary Table 4). Results were quantified relative to the input and the amount of irrelevant IgG immunoprecipitated in each condition.

For ChIP-seq analysis, two parallel ChIPs were performed and mixed after elution with nuclease-free water. The NEBNext Ultra DNA library Prep Kit for Illumina was used to prepare the libraries and samples were sequenced using Illumina HiSeq 2500 system.

MTT assays. Cells previously infected and selected for 48 h with puromycin were counted with Neubauer's Chamber and 10,000 cells/condition were seeded in 96-well plates by triplicate.

MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) assays were performed by adding 0.5 mg of MTT (Sigma) per mL of Dulbecco's modified Eagle's medium (DMEM) without fetal bovine serum (FBS) for 3 h at 37 °C to determine the percentage of viable cells. Cells were solubilized with dimethyl sulfoxide-isopropanol (1:4). The absorbance of insoluble formazan (purple) at 590 nm, which is proportional to the number of viable cells, was then determined. Cell viability was quantified during four consecutive days.

Migration and invasion. For migration experiments, control (Irrelevant shRNA) and KD (shRNA_LB1) NMuMG cells were treated with TGF-β. After 24 h, 50,000 cells were resuspended in DMEM 0.1% FBS–0.1% bovine serum albumin, reseeded on a transwell filter chamber (Costar 3422) and incubated for 6–8 h. For invasion assays, cells were placed in Matrigel-coated transwell filter (BD356234) and incubated for 12–16 h. In both cases, DMEM with 10% FBS was added to the lower chamber and used as a chemoattractant. Non-migrating and non-invading cells were removed from the upper surface of the membrane, whereas cells that adhered to the lower surface were fixed with paraformaldehyde 4% for 15 min. Nuclei were stained with PBS-DAPI (4′,6-diamidino-2-phenylindole) (0.25 μg/mL). DAPI-stained nuclei were counted in four fields per filter by ImageJ software.

Fluorescence recovery after photobleaching. NMuMG cells were first transfected with human mCerulean-Lamin B1-10, which was a gift from Michael Davidson (Addgene plasmid #55380)[55] using *Trans*IT-X2® Dynamic Delivery System (Mirus Bio LLC). After 8 h, transfected cells were infected using irrelevant shRNA and human shRNA LB1 and selected using puromycin (1 μg/mL) for 24 h. Transfected and infected cells were seeded in 35 mm MatTek dishes. After 24 h, Leica TCS SP5 confocal system with the LAS-AF application wizard was used to perform FRAP experiment. Cells were kept in a fully incubated (CO_2 and 37 °C) chamber, while imaging with a × 63, 1.4 objective and 458 nm laser line of the Argon laser for excitation. Selected nuclear areas were bleached for five times using maximum laser intensity and 100 frames after the photobleach were collected, with 370 ms intervals, using the minimal laser power required. Before photobleaching, 10 frames were collected as an internal control of the experiment. Signal recovery was measured by ImageJ software using pre-bleached (pb), background (bg), and non-bleached (nb) areas (regions of interest, ROIs) to normalize the data. For every time point, the data was normalized according to the formula: $(\mathrm{ROI_b} - \mathrm{ROI_{bg}})/(\mathrm{ROI_{nb}} - \mathrm{ROI_{bg}})$ / $(\mathrm{pbRO_{Ib}} - \mathrm{pbROI_{bg}})/(\mathrm{pbROI_{nb}} - (\mathrm{pbROI_{bg}})$[56].

Statistical analysis. Statistical significance was assessed using a two-tailed unpaired Student's *t*-test. The symbols *, ** and *** indicates significant differences with $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively.

RNA sequencing. RNA-seq experiments were performed with two biological replicates of NMuMG cells (with or without shRNA of either control or KD) that were untreated or treated with TGF-β for 8 or 24 h. After RNA extraction with GenElute$^{\mathrm{TM}}$ Mammalian Total RNA Miniprep Kit (Sigma-Aldrich), samples were sequenced using the Illumina HiSeq 2500 system.

ATAC sequencing. The ATAC experiment was performed as described[57]. NMuMG cells were either untreated or treated with TGF-β for 8 or 24 h, and then collected and treated with transposase Tn5 (Nextera DNA Library Preparation Kit, Illumina). DNA was purified using MinElute PCR Purification Kit (Qiagen). All samples were then amplified by PCR using NEBNextHigh-Fidelity 2× PCR Master Mix (New Englands Labs) with primers containing a barcode to generate libraries. DNA was again purified using MinElute PCR Purification kit and samples were sequenced using Illumina HiSeq 2500 system.

Hi-C experiments. Hi-C libraries were generated from NMuMG cells (that were untreated or treated with TGF-β for 8 or 24 h) according the previously published Hi-C protocol, with minor adaptations[58]. Five million cells were crosslinked with 1% formaldehyde for 10 min at room temperature. Before permeabilization, cells were treated for 5 min with trypsin to obtain single cells. DNA was digested with 400 units of Dpn II and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated in a small volume (in situ Hi-C). Libraries were generated independently in the three conditions (e.g., treated with TGF-β for 8 or 24 h, or untreated), controlled for quality, and sequenced on an Illumina HiSeq 2000 sequencer.

Data availability. Sequencing samples (raw data and processed files) are available at NCBI GEO under the accession code GSE96033.

Received: 17 August 2017 Accepted: 2 August 2018
Published online: 24 August 2018

References

1. Taberlay, P. C. et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* **26**, 719–731 (2016).
2. Stuurman, N., Heins, S. & Aebi, U. Nuclear lamins: their structure, assembly, and interactions. *J. Struct. Biol.* **122**, 42–66 (1998).
3. Aebi, U., Cohn, J., Buhle, L. & Gerace, L. The nuclear lamina is a meshwork of intermediate-type filaments. *Nature* **323**, 560–564 (1986).
4. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
5. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
6. Gruenbaum, Y. & Medalia, O. Lamins: the structure and protein complexes. *Curr. Opin. Cell Biol.* **32**, 7–12 (2015).
7. Pickersgill, H. et al. Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nat. Genet.* **38**, 1005–1014 (2006).
8. Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* **43**, 630–638 (2011).
9. Sadaie, M. et al. Redistribution of the Lamin B1 genomic binding profile affects rearrangement of heterochromatic domains and SAHF formation during senescence. *Genes Dev.* **27**, 1800–1808 (2013).
10. Shah, P. P. et al. Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes Dev.* **27**, 1787–1799 (2013).
11. van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780–791 (2017).
12. Shimi, T. et al. The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes Dev.* **22**, 3409–3421 (2008).
13. Shimi, T. et al. Structural organization of nuclear lamins A, C, B1, and B2 revealed by superresolution microscopy. *Mol. Biol. Cell.* **26**, 4075–4086 (2015).
14. Turgay, Y. et al. The molecular architecture of lamins in somatic cells. *Nature* **543**, 261–264 (2017).
15. Gerace, L. & Blobel, G. The nuclear envelope lamina is reversibly depolymerized during mitosis. *Cell* **19**, 277–287 (1980).
16. Moir, R. D. et al. Review: the dynamics of the nuclear lamins during the cell cycle-- relationship between structure and function. *J. Struct. Biol.* **129**, 324–334 (2000).
17. Gruenbaum, Y. & Foisner, R. Lamins: nuclear intermediate filament proteins with fundamental functions in nuclear mechanics and genome regulation. *Annu. Rev. Biochem.* **84**, 131–164 (2015).
18. Gesson, K. et al. A-type lamins bind both hetero- and euchromatin, the latter being regulated by lamina-associated polypeptide 2 alpha. *Genome Res.* **26**, 462–473 (2016).
19. Lund, E. et al. Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes. *Genome Res.* **23**, 1580–1589 (2013).
20. Oldenburg, A. R. & Collas, P. Mapping nuclear lamin-genome interactions by chromatin immunoprecipitation of nuclear lamins. *Methods Mol. Biol.* **1411**, 315–324 (2016).
21. Naetar, N., Ferraioli, S. & Foisner, R. Lamins in the nuclear interior - life outside the lamina. *J. Cell Sci.* **130**, 2087–2096 (2017).
22. Thiery, J. P. & Sleeman, J. P. Complex networks orchestrate epithelial-mesenchymal transitions. *Nat. Rev. Mol. Cell Biol.* **7**, 131–142 (2006).
23. Nieto, M. A. The ins and outs of the epithelial to mesenchymal transition in health and disease. *Annu. Rev. Cell. Dev. Biol.* **27**, 347–376 (2011).
24. Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev. Cell.* **14**, 818–829 (2008).
25. Millanes-Romero, A. et al. Regulation of heterochromatin transcription by Snail1/LOXL2 during epithelial-to-mesenchymal transition. *Mol. Cell* **52**, 746–757 (2013).
26. McDonald, O. G., Wu, H., Timp, W., Doi, A. & Feinberg, A. P. Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nat. Struct. Mol. Biol.* **18**, 867–874 (2011).
27. Miettinen, P. J., Ebner, R., Lopez, A. R. & Derynck, R. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J. Cell Biol.* **127**, 2021–2036 (1994).
28. Manilal, S., Nguyen, T. M., Sewry, C. A. & Morris, G. E. The Emery-Dreifuss muscular dystrophy protein, emerin, is a nuclear membrane protein. *Hum. Mol. Genet.* **5**, 801–808 (1996).
29. Lund, E. G., Duband-Goulet, I., Oldenburg, A., Buendia, B. & Collas, P. Distinct features of lamin A-interacting chromatin domains mapped by ChIP-sequencing from sonicated or micrococcal nuclease-digested chromatin. *Nucleus* **6**, 30–39 (2015).
30. Frenster, J. H., Allfrey, V. G. & Mirsky, A. E. Repressed and active chromatin isolated from interphase lymphocytes. *Proc. Natl Acad. Sci. USA* **50**, 1026–1032 (1963).
31. Meuleman, W. et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* **23**, 270–280 (2013).
32. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
33. Hill, C. S. Transcriptional Control by the SMADs. *Cold Spring Harb. Perspect. Biol.* **8** https://doi.org/10.1101/cshperspect.a022079 (2016).
34. Seoane, J. & Gomis, R. R. TGF-beta family signaling in tumor suppression and cancer progression. *Cold Spring Harb Perspect Biol* **9**, pii: a022277 https://doi.org/10.1101/cshperspect.a022277 (2017).
35. Lowery, J., Kuczmarski, E. R., Herrmann, H. & Goldman, R. D. Intermediate filaments play a pivotal role in regulating cell architecture and function. *J. Biol. Chem.* **290**, 17145–17153 (2015).
36. Rowley, M. J. & Corces, V. G. The three-dimensional genome: principles and roles of long-distance interactions. *Curr. Opin. Cell Biol.* **40**, 8–14 (2016).
37. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
38. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
39. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).
40. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
41. Li, L. et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
42. Broers, J. L. et al. Dynamics of the nuclear lamina as monitored by GFP-tagged A-type lamins. *J. Cell Sci.* **112**(Pt 20), 3463–3475 (1999).

43. Shimi, T. et al. The role of nuclear lamin B1 in cell proliferation and senescence. *Genes Dev.* **25**, 2579–2593 (2011).
44. Dechat, T., Gesson, K. & Foisner, R. Lamina-independent lamins in the nuclear interior serve important functions. *Cold Spring Harb. Symp. Quant. Biol.* **75**, 533–543 (2010).
45. Moir, R. D., Yoon, M., Khuon, S. & Goldman, R. D. Nuclear lamins A and B1: different pathways of assembly during nuclear envelope formation in living cells. *J. Cell Biol.* **151**, 1155–1168 (2000).
46. Dechat, T., Adam, S. A. & Goldman, R. D. Nuclear lamins and chromatin: when structure meets function. *Adv. Enzym. Regul.* **49**, 157–166 (2009).
47. Pegoraro, G. et al. Ageing-related chromatin defects through loss of the NURD complex. *Nat. Cell Biol.* **11**, 1261–1267 (2009).
48. Dahl, K. N., Engler, A. J., Pajerowski, J. D. & Discher, D. E. Power-law rheology of isolated nuclei with deformation mapping of nuclear substructures. *Biophys. J.* **89**, 2855–2864 (2005).
49. Lammerding, J. et al. Lamin A/C deficiency causes defective nuclear mechanics and mechanotransduction. *J. Clin. Invest.* **113**, 370–378 (2004).
50. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852 e837 (2017).
51. Jung, Y. H. et al. Chromatin states in mouse sperm correlate with embryonic and adult regulatory landscapes. *Cell Rep.* **18**, 1366–1382 (2017).
52. de Leeuw, R., Gruenbaum, Y. & Medalia, O. Nuclear lamins: thin filaments with major functions. *Trends Cell Biol.* **28**, 34–45 https://doi.org/10.1016/j.tcb.2017.08.004 (2017).
53. de Las Heras, J. I. et al. Tissue-specific NETs alter genome organization and regulation even in a heterologous system. *Nucleus* **8**, 81–97 (2017).
54. Herranz, N. et al. Polycomb complex 2 is required for E-cadherin repression by the Snail1 transcription factor. *Mol. Cell. Biol.* **28**, 4772–4781 (2008).
55. Rizzo, M. A., Davidson, M. W. & Piston, D. W. Fluorescent protein tracking and detection: fluorescent protein structure and color variants. *Cold Spring Harb. Protoc.* **2009**, pdb top63 https://doi.org/10.1101/pdb.top63 (2009).
56. Nissim-Rafinia, M. & Meshorer, E. Photobleaching assays (FRAP & FLIP) to measure chromatin protein dynamicsin living embryonic stem cells. *J. Vis. Exp.* pii: 2696 https://doi.org/10.3791/2696 (2011).
57. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
58. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

Acknowledgements

We thank J. Querol for technical assistance and V.A. Raker for manuscript editing. This work was supported by grants from the Instituto de Salud Carlos III (ISCIII) FIS/FEDER (PI15/00396; CPII14/0006), Ministerio de Economía y Competitividad (MINECO) (SAF2013-40922-R1; FPU14/0407; BFU2016-75008-P), Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional-FEDER (SAF2016-76461-R), Generalitat de Catalunya (2014 SGR 32), Fundació FERO, Fundació La Marató TV3, and La Caixa Foundation. We also thank the Advanced Light Microscopy Unit at the CRG for their assistance and the Cellex Foundation for providing research facilities and equipment. M.A.M.-R. acknowledges funding from the European Research Council under the 7th Framework Program (FP7/2010-2015, ERC grant agreement 609989), the European Union's Horizon 2020 research and innovation program (agreement 676556), the Spanish Ministry of Economy and Competitiveness (BFU2013-47736-P), and the Centro de Excelencia Severo Ochoa 2013-2017 (SEV-2012-0208) to the CRG.

Author contributions

L.P.-R., V.D.C., A.I., G.S.-B., and J.P.C.-C. performed the experiments. E.B., M.A.M.-R., S.G., and L.N. analyzed the data with contributions from L.D.C, F.L.D., Y.C., and A.G.H. L.P.R. and S.P. designed the experiments and wrote the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05912-z.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.