CaTCHing the functional and structural properties of chromosome folding

Inauguraldissertation
Zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Yinxiu Zhan

von Italien

Basel, 2020

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch
Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung-Nicht kommerziell 4.0 International Lizenz

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Attila Becskei
Dr. Luca Giorgetti
Prof. Nils Blüthgen

Basel, 15/10/2019

Prof. Dr. Martin Spiess
(Dean of Faculty)

Table of Contents

List of abbreviations......2

Summary......3

1. Introduction......5

1.1 Transcriptional control......5

1.2 First order of genome folding: nucleosomes......7

2. Methods to study higher order genome folding......8

2.1 The microscopy era......9

2.2 The 3C era......9

2.3 The hierarchy of chromatin folding......12

2.4 Chromatin conformation and promoter-enhancer communication......14

2.5 Models of chromatin folding......15

3. Aim of the thesis......16

4. Results......18

Chapter I: Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes......18

Chapter II: DamC reveals principles of chromatin folding in vivo without crosslinking and ligation......48

Chapter III: Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers......83

Discussion......94

Acknowledgement......98

References......99

List of abbreviations

3C	Chromosome conformation capture
3D	Three-dimensional
4C	Chromosome conformation capture-on-chip
5C	Chromosome conformation capture carbon copy
CaTCH	Caller of Topological Chromosomal Hierarchies
CTCF	CCCTC-binding factor
Dam	Deoxyadenosine methylase
DNA	DeoxyriboNucleic Acid
E. coli	Escherichia coli
ESC	Embryonic stem cell
FISH	Fluorescence in situ hybridization
mRNA	messenger RNA
H3K27ac	Acetylation on lysine 27 of histone 3
H3K9ac	Acetylation on lysine 9 of histone 3
H3K4	Histone 3 lysine 4
H3K36	Histone 3 lysine 36
H3K27	Histone 3 lysine 27
Hbb	β-globin gene
HP1	Heterochromatin-like protein 1
LAD	Lamin associated domain
NPC	Neural progenitor cell
PcG	Polycomb-group
PTM	Post-translationally modification
RI	Reciprocal insulation
RNA	RiboNucleic Acid
Shh	Sonic hedgehog
SINE	Short interspersed nuclear elements
SNP	Single nucleotide polymorphism
TAD	Topologically associating domain
TF	Transcription factor
TSS	Transcription start sites

Summary

Proper development requires that genes are expressed at the right time, in the right tissue, and at the right transcriptional level. In metazoans, this involves long-range *cis*-regulatory elements such as enhancers, which can be located up to hundreds of kilobases away from their target promoters. How enhancers find their target genes and avoid aberrant interactions with non-target genes is currently under intense investigations. The predominant model for enhancer function involves its direct physical looping between the enhancer and target promoter. The three-dimensional organization of chromatin, which accommodates promoter-enhancer interactions, therefore might play an important role in the specificity of these interactions. In the last decade, the development of a class of techniques called *chromosome conformation capture* (3C) and its derivatives have revolutionized the field of chromatin folding. In particular, the genome-wide version of 3C, Hi-C, revealed that mammalian chromosomes possess a rich hierarchy of folding layers, from multi-megabase compartments corresponding to mutually exclusive associations of active and inactive chromatin to topologically associating domains (TADs), which reflect regions with preferential internal interactions. Although the mechanisms that give rise to this hierarchy are still poorly understood, there is increasing evidence to suggest that TADs represent fundamental functional units for establishing the correct pattern of enhancer-promoter interactions. This is thought to occur through two complementary mechanisms: on the one hand, TADs are thought to increase the chances that regulatory elements meet each other by confining them within the same domain; on the other hand, by segregation of physical interactions across the boundary to avoid unwanted events to occur frequently.

It is however unclear whether the properties that have been attributed to TADs are specific to TADs, or rather common features among the whole hierarchy. To address this question, I have implemented an algorithm named *Caller of Topological Chromosomal Hierarchies* (CaTCH). CaTCH is able to detect nested hierarchies of domains, allowing a comprehensive analysis of structural and functional properties across the folding hierarchy. By applying CaTCH to published Hi-C data in mouse embryonic stem cells (ESCs) and neural progenitor cells (NPCs), I showed that TADs emerge as a functionally privileged scale. In particular, TADs appear to be the scale where accumulation of CTCF at domain boundaries and transcriptional co-regulation during differentiation is maximal. Moreover, TADs appear to be the folding scale where the partitioning of interactions within transcriptionally active domains (and notably between active enhancers and promoters) is optimized.

3C-based methods have enabled fundamental discoveries such as the existence of TADs and CTCF-mediated chromatin loops. 3C methods detect chromatin interactions as ligation products after crosslinking the DNA. Crosslinking and ligation have been often criticized as potential sources of experimental biases, raising the question of whether TADs and CTCF-mediated chromatin loops actually exist in living cells. To address this, in collaboration with Josef Redolfi, we developed a new method termed 'DamC' which combines DNA methylation with physical modeling to detect chromosomal interactions in living cells, at the molecular scale, without relying on crosslinking and ligation. By applying DamC to mouse ESCs, we provide the first *in vivo* and crosslinking- and ligation-free validation of chromosomal structures detected by 3C-methods, namely TADs and CTCF-mediated chromatin loops.

DamC, together with 3C-based methods, thus have shown that mammalian chromosomes possess a rich hierarchy of folding layers. An important challenge in the field is to understand the mechanisms that drive the establishment these folding layers. In this sense, polymer physics represent a powerful tool to gain mechanistic insights into the hierarchical folding of mammalian chromosomes. In polymer models, the scaling of contact probability, i.e. the contact probability as a function of genomic distance, has been often used to benchmark polymer simulations and test alternative models. However, the scaling of contact probability is only one of the many properties that characterize polymer models raising the question of whether it would be enough to discriminate alternative polymer models. To address this, I have built finite-size heteropolymer models characterized by random interactions. I showed that finite-size effects, together with the heterogeneity of the interactions, are sufficient to reproduce the observed range of scaling of contact probability. This suggests that one should be careful in discriminating polymer models of chromatin folding based solely on the scaling.

In conclusion, my findings have contributed to achieve a better understanding of chromatin folding, which is essential to really understand how enhancers act on promoters. The comprehensive analyses using CaTCH have provided conceptually new insights into how the architectural functionality of TADs may be established. My work on heteropolymer models has highlighted the fact that one should be careful in using solely scaling to discriminate physical models for chromatin folding. Finally, the ability to detect TADs and chromatin loops using DamC represents a fundamental result since it provides the first orthogonal *in vivo* validation of chromosomal structures that had essentially relied on a single technology.

1. Introduction

Diversity in biology is really fascinating especially when we think that all the information is stored in a universal code, the DNA sequence. It is astonishing how a simple biological alphabet like that of the four nucleotides can give rise to such a large number of species, each of which is able to transmit the syntax so precisely to their progeny to produce a faithful copy of themselves.

Most species are made of a single cell, but some of them, like humans, are made of a lot (really a lot, 10^{13}!) of cells. In humans, there are more than 200 cell types, each of which has the same DNA sequence, the same genetic information. It is incredible that a single cell (totipotent cell) can give rise to all these different types of cells using the same potential genetic information. How is this achieved? The understanding of how the genetic information is accessed and used is fundamental to shed light on this question.

The carrier of the genetic information is a long double-helix molecule called ***D****eoxyriboNucleic Acid* (**DNA**). The DNA-fiber is made of four units called nucleotides. Each nucleotide shares a common backbone made of a five-carbon sugar and three phosphate groups. There can be four types of nitrogenous bases attached to the backbone, resulting in four types of nucleotides called Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). The mechanisms that describe how genetic information flows from DNA into functional proteins, which are the directors of the biochemical processes fundamental for survival of the cell and generation of progeny, represents the central dogma in molecular biology. At the time when it was coined, the central dogma defined the flow of genetic information as a two-steps process. In the first step, called transcription, DNA sequences are transcribed to produce ***R****iboNucleic Acid* (RNA) molecules; some of the RNA molecules (called messenger RNAs, or mRNAs) are then translated (translation) into proteins (the second step).
As our knowledge of molecular biology and, in particular, of gene control expanded, it became increasingly clear that the flow of genetic information is not as linear as initially thought, but instead depends on an intricate network of feedback and feedforward loops regulating both transcription and translation. Interestingly, despite the fact that transcription and translation of DNA into functional proteins are very conserved processes across species and cell types, these multi-layer mechanisms that control them vary hugely; these differences are the keys to the doors of diversity in biology.

1.1 Transcriptional control

If we were asked to say which is the more complex organism, the human or the fruit fly *Drosophila Melanogaster*, the answer would certainly be the human. But what makes humans a more sophisticated organism than drosophila? Is it the genome size, the number of genes, or something else? Looking in nature, we can see that neither the genome size, nor the number of genes correlates with organism complexity[1]. For instance, we can find similar species that differs up to eight-fold in genome size[2,3], which clearly provides evidence against the genome size hypothesis. What about the number of genes? Does the complexity of an organism scale linearly with the number of genes? When genome sequences became available at the beginning of the new millennium, it became clear that also the number of genes does not represent a good measure of complexity[4]. Indeed, with the assembly of the human genome, it was shown that less than 2% of the genome corresponds to protein coding regions, resulting in roughly 30.000 genes[5], which was in the same order of magnitude of the small flowering plant, Arabidopsis thaliana[4]. However, the Human genome project also showed that the human genome is mainly made up of repeats (around 50%) and protein binding DNA sequences that control the timing and the level of gene expression called "transcriptional regulatory sequences" (around 20%). The massive presence of transcriptional regulatory sequences reinforced the idea that the combinatorial networks of transcriptional control may be directly related to complexity[6]. Indeed, a small change in gene number or regulatory elements can potentially lead to an enormous increase in the number of possible interconnections in the gene-regulatory network[4].

The key concepts of transcriptional control were established in pioneering work in bacteria by Monod and colleagues where it was discovered that the binding of transcription factors to specific DNA sequences at control elements (cis-elements) plays a fundamental role in the recruitment and regulation of the transcriptional machinery[7]. Further studies in eukaryotic cells showed that an important class of these cis-elements, called enhancer elements, play a central role in the process of transcription in eukaryotes[8,9]. The importance of enhancers for normal development is highlighted by genome-wide association studies showing that disease-associated single nucleotide polymorphisms (SNPs) often co-localize with these noncoding regulatory sequences[10]. Furthermore, chromosomal rearrangements affecting the regulatory network of target genes were shown to be able to induce congenital diseases[11].

Enhancers are DNA elements that contain binding sites for transcription factors (TFs), whose combinatorial binding can lead to a precise pattern of transcriptional activity[9]. For example, the combinatorial binding of activators and repressors to enhancers of the specific class of genes called "gap" genes gives rise to the stripe patterns during the early segmentation of the Drosophila melanogaster embryo[12]. Different patterns of TF binding, triggered by transcriptional networks, environmental cues, together with the intrinsic stochasticity of biochemical reactions[13,14], can lead to the activation of alternative genetic programs which, in turn, give rise to different cell types. TF binding can be influenced by several factors: cooperativity between TFs, competition with antagonists such as histones (the proteins around which DNA is wrapped, see next section), sequence specificity, motif affinity. For example, protein-protein interactions can lead to a non-linear relationship between TF occupancy and concentration, typical of cooperative binding. Nucleosomes, the basic unit of DNA folding (see next section), can compete with TFs to access the DNA[15]. TF binding is of course essential for the enhancer activity, but it is not sufficient to regulate the gene transcription as shown in a recent study where only 10-25% of eukaryotic binding events were found to be functional[16].

Over the last decade, it has become more and more evident that chromatin folding in the nucleus plays a crucial role in enhancer activity. In vertebrates, predictions based on chromatin state (defined by the set of histone post-translational modifications (see below) and TF binding) as well as genetic experiments have shown that enhancers are often located tens or even hundreds of kilobases away from their target promoters[17,18]. This raised the questions of how enhancers find their target genes and avoid aberrant interaction with non-target genes, since very often they bypass more proximal genes to interact with their target genes[19,20]. The current predominant model for enhancer function involves the direct interactions between enhancer elements and the region of the gene where the transcriptional machinery is assembled, i.e. the promoter of the gene[21] (Figure 1). The three-dimensional organization of chromatin (the DNA and bound proteins complex), which accommodates promoter enhancer interactions, therefore might play an important role in the specificity of these interactions.

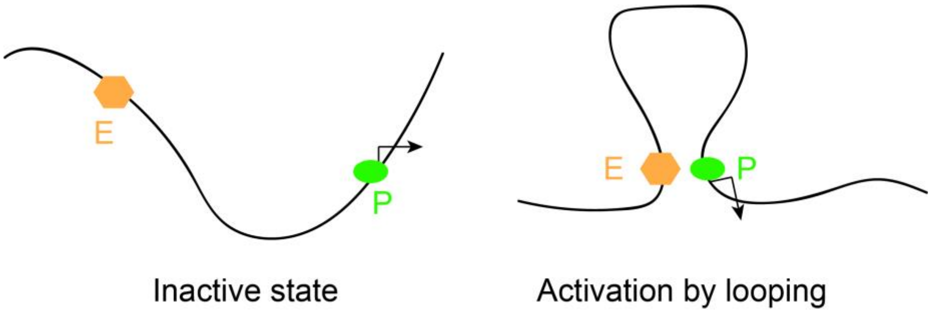


Figure 1: Looping model for promoter-enhancer interaction. On the left, the promoter is in an inactive state where there is no promoter-enhancer physical interaction. On the right, the promoter is activated by the enhancer physically looping to the promoter.

1.2 First order of genome folding: nucleosomes

The human genome, if laid end-to-end, would be approximately two meters long; yet the cell nucleus, where the genome is contained, is only few microns in diameter, roughly five orders of magnitude smaller. Packing such a long fiber in a such a small nucleus is a tremendously complex task and it is accomplished by a series of very specialized proteins that bind to the DNA and fold it. To make it even more complicated, the folding cannot be random since it has to allow the quick access of the right portion of the genome when needed. One of the first pieces of evidence of the structural compartmentalization of chromatin came from a study by Emil Heitz where he found that chromatin regions either stained dark (condensed) or light (decondensed) in the nucleus during interphase[22]. Even though it was not known what drove the differential condensation of chromatin, these findings already suggested that the folding of chromatin could not be stochastic and in particular different types of nuclear compartments exist. Later, it was shown that the lightly packed chromatin, termed as euchromatin, associates with gene-rich, transcriptionally active regions, while the tightly packed chromatin, called heterochromatin, corresponds to transcriptional repression and gene-poor regions[23].

The basic repeating unit of chromatin folding, the nucleosome, was discovered in 1974 by electron microscopy of chromatin obtained from interphase nuclei lysed in water[24]. Under the electron microscope, chromatin fibers appeared as arrays of spherical particles (nucleosomes) connected by filaments (linker DNA). Subsequent studies showed that each nucleosome core consists of ~147bp of DNA wrapped around an octamer of proteins called the histone[25], providing the basic unit of chromatin folding. The histone octamer consists of two copies each of the core histones H2A, H2B, H3, and H4; DNA wrapped around nucleosomes represents the first level in chromatin packaging, which effectively shortens the length of chromosomes by 7-fold[26]. Importantly, nucleosomes are not only means to compact DNA, but they also play a critical role in transcriptional regulation for instance by limiting the accessibility of the wrapped DNA[27] through their post-translational modifications (PTMs).

Histone modifications were discovered in the pioneering studies by Mirsky and colleagues in the early 1960s[28]. Histone PTMs are reversible: the enzymes that add the modifications are called `writers`, while the enzymes that remove the modifications are called `erasers`. Different histone PTMs are added/removed by different writers and erasers. Nowadays, many PTMs have been characterized including phosphorylation, ubiquitination, ADP-ribosylation and many others[29], but for the sake of brevity, I will discuss only lysine acetylation and methylation.

Lysine acetylation is the process where a negatively charged acetyl-group is covalently added to the lysine. This negative charge reduces the lysine's positive charge, weakening the interaction between DNA and histones, making DNA more accessible to functional proteins.

Indeed, acetylation is a PTM often associated with increases in DNA accessibility and transcriptional activity: examples of such acetylation are H3K27ac (acetylation on lysine 27 of histone 3) and H3K9ac.

Methylation instead is a process where a neutral methyl-group is added to the lysine, thus not altering the lysine charge. There are three lysine methylation states: mono-, di- and trimethylation, none of which changes the charge of histones. Unlike acetylation, which corresponds to active chromatin states, histone lysine methylations can confer active or repressive chromatin states depending on their positions and methylation states. For example, H3K4 and H3K36 methylation is found to mark active transcription, whereas H3K9 and H3K27 methylation is associated with silent chromatin states[30].

There are two main proposals for how histone PTMs can influence transcriptional activity. On the one hand, chromatin packing can be directly altered by changing the electrostatic interaction between histones and DNA through PTMs, thus, affecting the accessibility of DNA sequences to transcription factors; on the other hand, an increasing body of evidence suggests that histone PTMs can serve as binding surfaces for the association of effector proteins (‘readers’), such as chromatin remodelers, histone chaperones, DNA/histone-modifying enzymes and general transcription factors[31]. For example, it has been shown that the general transcription factor IID (TFIID) binds to H3K4me3 through its PHD domain-containing TAF3 subunit, resulting in more efficient preinitiation complex formation[32]. H3K9 methylation has been shown to promote transcriptional repression through the binding of the heterochromatin-like protein 1 (HP1) which, in turn, recruits chromatin condensation factors such as H3K9 methyltransferases and DNA methyltransferases[33]. Importantly the existence of histone modification readers led to the ‘histone code’ hypothesis, where specific histone tail modifications (a histone `language') serve to recruit other proteins. According to this hypothesis, the biological function of combinations of PTMs is mainly due to the protein complexes that recognize this code.

The histone code adds another layer of complexity that the cells can use to finetune their gene expression programs. In most cases, this code cannot be directly interpreted, as several histone modifications seem to have both a transcriptionally positive and a negative behavior depending on the genomic and regulatory context[34].

Transcriptional responses have been tightly linked with nucleosome organization, especially at promoters and enhancers, as nucleosomes have been classically thought to critically affect transcription factor binding[35,36]. PTMs play an important role in nucleosome organization by direct or indirect recruitment of chromatin remodelers which can modify chromatin accessibility for transcription factors[37]. This is not a unidirectional process, since TF binding can also lead to the recruitment of chromatin modifying enzymes, which, by adding or removing PTMs, can recruit other TFs. This complex network of feedback and feedforward loops between chromatin context (PTMs, chromatin accessibility) and TFs binding represents one of the fundamental mechanisms of transcriptional control. Recently, the higher order folding of chromatin, which accommodates interactions between different regulatory elements (such as the bridging between enhancers and promoters), started to emerge as another layer of control that cells have to finetune their expression programs. In the next section, I will focus on the methods used to study higher order chromatin folding.

2. Methods to study higher order genome folding

Our current view of genome folding is mainly based on two complementary classes of techniques. On the one hand, a variety of microscopy techniques, such as DNA *fluorescence in situ hybridization* (FISH), are currently used to directly visualize the proximity between

DNA segments. The power of DNA FISH lies in its ability to give single-cell information, while being limited in throughput and resolution. It is therefore unclear whether it uncovers general principles of nuclear organization or the behavior of specific individual genes. On the other hand, population based biochemical approaches such as 3C and its derivatives infer DNA proximity by quantifying the frequencies of contacts between DNA. 3C and its derivatives allow simultaneous detection of multiple and genome-wide chromosomal interactions, but they are limited to populations of cells and so do not provide single cell information. I will first review the fundamental discoveries in chromatin folding made by DNA FISH and then focus on the 'revolution' of 3C methods in studying chromatin folding.

2.1 The microscopy era

Until the development of biochemical techniques such as 3C and its high throughput derivatives, the main technique used to study genome folding was DNA FISH. DNA *In Situ Hybridization* is based on the concept that nucleotide sequences could hybridize to complementary sequences and form more stable complexes compared to sequences that were not complementary. DNA FISH, thus, relies on delivering complementary probes labeled with a fluorochrome to target genomic DNA. The fluorescently labeled regions can then be visualized using a fluorescence microscope. Key features in DNA FISH are sensitivity and resolution. Sensitivity refers to the ability of the microscope to detect weak signals, therefore determining the size of the probe you need (large probes give stronger signals leading to higher sensitivity). Sensitivity is directly linked to spatial (and therefore genomic) resolution, that is the ability to distinguish two genomic loci along the chromatin. Good sensitivity comes at the expense of resolution; thus, it is not surprising that FISH led to the discovery of low-resolution nuclear sub-structures, such as chromosome territories. The concept of chromosome territories goes back to the end of 19th century, when scientists started proposing the idea that chromosomes may occupy certain nonoverlapping areas of the nucleus; chromosome territories could be unequivocally detected only a century later with the development of DNA FISH. The painting of all the human chromosomes showed that chromosomes are largely confined to chromosome territories[38–40], which intermingle only to a limited extent.

Another fundamental discovery made by FISH was the functional positioning of genomic loci with respect to nuclear compartments (such as the nuclear periphery). Using fluorescence in situ hybridization, the lab of Wendy Bickmore showed gene-dense regions of the human genome are preferentially found in the nuclear interior, while the gene-poor regions are located progressively towards the nuclear periphery[41]. Thus, irrespective of its limitations, DNA FISH has nevertheless led to many fundamental discoveries, such as the existence of chromosomal territories and the preferential radial positioning of genomic loci within the nucleus. These discoveries already hinted towards a functional role of higher order chromatin folding that would later be confirmed by the advent of *chromosome conformation capture* methods.

2.2 The 3C era

The *Chromosome Conformation Capture (3C)* technique was invented by Job Dekker almost 20 years ago[42] to study the interactions of specific loci at high resolution. The technology relies on the simple idea that digestion and re-ligation fixed DNA, followed by the quantification of ligation junctions, could allow the quantification of chromosomal interaction frequencies. Briefly, in 3C a population of cells are treated with formaldehyde that creates covalent bond, thus 'freezing' the interactions between genomic loci (Figure 2).

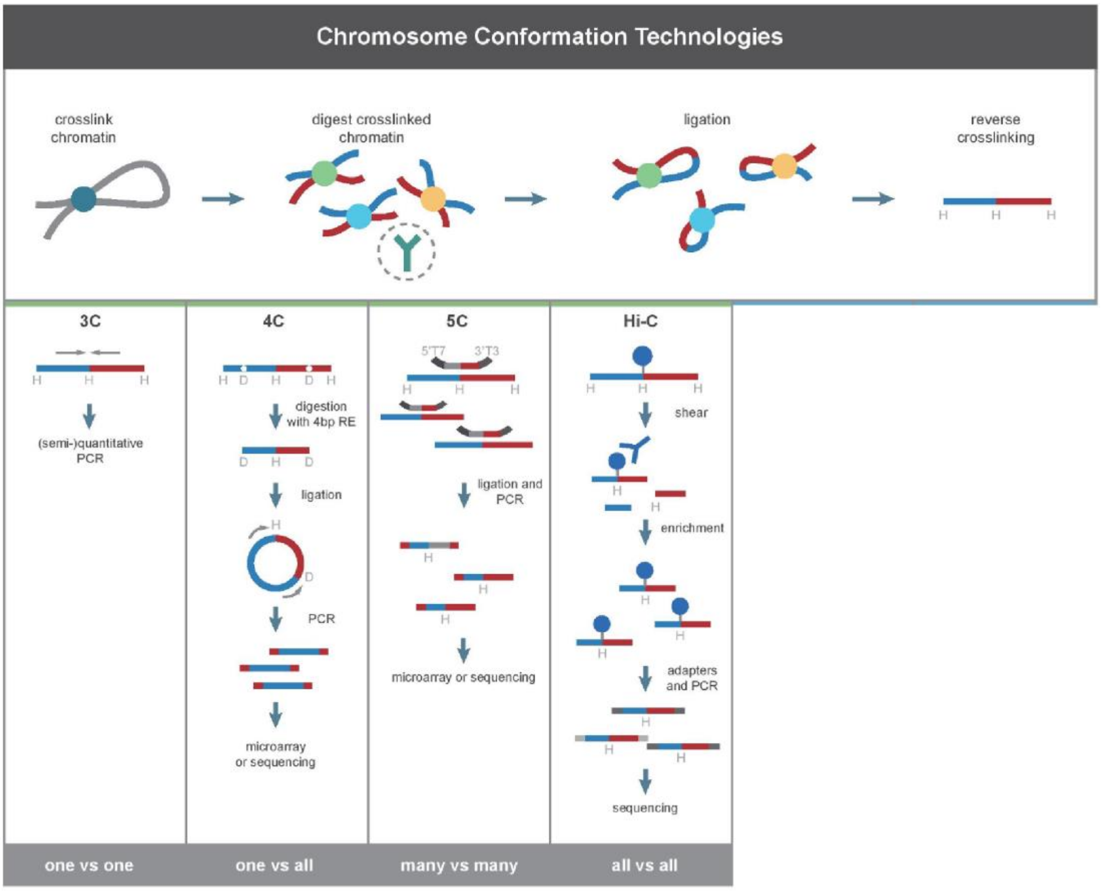


Figure 2: Overview of 3C methods (adapted from De Wit, de Laat, 2012)

The crosslinked DNA is then digested using restriction enzymes which cut across the genome at specific sites ('motifs'). The choice of the restriction enzyme dictates the resolution of the 3C experiment: an enzyme that recognize a motif of four base-pairs (4bp-cutter) gives a higher resolution than that of a 6bp-cutter since it cuts more frequently. The sticky ends of the digested fragments are then re-ligated in diluted conditions to favor ligation of cross-linked DNA fragments. Although proximity ligation had earlier been used to detect DNA interactions in non-crosslinked cells[43], a key step 3C was the introduction of formaldehyde cross-linking that boosted the efficiency and robustness of proximity ligation reactions. The quantification of a chromosomal interaction is made by measuring the number of ligation events. In 3C, this is done by polymerase chain reaction (PCR) amplification of selected ligation junctions ("one versus one"), thus, determining whether specific loci would interact more than others. In the original study, performing 3C in yeast revealed that chromosome 3 possesses a contorted ring structure[42]. Despite its limitation in high throughput, 3C has been instrumental in detecting promoter-enhancer specific interactions at the β-globin locus[44], as well as between regulatory sequences and genes at other loci[45]. However, the size of the genome and the related possible number interactions made the PCR based detection impracticable for large-scale mapping of chromosomal interactions.

Over the years, many additional modifications have been introduced to 3C to enhance the resolution and the detection efficiency of chromosomal interactions. Surfing on rapid advances in DNA sequencing technologies, 3C developed into genome-scale methods with the adoption of micro-array and high-throughput DNA sequencing as ways to measure the frequency of proximity ligation products.

The first variant of 3C, called *chromosome conformation capture-on-chip* (**4C**) was introduced in 2006[46]. In 4C, the ligated 3C template is further digested and re-ligated to create small circular pieces of DNA (Figure 2). By using primers for the fragment of interest (called the viewpoint), inverse PCR specifically amplifies all sequences ligated to this chromosomal site. The 4C library then can be analyzed by microarrays or by deep sequencing using Next Generation Sequencing (NGS) methods. 4C opened the door to studying interactions between a single locus (called the viewpoint) and the rest of the genome and thus it is known as a "one versus all" strategy to study chromosomal interactions. However, it was not suitable for studying the conformation of entire domains or chromosomes at high resolution.

The *chromosome conformation capture carbon copy* (**5C**) method was designed to overcome this limitation as it can detect up to millions of 3C ligation junctions between many restriction fragment pairs simultaneously[47]. In 5C, the ligated 3C template is hybridized to a set of oligonucleotides, covering a particular genomic region of interest (Figure 2). Oligonucleotides are designed to cover the restriction site of each fragment in the region of interest. Primers located next to each other across the 3C junction are next ligated together, generating the 5C library. The 5C library is amplified and then quantified by high-throughput sequencing. 5C allowed the high-resolution mapping of chromosomal interactions at large genomic regions, thus it is very often known as the "many-versus-many" strategy. 5C, together with Hi-C (see below) led to the discovery of a fundamental level of organization in mammalian chromosomes called *Topologically Associating Domains*[48–50] (TADs) that have been extensively studied over the last years as it was suggested that they represent a scaffold for promoter-enhancer communication.

The game changer technique in the field of chromatin folding is the development of a genome-wide chromosome conformation capture method called **Hi-C**[51], that uses high-throughput sequencing to directly quantify proximity ligation products in purified 3C libraries and therefore can be used to assess the spatial organization of an entire genome (thus the name "all versus all" technique). The procedure of Hi-C is very similar to 3C, with only a key adjustment, the biotin labeling of the digested fragment ends before re-ligation (Figure 2). Biotin fill-in is an essential step in Hi-C as it allows you to enrich the samples for DNA sequences containing the informative ligation junctions before quantification of chromosomal interactions in a genome wide manner using massive deep sequencing. The power of Hi-C resides in its ability to convert the information contained into the entire linear genomic sequence into a two-dimensional interaction matrix which represents the fraction of cells where any pair of genomic loci where found in spatial proximity (Figure 3). This interaction matrix is normally visualized as heatmap where the color-code corresponds to interaction frequency (Figure 3).

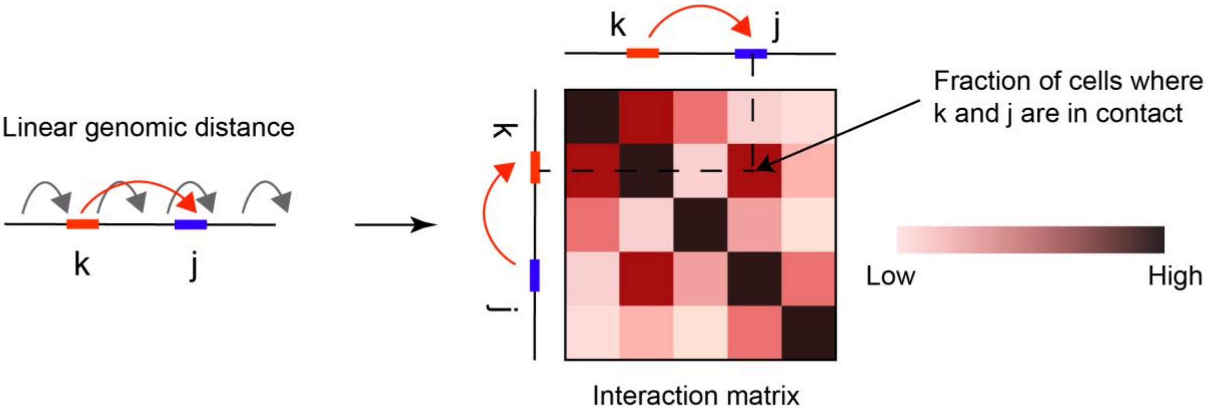


Figure 3: Visualisation of Hi-C interaction matrix

The C-techniques really paved the way for a more mechanistic understanding of genome folding and how it relates to biological function. In particular it was shown that mammalian chromosomes possess a rich hierarchy of structural layers[52]. In the next sections, I will review the hierarchical organization of chromatin fibers in mammals with particular focus on compartments, TADs and loops.

2.3 The hierarchy of chromatin folding

The first Hi-C study was performed on two human cell lines giving relatively coarse-grained (~1Mb resolution) views of genome topology[51]. Despite the low resolution, this first study gave several insights into the general properties of chromosomal folding. The existence of chromosomal territories was confirmed as Hi-C captured more intra-chromosomal contacts than interactions between chromosomes, even for loci hundreds of megabases apart on a given chromosome[51]. Looking at the interactions in *cis,* the resulting Hi-C contact matrices display a checkerboard like contact patterns (Figure 4) suggesting preferential interactions across large distances along the genome[51]. This interaction pattern is the result of the segregation of the genome into two types of multi-megabase compartments, called “A” and “B” compartments. A compartments, that interact preferentially with other A compartments, generally include regions that are enriched in genes, active histone modifications and transcriptional activity. B compartments, in contrast, interact preferentially with B compartments and include gene-poor regions, enriched in histone modifications associated with a transcriptionally repressed state[51]. B compartments were also found to be highly correlated with Lamin Associated Domains (LADs), consistent with the fact that LADs have been associated with gene repression[53]. The segregation of chromosome territories into A and B compartments has been observed for all mammalian cell types examined and has been also shown to be present in single cells[54,55]. The position of A and B compartments has been shown to vary during differentiation consistent with gene expression changes[56].

As the resolution increased with increasing sequencing depth, Hi-C and 5C experiments in mammals (mouse and human) and flies (Drosophila melanogaster) revealed that chromosomal compartments are partitioned into contiguous sub-megabase regions, called topologically associating domains (TADs) [48–50]. TADs correspond to genomic regions that interact more frequently within themselves than with neighboring regions and appear as squares along the diagonal in a Hi-C or 5C heatmap (Figure 4). TADs have been shown to be conserved both during differentiation and evolution [48,57]. Intra-TAD interactions, however, in some domains were strongly altered during differentiation and the direction of these changes correlated positively with an open chromatin state[58], suggesting that they might represent the building blocks of chromatin folding and gene regulation. Boundaries of TADs have been found to be enriched in active histone modifications, transcription start sites (TSSs), housekeeping genes, short interspersed nuclear elements (SINEs) and the architectural proteins CTCF and cohesin [48,59]. In mammals, replication of the genome occurs in units of 400–800 kilobases, termed replication domains[60,61]. Replication domain boundaries have been also shown to have an almost one-to-one correspondence with TAD boundaries[62]. The existence of TADs has been confirmed using FISH. Indeed, it has been shown that hybridization signals from probe pools entirely located within one TAD intermingle with each other to a greater extent than probe pools that span across TAD boundaries[57]. Further increases in sequencing depth have shown that TADs are partitioned into smaller domains termed sub-TADs or contact domains[63–65], a great fraction of which (~40%) are delimited by so-called chromatin loops (Figure 4), which occur when stretches of genomic sequence that lie on the same chromosome are in closer physical proximity to each other than to intervening sequences.

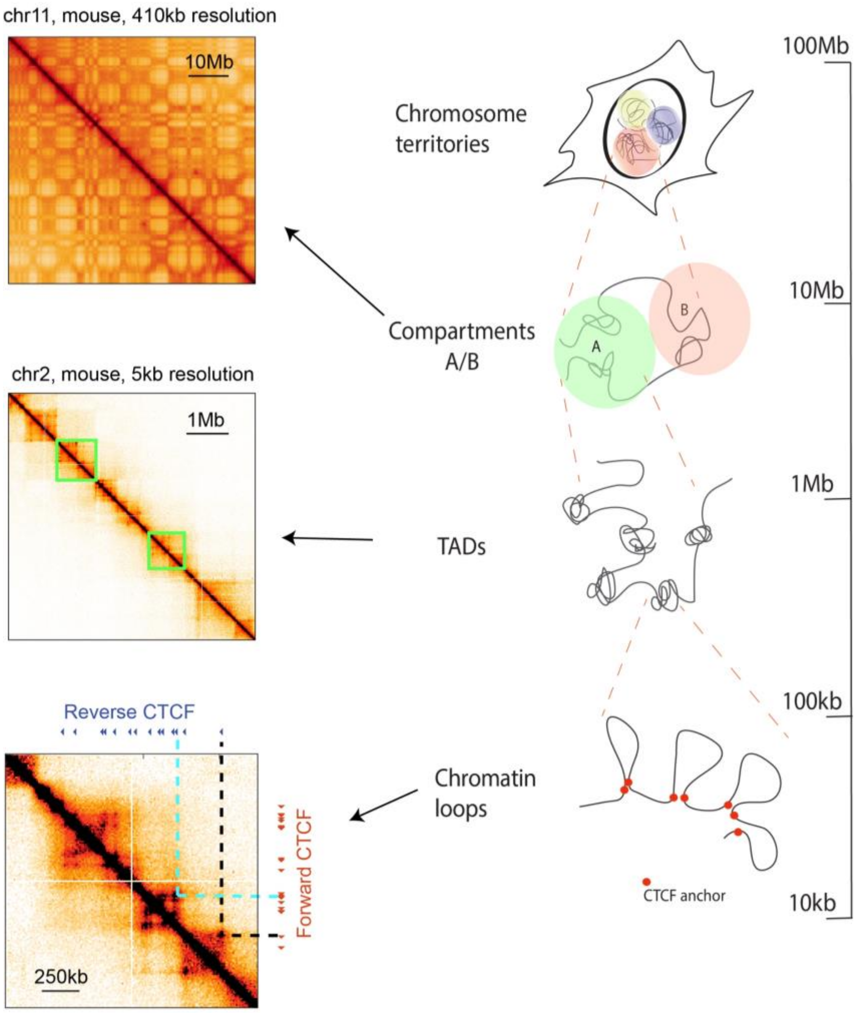


Figure 4: Hierarchical organisation of chromatin revealed by Hi-C. Top: segregation of the genome into two types of multi-megabase compartments, called "A" and "B" compartments. Middle topologically associating domains corresponding to genomic regions that interact more frequently within themselves than with neighboring. Bottom: chromatin loops, associated with convergent CTCF motif orientation at anchor sites.

Studies based on 3C-methods, thus, have shown that mammalian chromosomes possess a rich hierarchy of folding layers. An important question is how these folding layers are established. Recent studies have revealed that chromosome folding is driven by at least two independent mechanisms. On the one hand the mutually exclusive association between transcriptionally active and inactive chromatin give rise to the A and B compartments. Recent high resolution Hi-C data have suggested that these compartmental associations occur also at the level of genes resulting in the so-called 'compartmental domains' which often correspond to associations between active genes[66]. On the other hand, architectural proteins have been shown to play a major role in establishing TADs and chromatin loops. In line with this, the mediator complex, that promotes the assembly of transcription machinery[67], has been shown to be involved in promoter-enhancer chromatin loops[68,69]. Polycomb-group (PcG) proteins, that play an essential roles in gene silencing[70], have also been shown to mediate chromatin loops between polycomb-bound promoters[71].

The most studied architectural proteins are cohesin and CTCF. Cohesin is a ring-shaped protein complex that has been shown to be involved in sister chromatid cohesion and genome stability[72]. CTCF is a zinc-finger protein, known also as CCCTC-binding factor, which recognizes a specific non-palindromic motif, and was originally characterized as an insulator protein, capable of restricting enhancer-promoter interactions in their endogenous environment[73]. CTCF and cohesin have been found to be enriched at TAD boundaries and at almost all anchor sites of chromatin loops[65]. Moreover, CTCF sites at anchors of chromatin loops occur mostly in a convergent orientation, suggesting that, not only binding, but also CTCF orientation plays an important role in chromatin loop formation. Recently, the involvement of CTCF and cohesin in promoting the formation of TADs and chromatin loops has been demonstrated by using global depletion experiments which led to loss of TADs and chromatin loops[74–77]. Moreover, targeted deletions/inversions of CTCF sites resulted in loss of looping interactions[78–80]. According to a highly influential hypothesis, which is supported by recent in vitro evidence[81], the formation of chromatin loops and TADs might be driven by loop extrusion, where an extruder motor protein complex (most probably cohesin) extrudes chromatin loops until it is blocked by CTCF bound to DNA in a defined orientation[82](Figure 5). Many more studies are still needed to elucidate clearly the mechanisms that drive the hierarchical folding of chromatin. However, irrespective of the mechanisms, several studies have suggested that genome organization plays an essential role in establishing the correct pattern of interaction between promoters and enhancers.

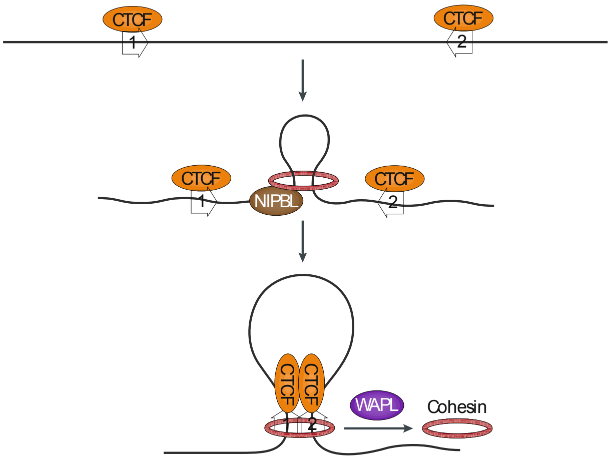


Figure 5: Loop extrusion model that postulates that the cohesion ring complex extrudes loops until it is blocked by CTCF bound in a convergent orientation (adapted from Schoenfelder and Fraser, 2019)

2.4 Chromatin conformation and promoter-enhancer communication

Proper development requires the tight control of gene expression in time and space. Enhancers play a key role in ensuring the correct spatio-temporal expression of genes, mainly by engaging in physical contact with the promoter of the target genes; yet the principles of enhancer function and mechanisms of promoter-enhancer communication are still poorly understood. In the last years, our understanding of promoter-enhancer communication has deepened considerably thanks to the development of technologies that have allowed the genome-wide mapping of enhancer-promoter contacts at high resolution[20,83] and the genome engineering of enhancer-promoter contacts. Although enhancer action that does not involve physical contact with the target promoter might exist, there is compelling evidence to support the promoter-enhancer physical contact model as the dominant mode of enhancer action. Indeed, it has been shown that forcing a loop between the mouse β-globin (*Hbb)* and its enhancer led to transcriptional activation of *Hbb* gene, demonstrating that direct promoter-enhancer looping can induce gene activation[84]. Simultaneous visualization of promoter-enhancer proximity and transcription in living cells showed that continuous physical proximity between the enhancer and its target promoter is required for gene activation in Drosophila Melanogaster, supporting the looping model[85]. These and several other studies show that the dominant model of enhancer action is through direct physical looping on promoters. In this context, the three-dimensional (3D) configuration of the genome is important because it must accommodate the physical contacts between promoters and distant enhancers. In particular, TADs have been proposed as microenvironments for establishing the correct interactions patterns between enhancers and promoters: on the one hand, by increasing the chances that regulatory elements meet each other in the 3D space within a single domain, and on the other hand, by segregating physical interactions across boundaries. In line with this, simultaneous random insertion of hundreds of reporter genes resulted in the same pattern of expression of reporter genes within the same TAD in contrast with reporter genes in adjacent TADs[86]. This is consistent with the enhancer action being confined within TADs. A study involving one of the paradigms of long-range cis-regulation, the sonic hedgehog (*Shh*) promoter and its corresponding limb-specific enhancer (also referred to as the ZRS for 'zone of polarizing activity regulatory sequence') also supports this view. Indeed, engineered chromosomal rearrangements that change *Shh-ZRS* genomic distance without affecting TAD boundaries had only a mild effect on *Shh* expression, while disruption of TADs by genomic inversion resulted in loss of *Shh* expression[87]. In addition, deletion of TAD boundaries have been shown to lead to ectopic interactions between enhancers and promoters in the adjacent domains, which has been linked to genetic diseases and oncogene activation[88–90]. Finally, it has been shown that transcriptional coregulation of neighboring genes is favored within TADs during differentiation and upon transcriptional responses to external stimuli[57,91]. Thus, a plethora of studies have provided evidence on the fundamental role of three-dimensional folding of genome and, in particular, TADs to ensure the correct pattern of interactions between promoters and enhancers that is essential for proper development.

2.5 Models of chromatin folding

The development of 3C-techniques and improvement in imaging have enhanced our understanding of chromatin folding, revealing the existence of different folding layers such as intra-chromosomal compartments, TADs and chromatin loops. Yet, the mechanisms of how these layers are established are still not completely understood. Building on the findings of 3C methods, polymer models have been a powerful tool to help uncover mechanisms that might shape genome folding.

Polymer models describes properties of a class of macromolecules called polymers that are made of many repeated units called monomers[92]. What makes polymers so interesting and powerful as a model system is that they typically show universal behavior independent of the chemical details of each monomer. For instance, all polymers are flexible at large enough length scale with respect to the polymer persistence length[93]. Flexibility at large length scale implies that there exist an infinite number of possible configurations that occur with similar probabilities, making only statistical quantities averaged over many different configurations of interest. Typical statistical quantities include the root mean square end-to-end distance as a function of the polymer length and the spatial distances between monomers. With the advent of Hi-C, the most used statistical quantity became the scaling of contact probability, which describes how the probability $P(|i-j|)$ that any two monomers are in contact depends on the monomers distance $|i-j|$ along the chain. Indeed, Hi-C gave direct access to scaling of contact probability, provided that crosslinking frequencies are proportional to absolute chromosomal contact probability. In most polymer models, the scaling of the contact probability can be described by a power law:

$$P(|i-j|) \sim \frac{1}{|i-j|^{\alpha}}$$

where α is called the *scaling exponent*.

A polymer model is defined by the interaction energies between monomers. A polymer model with the same monomer interaction energies is called a homopolymer, whereas a polymer model with different monomer interaction energies is called a heteropolymer. The statistical properties of several equilibrium homopolymer models have been characterized analytically[94,95]. For instance, the ideal chain, that corresponds to a homopolymer where monomers do not interact, is characterized by a scaling exponent of $\alpha = 1.5$; in contrast, the equilibrium globule homopolymer, where attraction between monomers dominates over excluded-volume interaction that accounts for the fact that two monomers cannot occupy the same positions, is characterized by a scaling exponent of $\alpha = 0$ for large distances between monomers.

Polymer models have been widely applied to describe the folding of chromatin fibers. Early studies based on microscopy have shown that the simple ideal chain or equilibrium globule could not capture the folding characteristics of chromatin, such as the presence of chromosomal territories[92,96]. Building on the emergence of the hierarchical folding structure of chromatin based on 3C methods (in particular Hi-C), many hypothesis-driven polymer models have been proposed to better understand the mechanisms that could give rise to these structures[97,98], using the scaling exponent as a benchmark for polymer simulations. Among these models, the loop extrusion model became very popular in recent years as it is able to reproduce several key observations. For example, the dependence of CTCF-associated chromatin loops on reciprocally orientated CTCF sites (which cannot be explained by direct looping) and the formation of the so-called chromatin "stripes" corresponding to a loop anchor interacting with entire domains at high frequency. The loop extrusion model suggests that the architectural proteins CTCF and cohesin play an essential role in the formation of chromatin structures at the sub-megabase scale, such as TADs, chromatin loops and chromatin stripes. Under the loop extrusion hypothesis, cohesin will bind chromatin and randomly extrude chromatin loops until it is blocked by CTCF bound in a defined orientation. Despite the power of hypothesis-driven methods to elucidate mechanisms of chromatin folding, the "risk" of these methods is that they account only for explicit hypotheses, completely ignoring other factors that might be important.

An alternative modeling strategy is to infer the model from the Hi-C experimental data without any prior assumptions of the mechanisms[99–101]. These agnostic approaches, whose goal is to provide unbiased and realistic reconstructions of chromatin conformations that would give rise to the Hi-C interaction matrix, have provided key insights into chromatin folding, such as the high cell-to-cell variability, notably at the scale of TADs.

3. Aim of the thesis

As outlined above, chromatin conformation plays an essential role in controlling gene expression by promoting the correct pattern of interactions between regulatory sequences such as enhancers and promoters. The development of 3C methods boosted our capability to study chromatin folding and have revealed that mammalian chromosomes possess a rich hierarchy of structural layers. Among this hierarchy, TADs have been extensively studied since they are thought to play an essential role in promoting the correct interactions between promoters and enhancers. It is, however, unclear whether the functional properties that have been attributed to TADs are specific to the folding layer of TADs themselves, and if so, why those properties emerge at this particular folding scale. As reported in Chapter I, I set out to perform a comprehensive analysis that considers all the folding layers in the hierarchy simultaneously and compares them to one another in terms of their functional and physical properties. This could clarify whether and, if so, why functional and/or structural properties are specific to TADs.

A fundamental question in chromatin field is how chromosomal interaction frequencies within and across TADs are 'read' by enhancer-promoter pairs. Do absolute interaction frequencies matter most in determining enhancer-promoter functionality, or their relative changes? Addressing these questions requires measuring chromosomal interactions with quantitative methods on the molecular level. A major limitation of 3C-based techniques is however that crosslinking and ligation are sources of experimental biases, which very often raised the question of whether the structures detected in 3C (namely TADs and chromatin loops) exist *in vivo*[102–105]. As presented in Chapter II, in collaboration with Josef Redolfi, we developed a method named DamC that allows the detection of chromosomal interactions at molecular scale and in living cells without crosslinking and ligation.

The overarching theme of my PhD has been to develop tools leading to a more quantitative understanding of chromatin organization, which I find absolutely fundamental to enhance our understanding how promoters and enhancers communicate. Thanks to DamC, we could prove that 3C techniques do not significantly distort the detection of chromosomal interactions. This is important because Hi-C data are routinely interpreted as being proportional to absolute chromosomal contact probabilities, and used to benchmark polymer models of chromosome structure. In particular, scaling properties of Hi-C data as a function of genomic distances are considered a hallmark of the mechanisms giving rise to structures observed in Hi-C. However, in contrast with classical equilibrium homopolymers where there is a one-to-one correspondence between scaling and polymer model, the heterogeneity in interactions in heteropolymer models might lead to a wild range of scaling behavior, including the one typical of homopolymers. This would suggest that scaling cannot be used alone as hallmark for polymer models. To study the general scaling properties of heteropolymers, as illustrated in Chapter III, I used heteropolymers with random gaussian interactions as model system and showed that finite-size effect, together with heterogeneity in interactions between monomers, can reproduce the range of scaling values detected in Hi-C, suggesting that caution is needed in using the scaling to discriminate alternative physical models.

4. Results

Chapter I: Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes

Yinxiu Zhan, Luca Mariani, Iros Barozzi, Edda G. Schulz, Nils Blüthgen, Michael Stadler, Guido Tiana, Luca Giorgetti

I wrote the code, performed all analyses, and wrote the paper together with Luca Giorgetti.

Summary

3C methods revealed that the folding of mammalian genomes is hierarchical with TADs being the most studied folding layer. Many functional properties have been attributed to TADs, but whether these properties are specific to TADs remains an open question. In this study we showed through an unbiased comparative analysis across the whole hierarchy that TADs emerge as a functionally privileged scale where the tendency of genes to be coregulated during differentiation is maximal; moreover, the scale of TADs maximizes CTCF clustering at domain boundaries, and optimizes promoter-enhancer interactions.

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes

Yinxiu Zhan,[1,2] Luca Mariani,[3,9] Iros Barozzi,[4] Edda G. Schulz,[3,10] Nils Blüthgen,[5,6] Michael Stadler,[1,7] Guido Tiana,[8] and Luca Giorgetti[1]

[1]Friedrich Miescher Institute for Biomedical Research, Basel, CH-4058, Switzerland; [2]University of Basel, CH-4003 Basel, Switzerland; [3]Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, 75248 Paris Cedex 05, France; [4]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [5]Institute of Pathology, Charité -Universitätsmedizin Berlin, 10117 Berlin, Germany; [6]Interdisciplinary Research Institute for the Life Sciences, Humboldt University, 10115 Berlin, Germany; [7]Swiss Institute of Bioinformatics, CH-4058 Basel, Switzerland; [8]Department of Physics and Center for Complexity and Biosystems, University of Milano and Istituto Nazionale di Fisica Nucleare, 20133, Milano, Italy

Understanding how regulatory sequences interact in the context of chromosomal architecture is a central challenge in biology. Chromosome conformation capture revealed that mammalian chromosomes possess a rich hierarchy of structural layers, from multi-megabase compartments to sub-megabase topologically associating domains (TADs) and sub-TAD contact domains. TADs appear to act as regulatory microenvironments by constraining and segregating regulatory interactions across discrete chromosomal regions. However, it is unclear whether other (or all) folding layers share similar properties, or rather TADs constitute a privileged folding scale with maximal impact on the organization of regulatory interactions. Here, we present a novel algorithm named CaTCH that identifies hierarchical trees of chromosomal domains in Hi-C maps, stratified through their reciprocal physical insulation, which is a single and biologically relevant parameter. By applying CaTCH to published Hi-C data sets, we show that previously reported folding layers appear at different insulation levels. We demonstrate that although no structurally privileged folding level exists, TADs emerge as a functionally privileged scale defined by maximal boundary enrichment in CTCF and maximal cell-type conservation. By measuring transcriptional output in embryonic stem cells and neural precursor cells, we show that the likelihood that genes in a domain are coregulated during differentiation is also maximized at the scale of TADs. Finally, we observe that regulatory sequences occur at genomic locations corresponding to optimized mutual interactions at the same scale. Our analysis suggests that the architectural functionality of TADs arises from the interplay between their ability to partition interactions and the specific genomic position of regulatory sequences.

[Supplemental material is available for this article.]

Characterizing the three-dimensional organization of chromosomes in mammalian cells is a central challenge, especially in light of determining how regulatory sequences such as enhancers and promoters interact and ensure precise control of gene expression during development. Methods based on chromosome conformation capture (3C) and notably 4C, 5C, and Hi-C, which measure physical interaction frequencies of genomic loci in the three-dimensional nuclear space, have revealed that mammalian chromosomes possess a rich hierarchy of structural layers (Gibcus and Dekker 2013). Each chromosome is partitioned in multi-megabase 'A' and 'B' compartments, reflecting the associations of alternating large regions of active and inactive chromatin (Lieberman-Aiden et al. 2009). Compartments are further subdivided into topologically associating domains (TADs), contiguous sub-megabase genomic regions within which the chromatin fiber preferentially associates (Dixon et al. 2012; Nora et al. 2012), which are further partitioned into smaller substructures and 'contact domains' (Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014). Finally, as a further level of complexity, TADs also interact with each other into "meta-TAD" trees that extend up to several Mb (Fraser et al. 2015). Given the cell population-averaged nature of 3C-based experiments, the observed nested hierarchies of interaction domains may arise as statistical patterns resulting from an average over millions of alternative conformations of the chromatin fiber (Fudenberg and Mirny 2012; Giorgetti et al. 2014; Junier et al. 2015).

Although more than one mechanism might give rise to TADs and sub-TAD structures, CTCF (CCCTC-binding factor) and the cohesin complex appear to be largely responsible for the establishment and maintenance of TADs and sub-TAD structures and

---

Present addresses: [9]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [10]Max-Planck-Institute for Molecular Genetics, Otto-Warburg-Laboratory, Max Planck Research Group–Regulatory Networks in Stem Cells, 14195 Berlin, Germany
Corresponding author: luca.giorgetti@fmi.ch

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.212803.116.

© 2017 Zhan et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

boundaries. Indeed, CTCF and cohesin are enriched at TAD boundaries (Dixon et al. 2012; Van Bortle et al. 2014), but they also bind pervasively within TADs and are involved in the formation of sub-TAD structure (Rao et al. 2014; de Wit et al. 2015; Sanborn et al. 2015), although the molecular mechanisms that lead to structure formation are unclear (Merkenschlager and Nora 2016). In addition, open chromatin and active transcription positively correlate with the presence of TADs and sub-TAD structure (Hou et al. 2012; Phillips-Cremins et al. 2013; Ulianov et al. 2015), and active histone modifications are enriched at TAD boundaries (Dixon et al. 2012), suggesting that interactions between active regulatory sequences may contribute to establish chromosomal architecture. However, transcription does not seem to be strictly needed for maintaining TAD boundaries (Nora et al. 2012).

Irrespective of the mechanisms underlying their formation, genetic evidence suggests that TADs contribute to establish correct interaction patterns between enhancers and promoters (Symmons et al. 2014; Lupiáñez et al. 2015; Franke et al. 2016). Consistent with this, transcriptional coregulation of neighboring genes is favored within TADs during differentiation (Nora et al. 2012) and upon transcriptional responses to external stimuli (Le Dily et al. 2014). TADs are thought to act, on the one hand, by increasing the chances that regulatory elements meet each other in the three-dimensional space within a single domain, and on the other hand, by segregating physical interactions across boundaries, thus decreasing the probability that deleterious interactions occur. Hence, the degree to which each TAD is insulated with respect to its neighbors may be an important parameter in the establishment of the correct regulatory connections. It is, however, unclear whether the functional attributes that have been observed at the level of TADs (namely the ability to constrain enhancer-promoter interactions and promote transcriptional coregulation) are specific to the folding layer of TADs themselves, and if so, why those properties emerge at this particular folding scale.

A comprehensive analysis that considers all previously identified topological levels simultaneously and compares them to one another in terms of their functional and physical properties is currently lacking. A small number of algorithms that identify hierarchies of topological domains are available (Filippova et al. 2014; Lévy-Leduc et al. 2014; Shin et al. 2015; Weinreb and Raphael 2015; Chen et al. 2016; Shavit et al. 2016). However, none of them provides a quantitative description of how the various layers of domains differ from one another. In addition, these algorithms define hierarchies of interaction domains depending on one or more parameters that do not have a clear biological or structural interpretation. To overcome these limitations, we developed a novel algorithm called CaTCH (Caller of Topological Chromosomal Hierarchies) that identifies nested topologies of structural domains in Hi-C data sets based on a single parameter, the reciprocal physical insulation between domains, which is a simple and biologically relevant measure. Here, we describe the CaTCH algorithm and report the results of comparing the structural and functional properties of domains across the folding hierarchy of the mouse genome.

Results

CaTCH: an algorithm to detect and stratify nested hierarchies of topological domains

In order to comprehensively describe the multiscale organization of chromosomal folding hierarchies, we developed an algorithm that segments Hi-C interaction maps into multiple alternative sets of domains and stratifies them according to a single parameter. We adopted a thermodynamic interpretation of Hi-C data sets (Fudenberg and Mirny 2012) in which the Hi-C signal between a pair of loci is proportional to the *probability* of detecting them in proximity across the cell population. For any pair of adjacent chromosomal domains A and B, we then defined their reciprocal insulation (RI) as

$$RI(A, B) = [P_{\text{in}}(A) + P_{\text{in}}(B) - P_{\text{out}}(A, B)]/[P_{\text{in}}(A) + P_{\text{in}}(B)] \times 100, \quad (1)$$

where P_{in} and P_{out} are the average Hi-C counts within a domain and across two adjacent domains, respectively (Fig. 1A; see Methods section). Small (large) values of RI thus correspond to domains that are poorly (strongly) insulated from their first neighbors. For example, 70% reciprocal insulation means that the average Hi-C counts across the boundaries of two adjacent domains are 70% smaller than the average counts within the two domains.

Given a certain degree of reciprocal insulation, the algorithm merges all consecutive domains whose reciprocal insulation is lower than the chosen threshold (Fig. 1B; see Methods section), similarly to what is commonly performed by agglomerative hierarchical clustering (Hastie et al. 2009). Thus, for any reciprocal insulation threshold, detected domains are *at least* insulated by the threshold value. By smoothly increasing the threshold on the insulation, the algorithm detects a set of domains that are increasingly more insulated, larger, and containing previous domain layers. This results in a nested hierarchy of differentially insulated domains (Fig. 1C). We dubbed this algorithm CaTCH, for Caller of Topological Chromosomal Hierarchies.

A key property of CaTCH is that it does not rely on the tuning of any free parameter to identify one particular folding scale. The only parameter in the algorithm is the reciprocal insulation threshold itself, which is systematically varied to define and stratify the entire hierarchy of domains, rather than tuned to identify a single domain set. Moreover, unlike parameters in existing approaches to identify multiscale domain structures in Hi-C data sets (Filippova et al. 2014; Lévy-Leduc et al. 2014; Shin et al. 2015; Weinreb and Raphael 2015; Chen et al. 2016; Shavit et al. 2016), the reciprocal insulation is a biologically relevant measure estimating how efficiently a domain is physically insulated from its immediate neighbors. CaTCH is provided as an R package at https://github.com/zhanyinx/CaTCH_R (source code can be found in Supplemental Methods).

Sub-TAD contact domains, TADs, and compartments emerge at different levels in the folding hierarchy

We first applied CaTCH to published Hi-C data sets from female mouse ESCs (Giorgetti et al. 2016) binned at 20-kb resolution. As expected, when increasing the reciprocal insulation parameter, the algorithm detected increasingly larger and fewer topological domains (Fig. 1C), with 5% changes in reciprocal insulation translating into ∼30% changes in the number and size of domains (Supplemental Fig. S1a). We found a similar trend when analyzing other cell types, notably neural precursor stem cells (NPCs) derived from the same ESC line (Giorgetti et al. 2016) and the mouse B-cell lymphoma CH12 cell line (Supplemental Fig. S1b; Rao et al. 2014). In ESCs, below 40% reciprocal insulation domains are too small (<100 kb on average) to be characterized with data at 20-kb resolution. At higher insulation values, however, we detected domains

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

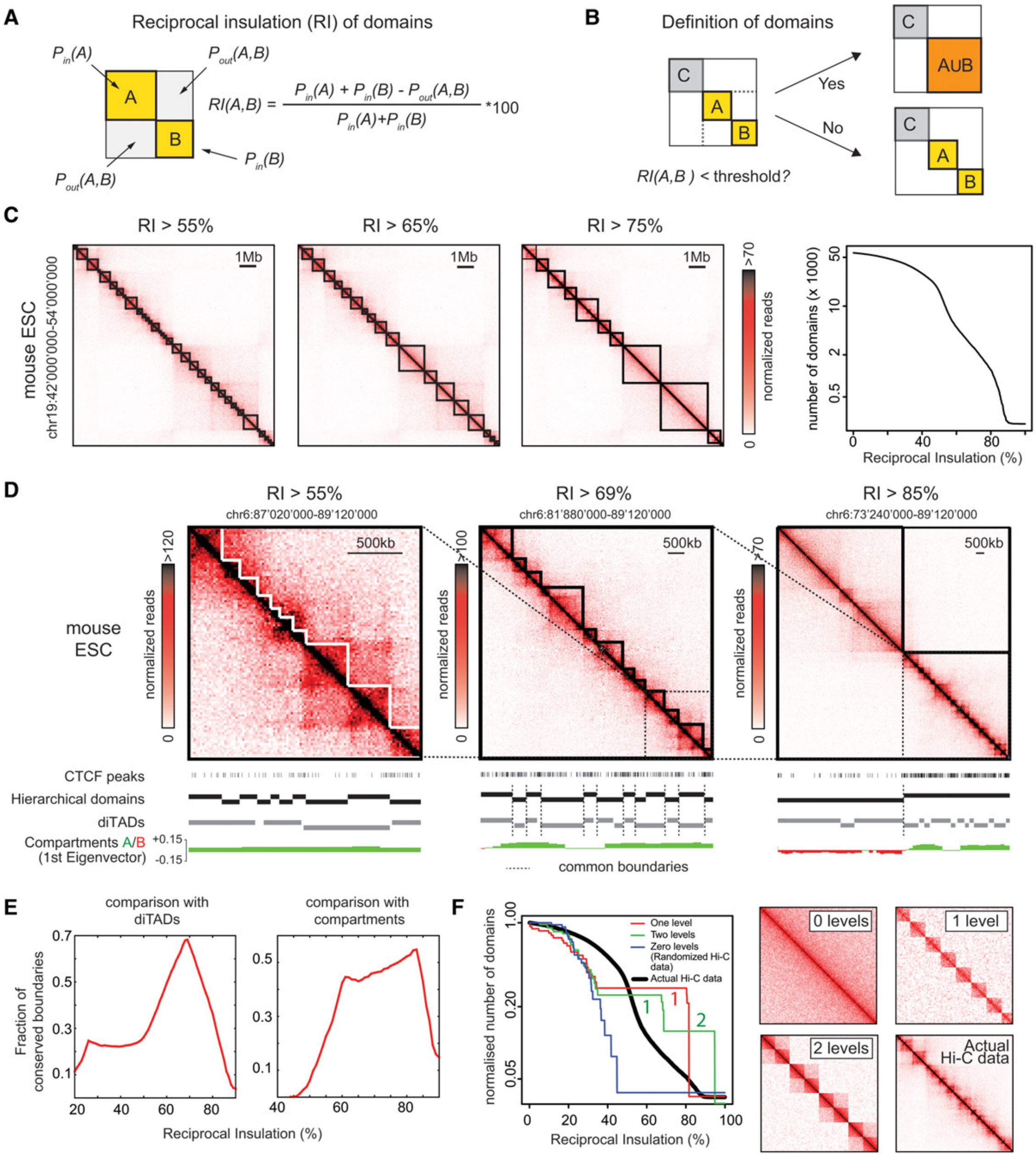


Figure 1. Schematic description of reciprocal insulation and the domain-calling algorithm. (*A*) Schematic representation of reciprocal insulation (RI) between two fictitious domains A and B in Hi-C data. (*B*) The CaTCH algorithm merges two adjacent domains if their reciprocal insulation is smaller than a given threshold. (*C*) (*Left* three panels) Examples of sets of domains defined in mouse ESCs Hi-C data (20-kb binning) imposing different threshold on RI. (*Right*) Number of domains detected in ESC as a function of RI. (*D*) Sub-TAD contact domains (*left*), directionality index-based TADs (*middle*), and A/B compartments (*right*) are identified at different RI values. (*E*) Fraction of boundaries of diTAD (*left*) and compartments (*right*) overlapping with boundaries of domains identified by CaTCH as a function of RI. (*F*) (*Left*) Number of domains detected by CaTCH as a function of RI in the real genome (black line), or in computationally generated contact maps with zero (blue), one (red), or two preferential folding levels (green). The corresponding heat maps are shown in the four *right* panels. Numbers of domains were normalized to the initial step (0% insulation) to allow comparison.

with a size (180 kb on average) in the range of sub-TAD structures and 'contact domains' identified in previous studies (Fig. 1D, left; Supplemental Fig. S1c; Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014). More than 60% of domain boundaries identified at 55% reciprocal insulation contain at least a CTCF peak identified in a published ChIP-seq data set (Cheng et al. 2014), consistent with the notion that sub-TAD structures are highly correlated with CTCF binding (Phillips-Cremins et al. 2013). In addition, although the resolution of the Hi-C data set is not high enough to distinguish most of the CTCF-associated 'loop' signals as in Rao et al. (2014), we noticed that ∼45% of domains at this scale have at least one CTCF peak at both boundaries (Supplemental Fig. S1d). Of the CTCF-delimited domains, however, only 35% had convergent CTCF sites (compared to 98% of 'loop

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

domains,' defined as contact domains with strong interaction between boundaries in Rao et al. 2014). This is largely due to the fact that the domains identified in the latter study are a subset of domains detected by CaTCH at 55% reciprocal insulation (see below); however, a direct comparison between the two domain sets is not possible, due to the lack of ESCs Hi-C data sets in the study by Rao et al. (2014).

To determine the actual overlap between domains identified by CaTCH and contact domains described in Rao et al. (2014), we analyzed the 10-kb-resolution Hi-C data that were obtained in CH12 cells in the same study. Maximal overlap between the two domain sets occurred at 62% reciprocal insulation in CH12 (Supplemental Fig. S1e), where 78% of boundaries of previously identified contact domains are also detected by CaTCH. However, CaTCH detects more domains than those identified in Rao et al. (2014) (Supplemental Fig. S1f), which explains the lower proportion of domains delimited by convergent CTCF sites in our data set. Thus, sub-TAD contact domains are detected by CaTCH as relatively lowly insulated regions.

We next sought to identify the scale in the folding hierarchy where domains detected by CaTCH most closely resemble TADs. Since directionality index analysis (Dixon et al. 2012) has been used to define TAD boundaries in a number of previous studies, here we adopted this benchmark definition of TADs and refer to these domains as 'diTADs' (directionality index TADs). It is important to point out that the set of diTADs identified in a Hi-C experiment depends on the value of two tunable parameters, one setting a limit to the maximal genomic distance over which Hi-C interactions are evaluated (Supplemental Fig. S1g) and the other defining the minimum acceptable size of domains. We set these parameters to 2 Mb and 80 kb, respectively, as used previously (Dixon et al. 2012), to build a reference set of diTADs. This resulted in the identification of 2220 diTADs with a median size of 840 kb, compatible with earlier analyses in mouse ES cells (Dixon et al. 2012). The best overlap between hierarchical domains detected by CaTCH and diTADs occurred at around 69% reciprocal insulation (Fig. 1D, center), where ∼70% of diTAD boundaries coincide with hierarchical domain boundaries (Fig. 1E, left) and their size distributions are very similar (Supplemental Fig. S1h). Domains detected by our algorithm at this scale are slightly (although not significantly) smaller than diTADs (median size 760 kb vs. 840 kb) (Supplemental Fig. S1h,i). Most (74%) CaTCH boundaries not corresponding to TADs indeed divide diTADs in smaller domains (Supplemental Fig. S1j). Thus, diTADs are detected by CaTCH as domains that are more robustly insulated than contact domains.

At even higher reciprocal insulation, hierarchical domains detected by CaTCH correspond to regions of increasingly longer-range associations between TADs, in the range of meta-TADs described in Fraser et al. (2015), themselves contained into even larger domains occurring at very high insulation (around 85%) (Fig. 1D, right). These domains largely overlap with consecutive stretches of genomic sequence belonging to either the 'A' or 'B' compartments (Lieberman-Aiden et al. 2009), as detected by eigenvector analysis (Imakaev et al. 2012) on the same ESCs Hi-C data set (Fig. 1D, right, 1E, right). Consistent with the notion that A/B compartments represent predominantly active/inactive chromatin, using publicly available ChIP-seq data sets in ESCs (Supplemental Table S1), we found that the difference in histone modification patterns within vs. across domain boundaries is maximized at this scale (Supplemental Fig. S1k).

Thus, CaTCH identifies a continuous spectrum of nested self-interacting chromosome domains, stratified as contiguous genomic regions with differential reciprocal insulation levels. Previously described chromosomal structures such as sub-TAD contact domains, TADs, and groups of TADs emerge at different scales within the nested folding hierarchy and are characterized by increasing reciprocal insulation levels.

A continuous nested hierarchy of topological associating structures

We then sought to determine whether one or more privileged reciprocal insulation levels exist among the folding hierarchy and correspond to any of the previously reported folding layers. If such level(s) existed, some simple fundamental quantities, such as the number or size of domains detected by the CaTCH algorithm, would have a discontinuous behavior as a function of the reciprocal insulation parameter. To exemplify this concept, we computationally generated simplified control contact maps by artificially imposing the presence of zero, one, or two scales of domains, separated by sharp transitions in contact probabilities between consecutive layers (see Methods section; Fig. 1F). For these controls, CaTCH detected a number of plateaus in the size (or number) of domains equal to the number of distinct hierarchical levels (Fig. 1F, left), irrespective of the genomic size of the domains (Supplemental Fig. S1l). Compared to these controls, the ESC genome does not exhibit any structurally privileged scale, at least for domains defined using reciprocal insulation as a measure (black line in Fig. 1F), irrespective of whether the entire genome is considered or select regions that belong to either the A (active) or B (inactive) compartment (Supplemental Fig. S1b). A similar trend can be observed in NPCs and CH12 cells (Supplemental Fig. S1b), suggesting that no obvious privileged structural scale exists in ESCs and differentiated cell types. As a notable consequence, TADs do not appear as a natural intrinsic structural scale in the nested hierarchy of domains. This prompted us to investigate whether functional properties that have been previously attributed to TADs specifically emerge at the TAD scale or are rather widespread among the folding hierarchy.

Enrichment in active histone marks is maximized at the scale of TADs

TAD boundaries have been shown to be enriched in histone modifications associated with active transcription (Dixon et al. 2012). We therefore analyzed publicly available ChIP-seq data sets in ESCs (Supplemental Table S1) and computed the enrichment for distinct histone marks at the boundaries of the domains across all the scales in the folding hierarchy. Marks associated with active transcription showed a steady increase in enrichment as a function of reciprocal insulation and reached a plateau at the level of diTADs (∼69%) (Supplemental Fig. S2a). Thus, although active histone marks show widespread enrichment across the folding hierarchy, they are maximally enriched at the scale of TADs and TAD aggregates (meta-TADs and compartments). Consistent with previous results (Dixon et al. 2012), the H3K9me3 repressive mark was found depleted at many levels in the folding hierarchy and notably at the level of diTADs (Supplemental Fig. S2a).

CTCF clustering at boundaries is maximized at the scale of TADs

Consistent with its putative role in establishing and/or maintaining chromosomal structure, CTCF is enriched at boundaries of contact domains (Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014), TADs (Dixon et al. 2012), and meta-TAD

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

trees (Fraser et al. 2015). We therefore computed the enrichment in the number of CTCF ChIP-seq peaks (Cheng et al. 2014) at domain boundaries at all folding scales. As expected, CTCF binding is enriched at boundaries of every level across the folding hierarchy; however, CTCF enrichment is maximized at the scale of TADs, and in particular at ∼65% reciprocal insulation (Fig. 2A) corresponding to domains that are slightly less insulated than diTADs detected using standard directionality index parameters. Identical results were found by using the input-normalized CTCF ChIP-seq signal per boundary, rather than the number of ChIP-seq peaks (Supplemental Fig. S2b). We noticed that maximal CTCF enrichment is due to both a maximal number of boundaries containing at least one CTCF peak and a maximal average number (∼1.9) of CTCF peaks per boundary (Supplemental Fig. S2c), which are mostly found within the 40 kb upstream of or downstream from the boundary (Supplemental Fig. S2d).

Domains at 65% reciprocal insulation are 20% smaller compared to 'standard' diTADs (600 kb vs. 840 kb median size) (Fig. 2B) and frequently originate from the splitting of one diTAD into two or more smaller domains (Supplemental Fig. S2e). The majority (∼70%) of these 'new' boundaries have at least one occupied CTCF site, which explains the slightly higher enrichment in CTCF compared to standard diTADs. However, by systematically varying the values of parameters in the directionality index algorithm, we identified alternative sets of diTADs where CTCF enrichment is higher than standard diTADs and comparable to (although slightly lower than) 65% RI domains (Supplemental Fig. S2f). Importantly, these alternative directionality index domains correspond to domains detected by CaTCH in the 65%–70% reciprocal insulation range (Supplemental Fig. S2f, arrows). This confirms that the TAD scale is characterized by maximal CTCF enrichment at boundaries compared to other folding levels. We will hereafter refer to domains identified by CaTCH at 65% minimal reciprocal insulation simply as TADs, since they constitute the set of domains with maximal CTCF enrichment.

Reciprocal orientation of CTCF binding sites has been shown to be highly predictive of strong long-range 'looping' interactions (Rao et al. 2014; de Wit et al. 2015; Guo et al. 2015; Vietri Rudan et al. 2015). We therefore assessed the orientation of the two most internal CTCF motifs on either side of each domain and found that, at the scale of TADs, the fraction of domains where CTCF motifs were convergent was maximal (Supplemental Fig. S2g, left), with ∼22% of domains possessing convergent binding sites. Thus, both CTCF clustering and head-to-head orientation of the most internal CTCF motifs are maximized at the scale of TADs. Using available CTCF ChIP-Seq data sets (Phillips-Cremins et al. 2013; Cheng et al. 2014), we found that, in both NPCs and CH12 cells, CTCF enrichment at boundaries showed a similar trend as in ESCs, with a peak around 58% and 82% reciprocal insulation in NPCs and CH12, respectively (Fig. 2C). The fraction of domains with convergent CTCF motifs peaked at the same *RI* values (Supplemental Fig. S2g). Importantly, despite the difference in absolute reciprocal insulation values, the number and size of domains at maximal CTCF enrichment were extremely

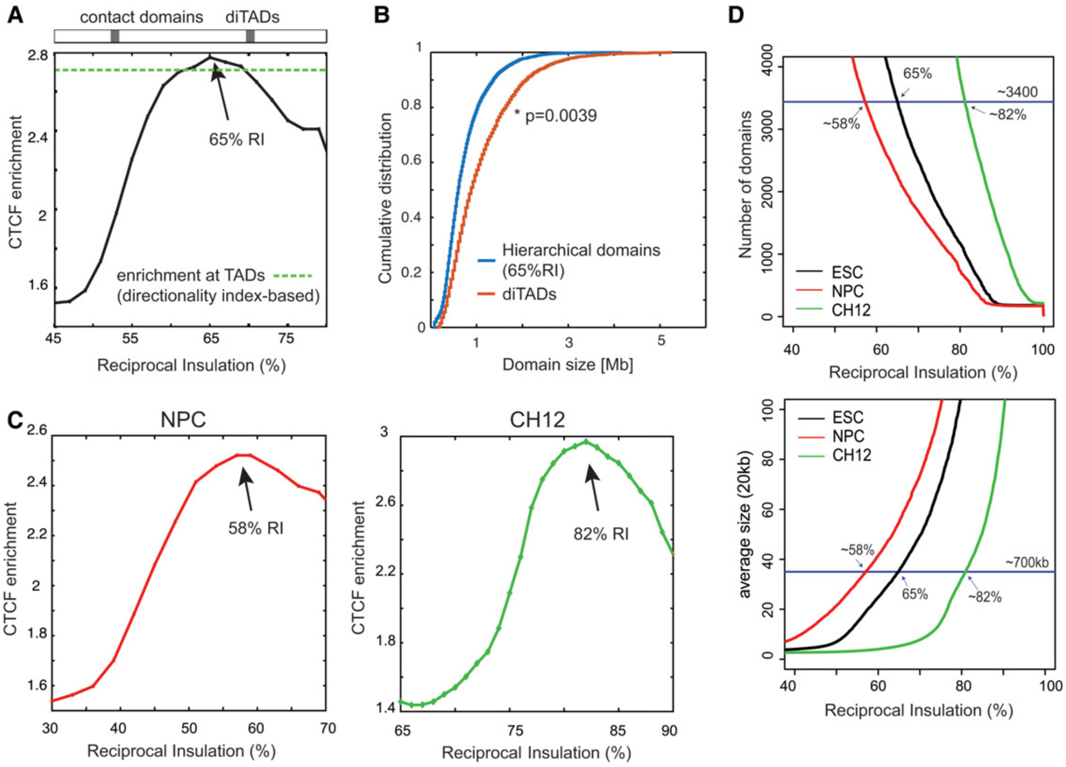


Figure 2. CTCF clustering at domain boundaries is maximal at the scale of TADs. (*A*) CTCF enrichment at domain boundaries is widespread among the folding hierarchy in mouse ES cells. However, maximal enrichment occurs at 65% RI, where it is slightly higher compared to TADs identified by directionality index analysis (diTADs). (*B*) Domains at 65% are slightly smaller than diTADs identified on the same data set. (*C*) CTCF enrichment at domain boundaries in NPCs and CH12 cells shows a similar trend as in ESCs, with maxima located at 58% and 82% RI in NPCs and CH12, respectively. (*D*) The number and size of domains defined by maximal CTCF enrichment at boundaries are similar in ESCs, NPCs, and CH12 cells.

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

similar across the three cell types (Fig. 2D). In addition, conservation of boundaries across the three cell types was also found to be maximal at the same scale, with ∼70% of boundaries conserved between any two cell types (Supplemental Fig. S2h). Thus, the scale of TADs appears in the entire folding hierarchy not only as the domain scale that maximizes CTCF enrichment at boundaries, but also as the scale where domains are most conserved across cell types.

We next sought to determine any confounding effect on the determination of the optimal *RI* value due to experimental factors, such as different sequencing depth of Hi-C libraries or different versions of the Hi-C protocol. To study the effect of sequencing coverage, we performed CaTCH and CTCF enrichment analysis on a down-sampled ESC Hi-C data set obtained by reducing by half the total number of sequenced reads. CTCF enrichment at domain boundaries was maximized at very similar reciprocal insulation value as in the full data set (67% vs. 65%) (Supplemental Fig. S2i), largely corresponding to the same set of domains (Supplemental Fig. S2k). Thus, sequencing depth is not likely to have a strong impact on the reciprocal insulation values where TADs appear. Next, to understand the impact of using Hi-C data sets obtained using different experimental protocols, we performed a comparative analysis of two data sets obtained in mouse fetal liver cells (Nagano et al. 2015) using either the 'dilution' (Lieberman-Aiden et al. 2009; Belton et al. 2012) or the 'in situ' ligation protocols (Nagano et al. 2013; Rao et al. 2014). Using a published CTCF data set (Cheng et al. 2014), we found that maximal CTCF enrichment occurred at different reciprocal insulation values (Supplemental Fig. S2j,k), with the dilution protocol leading to smaller values compared to the in situ protocol (70% vs. 77%). This is consistent with the lower insulation values where TADs appear in NPC and ESC (where Hi-C was performed with the dilution protocol [Giorgetti et al. 2016]) compared to CH12 cells (in situ protocol) and is compatible with the previous observation that the in situ protocol leads to sharper TAD boundaries (Nagano et al. 2015). These results point at Hi-C protocol variants as a main determinant of reciprocal insulation and suggest that the scale of TADs occurs in the 58%–70% range (64% ± 6%) in dilution Hi-C data sets and in the 77%–82% range (80% ± 3%) in the in situ experiments that were analyzed.

Transcriptional coregulation during differentiation is maximal at the scale of TADs

Motivated by the finding that CTCF and active histone marks enrichment at boundaries is maximal at the scale of TADs, we set out to determine whether domains at this scale encompass maximally coregulated genes, which is a further important functional attribute proposed for TADs (Nora et al. 2012; Le Dily et al. 2014). For this, we performed strand-specific RNA-seq on total RNA from the ESC and NPC lines in which the Hi-C had been performed (Giorgetti et al. 2016). Strand specificity allowed us to unambiguously assign up- or down-regulated transcripts in the case of two overlapping transcriptional units. For all levels in the folding hierarchy, we then set out to determine how many domains are transcriptionally coregulated during the differentiation from ESCs to NPCs.

We defined a domain to be down- (up-) coregulated at the empirical $P \leq 0.05$ level if the number of down- (up-) coregulated genes in the domain is larger than in 95% of cyclically permutated genomes (see Methods section). For each insulation level and the corresponding domain set, we then calculated a Z-score as the difference between the number of coregulated domains that were observed in the real genome and the mean number of coregulated domains detected in 2000 randomizations of the genome (Fig. 3A; see Supplemental Methods), weighted by its standard deviation. Interestingly, at all insulation levels, the subset of domains that we detected to be up- or down-regulated at the level of $P \leq 0.05$ show maximal transcriptional changes during development (Fig. 3B; Supplemental Fig. S3a). Thus, domains with a high level of transcriptional coregulation largely overlap with those where the most dramatic changes in gene expression occur during differentiation.

At the level of TADs in ESC (65% insulation), we detected 114 coregulated domains, accounting for ∼4% of the total number of TADs and ∼10% of those exhibiting expression changes during differentiation (≥2 up- or down-regulated genes). This represents a >2.5-fold enrichment relative to the values expected in randomized genomes. Moreover, the number of coregulated TADs (65% reciprocal insulation) is very similar to that observed at the level of TADs in the context of the acute transcriptional response to progesterone in a human breast cancer cell line (Le Dily et al. 2014).

For genes that are down-regulated during differentiation, the Z-score is maximum at the scale of TADs (Fig. 3C). To check the robustness of the analysis against stochastic fluctuation of expression changes, we studied the behavior of Z-scores upon randomly reshuffling ($n = 1000$) 10% of genes. For 66% of these partially reshuffled genomes, the maximum Z-score was found to be located within a 4% interval around 65% reciprocal insulation (63%–66%) (Supplemental Fig. S3b), supporting the robustness of the result. This analysis suggests that TADs in ESCs constitute a functionally privileged scale, maximizing the coregulation of genes that are down-regulated during the differentiation into neural precursor stem cells.

The behavior of up-regulated genes was remarkably different, with low (if any) enrichment in transcriptional coregulation within domains below 75% reciprocal insulation (Fig. 3D) and maximal enrichment at the scale of A/B compartments (>80%). We reasoned that this could be due to the fact that not all TADs identified in ESCs are predictive for transcriptional coregulation of genes that become activated during differentiation. We thus performed the same analysis on domains identified in NPCs and found that coregulation of both down- and up-regulated genes is maximized within domains defined in NPCs around 58% reciprocal insulation (Fig. 3E,F). This is the set of TADs defined in NPCs by maximal CTCF clustering at their boundaries (see Fig. 2C). We verified that these results are not affected by the presence of an inactive X chromosome in NPCs, as maximal coregulation was observed at the scale of TADs even when expression changes of X-linked genes (excluding genes that escape X inactivation in this clone [Giorgetti et al. 2016]) were corrected to account for their monoallelic expression in NPCs (Supplemental Fig. S3c; Supplemental Methods).

Thus, TADs defined in the initial developmental stage (ESCs) are the scale where transcriptional coregulation of down-regulated genes is maximal, whereas the set of domains that better favors the coregulation of up-regulated genes corresponds to TADs defined in the final state (NPCs). This can be largely explained by the fact that, although most TAD boundaries (∼70%) are conserved between ESCs and NPCs, a significant fraction (∼30%) is not. In particular, although most up-regulated TADs are conserved and detected in both ESCs and NPCs, we detected 20% more up-regulated TADs in NPCs than in ESCs, corresponding to domains that

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

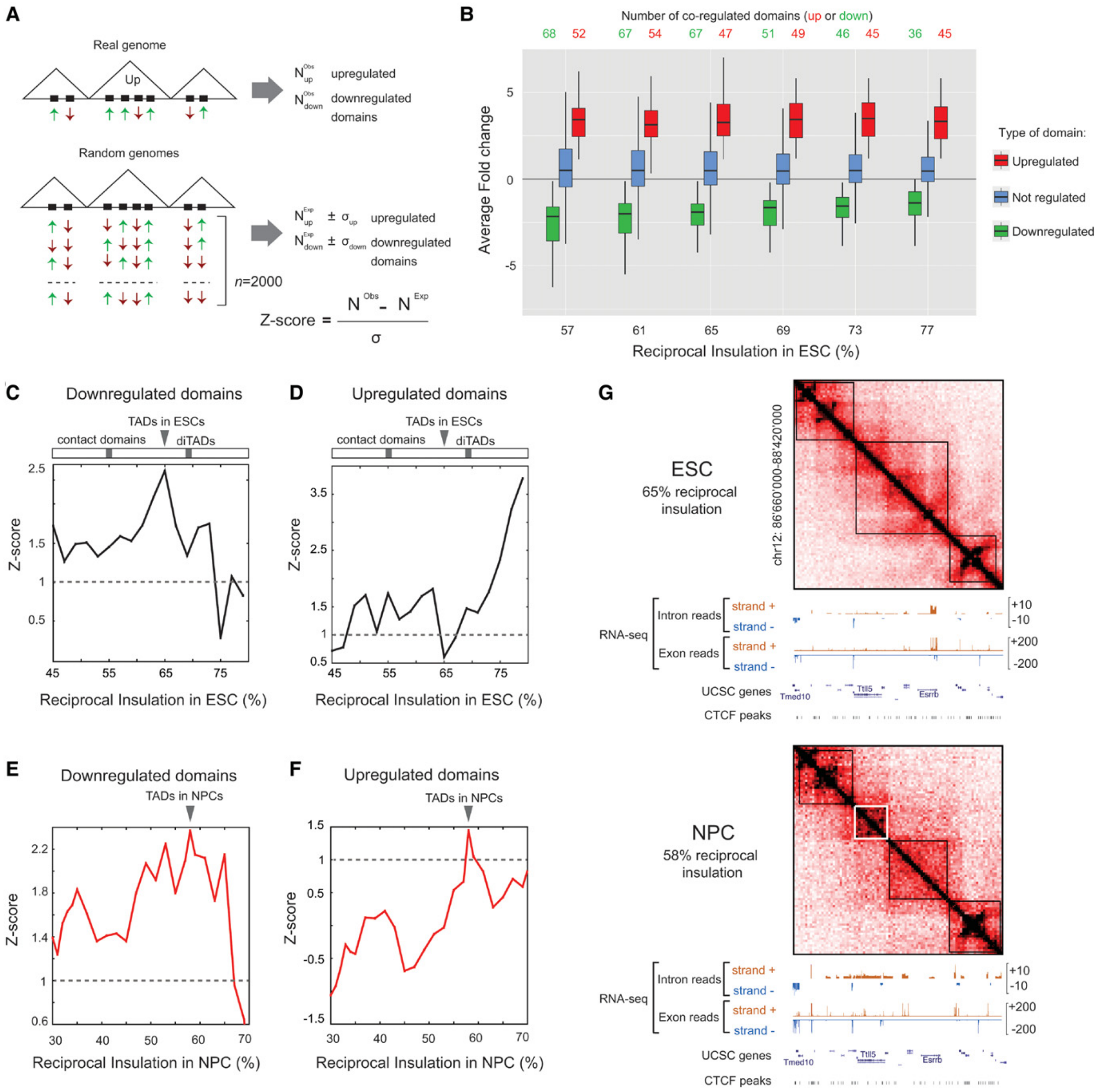


Figure 3. Transcriptional coregulation defines a functional privileged scale. (*A*) Schematic representation of the definition of statistical enrichment in the number of coregulated domains. A domain is down- (up-) coregulated if its number of down- (up-) coregulated genes is larger than in 95% of cyclic permutated genomes (empirical $P \leq 0.05$). A *Z*-score is calculated as the difference between the number of coregulated domains detected in the real genome (N^{obs}) and the mean number of coregulated domains detected in 2000 randomized genomes (N^{exp}), weighted by its standard deviation σ^{exp}. (*B*) Distribution of average fold changes in expression level for domains at different RI values. For each RI value, the number of domains that are either up- or down-regulated during differentiation (at the $P \leq 0.05$ level) is also shown in the *upper* part of the graph. Box: 25%–75% range (black line: median). (*C*) The statistical enrichment in the number of down-regulated domains is plotted as a function of the RI threshold. Transcriptional coregulation is significant at any level below ~70% RI but maximal at 65%. (*D*) Same as panel *C* for up-regulated domains. (*E*) Same analysis as in panel *C* when using domains based on Hi-C data in NPCs. (*F*) Same as panel *E* for up-regulated domains. (*G*) Example of domains that were created de novo during differentiation and detected only in the set of NPC TADs (58% RI).

were defined de novo during differentiation in parallel with a significant increase in their genomic activity (Fig. 3G; Supplemental Fig. S3d). This might suggest that TADs in NPCs are more predictive of transcriptional coregulation of up-regulated genes because domains that are transcriptionally active in the final state of differentiation can only be precisely detected when they are active (i.e., in the final state) but do not appear as defined in the set of TADs in the initial state. These domains represent extreme cases that

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

illustrate that increased genomic activity can be associated with increased structural complexity, and in this case, with de novo formation of local structures. This is reminiscent of what was observed on the inactive X chromosome (Giorgetti et al. 2016), where the presence of TAD-like structures is only observed in the context of gene activation.

Enhancer-promoter communication is optimized at the scale of TADs

The finding that TADs emerge as the folding scale that maximizes transcriptional coregulation but are not an intrinsically defined structural level (cf. Fig. 1) prompted us to ask whether TADs specifically favor enhancer-promoter communication in ESCs. We analyzed available ChIP-seq data sets (Supplemental Table S1) to identify enhancers based on H3K27ac, H3K4me1, and H3K4me3 patterns (see Supplemental Methods); active promoters were identified using the strand-specific total RNA-seq data sets generated in this study.

To check whether the presence of domains at each level in the hierarchy corresponds to gain (or loss) in interactions, we considered pairs of Hi-C bins containing enhancers and promoters. We then calculated the ratio between their Hi-C counts vs. the genomic average for loci separated by the same genomic distance (Fig. 4A). We observed substantial enrichment in interactions between enhancers and promoters within the same (active) domain up to ~65% reciprocal insulation (Fig. 4B, red curve). Thus, TADs appear to be within the uppermost scales in the folding hierarchy where enhancer-promoter contacts are maximally enriched within domains. On the other hand, enhancer-promoter interactions are also enriched *across* boundaries with the two neighboring domains, until slightly below the scale of TADs (Fig. 4C, red curve). This reflects the fact that domains up to TADs are detected as increasingly bigger units, which are defined by the union of smaller subdomains found at lower insulation values where enhancer-promoter interactions are strongly enriched (cf. Fig. 4B). At higher reciprocal insulation, interactions across boundaries are depleted. CTCF-bond loci showed a similar pattern, with even higher levels of enrichment within domains and lower enrichment across domain boundaries (Fig. 4B,C, black curves). This result is obtained irrespective of the reciprocal orientation of pairs of CTCF motifs, although enrichments are globally higher for convergent CTCF sites (Supplemental Fig. S4a). Importantly, when we considered all pairs of loci within the same active domains where enhancers and promoters were identified, or random interactions drawn from the same distribution of distances as enhancer-promoter pairs, we observed a much lower increase in interactions inside domains (Fig. 4B, green and blue curves). Moreover, interactions across domains were also depleted at low insulation levels (Fig. 4C). We obtained very similar results in NPCs and the CH12 cell line (Supplemental Fig. S4b,c).

Thus, TADs occur in the folding range where enhancer-promoter communication might be ‘optimal,’ i.e., enhancer-promoter contacts are maximally enriched within domains but begin to be depleted across domain boundaries.

The local complexity in chromosomal folding correlates with transcriptional activity and CTCF binding

We next used the CaTCH algorithm to quantify local chromosome folding complexity within each TAD and correlate it to the level of local transcriptional activity. To this aim, we first computed the number of hierarchical sublevels that can be identified within a domain (see Methods) as a measure for local folding complexity. We then used the RNA-seq profiles to assign transcripts to domains based on the genomic position of their promoters. We did not limit our analysis to the exonic signal (corresponding to mature mRNA), but we also considered the intronic reads, the latter being a more reliable measure of transcriptional activity (see Methods section). We found that at the level of single TADs, a quantitative correlation exists between the number of sublevels and both total (exonic) and unspliced mRNA reads per domain (Fig. 5A,B;

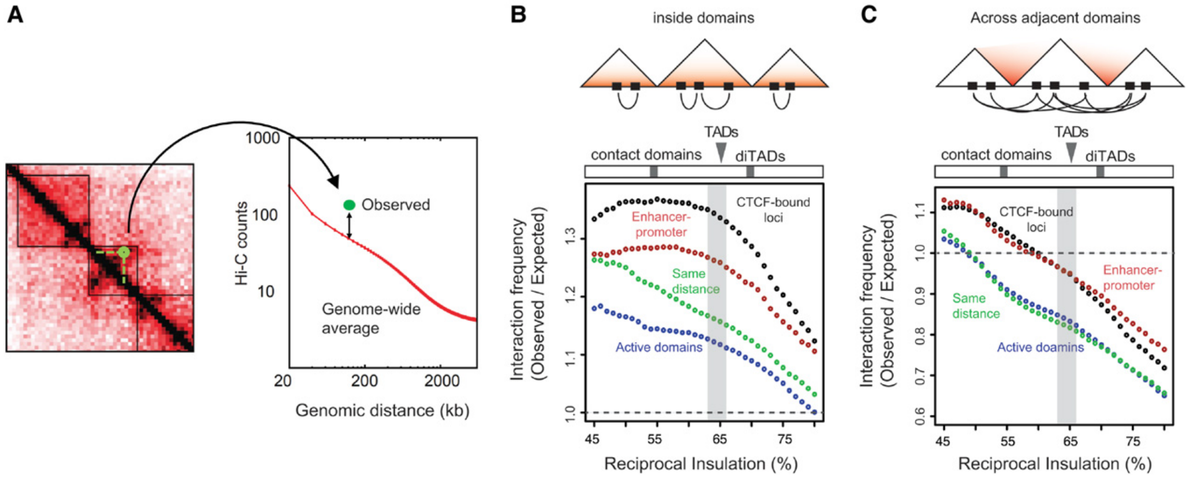


Figure 4. TADs define a scale where promoter-enhancer communication is optimal in ESCs. (*A*) Schematics of contact enrichment analysis. For each pair of loci, we calculated the ratio between observed Hi-C counts and the genome-wide average counts for loci located at the same genomic distance. (*B*) Enrichment in interactions between pairs of loci belonging to the same domain, as a function of reciprocal insulation. Colors refer to random loci within active TADs (blue), enhancer-promoter pairs (red), random loci with the same distance distribution as enhancer-promoter pairs (green), and CTCF-containing loci (black). Median enrichment over all pairs of considered loci are plotted. Gray shaded area indicates the 63%–66% confidence interval where maximal coregulation of genes occurs in partially reshuffled genomes (cf. Supplemental Fig. S3b). (*C*) Enrichment (or depletion) in interactions between pairs of loci, defined as in panel *A* but located across consecutive domains. Gray shaded area as in panel *B*.

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

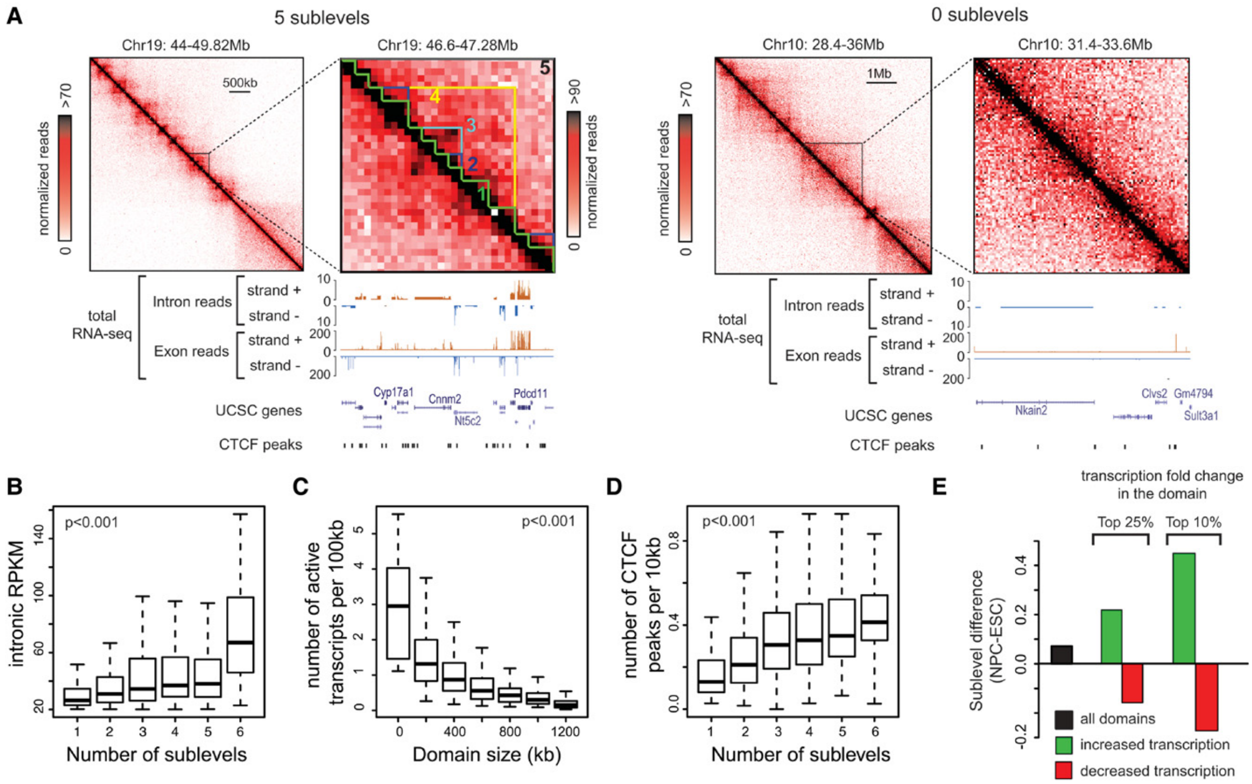


Figure 5. Local (changes in) folding complexity correlate with transcriptional activity in ESCs and during differentiation. (*A*) Examples of regions with different levels of local folding complexity and correlated transcriptional activities. (*B*) The number of sublevels in a domain correlates with the transcriptional activity within the domain (shown for domains at 65% RI in ESC). *P*-value: Student's *t*-test associated to Spearman's correlation coefficient. (*C*) Smaller domains tend to be denser in actively transcribed genes and therefore globally more active than larger domains (shown for domains at 65% RI). (*D*) The number of sublevels in a domain correlates with the density of CTCF-bound sites within the domain (shown for domains at 65% RI). (*E*) Local changes in transcriptional activities during differentiation from ESCs to NPCs correspond to changes in local hierarchical complexity (see Methods). Differences in the number of hierarchical sublevels are shown for the 25% and 10% most up- or down-regulated domains identified at 65% RI in ESC.

Supplemental Fig. S5a). The number of sublevels also correlates with the mean transcriptional level per gene (Supplemental Fig. S5b) and with the number of transcribed promoters in the domain (Supplemental Fig. S5c). We also observed that smaller TADs tend to be denser in actively transcribed genes (Fig. 5C) and are globally more active than larger domains (Supplemental Fig. S5d). In addition, the number of sublevels correlates with the density of CTCF ChIP-seq peaks *within* the domain (Fig. 5D).

These observations would predict that, during differentiation from ESCs to NPCs, local changes in transcriptional activities should correspond at least in part to changes in local folding complexity. To verify this hypothesis, we considered the set of TADs defined in ESCs and studied the changes in the number of sublevels in the same regions in NPCs. We found indeed that domains where transcriptional activity increases during differentiation tend to increase their internal structural complexity and vice versa, as exemplified by the average change in the number of sublevels in the most dynamic TADs (Fig. 5E; Supplemental Fig. S5e).

Finally, given that the local transcriptional activity and CTCF occupancy modulate folding complexity within single domains, we reasoned that sharp transitions in these quantities across domain boundaries could also contribute to domain segregation. By definition, each domain level in the folding hierarchy (including TADs) is defined by the *minimal* reciprocal insulation of its constituent domains. Thereby, each TAD in ESCs is insulated from its neighbors by at least 65%. Interestingly, we found that, at the level of single TADs, reciprocal insulation correlates with the difference in transcriptional activity and CTCF occupancy *within* vs. *across* its borders (Supplemental Fig. S5f,g). Similar results were found when considering all other levels in the hierarchy, either at lower or higher levels of insulation compared to TADs (Supplemental Fig. S5f). Thus, sharper transitions in the genomic density of CTCF binding sites and transcribed genes correspond to stronger boundaries between adjacent domains.

Discussion

Determining how enhancers exert their regulatory functions on distal promoters critically depends upon our level of understanding of the three-dimensional organization of chromatin. Several studies have provided evidence on the fundamental role of compartmentalization into TADs to instruct enhancer-promoter communication (Nora et al. 2012; Symmons et al. 2014; Lupiáñez et al. 2015; Franke et al. 2016), but they remain elusive on what makes TADs 'special' compared to other chromosomal folding layers, such as sub-TADs and notably contact domains or meta-TADs. In this study, we present a new domain-calling algorithm that is able to segment Hi-C interaction maps into nested sets of

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

topologically associating domains, based on their reciprocal physical insulation. Our approach to partition the genome into nested sets of domains has two main advantages over existing hierarchical TAD callers (Filippova et al. 2014; Lévy-Leduc et al. 2014; Shin et al. 2015; Weinreb and Raphael 2015; Chen et al. 2016; Shavit et al. 2016): (1) The CaTCH algorithm does not rely on any free parameters, except reciprocal insulation itself that is used to stratify the domains; (2) Unlike other methods that identify hierarchies of domains, where parameters have an unclear structural or biological interpretation, reciprocal insulation estimates how well a domain is segregated from its neighbors. CaTCH is fast and requires less computing power: Identifying a whole hierarchy of domains on a single 100-Mb chromosome takes <4 min on a single CPU, starting from mouse Hi-C data at 20-kb resolution. We note that reciprocal insulation is conceptually similar to the 'local contrast' measure introduced in Van Bortle et al. (2014); here, however, we used the parameter to define a full hierarchical tree of domains, rather than employing it to characterize the strength of boundaries of a given set of domains.

By applying CaTCH to published Hi-C data sets, we were able to show that previously reported topological structures are detected by the algorithm as differentially insulated levels within a continuous hierarchy of nested folding layers (Fig. 1). This gave us the possibility to compare all levels simultaneously in terms of their structural and functional properties. Based on purely structural characteristics of the domains detected over the entire mouse genome, we found that none of these sets constitutes an intrinsically privileged scale. However, we observed that the scale of TADs emerges as a privileged functional one, where fundamental properties previously associated with TADs and notably related to their role in long-range transcriptional regulation are maximized.

CTCF clustering at domain boundaries has been repeatedly reported as one of the hallmarks of topological domains across species (Dixon et al. 2012; Sexton et al. 2012; Van Bortle et al. 2014; Vietri Rudan et al. 2015). In agreement with that, we show that maximal CTCF clustering at boundaries is highly predictive of the set of domains with the most conserved boundaries across cell types (Fig. 2). In fact, finding hierarchical levels with ∼3400 domains seems to provide a sufficient operational criterion to identify the TAD scale when using CaTCH (Fig. 2D), even in the absence of matched CTCF ChIP-seq data sets.

The resolution of our data set (20 kb) does not enable the detecting of looping interactions between single CTCF sites that can be found in very high-resolution Hi-C (Rao et al. 2014) or ChIA-PET experiments (Tang et al. 2015), and it is therefore not possible to assess the precise reciprocal orientation of CTCF site clusters that occur within domain boundaries. However, between 15% and 22% of the most internal CTCF site pairs at the boundaries of TADs are convergent, which represents a maximum across the entire folding hierarchy (Supplemental Fig. S2f).

Although boundary-associated CTCF might play an important role in defining domains and in particular TADs, CTCF also pervasively binds within domains. Within a given hierarchical level and TADs in particular, domains that are more reciprocally insulated tend to have a higher imbalance in the number of CTCF-bound sites within vs. across their boundaries. Notably, regions that are highly bound by CTCF and are flanked by low-occupancy domains are highly insulated from the flanking regions (see, for example, Supplemental Fig. S5g, right). In addition, the density of CTCF-bound sites within a domain correlates with the hierarchical complexity of topological domains at all scales, including TADs (Fig. 5). Together with the fact that the hierarchical complexity also correlates with the overall transcriptional activity of a domain, this is in line with earlier findings that sub-TAD structures are strongly associated with CTCF-bound sites and active regulatory sequences (Phillips-Cremins et al. 2013). However, our results also suggest that interactions mediated by CTCF (and possibly additional factors associated with active regulatory sequences) *within* transcriptionally active domains play an important role in modulating the strength of boundaries between adjacent domains. Strong asymmetry in CTCF occupancy and transcriptional activity across boundaries can arise as a consequence of marked transitions in gene density and/or number of regulatory sequences. In addition, asymmetry can occur corresponding to cell-type–specific transitions in genomic activity between adjacent TADs (cf. Supplemental Fig. S5g, right panel). This in turn might be driven by transitions in the enrichment for cell-type–specific regulatory sequences (such as binding sites for lineage-determining transcription factors) across the boundary between the two TADs.

TADs appear in the uppermost layers in the folding hierarchy where interactions *within* active domains specifically, and between enhancers and promoters in particular, are strongly enriched compared to the genome-wide average interactions (Fig. 4). On the other hand, interactions *across* the boundaries of active TADs start to be depleted as compared to genome-wide averages. TADs thus appear to belong to the domain scale where a trade-off is established between maximizing interactions within the interior of domains and not enriching interactions across domain boundaries. In this light, it is remarkable that TADs emerge as the set of domains where the coregulation of genes during differentiation is maximal (Fig. 4). Although the precise mechanisms that govern enhancer action on promoters is still unknown, it is tempting to speculate that rather than absolute interaction frequency, the balance between interactions within and across domains determines the genomic range of action of enhancers, and this could contribute at least in part to establishing higher transcriptional coregulation at the level of TADs.

Methods

Hi-C data sets

ESCs and NPCs Hi-C data sets were obtained from Giorgetti et al. (2016). Reads from 129Sv and Cast/EiJ alleles were combined to increase coverage, and data were binned at 20-kb resolution. CH12 data are from Rao et al. (2014), binned at 10 kb. Mouse fetal liver Hi-C data are from Nagano et al. (2015), binned at 25 kb. ESC, NPC, and liver Hi-C were normalized with iterative correction (Imakaev et al. 2012). CH12 data were normalized with the VC-SQRT method (Rao et al. 2014).

The CaTCH algorithm

The algorithm takes a normalized Hi-C matrix as an input, binned at an arbitrary resolution r. The genome is first partitioned into domain seeds of size 2*r, which are progressively merged into larger domains. Reciprocal insulation (RI) is defined as in Eq. (1) in the main text. Given a threshold on RI, two consecutive domains are merged into one if their RI is smaller than the threshold. Increasing the RI threshold from 0% to 100% in steps of 0.1% results in increasingly larger domains. To lose memory of the initial partitioning of the genome into domain seeds, small shifts (two genomic bins) in domain boundaries are allowed at each step. Finally, to avoid that the discrete increase in RI threshold (0.1% steps) results in a final domain tree that depends on the order of mergings and is therefore not unique, we impose a rule on merging

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

order: A domain can be merged with either the one that precedes or the one that follows it along the genome; the pair with lowest RI is merged first (see Supplemental Methods).

Computationally generated contact maps with preferential folding levels

Control contact maps with one or two folding levels were created by generating a power law decreasing contact map for each level, to which Gaussian random noise is added (see Supplemental Methods). The contact map with zero folding layers was generated by replacing the actual Hi-C counts in the contact map for Chr 19 in ESCs with the average genome-wide counts for loci with the same genomic distance and adding Gaussian noise.

Cell culture

Culture of the female mouse ES cell line F121.6 (129Sv-Cast/EiJ) and NPC clone analyzed in Giorgetti et al. (2016) was performed as previously described (Gendrel et al. 2014; Giorgetti et al. 2016). All cell lines used in this study were characterized for absence of mycoplasma contamination.

RNA-seq data and analysis and other analyses of genomic data

Strand-specific total RNA-seq libraries from two biological replicates of ESCs and NPCs were prepared with the ScriptSeq v2 kit (Illumina) and sequenced on an Illumina HiSeq 2000 for a total of ∼30 million uniquely aligned reads per sample. Samples were aligned to mouse mm9. For details on the RNA-seq and ChIP-seq analysis, CTCF motif assignment, and enhancer calling, please refer to Supplemental Methods.

Definition of hierarchical sublevels

A subregion within a domain at any scale in the folding hierarchy was defined as a sublevel if it is detected as a domain over more than >5% of the preceding reciprocal insulation thresholds. *P*-values in transcription and CTCF content vs. number of sublevels (Fig. 5) were obtained using the function cor.test in R (Spearman method) and represent the results of Student's *t*-tests on the Spearman's correlation coefficient.

Analysis of structural reorganization during differentiation

We focused on TADs defined in ESCs and defined the number of sublevels detected in NPCs in the corresponding regions, using NPC domains below 58% since those are the domains that best match domains at 65% in ESCs (see Supplemental Fig. S2g). We estimated the local amount of structural reorganization as the change in the number of sublevels between ESCs and NPCs.

Analysis of enhancer-promoter interactions

Genomic 20-kb (ESCs and NPCs) and 10-kb (CH12) bins were assigned to 'enhancer,' 'promoter,' or 'CTCF' categories if they contain at least one of these elements (see Supplemental Methods). If a bin shows multiple classifications, it was assigned to all categories. In the analysis for Figure 4, in order to avoid including undersampled interactions due to limited Hi-C coverage at large genomic distances, we only considered pairs of loci separated by <2 Mb in ESCs and NPCs, and 1 Mb in CH12 cells. Cutoffs were chosen to exclude genomic distances where average Hi-C counts are dominated by experimental noise (Supplemental Fig. S4d).

Data access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE84724. CaTCH is provided as an R package at https://github.com/zhanyinx/CaTCH_R. Source code can be found in Supplemental Methods.

Acknowledgments

Research in the Giorgetti lab was supported by the Novartis Research Foundation. Initial analyses for this study were conceived and established in Edith Heard's laboratory, Institut Curie and PSL (Paris), where L.M.'s salary was paid for by an EMBO ASTF 563-2012 fellowship to L.M. and SyBoSS 62012 grant to Edith Heard. We thank Federico Comoglio for assistance on code development and for critically reading the manuscript, Stéphane Thiry and Tim Roloff for assistance on RNA-seq library preparation and sequencing, Edith Heard for critically reading the manuscript, NIBR computing resources and Stefan Grzybek for help with cluster and server supports, and Mikael Attia for cell culture. We acknowledge The ENCODE Project Consortium and in particular the Ren and Hardison laboratories for ChIP-Seq data sets in ESC and CH12, and the Myers laboratory for ChIP-seq data sets in fetal liver cells.

Author contributions: Y.Z. wrote the code and performed all analyses with assistance from G.T., L.M., and M.S.; L.M., E.G.S., and N.B. set up the preliminary analyses and discussed the results; I.B. performed enhancer calling and discussed the results; G.T. assisted with data analysis. L.G. prepared RNA samples. L.G. designed the study. L.G. and Y.Z. wrote the paper.

References

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58:** 268–276.

Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. 2013. Clustering of tissue-specific sub-TADs accompanies the regulation of *HoxA* genes in developing limbs. *PLoS Genet* **9:** e1004018.

Chen J, Hero AO, Rajapakse I. 2016. Spectral identification of topological domains. *Bioinformatics* **32:** 2151–2158.

Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515:** 371–375.

de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, Splinter E, Wijchers PJ, Krijger PHL, de Laat W. 2015. CTCF binding polarity determines chromatin looping. *Mol Cell* **60:** 676–684.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485:** 376–380.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* **9:** 14.

Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, et al. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538:** 265–269.

Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DC, Aitken S, et al. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11:** 852.

Fudenberg G, Mirny LA. 2012. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev* **22:** 115–124.

Gendrel A-V, Attia M, Chen C-J, Diabangouaya P, Servant N, Barillot E, Heard E. 2014. Developmental dynamics and disease potential of random monoallelic gene expression. *Dev Cell* **28:** 366–380.

Gibcus JH, Dekker J. 2013. The hierarchy of the 3D genome. *Mol Cell* **49:** 773–782.

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E. 2014. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157:** 950–963.

Downloaded from genome.cshlp.org on July 25, 2019 - Published by Cold Spring Harbor Laboratory Press

Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature* **535:** 575–579.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, 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.

Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning*. Springer, New York. http://link.springer.com/10.1007/978-0-387-84858-7.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol Cell* **48:** 471–484.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9:** 999–1003.

Junier I, Spill YG, Marti-Renom MA, Beato M, le Dily F. 2015. On the demultiplexing of chromosome capture conformation data. *FEBS Lett* **589:** 3005–3013.

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

Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. 2014. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30:** i386–i392.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161:** 1012–1025.

Merkenschlager M, Nora EP. 2016. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu Rev Genomics Hum Genet* **17:** 17–43.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502:** 59–64.

Nagano T, Várnai C, Schoenfelder S, Javierre B-M, Wingett SW, Fraser P. 2015. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* **16:** 175.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485:** 381–385.

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, et al. 2013. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153:** 1281–1295.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, 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* **112:** E6456–E6465.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148:** 458–472.

Shavit Y, Walker BJ, Lio' P. 2016. Hierarchical block matrices as efficient representations of chromosome topologies and their application for 3C data integration. *Bioinformatics* **32:** 1121–1129.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. 2015. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44:** e70.

Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24:** 390–400.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163:** 1611–1627.

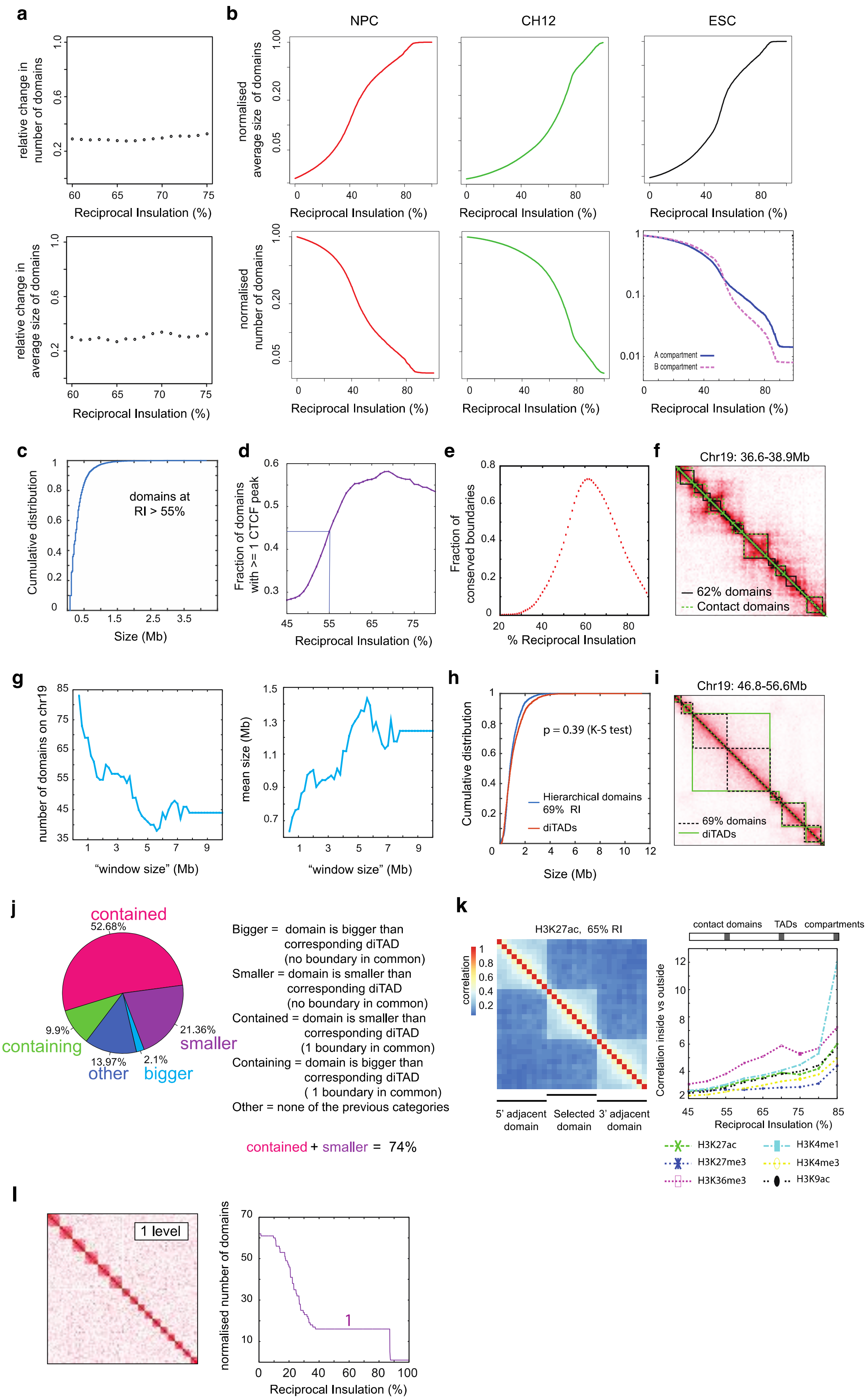
Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, et al. 2015. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* **26:** 70–84.

Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, Corces VG. 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol* **15:** R82.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10:** 1297–1309.

Weinreb C, Raphael BJ. 2015. Identification of hierarchical chromatin domains. *Bioinformatics* **32:** 1601–1609.

Received July 15, 2016; accepted in revised form January 4, 2017.



Bigger = domain is bigger than
corresponding diTAD
(no boundary in common)

Smaller = domain is smaller than
corresponding diTAD
(no boundary in common)

Contained = domain is smaller than
corresponding diTAD
(1 boundary in common)

Containing = domain is bigger than
corresponding diTAD
(1 boundary in common)

Other = none of the previous categories

contained + smaller = 74%

Supplemental Figure S1.

a. Changes in the number (top) and the mean size (bottom) of domains between when the threshold RI value is increased by 5% with respect to the x coordinate of the data point (i.e. the point at 60% represent changes when increasing from 60% to 65% etc).

b. Normalized mean size (upper panels) and number (lower panels) of domains as a function of the minimal RI threshold used to define them (semi-log scale), in NPC, CH12 and ESC. The normalized number of domains detected in regions belonging exclusively to either the A and B compartment is shown for ESCs. Normalized number of domains genome-wide for ESC is plotted in main Figure 1c.

c. Cumulative distribution of the size of domains at 55% (median 180 kb).

d. Fraction of domains with at least one CTCF bound site at both boundaries, as a function of the RI used to define domains. Straight lines indicate domains whose size is most similar to contact domains (7).

e. Fraction of boundaries of contact domains in CH12 cells (7), which are detected by the CaTCH algorithm applied to same Hi-C dataset, as a function of the RI used to define hierarchical domains. Maximal overlap is found at 62% RI.

f. Example of domains at 62% of RI in CH12 (black lines) and contact domains (7) (green lines) for a small region of chromosome 19 in CH12 cells.

g. Number (left panel) and mean size (right panel) of domains detected by the directionality index method (diTADs) (7) as a function of the “window size” parameter required by the method. No intrinsic rule allows to choose any window size below 8 Mbps from these plots.

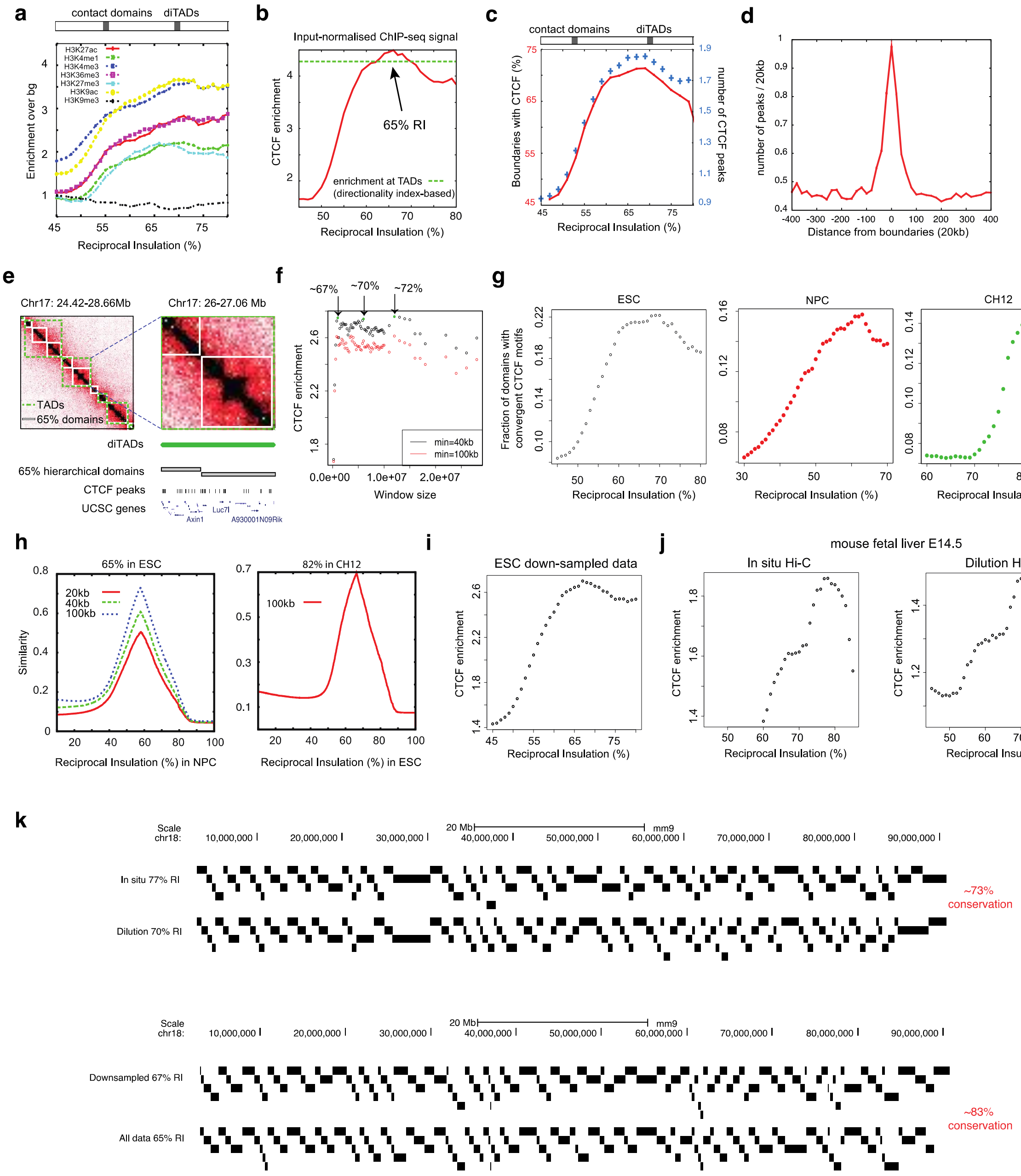
h. Cumulative distribution of domains size for diTADs (red curve) and domains at 69% RI (blue curve). KS-test shows that these two sets of domains are not significantly different.

i. Example of domains at 69% of RI (black squares) and diTADs (green squares) for a small region of chromosome 19 in ESCs. The CaTCH algorithm is more sensitive to small variations in contact probabilities within large, uniform domains.

j. Pie chart representing the properties of non-conserved domains between diTADs and the set of domains at 69%. Most domains at 69% divide a diTAD into smaller domains.

k. Enrichment in correlation of histone marks within vs. across domains (7). Correlation is maximal at the scale of compartments, consistent with the notion that compartments A and B represent stretches of active and inactive chromatin respectively.

l. Right: Number of domains detected by CaTCH as a function of RI in computationally generated contact maps with one preferential folding level and different domain sizes. Left: The corresponding heat map.



Supplemental Figure S2.

a. Enrichment of histone modification at domains boundaries as a function of RI. All active marks reach maximal enrichment at TAD level. H3K9me3 is depleted everywhere in the hierarchy and in particular at the scale of TADs.

b. Enrichment in CTCF input-normalised ChIP-seq signal at the boundaries as a function of RI. Enrichment is maximum at ~65% RI and slightly higher than diTADs (green line).

c. The number of boundaries that contain at least one CTCF peak, and the number of CTCF peaks per boundary are both maximized around 65% RI.

d. Meta-boundary profile showing CTCF peak abundance in the genomic neighborhood of domains boundaries at 65% RI. Genomic coordinates were aligned to the position of boundaries.

e. Domains at 65% often split a single TAD into smaller domains.

f. CTCF enrichment at the boundaries of domains detected by the directionality index algorithm using different combinations of the ‘window size’ and ‘minimum size’ parameters, which determine the maximal genomic distance used in the computation of the directionality index and the minimum domain size, respectively. Arrows indicate examples of domains where CTCF enrichment is maximal and the corresponding RI value with maximal boundary overlap.

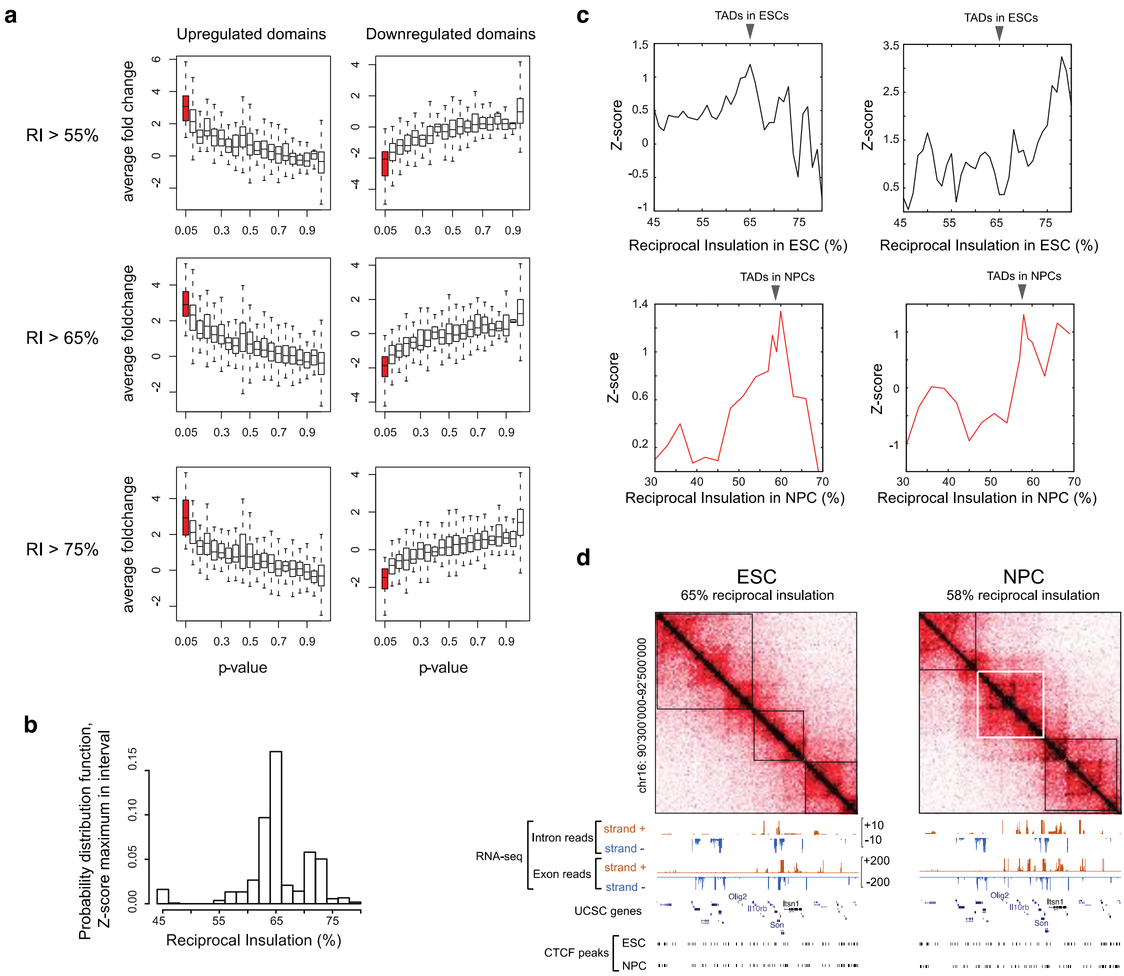
g. Fraction of domains with convergent CTCF motifs at the boundaries as a function of RI. At 65% RI, according to our criterion (see Methods), ~22% of domains possess convergent CTCF sites in ESC (left panel). At ~82% RI ~14% of domains possess convergent CTCF sites in CH12 (right panel).

h. Fraction of conserved boundaries between domains defined in ESCs and either NPCs (left) or CH12 (right). In all cases, maximal conservation of boundaries occurs at the insulation value where maximal enrichment of CTCF at boundaries was found in the various cell types.

i. CTCF enrichment analysis at the domain boundaries as a function of RI, performed on 2X down-sampled dataset in ESC. Maximal enrichment occurs at 67% (cfr. **Figure 2a**).

j. CTCF enrichment analysis at domain boundaries as a function of RI performed on *in situ* (left) and *dilution* (right) Hi-C libraries from mouse E14.5 fetal liver cells (Nagano 2015). Maximal enrichment occurs at different RI for the two protocols (77% for *in situ* and 70% for *dilution* Hi-C).

k. Examples of domains at the scale where CTCF enrichment is maximal. Upper panel shows the domains for *in situ* (77% RI) versus *dilution* (70%) protocols in fetal liver cells. Lower panel shows the domains for the full ESC dataset versus the down-sampled dataset.



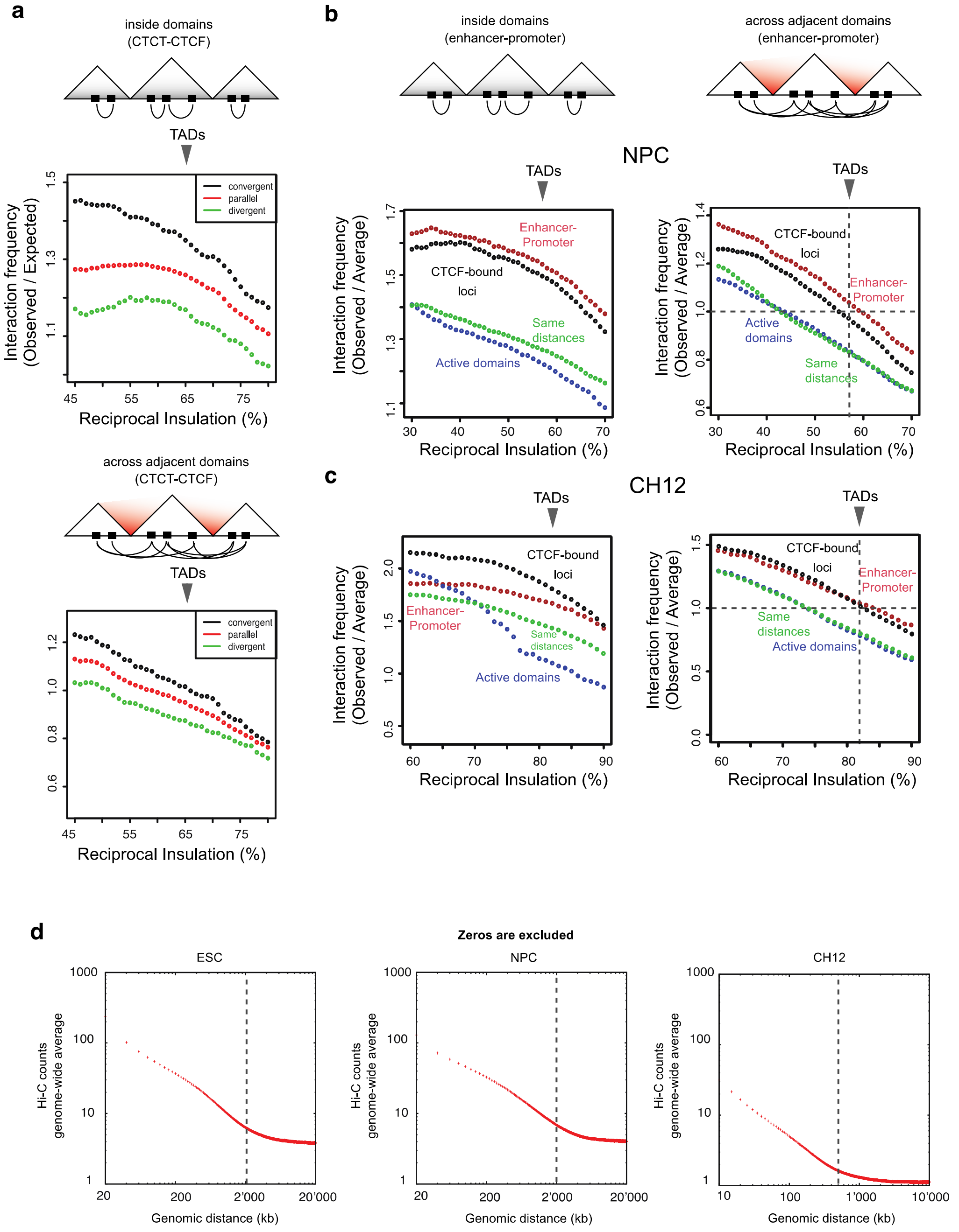
Supplemental Figure S3.

a. At all insulation levels, the subset of domains that we detected to be up- (left panels) or downregulated (right panels) at the p<=0.05 level (red rectangle) are those where the mean fold change of expression level within the domain is maximal. Shown for 3 RI value (55% top, 65% center, 75% bottom).

b. Density histogram showing the RI value where the maximum Z-score for coordinately down-regulated genes occurs when randomly reshuffling the fold-changes of 10% of genes. For 66% of the partially reshuffled genomes, the Z-score maximum was found to be located within a 4% interval around 65% reciprocal insulation (63%-66%).

c. Analysis of transcriptional co-regulation with a correction applied to genes on chrX to account for the inactive X in NPCs (see Supplemental Methods). Maximal transcriptional co-regulation consistently occurs at the scale of TADs.

d. Example of domains that were *de novo* created during differentiation and thus detected using domains based on Hi-C data in NPCs.



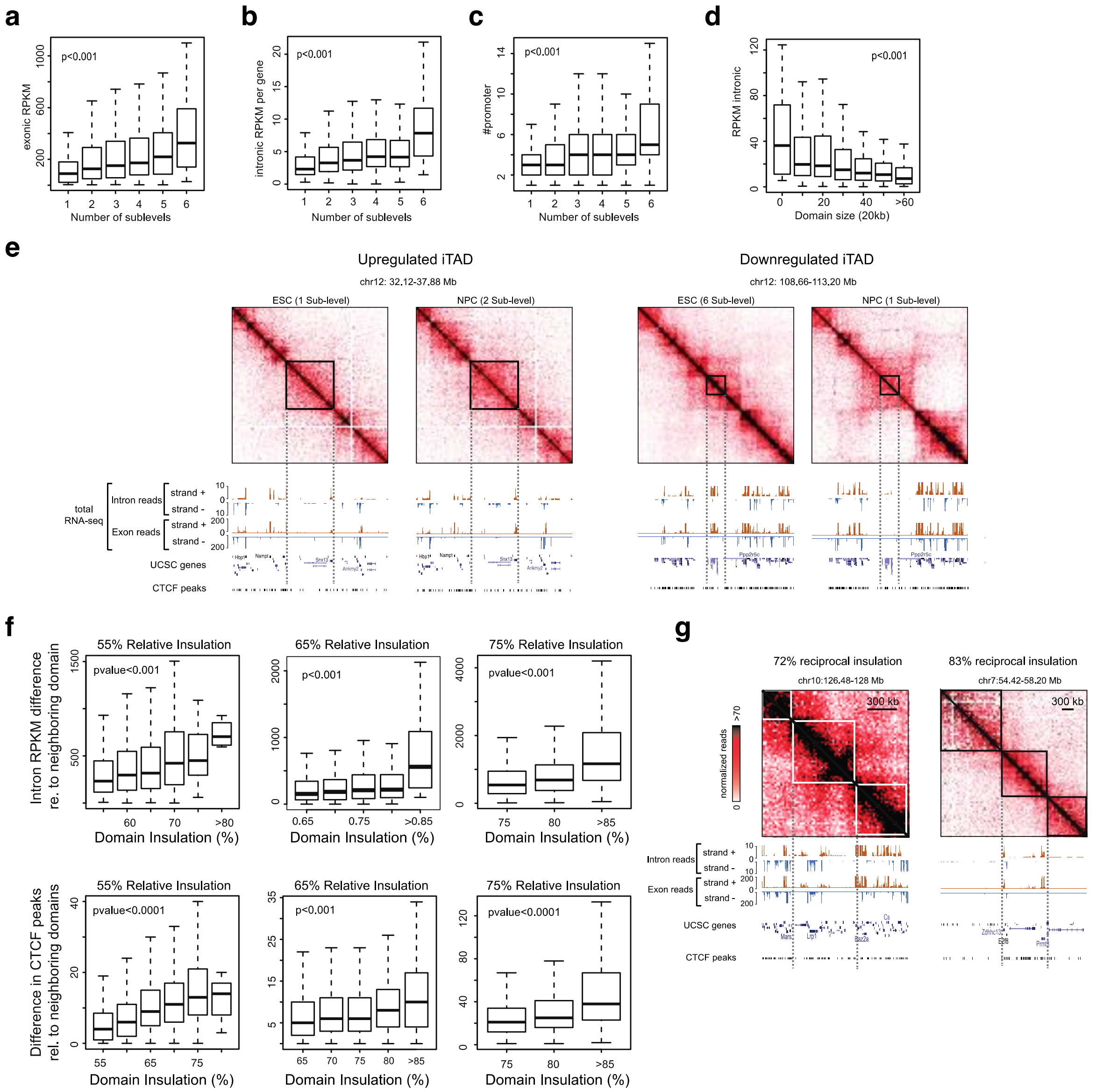
Supplementary Figure 4.

a. Enrichment in interactions between pairs of loci bound by CTCF and belonging to the same domain or across the boundary with the neighboring domain, as a function of reciprocal insulation in ESCs. Pairs of CTCF sites occur at convergent, divergent or parallel CTCF DNA binding motifs are plotted separately.

b. Enrichment in interactions between pairs of loci belonging to the same domain or across the boundary with the neighboring domain, as a function of reciprocal insulation in NPCs. Colors refer to random loci within active TADs (blue), enhancer-promoter pairs (red), random loci with the same distance distribution as enhancer-promoter pairs (green) and CTCF-containing loci (black). Median enrichments over all pairs of considered loci are plotted.

c. Same as in panel b, for CH12 cells.

d. Scaling of contact probabilities as a function of genomic distance. Genome-wide average Hi-C contacts are plotted for the three cell types analyzed in this work. Zeroes were not included in the averages to highlight the regime where random sequencing noise sets in at large genomic distances. Black lines indicate the cutoff distance used in the analysis for Figure 4 (2 Mb for ESCs and NPCs, 1 Mb for CH12).



Supplemental Figure S5.

a. The number of sub-levels of a domain correlates with the average transcriptional activity measured by exonic RPKM (shown here for TADs).

b. Same as in a but using intronic RPKM per gene.

c. The number of sub-levels of a domain correlates with the number of active promoters within the domain (shown for iTADs). p-value: t-test associated to Spearman correlation.

d. More active domains are smaller than inactive domains (shown here for TADs at 65% RI), measured by intronic RPKM. p-value: t-test associated to Spearman correlation.

e. Example of domains (highlighted by black squares) where transcription and number of sub-levels increase (left) or decrease (right) during differentiation.

f. The absolute value of the reciprocal insulation of a domain correlates with the level of changes in expression level (upper panels) and in the number of CTCF peaks (lower panels) with the adjacent domains (shown for domains at 55% RI (left panels), 65% RI (middle panels) and 75% RI (right panels)).

g. Example of domains at 65% RI that are highly (right) or lowly (left) insulated from neighboring domains, and have dissimilar or similar transcriptional levels and number of CTCF peaks, respectively, compared to neighboring domains.

Zhan *et al.*, 'Reciprocal insulation analysis of Hi-C data show that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes'

Supplemental Methods

Hi-C datasets

ESCs and NPCs Hi-C datasets were obtained in Ref. (Giorgetti et al. 2016). Reads from 129Sv and Cast/EiJ alleles were combined to increase read depth, and data were binned at 20 kb resolution. CH12 data are from Rao *et al.* (Rao et al. 2014), binned at 10 kb. Mouse fetal liver Hi-C data are from Nagano *et al.* (Nagano et al. 2015), binned at 25 kb. ESC, NPC and liver Hi-C were normalized with iterative correction (Imakaev et al. 2012). CH12 were normalized with the VC-SQRT method (Rao et al. 2014).

Domain-calling algorithm

The CaTCH algorithm takes as an input a normalized Hi-C matrix, binned at an arbitrary resolution r. The genome is first partitioned into seeds of domains of size 2*r, which are then progressively merged into large domains. Merging of two consecutive domains A and B is determined by the reciprocal insulation (RI) measure:

$$RI(A,B)=[\,P_{in}(A)+P_{in}(B)-P_{out}(A,B)\,]/\,[P_{in}(A)+P_{in}(B)]*100 \qquad (1)$$

Where P_{in} and P_{out} are the average Hi-C counts within domains A and B, and across their boundary respectively (see **Figure 1a** in the main text).

A threshold on RI is then defined, and any two consecutive domains whose RI is below the threshold are merged in a single domain. The threshold is progressively increased from 0% to 100% in steps of 0.1%, resulting in increasingly larger domains. The fact that only consecutive domains can be merged ensures that the overall organization of the domains is tree-like, excluding the possibility of interactions between distant domains. This could be observed otherwise by imposing a different distance based on the Hi-C map, which is not strictly ultrametric. In order to lose dependency on the initial partitioning of the genome in the final determination of domain boundaries, we allowed small shifts in the boundaries of domains (2 genomic bins) at each step. Note that the domains identified by CaTCH do not depend on bin size, provided the domain is larger than the genomic bin.

Since the increase of the threshold is discrete, the above procedure undergoes the risk of being dependent on the order of mergings, which would result in a non-unique tree. To overcome this problem, we set a specific rule on the matching order. Namely, if a domain can be merged with either the preceding or the following along, the pair that has the lowest RI is merged first. This is in fact equivalent to merging domains according to their order along the chromosomes, and increasing smoothly (rather than in discrete steps) the threshold on the reciprocal insulation value. Indeed, smoothly increasing the threshold corresponds to cutting the hierarchical tree densely enough that one is able to always merge the domains with the lowest RI.

Computationally generated contact maps with preferential folding levels

To generate contact maps characterized by one or two preferential folding levels, we generated a contact map for each individual level (where contact probabilities decrease as a power law with increasing genomic distance), to which a weak background Gaussian noise was added. For example, to generate a pseudo-genome with two folding levels (see right panel in main **Figure 1f**), we first generated a uniform (power-law decaying) contact map with Gaussian noise. Then, we partitioned the matrix into a set of small domains **d1**={$d1_i$} (smallest squares along the diagonal in **Figure 1f**). The first folding level was generated within this set of domains by adding a new power-law decreasing interaction pattern. We then merged pairs of adjacent domains (e.g. $d1_1$ with $d1_2$; $d1_3$ with $d1_4$ and so on) leading to a second set of domains **d2**=$\{d2_i\}$ to which the same power-law decreasing interaction pattern was added. The contact map with no folding layer was generated by replacing the actual Hi-C counts in the contact map for chr19 in ESCs with the average genome-wide counts for oci with the same genomic distance, and adding Gaussian noise.

CTCF motif analysis

We called CTCF peaks using macs2 (Zhang et al. 2008) using default parameters. We used the top 1000 high-significance peaks to define a CTCF position-weight matrix, resulting in a PWM that is indistinguishable from previous reports (Jaspar accession number MA0139.1; see Ref. (Mathelier et al. 2013)). We then used MEME tool (Bailey and Elkan 1994) with a custom background, which includes non-overlapping mappable sequences with the same distribution size of the top 1000 peaks, to perform *de novo* motif discovery. Finally, we used the motif identified within the top 1000 CTCF peaks called by macs2 to extract the position and directionality of CTCF-bound sites among all the peaks using the MAST tool (Bailey and Gribskov 1998).

Boundary conservation

In order to identify the fraction of boundaries that are conserved either between cell types or domain sets, we allowed a 40-kb tolerance in boundary conservation between contact domains and sets of domains in the hierarchy of CH12 cells; for comparison with compartments we allowed a tolerance of 750kb; for all the other comparisons, we allowed a tolerance of 100-kb.

Cell culture

The female mouse ES cell line F121.6 (129Sv-Cast/EiJ) was grown on mitomycin C-inactivated MEFs in ES cell media containing 15% FBS (Gibco), 10^{-4}M ß-mercaptoethanol (Sigma), and 1000U/ml of leukaemia inhibitory factor (LIF, Chemicon). Culture of the same NPC clone that was analyzed in wa(Giorgetti et al. 2016)s performed as previously described (Gendrel et al. 2014; Giorgetti et al. 2016). All cells used in this study were characterized for absence of mycoplasma contamination.

RNA-seq data, analysis and transcript annotation

After Trizol extraction, strand-specific total RNA-seq libraries from two biological replicates for both ESCs and NPCs were prepared with the ScriptSeq v2 kit (Illumina) and sequenced on an Illumina HiSeq 2000 for a total of ~30 million uniquely aligned reads per sample on average. Libraries were prepared in two technical replicates per biological replicate (technical replicates were pooled for subsequent analyses). All samples were aligned to mouse mm9 using QuasR (Gaidatzis et al. 2015) keeping uniquely mappable reads only. A complete list of all non-overlapping known genes from UCSC (Carlson M and Maintainer BP. *TxDb.Mmusculus.UCSC.mm9.knownGene: Annotation package for TxDb object(s)*. R package version 3.2.2) was used to quantify both exonic and intronic transcription. Levels were estimated by separately aligning the reads to exonic and intronic regions and quantifying RPKMs as

$$RPKM=M/(N*L)*1'000*1'000'000$$

Where M is the mapped reads to the genomic region, L is the length of the region (sum of all exons or introns for each gene) and N is the total number of mapped reads.

We used the DESeq2 package (Love et al. 2014) to perform differential gene expression analysis between ESCs and NPCs. Cutoff on q-value<=0.05 and on a fold-change larger than 3 were used to define differentially expressed genes.

ChIP-seq analysis

We analysed the available ChIP-seq datasets listed in **Supplemental Table S1**. Reads were aligned to mouse mm9 using (Gaidatzis et al. 2015) and only the uniquely mapped reads were kept for further analysis. Quantification of ChIP-seq signal was made using the csaw package (Lun and Smyth 2016), in particular using the function windowCounts with options dedup=T and minq=28. A window of 10 kb was used for quantification. If more than one replicate were available, all replicates were combined using the geometric mean of the mapped reads. Normalisation over input was performed as in (Perner et al. 2014) using a pseudo-count of 8. Peaks were called with macs2 (Zhang et al. 2008) using default parameters. A peak is assigned to a specific boundary if it belongs to the 40kb window centered on the boundary coordinate.

Transcriptional coregulation

To determine whether a domain is transcriptionally co-regulated during differentiation, a cyclic permutation of gene locations is performed. We defined a domain at any scale in the hierarchy to be co-regulated, if the number of co-regulated genes in the domain is larger than in 95% of the cyclic permutated genomes (empirical $p<=0.05$). For each insulation value, we calculated the number of domains (N_{obs}) that are up or down-regulated. In order to measure the statistical enrichment of N_{obs}, we calculated a Z-score as follows. We randomly reshuffled gene positions in the genome N=2000 times, and calculated the mean value (N_{exp}) and the standard deviation (σ) of the number of up- or down-domains (defined as described above) in the randomized genomes. The Z-score was defined as:

$$\text{Z-score} = (N_{obs} - N_{exp}) / \sigma$$

Enhancer calling

Enhancer regions were identified taking advantage of H3K27ac, H3K4me1, H3K4me3 and CTCF ChIP-Seq data (**Supplemental Table S1**) as follows. We used H3K27ac peaks (called with macs2 (Zhang et al. 2008) with qvalue <= 10E-8) as landmark regions. We then expanded the peak regions to +/-1kb and evaluated the ratio between H3K4me1 and H3K4me3 signal in these regions. Since the distribution of ratios is bimodal, we could define a list of regions with high H3K4me1 and low H3K4me3 (Heintzman et al. 2007). This allows us to distinguish enhancer regions (characterized by high H3K4me1 and low H3K4me3) from promoter regions (characterized by low H3K4me1 and high H3K4me3). From this list of regions, we finally defined enhancers by discarding those regions that overlap with conserved CTCF peaks conserved across ESCs and NPCs (putative insulators), and those that localize within +/-2.5kb from the gene promoters (putative core promoters and TSS-proximal, *cis*-acting regulatory elements).

Analysis of enhancer-promoter interactions

For each pair of genomic loci used in the analysis, we calculated the ratio between the observed Hi-C counts and the genome-wide average Hi-C count (including zeroes) for loci that are separated by the same genomic distance. The median ratios for interactions occurring within a domain, or across two adjacent domains, were used for plotting the curves in **Figure 4**. Similar results were found using mean values (data not shown). To avoid including under sampled interactions due to limited Hi-C coverage at large genomic distances, we only considered pairs of loci separated by less than 2Mb in ESCs and NPCs, and 1 Mb in CH12 cells. Cutoffs were chosen to exclude genomic distances where average Hi-C counts are dominated by experimental noise (**Supplemental Figure S4d**). Genomic 20-kb (ESCs and

NPCs) and 10-kb (CH12) bins were assigned to 'enhancer', 'promoter' or 'CTCF' categories if they contain at least one of these elements, identified as described before. If a bin shows multiple classifications, we assigned it to all the categories.

Correction to account for the presence of an inactive X chromosome in NPCs

The presence of an inactive X chromosome in the NPC sample we analyzed implies that only one copy of the genes on chromosome X is active (except the set of escape genes identified in the same NPC clone in (Giorgetti et al. 2016)). As a consequence, the expression level of a gene (excluding escapees) that increases by a factor 2 specifically on the active X during differentiation will be detected as unchanged in non-allelic RNA-seq data. To correct for this issue in the definition of down- and up- regulated chrX genes (except escapees), we introduced a modified criterion compared to autosomal genes:

$$FC < -log2(3) - log2(2) \qquad \text{for down} - \text{regulation}$$

$$FC > log2(3) - log2(2) \qquad \text{for up} - \text{regulation}$$

where the factor -log2(2) accounts for the twofold reduction in the detected expression level of genes on the active X in NPCs.

Correlation of histone marks within and across domains

To look at correlation of histone modification we proceeded as in (Rao et al. 2014). To briefly summarize the method, we divided each domain into 10 bins, where the bin size was a tenth of the size of the domain. For each domain and its corresponding adjacent domains we then recorded the mean value of the chromatin mark of interest for each bin.

This procedure yielded a matrix whose length was the number of domains, and whose width was 30. By calculating the correlation of the columns of this matrix, we obtain a 30x30 correlation matrix (**Supplemental Figure S1k**). This correlation matrix represents how correlated the chromatin marks are at any two loci within and across domains.

Source Code

```
#include <R.h>
#include <Rinternals.h>
#include <Rmath.h>
#define ND          1000
#define MINSIZE  15
#define MAXMOVE            3
#define MINDIST 1

float max(float a, float b){
            if(a>b) return a;
            else return b;
}

float min(float a, float b){
      if(a<b) return a;
      else return b;
}
//calculate total counts
float sum(int i, int j, unsigned short **mat)
{
            float x=0,h;
            int k,l;
```

```
        for (k=i;k<=j;k++)
                for (l=i;l<=j;l++)
                        x += mat[k][l];
        return x;
}

float dist(int i1, int j1, int i2, int j2, unsigned short **mat)
{
    float x=0,v=0,di=0,d1=0,d2=0;
    int k,l;
    for(k=i1;k<=j1;k++)
         for(l=i2;l<=j2;l++) if(k!=l && abs(l-k)>=MINDIST) x+=mat[k][l];
    v=(j1-i1+1)*(j2-i2+1)-1;

        for(k=i1;k<=j1;k++)
                for(l=i1;l<=j1;l++)     if(k!=l && abs(l-k)>=MINDIST) d1+=mat[k][l];
        for(k=i2;k<=j2;k++)
                for(l=i2;l<=j2;l++)     if(k!=l && abs(l-k)>=MINDIST) d2+=mat[k][l];
        di=(x/v)/((d1+d2)/((j1-i1+1)*(j1-i1)+(j2-i2+1)*(j2-i2)));
        return di;
}

SEXP catch(SEXP input)
{
        int i,j,k,id,joined,imin=99999,size=0,tot=0,appo=0;
        //float **insulation;
        float dt,p[ND+1],prevdist=0,newdist=0;
        int nrow,ncol;

        unsigned short **cfrom,**cto,*ncl;
        unsigned short **mat;
        SEXP out,attrib,prof,ncluster;
        FILE *fp;

        nrow = INTEGER(getAttrib(input, R_DimSymbol))[0];
        ncol = INTEGER(getAttrib(input, R_DimSymbol))[1];

    for(i=0;i<nrow;i++)
        for(j=0;j<2;j++) {
                        if((j==0 || j==1) && REAL(input)[i+2*nrow]!=-1){
                                if(REAL(input)[i+j*nrow]>size)  size=REAL(input)[i+j*nrow];
                                if(REAL(input)[i+j*nrow]<imin) imin=REAL(input)[i+j*nrow];
                        }
        }

        size++;
        mat = (unsigned short **) calloc(size,sizeof(unsigned short *));
    for (i=0;i<size;i++) mat[i] = (unsigned short *) calloc(size,sizeof(unsigned short));
    for (i=0;i<size;i++)
         for (j=0;j<size;j++) mat[i][j]=0;
        for(i=0;i<nrow;i++){
                if(REAL(input)[i+2*nrow]!=-1)
        mat[(int)REAL(input)[i+0*nrow]][(int)REAL(input)[i+1*nrow]]=(unsigned short) REAL(input)[i+2*nrow];
        }

        cfrom = (unsigned short **) calloc(ND+1,sizeof(unsigned short *));
        for (i=0;i<ND+1;i++) cfrom[i] = (unsigned short *) calloc(size,sizeof(unsigned short));
        cto = (unsigned short **) calloc(ND+1,sizeof(unsigned short *));
        for (i=0;i<ND+1;i++) cto[i] = (unsigned short *) calloc(size,sizeof(unsigned short));
        ncl = (unsigned short *) calloc(ND+1,sizeof(unsigned short));

        for (i=0;i<ND+1;i++) ncl[i]=0;
        for (i=0;i<(int)(size-imin)/(MINDIST+1);i++)
        {
                cfrom[0][i]=i*(1+MINDIST)+imin;
                cto[0][i]=(i+1)*(1+MINDIST)+imin-1;
        ncl[0]++;
    }

    Rprintf("Clustering on different thresholds: \n");
    for (id=1;id<=ND;id++)        // increasing threshold
    {
        dt = (float) (ND-id)/ND;
        if(id%100==0)      Rprintf("Relative Insulation: %f\n",1-dt);
        for (i=0;i<ncl[id-1];i++)        // run on clusters
        {
            joined=-1;

            cfrom[id][ncl[id]] = cfrom[id-1][i];
            cto[id][ncl[id]] = cto[id-1][i]
            for (k=i+1;k<ncl[id-1];k++)   // clusters to join previous
            {


                if ( dist(cfrom[id][ncl[id]],cto[id][ncl[id]],cfrom[id-1][k],cto[id-1][k],mat) >= dt )
                {
                    cfrom[id][ncl[id]] = cfrom[id-1][i];
                    cto[id][ncl[id]] = cto[id-1][k];

                    joined = k;
                }
                else break;
            }

            if (joined==-1)
            {
                cfrom[id][ncl[id]] = cfrom[id-1][i];
                cto[id][ncl[id]] = cto[id-1][i];
                ncl[id] ++;
            }
            else
            {
                i=joined+1;

                ncl[id] ++;
                i = joined;
            }
        }

        //movement
    for(i=0;i<ncl[id]-1;i++){
        //except last
            if((cto[id][i]-cfrom[id][i]>(2*MAXMOVE) && cto[id][i+1]-cfrom[id][i+1]>(2*MAXMOVE))
&& (cto[id][i]-cfrom[id][i]>MINSIZE || cto[id][i+1]-cfrom[id][i+1]>MINSIZE)){
                prevdist=dist(cfrom[id][i],cto[id][i],cfrom[id][i+1],cto[id][i+1],mat);
                    for(j=1;j<MAXMOVE;j++){
                    newdist=dist(cfrom[id][i],cto[id][i]+j,cfrom[id][i+1]+j,cto[id][i+1],mat);
                        if(newdist<prevdist){
                            prevdist=newdist;
                            cto[id][i]=cto[id][i]+j;
                            cfrom[id][i+1]=cfrom[id][i+1]+j;
                        }
               newdist=dist(cfrom[id][i],cto[id][i]-j,cfrom[id][i+1]-j,cto[id][i+1],mat);

             if(newdist<prevdist){
               prevdist=newdist;
               cto[id][i]=cto[id][i]-j;
               cfrom[id][i+1]=cfrom[id][i+1]-j;
             }



                }
```

```
                        }

            }

            }
            Rprintf("\n");

            PROTECT(ncluster=allocMatrix(REALSXP,ND+1,2));

            for (i=0;i<ND+1;i++) {
                        REAL(ncluster)[i+0*(ND+1)]=(float)i/ND;
                        REAL(ncluster)[i+1*(ND+1)]=ncl[i];
            }
        tot=0;
        for (i=1;i<=ND;i++)
             for (j=0;j<ncl[i];j++)  tot++;
            appo=0;
            PROTECT(out = allocMatrix(REALSXP, tot,3));

        for (i=1;i<=ND;i++)
             for (j=0;j<ncl[i];j++){
                    REAL(out)[appo+tot*0] =(float)i/ND;
                    REAL(out)[appo+tot*1] =cfrom[i][j];
                    REAL(out)[appo+tot*2] =cto[i][j];
                    appo++;
             }

            PROTECT(prof=allocVector(VECSXP,2));
            PROTECT(attrib=allocVector(STRSXP,2));
            SET_STRING_ELT(attrib,0,mkChar("clusters"));
            SET_STRING_ELT(attrib,1,mkChar("ncluster"));

            SET_VECTOR_ELT(prof,0,out);
            SET_VECTOR_ELT(prof,1,ncluster);
            setAttrib(prof, R_NamesSymbol,attrib);

            UNPROTECT(4);
            return prof;

}
```

References

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. pp. 28–36, AAAI Press, Menlo Park, California.

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.

Gaidatzis D, Lerch A, Hahne F, Stadler MB. 2015. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**: 1130–1132.

Gendrel A-V, Attia M, Chen C-J, Diabangouaya P, Servant N, Barillot E, Heard E. 2014. Developmental Dynamics and Disease Potential of Random Monoallelic Gene Expression. *Dev Cell* **28**: 366–380.

Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**: 575–579.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lun ATL, Smyth GK. 2016. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res* **44**: e45–e45.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C, Chou A, Ienasescu H, et al. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids* **42**: D142-7.

Nagano T, Várnai C, Schoenfelder S, Javierre B-M, Wingett SW, Fraser P. 2015. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* **16**: 175.

Perner J, Lasserre J, Kinkley S, Vingron M, Chung H-R. 2014. Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Res* **42**: 13689–13695.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **162**: 687–688.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Chapter II: DamC reveals principles of chromatin folding in vivo without crosslinking and ligation

Josef Redolfi*, Yinxiu Zhan*, Christian Valdes*, Mariya Kryzhanovska, Isabel Guerreiro, Vytautas Iesmantavicius, Tim Pollex, Ralph S. Grand, Eskeatnaf Mulugeta, Jop Kind, Guido Tiana, Sebastien Smallwood, Wouter de Laat, Luca Giorgetti

Redolfi, Zhan and Valdes contributed equally to this work.

I performed the analysis of all the data, wrote the physical model, and wrote the paper together with Josef Redolfi and Luca Giorgetti.

Summary:
Most of the current knowledge on chromatin folding has come from 3C methods, which detect chromatin interactions using crosslinking and ligation. Crosslinking and ligation have often been criticized as potential sources of experimental biases raising the question of whether TADs and chromatin loops really exist in living cells. In this study, we developed a crosslinking- and ligation-free method, named DamC, to detect chromosomal interactions at the molecular scale in living cells. DamC provides an orthogonal validation of chromatin structures such as TADs and CTCF-mediated chromatin loops in living cells. By combining DamC with genetic engineering, we showed that we can rewire genome architecture within TADs by inserting ectopic CTCF sites.

nature structural & molecular biology

ARTICLES
https://doi.org/10.1038/s41594-019-0231-0

DamC reveals principles of chromatin folding in vivo without crosslinking and ligation

Josef Redolfi[1,2,8], Yinxiu Zhan[1,2,8], Christian Valdes-Quezada[3,4,8], Mariya Kryzhanovska[1], Isabel Guerreiro[3,4], Vytautas Iesmantavicius[1], Tim Pollex[5], Ralph S. Grand[1], Eskeatnaf Mulugeta[6], Jop Kind[3,4], Guido Tiana[7], Sebastien A. Smallwood[1], Wouter de Laat[3,4] and Luca Giorgetti[1]*

Current understanding of chromosome folding is largely reliant on chromosome conformation capture (3C)-based experiments, where chromosomal interactions are detected as ligation products after chromatin crosslinking. To measure chromosome structure in vivo, quantitatively and without crosslinking and ligation, we implemented a modified version of DNA adenine methyltransferase identification (DamID) named DamC, which combines DNA methylation-based detection of chromosomal interactions with next-generation sequencing and biophysical modeling of methylation kinetics. DamC performed in mouse embryonic stem cells provides the first in vivo validation of the existence of topologically associating domains (TADs), CTCF loops and confirms 3C-based measurements of the scaling of contact probabilities. Combining DamC with transposon-mediated genomic engineering shows that new loops can be formed between ectopic and endogenous CTCF sites, which redistributes physical interactions within TADs. DamC provides the first crosslinking- and ligation-free demonstration of the existence of key structural features of chromosomes and provides novel insights into how chromosome structure within TADs can be manipulated.

Characterizing chromosome folding is fundamental to enhancing the understanding of gene expression and how it potentially constrains genome evolution. Chromosome conformation capture (3C) methods, and notably their high-throughput sequencing-based derivatives such as Hi-C, 5C and 4C[1], have greatly contributed to current understanding of genome architecture, revealing that chromosome folding is driven by at least two independent mechanisms. On the one hand, the mutually exclusive associations between transcriptionally active or inactive loci generate the so-called A and B compartments[2]. On the other hand, chromatin loops are formed between regulatory sequences and between convergent CCCTC-binding factor (CTCF) binding sites, the latter through cooperative action between cohesin and the DNA-binding protein CTCF[3]. The interplay between compartmentalization and CTCF–cohesin looping results in complex hierarchies of folding domains[4,5], among which topologically associating domains (TADs)[6–8] stand out as preferential functional units[9]. The involvement of CTCF in loop formation has been demonstrated using global depletion experiments[10,11], as well as targeted deletions and inversions of CTCF sites leading to loss of looping interactions[12–14]. The underlying mechanisms are, however, still incompletely understood. An influential hypothesis is that CTCF-mediated interactions occur as cohesin extrudes chromatin loops until it is blocked by CTCF bound to DNA in a defined orientation[15]. According to this hypothesis, ectopic insertion of CTCF sites could result in newly established loops on endogenous CTCF sites, depending on their mutual orientation. Whether this actually occurs and how it modifies interactions within TADs have, however, not been demonstrated.

In 3C, detection of spatial proximity relies on formaldehyde crosslinking followed by digestion and ligation of crosslinked chromatin[1]. Crosslinking and ligation are sources of potential experimental bias, raising the question of whether structures detected by 3C methods actually exist in living cells[16–19]. The frequency of 3C crosslinking is assumed to be proportional to absolute chromosomal contact probabilities, and is used to build mechanistic physical models of chromosome folding[20,21], including the loop-extrusion model[14,15,22,23]. However, formal proof of this assumption is missing. Independent techniques such as DNA fluorescence in situ hybridization[6,24], genome architecture mapping[25], native 3C[26] and split-pool recognition of interactions by tag extension[27] have also detected loops, TADs and compartments. Nevertheless, these methods still involve substantial biochemical manipulation of cells and employ either crosslinking or ligation.

An alternative approach to the study of chromosomal contacts without crosslinking and ligation is by recruitment of an ectopic DNA-modifying enzyme to specific genomic locations, and detection of chemically modified DNA at sequences that physically interact with the recruitment sites. Three previous studies have provided proof of principle for such an approach using a modified version of DNA adenine methyltransferase identification (DamID)[26,28,29]. In DamID, the bacterial DNA adenine methyltransferase Dam is fused to a DNA-binding protein resulting in adenine methylation within guanine-adenine-thymine-cytosine (GATC) motifs in the neighborhood of the protein–DNA binding sites[30]. Methylated GATCs (GmATC) are specifically digested by the DpnI restriction enzyme, allowing determination of DNA binding locations of the fusion protein after normalization for non-specific methylation by freely diffusing Dam. Methylation at distal chromosomal sites interacting with the viewpoint in three dimensions can also be observed[26,28,29] if interaction-specific methylation is considerably higher than non-specific methylation. However, previous studies detected methylated

---

[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Oncode Institute, Hubrecht Institute–KNAW, Utrecht, the Netherlands. [4]University Medical Center Utrecht, Utrecht, the Netherlands. [5]EMBL, Heidelberg, Germany. [6]Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands. [7]Università degli Studi di Milano and INFN, Milan, Italy. [8]These authors contributed equally: J. Redolfi, Y. Zhan, C. Valdes-Quezada. *e-mail: luca.giorgetti@fmi.ch

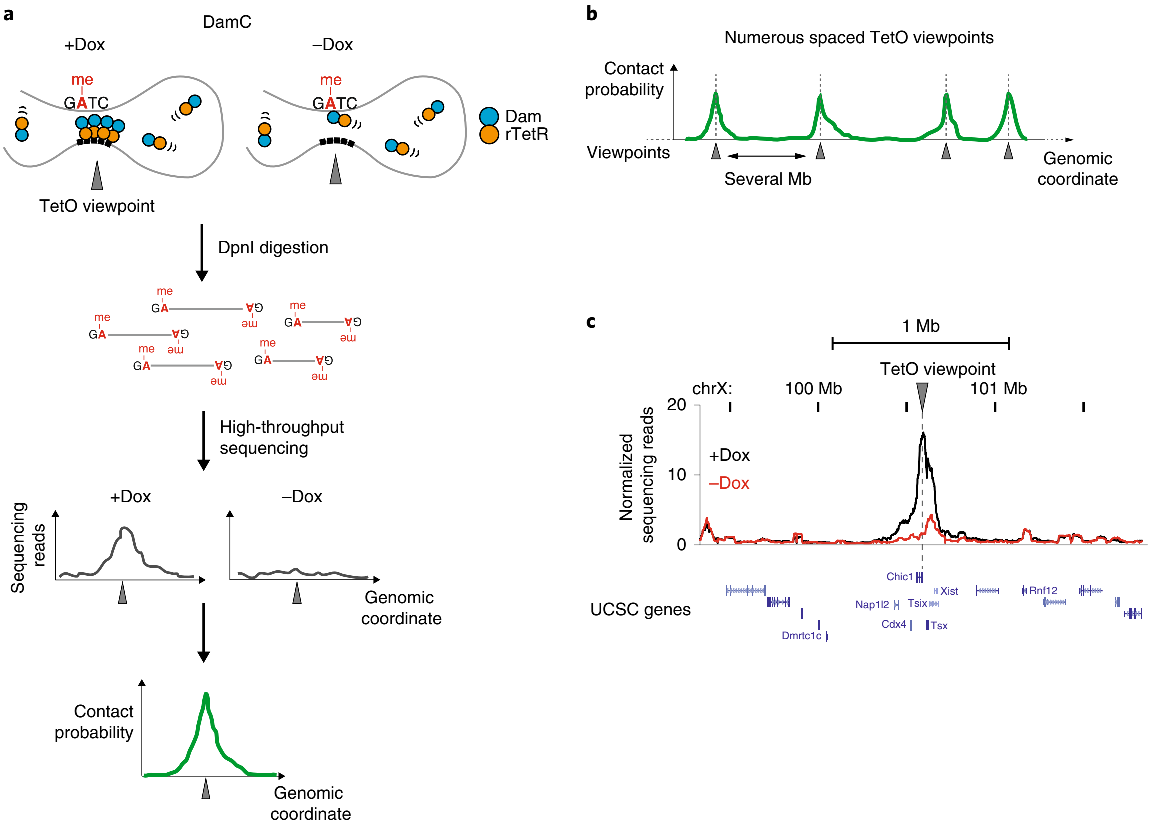


Fig. 1 | DamC: methylation-based measurement of chromosomal interactions. **a**, Scheme of DamC experiments. In the presence of doxycycline (+Dox), rTetR-Dam binds to a genomic viewpoint through a TetO array and methylates adenines in GATC sites that contact the viewpoint. In the absence of doxycycline (−Dox), only non-specific methylation (me) by freely diffusing rTetR-Dam occurs. Methylated GATCs can be detected by digestion of genomic DNA with DpnI and next-generation sequencing of the restriction sites. Correction for non-specific methylation allows extraction of contact probabilities with the TetO viewpoint. **b**, Insertion of multiple TetO arrays spaced by several megabases (Mb) allows detection of interaction from single viewpoints in parallel. **c**, Proof-of-principle experiment showing increased methylation in *cis* following the recruitment of rTetR-Dam to an array of 256 TetO sites in the 5′ UTR of the *Chic1* gene in the presence of Dox (black), compared to the −Dox control where rTetR-Dam is not recruited (red). UCSC, University of California, Santa Cruz. chr, chromosome.

DNA with semi-quantitative PCR readouts and analyzed interactions of one viewpoint with a limited number of restriction sites, similar to early 3C experiments[31]. This, and the lack of formal schemes available to convert methylation states into contact probabilities, have prevented these versions of DamID from reaching the resolution and throughput needed for the detection of TAD boundaries and CTCF loops. Thus to date no crosslinking- and ligation-free method is available to study chromosome interactions in the context of contacts made by all other surrounding genomic sequences. Remarkably, evidence for the existence of CTCF-associated loops is based exclusively on crosslinking methods.

Here we present DamC, a new modified version of DamID, coupled to physical modeling of DNA methylation kinetics. In DamC, Dam is recruited to ectopically inserted Tet operators (TetOs) through fusion to the reverse tetracycline receptor (rTetR). Methylated DNA is detected by high-throughput sequencing, allowing the identification of chromosomal contacts at high genomic resolution across hundreds of kilobases around viewpoints. Modeling of this process shows that experimental output in DamC is proportional to chromosomal contact probabilities, providing a theoretical framework for the interpretation of data.

Using DamC, we provide the first crosslinking- and ligation-free validation of structures identified by 3C methods. By comparing DamC with 4C sequencing (4C-seq) and Hi-C at hundreds of genomic locations in mouse embryonic stem cells (mESC), we confirm the existence of TADs and CTCF loops. We also show that the scaling of contact probabilities measured in DamC is the same as in 4C and Hi-C, providing evidence in favor of current interpretations of 3C-based data in terms of physical models of chromosome folding. We additionally demonstrate that ectopic insertion of CTCF sites can lead to the formation of new loops with endogenous CTCF-bound sequences and alter sub-TAD contacts. This shows that chromosome structure can be manipulated by the insertion of short ectopic sequences that rewire interactions within TADs.

Results

DamC: methylation-based detection of chromosomal contacts in vivo. Based on the results of previous studies[26,28,29], we reasoned that fusion of Dam to rTetR and insertion of an array of TetOs in the genome would ensure targeted, inducible recruitment of large numbers of Dam molecules to a specific genomic viewpoint in the presence of doxycycline (Dox) (Fig. 1a, left). In the absence of Dox, rTetR-Dam would not bind to the viewpoint, allowing accurate estimation of non-specific methylation (Fig. 1a, right) and precise background correction. Coupled to high-throughput sequencing, this strategy could provide 4C-like, 'one versus all' profiles[32] of contact probabilities from the TetO viewpoint (Fig. 1a) across large genomic distances and at high genomic resolution (one GATC every ~250 base pairs (bp) on average). Insertion of multiple TetO arrays separated by large genomic distances would allow investigation of chromosomal interactions in parallel from many viewpoints (Fig. 1b). We refer to this method as DamC.

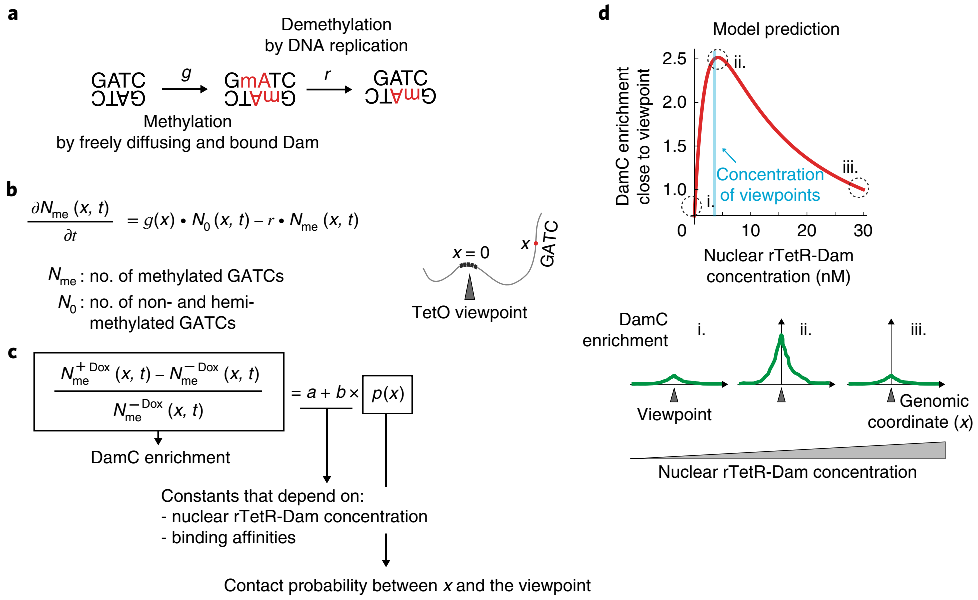

a
Demethylation by DNA replication
GATC GATC
g
GmATC GmATC
r
GATC GmATC
Methylation by freely diffusing and bound Dam
b
∂Nme (x, t) / ∂t = g(x) • N0 (x, t) – r • Nme (x, t)
Nme : no. of methylated GATCs
N0 : no. of non- and hemi-methylated GATCs
x = 0
x
GATC
TetO viewpoint
c
(Nme+Dox (x, t) – Nme−Dox (x, t)) / Nme−Dox (x, t) = a + b × p(x)
DamC enrichment
Constants that depend on:
- nuclear rTetR-Dam concentration
- binding affinities
Contact probability between x and the viewpoint
d
Model prediction
DamC enrichment close to viewpoint
2.5
2.0
1.5
1.0
0
10
20
30
i.
ii.
iii.
Concentration of viewpoints
Nuclear rTetR-Dam concentration (nM)
DamC enrichment
i.
ii.
iii.
Viewpoint
Genomic coordinate (x)
Nuclear rTetR-Dam concentration

Fig. 2 | Physical model of methylation dynamics. **a**, Unmethylated GATCs can be methylated by either freely diffusing or TetO-bound rTetR-Dam at rate *g*, and partially demethylated during DNA replication at rate *r*. Partially demethylated GATCs are inefficiently cut by DpnI and do not contribute to the DamC experiment. **b**, Model of methylation dynamics. The time (*t*) evolution of the number of methylated and non- and hemi-methylated GATCs located at genomic distance *x* from a TetO viewpoint is described in terms of ordinary differential equations governed by rates *g* and *r*. **c**, DamC enrichment at a generic location *x* is independent of time and is proportional to the contact probability between *x* and the viewpoint. Proportionality constants *a* and *b* are dependent on the nuclear rTetR-Dam concentration and the binding affinities to TetO and non-specific genomic sites. **d**, Model prediction using example parameter values (rTetR-TetO affinity, 5 nM; non-specific affinity, 80 nM, 600 TetO insertions corresponding to ~2 nM in a nuclear volume of ~490 fl, and contact probability of 0.5 corresponding to an interaction occurring in half of the cell population). The behavior of the curve is conserved across a wide range of physiologically relevant parameter values (see Supplementary Fig. 1b).

To test this approach, we employed female mESCs carrying an array of 256 TetOs at the 3’ end of the *Chic1* gene within the X inactivation center[33]. We transfected an X0 subclone of these cells with an rTetR-Dam expression plasmid and measured methylation after 24 h (ref. [34]). Quantification of methylated GmATCs by high-throughput sequencing revealed significantly higher methylation following Dox induction compared to the uninduced control over approximately 300 kb around the TetO viewpoint (Fig. 1c). Thus, targeted recruitment of Dam leads to increased methylation in *cis* over long genomic distances, consistent with previous observations using semi-quantitative methods for the detection of methylation[26,28,29]. Since methylation is determined by the interplay between methyltransferase activity and passive demethylation during DNA replication, we reasoned that it should be possible to model this process and derive chromosomal contact probabilities from sequencing readouts.

DamC enrichment is proportional to chromosomal contact probabilities. The methylation level of a single GmATC is determined by a dynamic interplay between methylation (by freely diffusing or TetO-bound Dam) and passive demethylation by DNA replication, when a fully methylated GmATC becomes two hemi-methylated sites that are essentially not detected in DamID[35] (Fig. 2a). To identify experimental quantities that are directly proportional to chromosomal contact probability, we generated a physical model describing the time evolution of methylation at an arbitrary genomic distance from the TetO viewpoint (Supplementary Note 1) through rate equations (Fig. 2b), which take into account the fact that methylation by TetO-bound Dam occurs only in the presence of Dox (Fig. 1a). Methylation rates are allowed to depend on local biases (for example, chromatin accessibility or mappability). Under the assumption that methylation is faster than demethylation[35] and bearing in mind the duration of an experiment (approx. 18 hours), we found that contact probabilities between the GATC site and the TetO viewpoint are directly proportional to a measurable quantity. This quantity, which we refer to as DamC enrichment, is simply the relative difference between methylation levels in the presence and absence of Dox (Fig. 2c). Thus, DamC can directly measure chromosomal contact probabilities.

For a given contact probability between the GATC site and the TetO viewpoint, the model predicts that DamC enrichment is dependent on: (1) the nuclear rTetR-Dam concentration, (2) the rTetR-Dam binding affinity for the TetO array and (3) the average non-specific binding affinity of rTetR-Dam for endogenous genomic sites (Fig. 2c). DamC enrichment is not dependent on local methylation biases and therefore should not be affected by differential accessibility or mappability, provided that interactions with the TetO viewpoint can increase methylation at the GATC site (that is, local methylation is not saturated in the absence of Dox). In real experiments, where binding affinities are fixed, the main determinant of DamC enrichment is the nuclear concentration of rTetR-Dam. In particular, DamC enrichment should be maximal when rTetR-Dam concentration is around the nuclear concentration of TetO viewpoints (Supplementary Fig. 1a), and negligible when the concentration is very high or very low (Fig. 2d). This is not dependent on the particular values of the affinity constants (Supplementary Fig. 1b), and implies that maximal DamC enrichment occurs at different Dam concentrations depending on the number of viewpoints. Thus, modeling predicts that accurate control of rTetR-Dam nuclear concentrations is needed to perform DamC with optimal signal-to-noise ratio.

DamC from hundreds of genomic viewpoints validates model predictions. To test model predictions and measure chromosomal interactions using DamC, we established mESCs allowing control of the rTetR-Dam nuclear concentration. We first created a stable cell

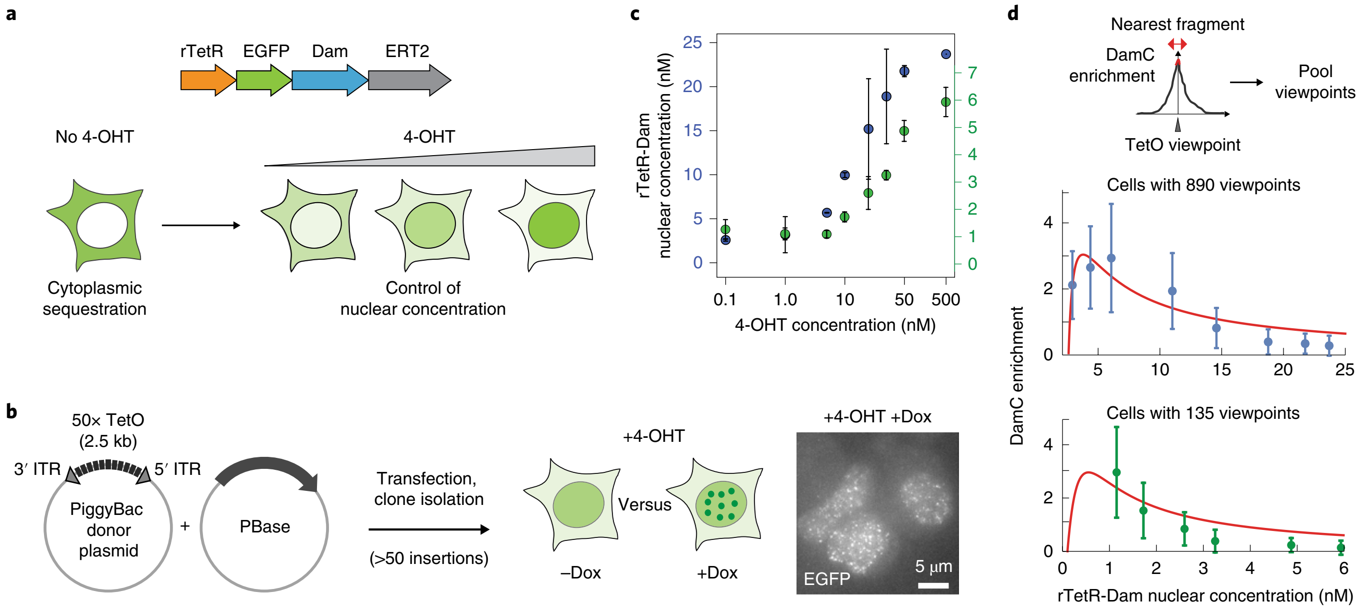


Fig. 3 | An inducible mESC line used to perform DamC and test the model predictions. **a**, mESCs expressing rTetR-Dam-EGFP-ERT2 allow control of the nuclear concentration of the fusion protein by changing the amount of 4-OHT in the culture medium. **b**, Nuclear concentration of the rTetR-Dam fusion protein as a function of 4-OHT concentration in the polyclonal population with 890 insertions (blue), and in the subclone with 135 insertions (green). Numbers of protein copies per nucleus were determined using mass spectrometry on nuclear extracts and divided by the average nuclear volume (~490 fl) as determined using DAPI staining (see Supplementary Fig. 2). Error bars are the s.d. of two biological replicates (independent cell cultures). **c**, Random integration of large numbers of 50× TetO platforms using the piggyBac transposon. Accumulation of EGFP signal to nuclear foci in the presence of Dox (right: max. intensity projection over ten *Z*-planes) indicates binding of rTetR-Dam to the arrays and allows selection of clones with large numbers of insertions. PBase: piggyBac transposase. **d**, Quantification of DamC experiments as a function of rTetR-Dam concentration in cells with 890 (upper panel, blue) and 135 (lower panel, green) TetO viewpoints. Blue data points, mean and s.d. from the 100 TetO viewpoints with highest enrichment. Green data points, mean and s.d. from 130 TetO viewpoints (five viewpoints were excluded due to absence of DamC signal). Red line, model fit to the experimental data.

line expressing rTetR fused with enhanced green fluorescent protein (EGFP), Dam and the mutant estrogen ligand-binding domain ERT2. ERT2 ensures cytoplasmic localization of the fusion protein in the absence of 4-hydroxy-tamoxifen (4-OHT), preventing constitutive GATC methylation. It also enables control of its nuclear level by changing 4-OHT concentrations in the culture medium (Fig. 3a), as confirmed by increasingly nuclear accumulation of EGFP after increasing 4-OHT doses (Supplementary Fig. 2a).

To measure chromosomal interactions in a wide variety of randomly selected genomic contexts in parallel, we further inserted arrays of 50 TetOs (each spanning approximately 2.7 kb) using the piggyBac transposon system[36] (Fig. 3b). This resulted in clonal mESC lines carrying at least 50 TetO arrays, judging from EGFP accumulation in nuclear foci in the presence of 4-OHT and Dox (Fig. 3b). We further selected one polyclonal population carrying a total of 890 TetO array insertions and one clonal line with 135 insertions (Supplementary Tables 1 and 2), as determined by mapping piggyBac insertion locations (see Methods). To quantitatively measure rTetR-Dam nuclear concentrations as a function of 4-OHT concentration, we analyzed nuclear protein extracts using mass spectrometry. Combining the proteomic ruler strategy with parallel reaction monitoring (PRM) (Methods, Supplementary Fig. 2b,c and Supplementary Table 3), we estimated that nuclear rTetR-Dam concentrations vary gradually between approximately 3 and 25 nM and between 1 and 6 nM, in the polyclonal and pure clonal lines, respectively, when increasing the 4-OHT concentration from 0.1 to 500 nM (Fig. 3c).

We performed DamC after treating cells overnight with different doses of 4-OHT in the presence or absence of Dox. Experiments were performed using a custom next-generation sequencing library preparation protocol that includes unique molecular identifiers (UMI) and increases the coverage of methylated GATC sites genome wide, thus maximizing proportionality between methylation levels and sequencing readout (Supplementary Fig. 2d,e and Methods). We quantified DamC enrichment in the immediate vicinity of TetO viewpoints, and plotted this as a function of rTetR-Dam concentration as quantified by mass spectrometry (Fig. 3d). For the polyclonal line, we considered the 100 insertions with highest signal-to-noise ratios corresponding to the most abundant insertions. In the pure subclone, all insertions except five showed similar enrichment levels, possibly as a consequence of recombination of the TetO array or high levels of transcription at the insertion point preventing TetO binding. These insertions were discarded from analysis and their coordinates are provided in Methods.

Consistent with model predictions, in the polyclonal mESC line maximum enrichment occurs at ~3 nM corresponding to ~860 viewpoints (Fig. 3d, upper panel). Model fitting returned an estimate of 0.4 nM for the specific rTetR-TetO binding constant, in the range of in vitro measurements[37], and of 17 nM for the average non-specific binding constant. Again, in line with the model, enrichment in the clonal line carrying ~sevenfold fewer viewpoints (130) was compatible with maximal enrichment occurring at ~sevenfold lower rTetR-Dam nuclear concentration (0.4 nM). These results provide a validation of the DamC model and support the interpretation of DamC enrichment in terms of contact probabilities. They additionally highlight that, in our experimental system, maximal DamC enrichment in cells with ~100 insertions is observed in a range of rTetR-Dam nuclear concentrations corresponding to 0.1–1 nM 4-OHT (Supplementary Fig. 2f). In the following analysis, reads from these two conditions were pooled to maximize read coverage.

DamC reveals the existence of TADs and loops in vivo. Under optimal 4-OHT concentrations (0.1–1 nM pooled), zooming into

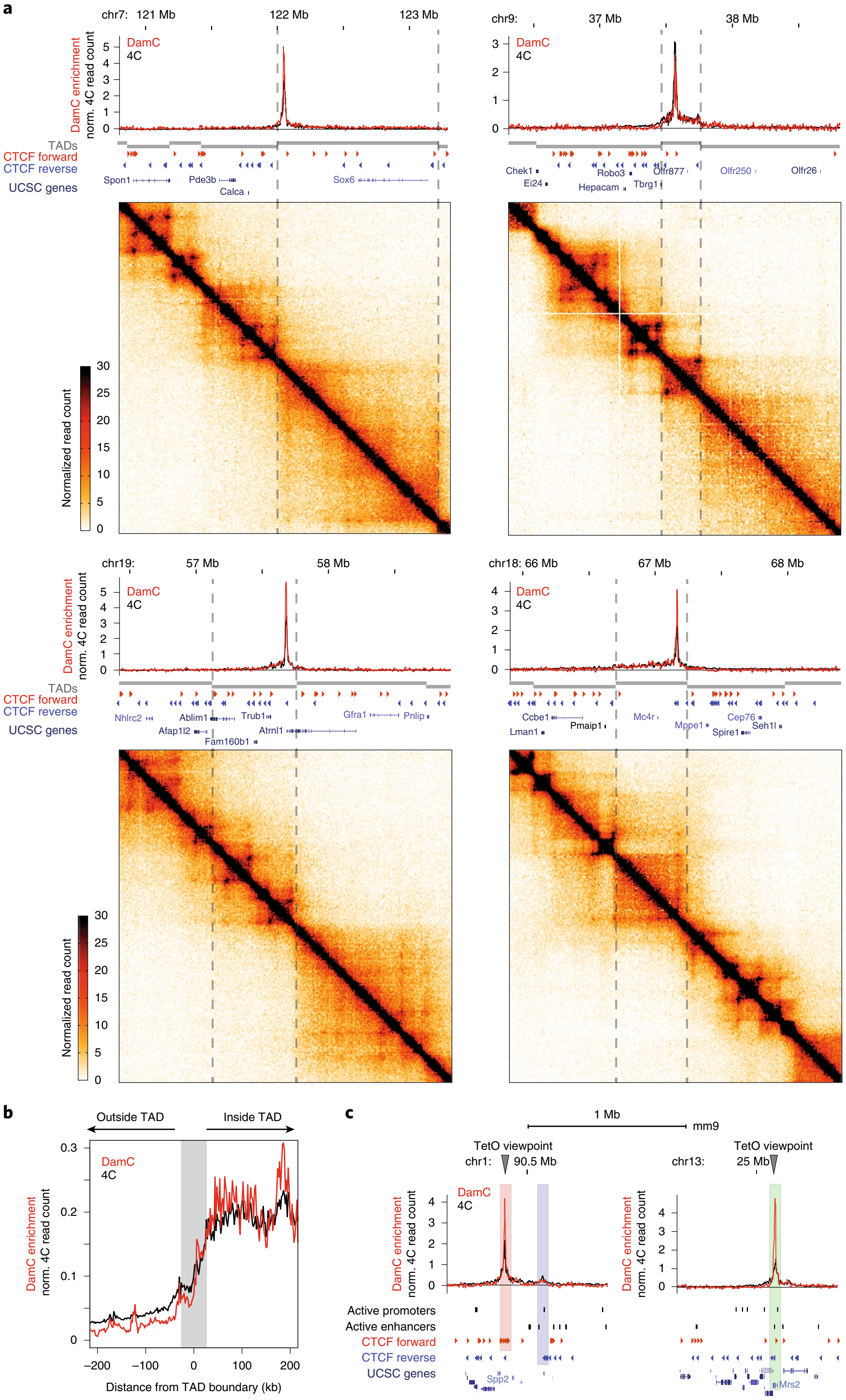


Fig. 4 | DamC confirms the existence of TAD boundaries and quantitatively correlates with 4C and Hi-C. a, Four representative DamC and 4C interaction profiles from the same piggyBac-TetO viewpoints, aligned with Hi-C experiments performed in the same cell line. Dashed lines mark TAD boundaries in mESC, detected using Caller of Topological Chromosomal Hierarchies (CaTCH)[9]. Hi-C data were binned at 10 kb resolution. DamC was performed using 0.1 and 1 nM 4-OHT (pooled). Data from two biological replicates were pooled for DamC, 4C and Hi-C. **b**, Aggregated plot over 130 TetO viewpoints showing DamC and 4C data aligned to TAD boundaries identified using CaTCH[9]. Gray shading: ±40 kb uncertainty on boundary definition[9]. **c**, Interaction profiles from viewpoints located <1 kb from a CTCF site (left) belonging to a cluster of forward sites (red shading) interacting with reverse CTCF sites (blue shading), and <1 kb from the active promoter of the *Mrs2* gene (right), highlighted in the green shaded area.

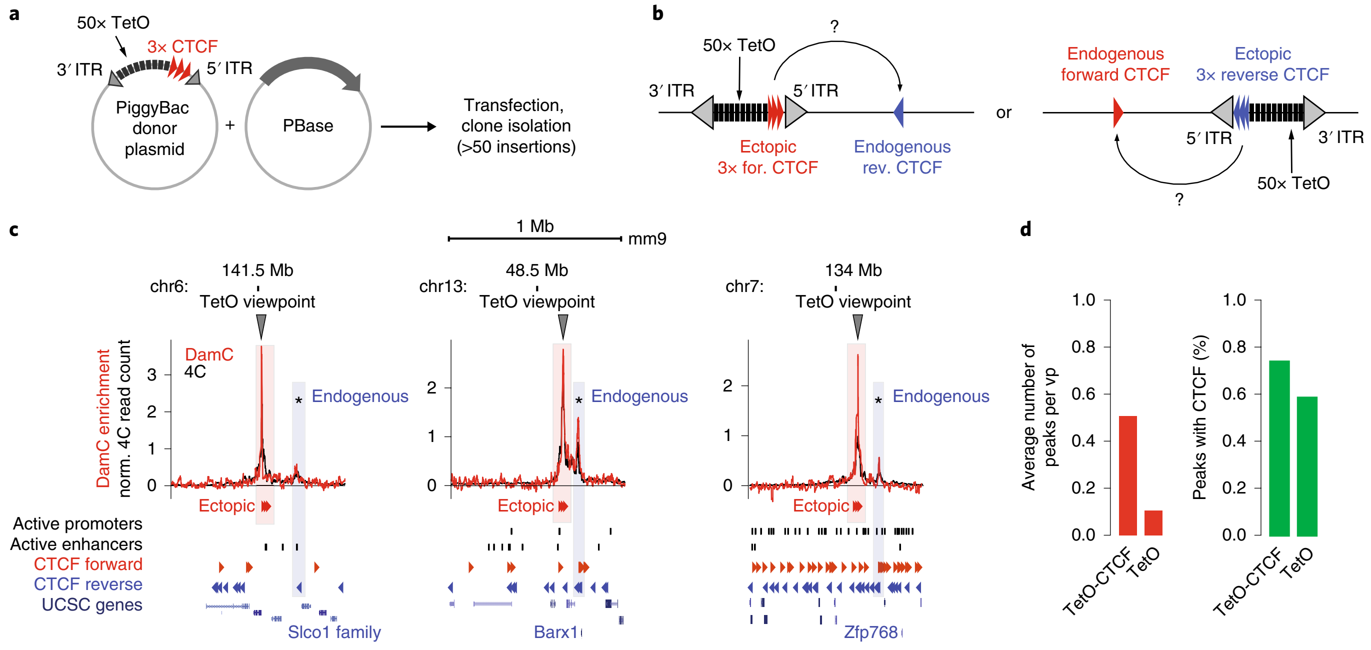


Fig. 5 | DamC-based detection of CTCF loops. a, Modified piggyBac strategy used to insert TetO viewpoints flanked by three CTCF sites oriented outwards. **b**, The TetO-CTCF cassette can insert in the genome in either direction and lead to the formation of interactions with either forward (for.) or reverse (rev.) endogenous CTCF sites. **c**, Three representative interaction profiles obtained using DamC and 4C from TetO-CTCF viewpoints. Asterisks indicate interactions identified by the PeakC algorithm that overlap with CTCF sites. Shaded boxes indicate overlap with the genomic positions of endogenous and ectopic CTCF locations. **d**. Left, average number of peaks per viewpoint (vp) detected by PeakC at least 20 kb away from the viewpoint in cells with TetO-CTCF or TetO-only insertions. Viewpoints landing within 1 kb from an endogenous CTCF site were excluded. Right, percentage of peaks containing a CTCF motif that is bound based on ChIP-seq data[11].

individual TetO viewpoints in the clonal line with 130 insertions revealed significant DamC enrichment over hundreds of kilobases around each viewpoint (Fig. 4a). Since biological replicates were highly correlated (Supplementary Fig. 3a), we analyzed merged data. DamC enrichment profiles showed remarkable agreement with 4C performed using the same TetO arrays as viewpoints and DpnII as the primary restriction enzyme (Fig. 4a and Methods). DamC enrichment was systematically concentrated within TAD boundaries detected in Hi-C (Fig. 4a) and steeply decayed across TAD boundaries by a factor of approximately 2, in excellent agreement with 4C (Fig. 4b). Only a minor fraction of TetO insertions occurred in close proximity (<1 kb) to either an active regulatory element or a CTCF site (Supplementary Fig. 3b). Also in these cases, DamC enrichment profiles highly overlapped with 4C (Fig. 4c) and revealed looping interactions between endogenous convergent CTCF sites (Fig. 4c, left), which was confirmed using the partner CTCF sites as reciprocal viewpoint in 4C (Supplementary Fig. 3c). The targeted TetO insertion at the 3' end of *Chic1* (ref. [33]) (Fig. 1c) allowed measurement of chromosomal interactions within the well-characterized *Tsix* TAD[6,38]. In accordance with 4C, DamC recapitulated the previously observed CTCF-mediated interactions between *Chic1*, *Linx* and *Xite/Tsix*[6], as well as the boundaries of the *Tsix* TAD (Supplementary Fig. 3d). Additional DamC and 4C profiles are plotted in Supplementary Fig. 4, and bedGraph tracks are available online (see Data availability).

We next investigated whether, despite evident global similarities, DamC and 4C showed local differences. We defined a deviation score measuring differences between DamC and 4C interaction profiles within windows of 20 DpnI/II restriction fragments (5 kb on average) (Supplementary Fig. 3e). Most dissimilar windows were enriched in active chromatin marks (Supplementary Fig. 3f), although local differences between DamC and 4C within these regions were relatively mild (Supplementary Fig. 3e). We reasoned that local discrepancies might be due to the fact that the methylation signal correlates highly with chromatin accessibility as measured by DNase I sensitivity (Supplementary Fig. 3g). Correction by non-specific methylation generally normalizes for chromatin accessibility in DamC enrichment unless GATC sites are highly methylated in the absence of Dox, thus preventing further increases when interacting with the TetO viewpoint in +Dox conditions. However, only 0.05% of GATC sites within DNase I hypersensitive regions were saturated (Methods), and masking of DNase I hypersensitive sites did not increase the overall similarity between DamC and 4C profiles (Supplementary Fig. 3h). Thus local differences between the two techniques are not due to saturated methylation levels, but may be due to experimental factors not described by the DamC model and thus not accounted for in the calculation of DamC enrichment.

In summary, crosslinking- and ligation-free measurements of contact probabilities using DamC quantitatively agree with 4C, confirm the existence of TAD boundaries and show that crosslinking and ligation do not greatly distort the detection of chromosomal interactions.

piggyBac-TetO insertions do not perturb chromosome structure. Next, we aimed to determine whether insertion of TetO/piggyBac cassettes per se could perturb local chromosome structure. We compared TetO insertion sites with the corresponding wild-type (WT) loci in Hi-C experiments (Supplementary Fig. 5a) using a modified version of the deviation score defined in Supplementary Fig. 3d that describes differences in virtual 4C profiles extracted from Hi-C data (Supplementary Fig. 5b). Deviation scores between WT cells and those carrying TetO arrays were similar to deviation scores between Hi-C replicates at random genomic locations, and significantly smaller than deviation scores between different WT loci (Supplementary Fig. 5b). Finally, 4C profiles obtained with and without TetO viewpoints were indistinguishable (Supplementary

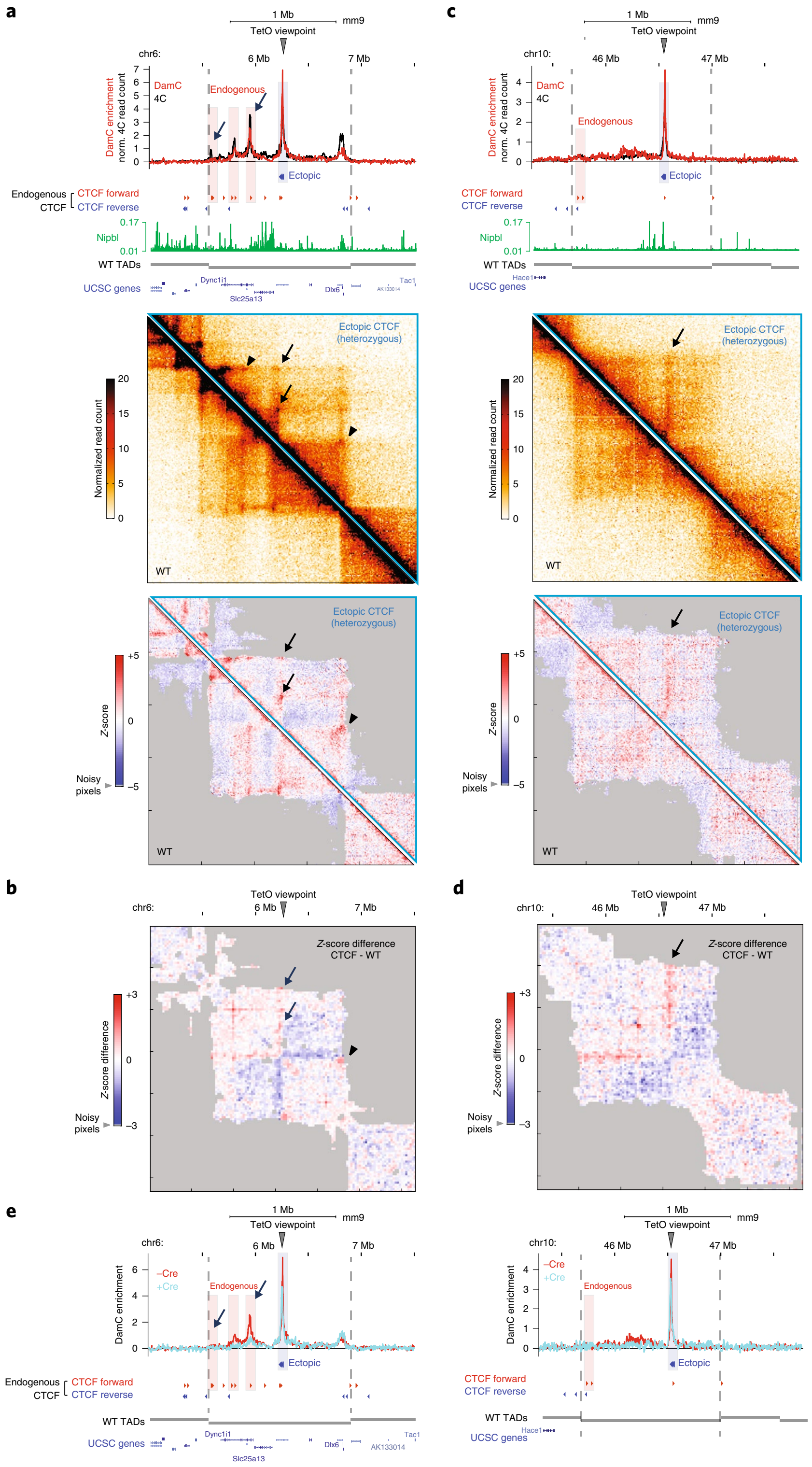


Fig. 6 | Ectopic CTCF insertion leads to the formation of new loops and stripes. a, TetO-CTCF insertion site giving rise to ectopic loops with convergently oriented endogenous CTCF sites. Top, interaction profiles measured with DamC and 4C are overlaid with the position of CTCF ChIP-seq sites from ref.[11] and Nipbl ChIP-seq data from ref.[42]. Middle, Hi-C data from ESC lines carrying either a heterozygous TetO-CTCF insertion or two WT alleles. Bottom, distance-normalized *Z*-scores, highlighting interactions that are either enriched (red) or depleted (blue) compared to the expected interaction frequency. Arrows, interactions between convergent CTCF sites established following CTCF insertion. Arrowheads, pre-existing interactions strengthened after CTCF insertion. Hi-C data are binned at 10 kb resolution. Data from two biological replicates (independent cell cultures) were pooled for DamC, 4C and Hi-C. **b**, *Z*-score difference between heterozygous CTCF and WT cells showing increased partitioning of interactions inside the TAD. Hi-C data were binned at 20 kb. Shaded areas correspond to 'noisy' interactions that did not satisfy a quality control filter based on their correlations with immediate nearest neighbors (see Methods). **c**, Same as **a** for an insertion on chromosome 10, occurring in proximity to an isolated cluster of Nipbl binding and giving rise to a stripe-like interaction pattern. **d**, *Z*-score differences for the locus shown in **c**. **e**, DamC interaction profiles from the same viewpoints as in **a** and **c**, before and after Cre-mediated excision of ectopically inserted CTCF sites (but not of the piggyBac cassette).

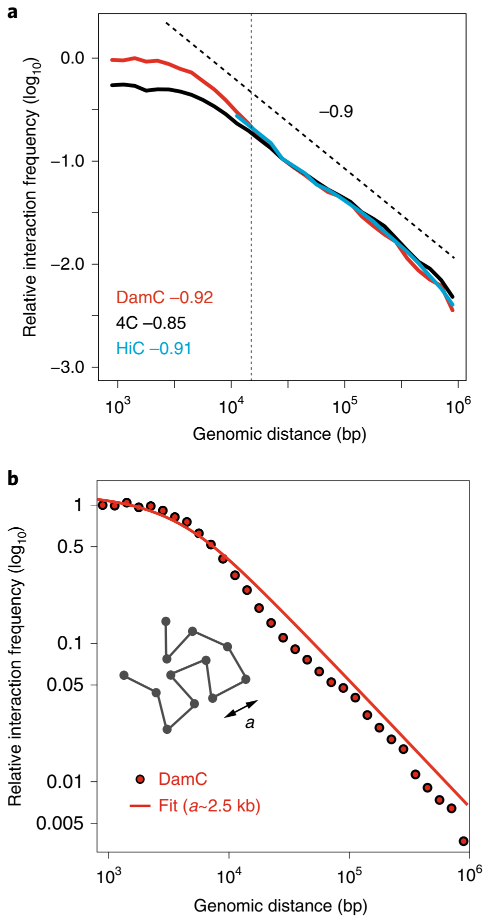


Fig. 7 | Scaling analysis of contact probabilities in vivo. a, Scaling of contact probabilities measured in DamC, 4C and Hi-C from all 130 TetO and 91 TetO-CTCF viewpoints. Power-law fitting was performed between 15 kb (dotted vertical line) and 1 Mb. **b**, Best fit of scaling behavior measured by DamC with a polymer model of persistence length *a* (see Supplementary Note 1). The best value of *a* extracted from the fit is ~2.5 kb.

Fig. 5c). Thus, piggyBac-mediated insertion of TetO arrays does not lead to measurable perturbations of chromosome structure.

In vivo detection and manipulation of CTCF-mediated interactions. Loops between convergent CTCF sites are a defining feature of chromosome architecture. However, it is unclear whether new loops can be established between endogenous and ectopically inserted CTCF sites. Early 3C observations suggested that ectopic sequences containing CTCF sites can change the surrounding chromosomal interactions[39,40]; however, experimental resolution in 3C did not allow us to resolve single CTCF loops, and inserted sequences contained additional regulatory regions. Since piggyBac-TetO constructs per se do not perturb chromosome structure, we further engineered them to insert ectopic CTCF sites in the genome and to detect the resulting structural modifications without confounding effects.

Starting from the founder rTetR-GFP-Dam-ERT2 mESC line described in Fig. 3a, we randomly introduced modified piggyBac cassettes where the TetO array is flanked by three CTCF sites oriented outwards (Fig. 5a). To test whether ectopically inserted CTCF sites could establish loops with endogenous CTCF sites (Fig. 5b), we selected one clone carrying 91 insertions for which we could map insertion positions and genomic orientations (Supplementary Table 4), and performed 4C and DamC with 0.1–1 nM 4-OHT.

Interaction profiles from TetO-CTCF viewpoints displayed prominent distal peaks (Fig. 5c) detected by both DamC and 4C. We used the PeakC algorithm, developed to analyze 4C profiles[41], to identify distal preferential interactions. Using stringent thresholds (Methods) and excluding viewpoints within 1 kb from an endogenous CTCF site (Supplementary Figs. 3b and 6a), we detected 38 specific interactions separated by at least 20 kb from single TetO-CTCF viewpoints (~0.5 distal peaks per insertion site on average, Supplementary Fig. 6b). Of those, 74% contained one or more bound CTCF sites based on chromatin immunoprecipitation sequencing (ChIP-seq) datasets in mESC[11], predominantly (79%) convergent with the ectopic CTCF insertion (Fig. 5d). As a comparison, in the cell line harboring TetO viewpoints without CTCF we detected only 0.1 peaks per insertion site (Supplementary Fig. 6b), of which 58% contained one or more bound CTCF sites (Fig. 5d). These correspond to endogenous CTCF loops since, in virtually all these cases, the TetO was located between 1 and 20 kb away from an endogenous CTCF. Thus, peaks in the TetO-CTCF line are likely to coincide with new loops established by ectopic CTCF sites. Insertions without distal peaks predominantly correspond to TetO-CTCF cassettes integrated either in CTCF 'deserts' or, conversely, close (<30 kb) to the nearest endogenous convergent CTCF site and in regions with many endogenous CTCF sites (Supplementary Fig. 6c), resulting in short-distance loops that are difficult to distinguish in 4C and DamC profiles. Additional TetO-CTCF DamC and 4C profiles are plotted in Supplementary Fig. 7, and bedGraph tracks are available online (see Data availability).

We then performed Hi-C in the TetO-CTCF line and compared it to the data obtained from TetO-only mESCs (see Fig. 4a), where insertion locations are different. Since TetO-CTCF insertions are heterozygous, the Hi-C readout is confounded by the presence of a WT allele. Nevertheless, in a fraction of insertions showing prominent distal CTCF peaks in 4C and DamC, we could detect the formation of new structures in Hi-C and, notably, new loops (Fig. 6a, arrows and Supplementary Fig. 6d), leading to increased partitioning of interactions within TADs and the appearance of sub-TAD boundaries (Fig. 6b). Ectopic CTCF insertion also reinforced pre-existing interactions between convergently oriented sites (Fig. 6a, arrowheads), possibly by bringing them closer due to the effect of the new loops. Even insertions without prominent distal CTCF peaks could be associated with new structures (Fig. 6c), reminiscent of stripes predicted by the loop-extrusion model[15] and recently observed in Hi-C data at endogenous locations[42]. Consistent with the loop-extrusion model interpretation, the stripe shown in Fig. 6c occurred at a location where the three ectopic CTCF sites landed close to a cluster of Nipbl sites, and far from the nearest convergent CTCF sites (~800 kb). Formation of an ectopic CTCF-associated stripe also resulted in modifications of intra-TAD chromosomal interactions (Fig. 6d).

Finally, to formally prove that new structures are induced by ectopic CTCF binding sites (rather than the piggyBac-TetO cassette), we removed the three CTCF sites by Cre-assisted recombination using two flanking LoxP sites (Supplementary Fig. 6e). DamC performed in one mESC clone, where CTCF sites had been excised at both loci shown in Fig. 6a,c (Supplementary Fig. 6f), revealed that removal of these sites led to loss of distal interactions (Fig. 6e).

In summary, DamC identifies chromatin loops formed through specific long-range chromatin interactions. Additionally, our data show that ectopically inserted CTCF sites can establish new loops with endogenous CTCF sites and stripes, leading to modified partitioning of interactions within TADs.

Quantitative properties of chromosome folding in vivo. Given the high similarity between DamC and 4C both with and without CTCF sites at the viewpoint, we next asked whether DamC and 3C-based techniques measured the same scaling of interaction probabilities. We pooled all viewpoints from TetO-only and TetO-CTCF lines and plotted the data as a function of genomic distance from the viewpoints. For distances between 15 kb and 1 Mb, fitting both DamC and 4C with a power law resulted in decay exponents around 0.9, in excellent agreement with Hi-C from the same cells and viewpoints (Fig. 7a), and in accordance with previous measurements in similar genomic ranges[14,43].

Below ~10 kb, both DamC and 4C showed a gentler decay as recently observed in Hi-C experiments on yeast chromosomes[44]. In ref. [44] this was attributed to crosslinking artifacts but DamC, showing the same behavior, argues against this explanation. The leveling-off of contact probabilities at short genomic distances can be explained in terms of a simple, coarse-grained polymer model with a persistence length of ~2.5 kb (Fig. 7b and Supplementary Note 1). We cannot formally rule out alternative explanations, such as experimental factors not accounted for by the DamC model and thus not normalized in the calculation of enrichment. One such scenario could be that the spacing between GATC sites imposes an effective capture range of a few kilobases, consistent with micrococcal nuclease-based Hi-C (Micro-C) experiments showing that yeast chromatin is flexible at lower scales[45]. However, in the absence of Micro-C measurements on mammalian chromatin, we can safely assume that DamC provides an upper limit to the persistence length of chromosomes in vivo of approximately 2.5 kb.

Discussion

In this work we provide in vivo, high-resolution, systematic measurements of chromatin contacts that require neither crosslinking nor ligation, using DamC. An essential feature of this method is that its experimental output is directly proportional to contact probabilities. This is supported by rigorous modeling of methylation kinetics (Fig. 2), providing a rational basis to quantitatively interpret sequencing results. Importantly, DamC confirms that contact frequencies fall across TAD boundaries by a factor of approximately 2, in accordance with 4C (Fig. 4b) and previous estimations based on Hi-C[15,46]. Such a modest decrease raises the question of how TAD boundaries can functionally insulate enhancers and promoters from a biophysical point of view, although they might represent an optimal compromise between enrichment and depletion interactions between regulatory sequences within and across boundaries, respectively[9].

DamC detects chromosomal contacts at short spatial distances, since GATC motifs can be methylated only if Dam directly binds DNA. We estimate a detection range of <10 nm, given that the expected physical size of the rTetR-EGFP-Dam-ERT2 fusion protein does not exceed 3 nm (ref. [47]). Decreases in interaction frequencies at TAD boundaries, as well as increases due to CTCF loops, therefore closely match what a promoter would 'experience' through its bound protein complexes. Interestingly DamC also picks up 'non-specific' interactions due to random collisions within the chromatin fiber to the same extent as 4C and Hi-C (Fig. 4a and Supplementary Fig. 4). Thus, random collisions do occur in vivo, despite not being detected in crosslinking-free analysis of chromosome folding using native 3C[26].

Scaling of crosslinking probabilities measured in Hi-C data is at the core of physical models developed to explain chromosome folding and infer its mechanistic determinants[15,38,48–50], including the highly influential loop-extrusion model. Importantly, DamC confirms scaling exponents measured in 4C and Hi-C (Fig. 7). Since DamC enrichment is proportional to actual short-range contact probabilities, our measurements provide strong evidence in favor of chromosome-folding models based on Hi-C. Scaling analysis at short genomic distances additionally suggests that mouse chromosomes may have a persistence length of approximately 2.5 kb, corresponding to ~40 nm assuming a linear density of ~60 bp nm^{-1} (ref. [38]).

The finding that loops can be established de novo following insertion of CTCF binding sites and can be detected in vivo (Figs. 5 and 6) confirms earlier reports[39,40] and argues that chromosome structure at the TAD level can be manipulated in a 'gain-of-function' manner by the addition of new structures. New structures formed following ectopic insertion of three CTCF sites can greatly modify intra-TAD interactions and may result in the formation of new boundaries within pre-existing TADs (Fig. 6b,d). Remarkably, we could detect only newly formed interactions within pre-existing TAD boundaries, possibly due to the fact that these boundaries are particularly enriched in clusters of CTCF sites[7,9] providing efficient barriers to loop extrusion.

One limitation of DamC is that it requires genetic manipulation for the insertion of genomic viewpoints and for stable expression of rTetR-Dam-ERT2, allowing accurate control of nuclear Dam concentration. This prevents consideration of DamC in its current form as an alternative to 3C-based methods in routine experimentation. However, DamC can be performed by transiently nucleofecting cells with a Dam-TetR expression plasmid, which ensures low expression levels (Fig. 1c). Future implementations based on TAL effector proteins (similar to TALE-ID[26]) or catalytically inactivated Cas9 could overcome the need for targeted insertion of TetO arrays.

The current TetR-based implementation of DamC may nevertheless be beneficial in situations where 4C cannot be used, notably to detect chromosomal interactions in a tissue-specific context by expressing the rTetR-Dam fusion under a tissue-specific promoter[51] and starting from small numbers of cells[52]. Contrary to 3C methods, where one ligation event per allele can be retrieved at most, in the course of a DamC experiment (~18 h) several GATCs might be contacted by a TetO viewpoint depending on the temporal dynamics of chromosome structure. Based on our previous measurements of the dynamics of the TetO array at the *Chic1* locus[53], as well as recent data from other chromosomal locations[54,55], several contacts may be created and disassembled in 18 h. If *n* GATC sites are methylated in this time window, DamC would in principle require *n* times fewer cells than 4C to build similar contact profiles. In this manuscript we analyzed ~10,000 cell equivalents per 4C and DamC experiment, but scaling down of cell numbers in DamC will be an interesting future development.

In summary, by coupling a methylation-based readout with physical modeling, DamC enables systematic and quantitative crosslinking- and ligation-free measurements of chromatin interaction frequencies. Our experiments provide an orthogonal validation of 3C-based findings, including TADs and both endogenous and ectopically induced CTCF loops, and demonstrate that 3C methods do not substantially distort the detection of chromosomal interactions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41594-019-0231-0.

Received: 14 November 2018; Accepted: 18 April 2019;
Published online: 27 May 2019

References

1. Denker, A. & de Laat, Wde The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–1382 (2016).
2. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. 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).
4. Norton, H. K. et al. Detecting hierarchical genome folding with network modularity. *Nat. Methods* **15**, 119–122 (2018).
5. Fraser, J. et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852–852 (2015).
6. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
7. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

8. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
9. Zhan, Y. et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* **27**, 479–490 (2017).
10. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl Acad. Sci. USA* **111**, 996–1001 (2014).
11. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e22 (2017).
12. de Wit, E. et al. CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**, 676–684 (2015).
13. Guo, Y. et al. CRISPR Inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
14. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
15. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
16. Gavrilov, A., Razin, S. V. & Cavalli, G. In vivo formaldehyde cross-linking: it is time for black box analysis. *Brief. Funct. Genom.* **14**, 163–165 (2015).
17. Gavrilov, A. A. et al. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res.* **41**, 3563–3575 (2013).
18. Williamson, I. et al. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* **28**, 2778–2791 (2014).
19. Belmont, A. S. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.* **26**, 69–78 (2014).
20. Fudenberg, G. & Mirny, L. A. Higher-order chromatin structure: bridging physics and biology. *Curr. Opin. Genet. Dev.* **22**, 115–124 (2012).
21. Tiana, G. & Giorgetti, L. Integrating experiment, theory and simulation to determine the structure and dynamics of mammalian chromosomes. *Curr. Opin. Struct. Biol.* **49**, 11–17 (2018).
22. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).
23. Nichols, M. H. & Corces, V. G. A CTCF code for 3D genome architecture. *Cell* **162**, 703–705 (2015).
24. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).
25. Beagrie, R. A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
26. Brant, L. et al. Exploiting native forces to capture chromosome conformation in mammalian cell nuclei. *Mol. Syst. Biol.* **12**, 891 (2016).
27. Quinodoz, S. A. et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744–757.e24 (2018).
28. Lebrun, E., Fourel, G., Defossez, P.-A. & Gilson, E. A methyltransferase targeting assay reveals silencer-telomere interactions in budding yeast. *Mol. Cell. Biol.* **23**, 1498–1508 (2003).
29. Cléard, F., Moshkin, Y., Karch, F. & Maeda, R. K. Probing long-distance regulatory interactions in the *Drosophila melanogaster* bithorax complex using Dam identification. *Nat. Genet.* **38**, 931–935 (2006).
30. Steensel, Bvan & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
31. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
32. van de Werken, H. J. G. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9**, 969–972 (2012).
33. Masui, O. et al. Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation. *Cell* **145**, 447–458 (2011).
34. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
35. Kind, J. et al. Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).
36. Cadiñanos, J. & Bradley, A. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res.* **35**, e87 (2007).
37. Kamionka, A., Bogdanska-Urbaniak, J., Scholz, O. & Hillen, W. Two mutations in the tetracycline repressor change the inducer anhydrotetracycline to a corepressor. *Nucleic Acids Res.* **32**, 842–847 (2004).
38. Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).
39. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl Acad. Sci. USA* **105**, 20398–20403 (2008).
40. Rawat, P., Jalan, M., Sadhu, A., Kanaujia, A. & Srivastava, M. Chromatin domain organization of the TCRb Locus and its perturbation by ectopic CTCF binding. *Mol. Cell. Biol.* **37**, e00557–16 (2017).
41. Geeven, G., Teunissen, H., de Laat, W. & de Wit, E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. *Nucleic Acids Res.* **46**, e91 (2018).
42. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178.e20 (2018).
43. Bonev, B. et al. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557–572.e24 (2017).
44. Scolari, V. F., Mercy, G., Koszul, R., Lesne, A. & Mozziconacci, J. Kinetic signature of cooperativity in the irreversible collapse of a polymer. *Phys. Rev. Lett.* **121**, 057801 (2018).
45. Hsieh, T.-H. S. et al. Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**, 108–119 (2015).
46. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).
47. Erickson, H. P. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol. Proced. Online* **11**, 32 (2009).
48. Brackley, C. A. et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.* **17**, 59 (2016).
49. 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 (2012).
50. Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. *PLOS Comput. Biol.* **4**, e1000153 (2008).
51. La Fortezza, M. et al. DamID profiling of dynamic Polycomb-binding sites in *Drosophila* imaginal disc development and tumorigenesis. *Epigenetics Chromatin* **11**, 27 (2018).
52. Tosti, L. et al. Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo. *Genome Res.* **28**, 592–605 (2018).
53. Tiana, G. et al. Structural fluctuations of the chromatin fiber within topologically associating domains. *Biophys. J.* **110**, 1234–1245 (2016).
54. Gu, B. et al. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050–1055 (2018).
55. Germier, T. et al. Real-time imaging of a single gene reveals transcription-initiated local confinement. *Biophys. J.* **113**, 1383–1394 (2017).

Acknowledgements

This work is dedicated to the memory of M. Dahan. Research in the Giorgetti laboratory is funded by the Novartis Foundation and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation (grant agreement no. 759366, 'BioMeTre'). The Kind laboratory was funded by the ERC (grant agreement no. 678423, 'EpiID') and EMBO (no. LTF 1214-2016 to I.G.). R.S.G. acknowledges support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 705354 and an EMBO Long-Term fellowship (no. ALTF 1086-2015). We would like to thank P. Cron for cloning TetO-piggyBac plasmids; S. Aluri and S. Thiry for assistance with high-throughput sequencing; M. Stadler for help with bioinformatics analysis; S. Grzybek and H.-R. Hotz for server supports; and E. Heard and R. Galupa (Institut Curie, PSL Research University) for kindly providing PGK cells. We are grateful to D. Schuebeler and R. Galupa for critically reading the manuscript, and to G. Fudenberg for useful comments on scaling behavior. We acknowledge The ENCODE Project Consortium and, in particular, the Ren and Hardison laboratories for ChIP-Seq datasets in ESC.

Author contributions

J.R. generated cell lines and performed DamC experiments. Y.Z. wrote the model with assistance from G.T. and analyzed the data. C.V.-Q. performed 4C in W.dL.'s laboratory. M.K. assisted with cell culture and DamC library preparation and performed Hi-C experiments. I.G. and J.K. helped with experimental design and data analysis. V.I. performed mass spectrometry experiments and analysis. T.P. provided constructs for initial experiments and discussed the data. R.S.G. provided CTCF site sequences and tested CTCF binding in preliminary experiments. E.M. contributed to design of the initial experiments. S.A.S. developed the DamC library preparation protocol and performed piggyBac insertion mapping experiments. L.G. designed the study and wrote the paper with J.R. and Y.Z. and input from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-019-0231-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.G.

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

Physical modeling. Detailed descriptions of the physical model of methylation kinetics in DamC, as well as of the polymer model with persistence length, are available as a separate file (Supplementary Note 1).

Cell culture and sample collection. All cell lines are based on feeder-independent PGK12.1 female mESC, kindly provided by Edith Heard's laboratory. The founder cell line in our study is an X0 subclone of the PGKT2 clone described in ref. [33], carrying the insertion of a 256× TetO array within the 3' UTR of the *Chic1* gene on chromosome X and the additional deletion of the *Linx* promoter[6]. Cells were cultured on gelatin-coated culture plates in DMEM (Sigma) in the presence of 15% fetal calf serum (Eurobio Abcys), 100 μM β-mercaptoethanol and 20 U ml^{-1} leukemia inhibitory factor (Miltenyi Biotec, premium grade) in 8% CO$_2$ at 37 °C. Cells were tested for *Mycoplasma* contamination once per month, and no contamination was detected. After insertion of the rTetR-Dam vector (see below), cells were cultured in the presence of 250 μg ml^{-1} hygromycin. To induce nuclear translocation of the rTetR-Dam fusion protein to the nuclei, mESC were trypsinized and directly seeded in culture medium containing 4-OHT, at the concentrations indicated in the main text, for 18 h. Binding of the Dam fusion protein to the TetO arrays was induced by the simultaneous addition of 2.5 μg ml^{-1} Dox.

Generation of cell lines expressing rTetR-Dam and carrying random insertions of TetO arrays. The rTetR-EGFP-Dam-ERt2 construct was cloned into a pBroad3 backbone (Invivogen) carrying a mouse Rosa26 promoter. We used a modified rTetR based on the rtTA-M2 transactivator in ref. [56], which has substantially decreased affinity for the Tet operator in the absence of Dox. The construct was randomly integrated in the PGKT2 X0 subclone by co-transfecting 5×10^5 cells with 3 μg pBROAD3-rTetR-ICP22-EGFP-EcoDam-Ert2 and 0.2 μg pcDNA3.1hygro plasmid using Lipofectamin 2000 (Thermo Fisher Scientific). After 10 days of hygromycin selection (250 μg ml^{-1}), one clone (No. 94.1) expressing low levels of EGFP was selected and expanded for subsequent experiments. To obtain large numbers of viewpoints for DamC experiments, stable random integrations of arrays of TetO sites were introduced in the No. 94.1 mESC clone using the piggyBac transposon system. A mouse codon optimized version of the piggyBac transposase[36] was cloned in frame with the red fluorescent protein tagRFPt (Evrogen) into a pBroad3 vector using Gibson assembly cloning (pBroad3_hyPBase_IRES_tagRFPt). Next, 5×10^5 cells were co-transfected with 0.2 μg pBroad3_hyPBase_IRES_tagRFPt and 1 μg of a piggyBac donor vector containing an array of 50 TetO binding sites using Lipofectamin 2000 (Thermo Fisher Scientific). Cells with high levels of RFP were fluorescence-activated cell sorted (FACS) two days after transfection and seeded at three serial 10× dilutions in 10-cm dishes to ensure optimal density for colony picking. To identify clones with high numbers of TetO integration sites, cells were screened for large numbers of nuclear EGFP accumulation foci using live-cell imaging (see below) in the presence of 500 nM 4-OHT and 2.5 μg ml^{-1} Dox. One polyclonal population (No. 94.1_2.7) and one subclone (No. 94.1_2.7_pureclone3) were further expanded.

To introduce CTCF binding sites flanking the TetO viewpoints, the piggyBac donor vector was modified as follows. Three CTCF binding motifs (TGGCCAGCAGGGGGCGCTG, CGGCCAGCAGGTGGCGCCA and CGACCACCAGGGGGCGCTG) were selected based on high CTCF occupancy in ChIP-seq experiments[11] and cloned into the piggyBac donor vector in an outward direction with respect to the TetO array, including 100 bp of their surrounding endogenous genomic sequence (chr8:13461990-13462089, chr1:34275307-34275419 and chr4:132806684-132806807, respectively). The three CTCF binding motifs were flanked by two LoxP sites for Cre-assisted recombination. We then co-transfected 5×10^5 No. 94.1 with 0.2 μg pBroad3_hyPBase_IRES_rfp and 1 μg of the modified piggyBac donor vector using Lipofectamin 2000. Cells with high levels of RFP were FACS sorted two days after transfection and seeded at three serial 10× dilutions in 10-cm dishes for colony picking. Clones with >50 integration sites were identified through accumulation of EGFP at nuclear TetO foci in the presence of 500 nM 4-OHT and 2.5 μg ml^{-1} Dox. One clone (No. 94.1_216_C3) was further selected for analysis.

Transient transfection. To transiently express rTetR-Dam for the proof-of-principle experiment in Fig. 1d, the PKGT2 X0 subclone was transiently transfected with pBroad3-rTetR-EGFP-Dam-ERt2 using the Amaxa 4D-Nucleofector X-Unit and the P3 Primary Cell 4D-Nucleofector X Kit (Lonza). Cells (5×10^6) were resuspended in 100 μl transfection solution (82 μl primary solution, 18 μl supplement 1, 2 μg pBroad3-rTetR-EGFP-Dam-ERt2) and transferred in a single Nucleocuvette (Lonza). Nucleofection was done using the protocol CG109. Transfected cells were directly seeded in pre-warmed 37 °C culture medium containing 10 nM 4-OHT ± 2.5 μg ml^{-1} Dox. Genomic DNA was collected 18 h after transfection. Sequencing libraries were prepared as previously described[34,57].

Mapping of piggyBac insertion sites. Genomic DNA (2 μg) was fragmented to an average of 500 bp by sonication (Covaris S220; duty cycle, 5%; peak power, 175 W; duration, 25 s). End-repair, A-tailing and ligation of full-length barcoded Illumina adapters were performed using the TruSeq DNA PCR-free kit (Illumina) according to the manufacturer's guidelines, with the exception that large DNA fragments were not removed. Libraries for each sample (750 ng) were pooled together, and fragments of interest were captured using biotinylated probes against the the piggyBac inverted terminal repeats (ITRs) sequence and the xGen Hybridization Capture kit (IDT) according to the manufacturer's protocol (probe concentration of 2.25 pmol μl^{-1}). Following capture, libraries were amplified for 12 cycles using Kapa Hi-fi polymerase and the following primers: 5'-AATGATACGGCGACCACCGAGAT, 5'-CAAGCAGAAGACGGCATACGAGA. Final libraries were purified using AMPure XP beads (1/1 ratio), quality controlled and sequenced on the NextSeq500 platform (paired-end 300 cycles mid-output) for a total of 8×10^8 paired-end reads per sample on average.

Capture probe sequences are as follows:

ITR3-1 [Btn]ATCTATAACAAGAAAATATATATATAATAAGTTATCACGTAAGTAGAACATGAAATAACAATATAATTATCGTATGAGTTAAATCTTAAAAGTCACGTAAAAGATAATCATGCGTCATTT,

ITR3-2 [Btn]TCCAAGCGGCGACTGAGATGTCCTAAATGCACAGCGACGGATTCGCGCTATTTAGAAAGAGAGAGCAATATTTCAAGAATGCATGCGTCAATTTTACGCAGACTATCTTTCTAGGGTTAA,

ITR5-1[Btn]TTAACCCTAGAAAGATAATCATATTGTGACGTACGTTAAAGATAATCATGCGTAAAATTGACGCATGTGTTTTATCGGTCTGTATATCGAGGTTTATTTATTAATTTGAA,

ITR5-2 [Btn]ATTAAGTTTTATTATATTTACACTTACATACTAATAATAAATTCAACAAACAATTTATTTATGTTTATTTATTTATTAAAAAAAAACAAAAACTCAAAATTTCTTCTATAAAGTAACAAA.

Genotyping of CTCF integration sites by PCR. We designed primers binding to endogenous genomic DNA sequence outside the piggyBac 3' ITR based on the genomic position of mapped piggyBac insertion sites. We then amplified the junction between the ITR and the genome using Phusion High-Fidelity DNA Polymerase (Thermo Scientific) with one genomic primer and a T7 promoter primer (5'TAATACGACTCACTATAGGG3') flanking the piggyBac CTCF integration cassette (see Supplementary Fig. 6e). PCR products were purified and Sanger sequenced. For the verification of CTCF integrations shown in Fig. 6 on chromosomes 6 and 10, the following genomic primers were used: Ch6_flxCTCF_11F (5'AGGCATTCTGTCCAACTGGT3') and Chr10_flxCTCF_13F (5'TGTTGAGCATCTATCACATTCCTTA3').

Excision of CTCF sites using Cre recombinase. To excise ectopically inserted CTCF sites from clone No. 94.1_216_C3, 5×10^5 cells were transfected with 0.5 μg pIC-Cre[58] using Lipofectamine 2000 (Thermo Fisher Scientific). After 4 days under G418 selection (300 μg ml^{-1}), single colonies were expanded and genotyped following the procedure described above.

Live-cell Imaging. Gridded glass-bottom dishes (Mattek) were coated with 2 μg ml^{-1} recombinant mouse E-cadherin (R&D Systems, No. 748-EC) in PBS at 4 °C overnight. Cells (5×10^5) were seeded in full medium, one day before imaging, supplemented with 4-OHT and Dox as indicated above. Cells were imaged with a Nikon Eclipse Ti-E inverted widefield microscope (Perfect Focus System with real-time drift correction for live-cell imaging) operating in highly inclined and laminated optical sheet (HILO) mode using a CFI APO total internal reflection fluorescence (TIRF) 100×/NA 1.49 oil objective (Nikon). A 488-nm, 200-mW Toptica iBEAM SMART laser was used as excitation source. Cells were maintained at a constant temperature of 37 °C and in 8% CO$_2$ within an incubation box. Images were collected with an Evolve 512 Delta EMCCD high-speed Camerang using Visiview (Visitron). Background subtraction (150-pixel rolling ball radius) and maximum intensity projections were performed in ImageJ.

Nuclear volume measurements. Cells (3×10^6) from mESC clone No. 94.1_2.7 were cultured in gelatin-coated 6-well plates in full medium, dissociated for 5 min at room temperature with Accutase (GIBCO) then centrifuged for 4 min at 950 r.p.m. and resuspended in 500 μl culture medium. Cell suspension droplets (25 μl) were spotted on coverslips previously coated with poly-L-lysine, adsorbed on ice for 5 min and washed gently once with 1× PBS. Cells were then permeabilized on ice for 5 min in 1× PBS and 0.5% Triton X-100, and coverslips were stored in 70% ethyl alcohol at −20 °C. Nuclei were counterstained with 0.2 mg ml^{-1} DAPI, and *Z*-stack images were acquired using a Zeiss Z-1 microscope equipped with a ×40 oil immersion lens (numerical aperture 1.3; voxel size $0.227\times0.227\times0.73\,\mu m^3$). *Z*-stacks were then deconvolved using Huygens software (20 iterations of the CMLE algorithm). To segment individual nuclei, we binarized DAPI images based on a single-intensity threshold based on the fact that image histograms of all *Z*-stacks were bimodal (threshold, 7,000 in 32-bit images). The volumes of binary three-dimensional (3D) objects were then calculated using the 3DObjectCounter plugin in FIJI/ImageJ, excluding objects on the edges of each *Z*-stack.

Preparation of nuclear extracts. Cell nuclei were extracted as previously described[59]. Briefly, 10^7 mESC were seeded in ES medium (see above) supplemented with the appropriate concentration of 4-OHT on a gelatin-coated

15-cm^2 dish. The next day, cells were harvested using trypsin and washed twice in ice-cold PBS. Next, cells were carefully resuspended in 500 µl ice-cold Buffer A1 (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.34 M sucrose, 10% glycerol, 0.1% Triton X-100, 1 mM DTT, 1 mM phenylmethanesulfonyl fluoride) to obtain nuclei. After incubation for 5 min on ice, extracted nuclei were washed twice with buffer A1.

Mass spectrometry. Nuclear extracts were dissolved in 400 µl 50 mM HEPES pH 8.5 in 8.3 M guanidine hydrochloride. All samples were heated at 95 °C for 5 min, sonicated using the Bioruptor sonication device and supplemented with 5 mM tris(2-carboxyethyl)phosphine and 10 mM chloroacetamide (CAA). To reduce sample complexity, lysates were diluted to 6 M guanidine hydrochloride and transferred onto 100 kDa molecular weight cutoff Amiconultra-0.5 centrifugal filter units. Samples were concentrated for 2 × 15 min at 14 kg followed by refill of the filter with 6 M guanidine hydrochloride in 50 mM HEPES pH 8.5 and 3 × 45 min at 14 kg, then followed by refilling of the filter with 1 M guanidine hydrochloride in 50 mM HEPES pH 8.5. For digestion, 10 µg Lys-C (Wako Chemicals) and 10 µg trypsin (Thermo Fisher Scientific) were added to each sample and incubation overnight at 37 °C. The following morning, an additional 10 µg of trypsin was added, with incubation for 3 h and acidification using trifluoroacetic acid.

To estimate nuclear proteins, copy numbers samples were desalted using SEP-PAK (Waters) and subjected to high-pH offline fractionation on a YMC Triart C18 0.5 × 250 mm^2 column (YMC Europe GmbH) using the Agilent 1100 system (Agilent Technologies). Ninety-six fractions were collected for each experiment and concatenated into 48 fractions as previously described[60]. For each liquid chromatography–tandem mass spectrometry (LC–MS) analysis, approximately 1 µg of peptides was loaded onto a PepMap 100 C18 2-cm trap (Thermo Fisher Scientific) using the Proxeon NanoLC-1000 system (Thermo Fisher Scientific). Online peptide separation was performed on a 15-cm EASY-Spray C18 column (ES801, Thermo Fisher Scientific) by applying a linear gradient of increasing acetonitrile (ACN) concentration at a flow rate of 150 nl min^{-1}. An Orbitrap Fusion Tribrid (Thermo Fisher Scientific) or an Orbitrap Fusion LUMOS Tribrid (Thermo Fisher Scientific) mass spectrometer was operated in data-dependent mode. The ten most intense precursor ions from the Orbitrap survey scan were selected for higher-energy collisional dissociation fragmentation and analyzed using the ion trap.

Mass spectrometry data processing. Maxquant v.1.5.3.8 was used to search raw mass spectrometry data using default settings[61,62] against the mouse protein sequences from Uniprot database (released April 2017). The label-free quantification (LFQ) algorithm was used for quantification. The protein groups table was loaded to Perseus software[63] (v.1.5.0.0) filtered for potential contaminants and reverse hits. Protein copy numbers per cell were calculated using the Protein ruler plugin of Perseus by standardization to the total histone MS signal[64]. LFQ values were normalized using the same normalization for all samples. To estimate cytoplasmic contamination 'GOCC slim name' annotations provided in Perseus were used. Exclusively cytoplasmic proteins were defined as those associated with the GOCC terms 'cytoplasm' or 'cytosol', and not associated with the terms 'nucleus', 'nuclear', 'nucleoplasm' or 'nucleosome'. Exclusively nuclear proteins were defined as those associated with the GOCC terms 'nucleus', 'nuclear', 'nucleoplasm' or 'nucleosome', and not associated with the terms 'cytoplasm' or 'cytosol'. Cytoplasmic contamination was estimated using a ratio of summed LFQ intensity between exclusively cytoplasmic proteins and exclusively nuclear proteins in samples with and without nuclear extraction.

PRM data acquisition and analysis. To select peptides for PRM assays, the rTetR-Dam-EGFP-ERT2 construct was enriched with ChromoTek GFP-Trap magnetic beads and analyzed using shotgun data-dependent acquisition LC–MS/MS on an Orbitrap Fusion Lumos platform as decribed above. For PRM analysis, the resolution of the orbitrap was set to 240 k full width at half maximum (at 200 m/z), the fill time was set to 1,000 ms and the ion isolation window was set to 0.7 Th. For LC–MS analysis of samples derived from a polyclon carrying 890 TetO array insertions, approximately 1 µg of peptides was loaded onto a PepMap 100 C18 2-cm trap (Thermo Fisher Scientific) using the Proxeon NanoLC-1000 system (Thermo Fisher Scientific). Online peptide separation was performed on the 15-cm EASY-Spray C18 column (ES801, Thermo Fisher Scientific) by applying a linear gradient of increasing ACN concentration at a flow rate of 150 nl min^{-1}. For LC–MS analysis of samples derived from a polyclon carrying 100 TetO array insertions, approximately 1 µg of peptides was separated online on a 50-cm µPACTM cartridge (PharmaFluidics) by applying a linear gradient of increasing ACN concentration at a flow rate of 300 nl min^{-1} with the Proxeon NanoLC-1000 system (Thermo Fisher Scientific). The acquired PRM data were processed using Skyline 4.135 (ref. [65]). The transition selection was systematically verified and adjusted when necessary to ensure that no co-eluting contaminant distorted quantification, based on traces co-elution (retention time) and the correlation between the relative intensities of the endogenous fragment ion traces and their counterparts from the library. As a loading control, the mean of total MS1 signal was estimated using RawMeat v.2.0b1007.

DamC library preparation. DamC experiments were based on a newly developed DamID-seq next-generation sequencing library preparation protocol to maximize the proportionality between methylation levels and sequencing readout (Supplementary Fig. 2d). One crucial issue in the calculation of enrichment, as shown in Fig. 2c, is that small fluctuations in –Dox methylation in the denominator can be amplified into large fluctuations in enrichment levels. GATC sites must therefore be equally and robustly represented in the DamID sequencing library, irrespective of their methylation level. From this perspective, the principal limitation of the original DamID protocol[57] for our present application is its dependence on the genomic distance between two GmATC sites, resulting in large adapter-ligated molecules and, as a consequence, in a strong bias towards densely methylated regions. In our optimized protocol, GmATC sites are sequenced independently of the neighboring GATC methylation status, resulting in an increase of ~30% in GmATC coverage at equivalent sequencing/read depth (Supplementary Fig. 2e). In addition, we introduced UMIs allowing a precise enrichment quantification after exclusion of PCR duplicates from the sequencing data.

Overall, the DamC library construction protocol can be divided in three parts: (1) ligation of UMI adapters with a 'one-tube' strategy, (2) integration of the second sequencing adapter, followed by (3) a final PCR amplification. Briefly, 3 × 10^6 cells were harvested using trypsin after 18 h induction with tamoxifen ± Dox. Genomic DNA was extracted using the Qiagen blood and tissue kit, adding 250 U RNaseA in step 1. Genomic DNA was eluted in 80 µl double-distilled H_2O. DNA concentration was measured using the Qbit DNA Broad Range kit. Genomic DNA (350-ng input) was treated with Shrimp Alkaline Phosphatase treatment (NEB, 1 U), followed by DpnI digestion (Thermo Fisher Scientific, 10 U), A-tailing (0.6 mM final dATP, 5 U Klenow exo-, Thermo Fisher Scientific) and UMI ligation (30 U T4 DNA ligase, PEG4000, Thermo Fisher Scientific), performed within the same tube and buffer (Tango 1×, Thermo Fisher Scientific) by heat inactivation of each enzymatic step followed by adjustment with the reagents required for the next step. UMI adapters were made by annealing the following oligos: 5′-AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATC*T and 5′-pGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT. Ligation reactions were treated with Exonuclease I (20 U, Thermo Fisher Scientific) then purified using AMPure XP beads (1/0.8 ratio, Agencourt), and the second sequencing adapter (5′ TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNN*N 3′, IDT) was tagged using heat denaturation and second-strand synthesis (5 U T4 DNA Polymerase, Thermo Fisher Scientific). The tagging reaction was purified using AMPure XP beads (1/1 ratio) followed by a final library amplification (12 cycles) using 1 U Phusion polymerase, 2 µl 10 µM DAM_UMIindex_PCR (5′ AATGATACGGCGACCACCGAGATCTACA*C 3′) and 2 µl 10 µM NEBnext indexed primer (NEB). Final libraries were purified using AMPure XP beads (1/1 ratio) and QCed using Bioanalyser and Qbit. DamC libraries were sequenced on a NextSeq500 (75 cycles single-end) with a custom-sequencing protocol (dark cycles at the start of read1 to 'skip' the remaining DpnI site TC sequence). Samples index was determined using index1 read, and UMI sequence using index2 read. Details on numbers of total and valid reads can be found in Supplementary Table 5.

4C-seq. Sample preparation for 4C was performed as previously described[66]. Briefly, 10^7 cells were crosslinked in 2% formaldehyde for 10 min and quenched with glycine (final concentration 0.125 M). Cells were lysed in 150 mM NaCl/50 mM Tris-HCl (pH 7.5)/5 mM EDTA/0.5% NP-40/1% Triton X-100. The first digest was performed with 200 U DpnII (NEB), followed by ligation at 16 °C with 50 U T4 DNA ligase (Roche) in 7 ml. Ligated samples were de-crosslinked with Proteinase K (0.05 µg ul^{-1}) at 65 °C, purified and digested with 50 U Csp6I (Thermo Fisher Scientific), followed by ligation with 100 U T4 DNA ligase in 14 ml and purification. The resulting products were used directly as a PCR template for genomic dedicated 4C viewpoints. Primers for PCR were designed using guidelines described previously[66]. We obtained the following read counts: for cell line No. 94.1_2.7 (135 TetO insertions only), 5.7 × 10^6 valid reads in total (+Dox, two replicates); for cell line No. 94.1_216_C3 (TetO-CTCF), 3.5 × 10^6 valid reads on average per sample; for the experiments shown in Supplementary Figs. 3c and 5c, we obtained an average of 3.2 × 10^6 reads per sample. Detailed numbers of total and valid reads can be found in Supplementary Table 5.

In vitro Cas9 digestion of 4C templates. For direct detection of chromosomal interactions from the genome-integrated TetO platform, viewpoint primers were designed for direct amplification from the DpnII fragments contained in the TetO sequence. The 2.7-kb TetO platform contains a total of 50x contiguous repeats of the same TetO DpnI/II viewpoint. To prevent PCR amplification and sequencing of TetO repeats due to tandem ligation of two or more TetO DpnII fragments in a given 4C circle, in vitro Cas9 digestion was performed on the 4C templates. Cas9 was targeted into the TetO repeats between viewpoint primers using a single-guide RNA. In vitro transcribed guide RNA template was obtained using the Megashortscript T7 transcription kit (Invitrogen); gRNA was purified with 4× AMPure purification (Agencourt). Purified Cas9 protein was kindly provided by N. Geijsen. Cas9 was pre-incubated with sgRNA for 30 min at 37 °C. Subsequently, 4C template DNA was added to the pre-incubated gRNA–Cas9 complex and incubated for 3–6 h at 37 °C for digestion. Cas9 was inactivated by incubation at 70 °C for 5 min.

Hi-C library preparation. A total of 6×10^6 mESC were harvested and diluted in 1× PBS to a final concentration of 1×10^6 cells ml^{-1}, then crosslinked with 1% formaldehyde and quenched with 0.125 M glycine for 5 min at room temperature. After two 1× PBS washes, cell pellets were obtained by centrifugation, snap-frozen and stored at −80 °C. Pellets were thawed on ice and resuspended in 500 ul lysis buffer (10 nM Tris-HCl pH 8.0, 10 nM NaCl, 0.2% NP-40, 1× Roche protease inhibitors) and left for 30 min on ice. Cells were then pelleted by centrifugation (954*g*, 5 min, 4 °C), washed once with 300 μl 1× NEB2 buffer and nuclei were extracted with 1 h incubation at 37 °C in 190 μl 0.5% sodium dodecyl sulfate and 1× NEB2 buffer. Sodium dodecyl sulfate was neutralized by dilution of the sample with 400 μl NEB2 buffer and the addition of 10% Triton X-100. After 15 min incubation at 37 °C, nuclei were pelleted, washed once in PBS and resuspended in 300 μl NEB2 buffer, then 400 U of MboI (NEB, 25,000 units ml^{-1}) were added and incubated at 37 °C overnight. The next day, nuclei were pelleted again, resuspended in 200 μl fresh NEB2 buffer and an additional 200 U MboI was added for two further hours before heat inactivation at 65 °C for 15 min. Next, 43 μl of end-repair mix (1.5 μl 10 mM dCTP, 1.5 μl 10 mM dGTP, 1.5 μl 10 mM dTTP, 37.5 μl 0.4 mM Biotin-11-dATP (Invitrogen)and 1 μl 50 U μl^{-1} DNA Polymerase I Large Klenow fragment (NEB) were added to the nuclear suspension, incubated at 37 °C for 45 min and heat inactivated at 65 °C for 15 min. The end-repair mix was exchanged with 1.2 ml ligation mix (120 μl 10× T4 DNA Ligase Buffer, 100 μl 10% Triton X-100, 6 μl 20 mg ml^{-1} BSA, 969 μl H_2O) plus 5 μl T4 ligase (NEB, 2,000 units ml^{-1}), and ligation was performed at 16 °C overnight. Nuclei were reconstituted in 200 μl fresh NEB2 buffer followed by RNA digestion in 0.5 mg ml^{-1} RNAse A for 10 min at 37 °C. Samples were de-crosslinked with Proteinase K at 65 °C overnight and DNA was purified using phenol/chloroform. The DNA sample (2 μg) was sonicated using Diagenode Bioruptor Pico. MyOne Streptavidin T1 (Life Technologies, No. 65601) magnetic beads were used to capture biotinylated DNA followed by A-tailing. Adapter ligation was performed according to NEB Next Ultra DNA Library prep kit instructions. Two independent PCR reactions with multiplex oligos for Illumina sequencing were performed and pooled for the final PCR cleanup by magnetic AMPure bead (Beckman Coulter) purification. The final libraries were eluted in nuclease-free water, QCed by Bioanalyzer and Qubit. HiC libraries were sequenced on an Illumina Nextseq500 platform (2 × 42 bp paired-end). We obtained an average of 3.5×10^8 valid reads per sample (TetO-only and TetO-CTCF cells, –Dox, two biological replicates each). Details on numbers of total and valid reads can be found in Supplementary Table 5.

Sequencing data processing and data analysis. *DamC analysis.* All samples were aligned to mouse mm9 using qAlign (QuasR package[67]) using default parameters. PCR duplicates were removed using a custom script. Briefly, reads were considered PCR duplicates if they mapped to the same genomic location and had the same 8-bp UMI sequence. We quantified the number of reads mapped to each GATC that could be uniquely mapped using qCount (QuasR package[67]). The query object we used in qCount was a GRanges object containing the uniquely mappable 76-mer GATC loci in the genome shifted upstream (plus strand) or downstream (minus strand) by five base pairs (three dark cycles + GA, see DamC library preparation paragraph in the Methods section). Each sample was then normalized to a common library size of 10 million reads and a pseudo-count of 0.2 was added. Before calculation of DamC enrichments, a running average over 21 restriction fragments was performed and the mean value was assigned to the central GATC. Enrichment was then calculated as in Fig. 2c: $E=([+\text{Dox}]-[-\text{Dox}])/[-\text{Dox}]$, where [+Dox] and [–Dox] are the normalized and running-averaged number of reads in the presence and absence of Dox, respectively. We define the DamC signal as saturated if it satisfies the following criteria: (1) it belongs to the highest 25% genome width in both + Dox and –Dox samples, and (2) the ratio between +Dox and –Dox methylation is close to 0.5—that is, it belongs to the [0.45, 0.55] quantile of all ratios genome wide. Coordinates of excluded viewpoints in the clonal cell line with TetO integrations are: chr6:25758950, chr8:26653938, chr8:96714938, chr11:33429300 and chr11:51411650.

4C analysis. Mapping of 4C reads was performed as described for DamC, with the exception of UMI de-duplication, since 4C libraries do not include UMIs and quantification was done by counting the reads mapped exactly to the GATC sites. The two restriction fragments immediately flanking the piggyBac-TetO cassette were excluded from subsequent analyses.

Hi-C analysis. Hi-C data were analyzed using HiC-Pro[68] v.2.7.10 with the -very-sensitive -end-to-end -reorder option. Briefly, reads pairs were mapped to the mouse genome (build mm9). Chimeric reads were recovered after recognition of the ligation site. Only unique valid pairs were kept. Contact maps at a given binning size were then generated after dividing the genome into equally sized bins and applying iterative correction[69] on binned data.

Fit of scaling plots. Average normalized Hi-C counts, DamC enrichment and 4C counts were calculated for all pairs of loci separated by logarithmically binned distance intervals. The binning size in logarithmic scale (base 10) was 0.1. Curves were fitted in log–log scale using the lm function in R.

Fitting the DamC model to DamC experiments as a function of 4-OHT concentration. DamC enrichment is dependent on the rTetR-TetO-specific and -non-specific dissociation constants, the concentration of TetO and the nuclear rTetR-Dam concentration (Supplementary Note 1). In addition, it is dependent on the actual contact probability between the genomic location where it is calculated and the TetO viewpoint. In Fig. 3d we calculated the DamC enrichment at the fragments closest to the 100 TetO viewpoint with higher signal to noise ratio in the polyclonal line. We assumed that the contact probability between the TetO array and the closest fragment is ~1, and fitted the model to the experimental data using the other parameters with the NonlinearModelFit function in Mathematica. The constraints that the dissociation constants and the concentration of TetO are positive were imposed. The goodness of the fit was evaluated using the adjusted R^2 (0.73). In the clonal line, we assumed that the specific dissociation rTetR-TetO does not change compared to the polyclonal line and, by setting the concentration of viewpoints to 135 per cell, we fitted the non-specific dissociation constant using the NonlinearModelFit function in Mathematica. Model fitting resulted in an estimate of 5 nM for the average non-specific binding constant accounting for rTetR and Dam interactions with GATC sites genome wide. The goodness of the fit was evaluated using the adjusted R^2 (0.68).

ChromHMM. To assign chromatin states, we used the ChromHMM software[70] with four states. We used histone modifications as in Supplementary Table 6. The four states correspond to active (enriched in H3K36me3, H3K27ac, H3K4me1 and H3K9ac), poised (enriched in H3K36me3, H3K27ac, H3K4me1, H3K9ac and H3K27me3), inert (no enrichment) and heterochromatic (enriched in H3K9me3) states.

Deviation scores. Given a set of restriction fragments (or genomic bins) $\{x_i\}$ belonging to a window $[a,b]$, the deviation score (Dev) is defined as

$$\mathrm{Dev}(a,b)=2\frac{\sqrt{\langle(\mathbf{f}-\mathbf{g})^2\rangle_{[a,b]}}}{\langle|\mathbf{f}|+|\mathbf{g}|\rangle_{[a,b]}}$$

where **f** and **g** are data vectors (for example, DamC enrichment, 4C or virtual 4C counts) and $\langle\rangle_{[a,b]}$ represents the average in the window $[a,b]$. If two profiles are identical in the window $[a,b]$, then the deviation score is zero; increasing deviation from zero indicates increasing dissimilarity.

PiggyBac-TetO integration site mapping. Paired-end reads (see Mapping of piggyBac insertion sites, above) were trimmed to 50 bp using a custom script. Read1 and Read2 were mapped separately to the piggyBac-TetO sequence using QuasR (qAlign). Only hybrid pairs with one of the reads mapping to array were kept. The second reads from hybrid pairs were mapped to the mouse genome (build mm9) using QuasR (qAlign). Reads were then piled up in 25-bp windows using csaw (windowCounts function). Integration sites can be identified because they correspond to local high-read coverage. Local coverage was calculated by resizing all non-zero 25-bp windows up to 225 bp (expanding by 100 bp upstream and downstream). Overlapping windows were then merged using reduce (from GenomicRanges), resulting in a set of windows $\{w_i\}$. The size distribution of w_i is multimodal, and only w_i from the second mode onward were kept. For each w_i we estimated the coverage c_i as the number of non-zero 25-bp windows. Only w_i where the coverage was >16 were considered. The exact positions of the integration sites were then identified with the center of w_i.

Determination of the orientation of TetO-CTFC insertions. To determine the orientation of ectopically inserted TetO-CTCF sites, we exploited the fact that the three CTCF sites are oriented within the piggyBac casette in the 3'–5' ITR direction. If the genomic position of the 5' ITR is upstream of the 3' ITR, then CTCF sites are in the reverse orientation (– strand) and vice versa. To determine the relative orientation of the 3' and 5' ITRs in the genome, we used only reads that run through the junction between the ITRs and the genome. More precisely, we extracted reads that contained an exact match to 30 bps of the ITRs (3' and 5' ITR separately), trimmed the ITR sequence and mapped the reads to the mouse genome using qAlign (from QuasR). We quantified the reads at single-base pair resolution using scanBam. Only integration sites where both 5' and 3' ITRs are mapped were kept. This resulted in nine integration sites (Supplementary Table 4).

Z-score analysis of Hi-C data. To identify and exclude 'noisy' interactions in Hi-C maps we used a custom algorithm named 'Neighborhood Coefficient of Variation' (van Bemmel et al., manuscript submitted). Since the chromatin fiber behaves as a polymer, the contact probability of a given pair of genomic loci *i* and *j* is correlated to that of fragments $i+N$ and $j+N$ if *N* is smaller (or in the order of) than the persistence length of the chromatin fiber. Hence, a given pixel in a Hi-C map can be defined as noisy if its numerical value is widely different from those corresponding to neighboring interaction frequencies. To operatively assess the similarity between neighboring interactions, we calculated the coefficient of variation within a 10 × 10 pixel square centered on every interaction and discarded all pixels whose coefficient of variation was larger than a certain threshold. Given that the distribution of the coefficient of variation of Hi-C samples in this study is multimodal, with the first component terminating around0.6, we set the coefficient of variation threshold to 0.6. Discarded interactions appear as gray pixels in the differential Hi-C maps. For differential analysis between TetO-CTCF and WT samples, we calculated the difference between distance-normalized *Z*-scores calculated for each individual map[71]. The *Z*-score is defined as (obs – exp)/stdev, where obs is the Hi-C signal for a given interaction and exp and stdev are the genome-wide average and standard deviation, respectively, of Hi-C signals at the genomic distance separating the two loci.

4C peak calling. To call specific interactions in 4C profiles, we used the peakC package[41] using the following parameters: qWr = 2.5 and minDist = 20,000. peakC was applied to two replicates of running averaged (21 fragments average) 4C profiles at single-fragment resolution. Peak regions were then extended 1 kb upstream and downstream. Overlapping peaks were merged.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequencing data from this study, including bedgraph files for the visualization of DamC and 4C profiles from all samples described in the manuscript, are available at the NCBI Gene Expression Omnibus with accession code GEO GSE128017. A University of California, Santa Cruz session containing all the DamC and 4C tracks used can be found at https://genome.ucsc.edu/s/zhan/DamC_publication_2019. The mass spectrometry proteomics data have been deposited with the ProteomeXchange Consortium via the PRIDE[72] partner repository with the dataset identifier PXD013507. Source data for Figs. 1 and 3–7 and Supplementary Figs. 1–3, 5 and 6 are available online.

Code availability

The custom-made codes used to analyze the data are available at https://github.com/zhanyinx/NMSB_2019_redolfi_et_al.

References

56. Urlinger, S. et al. Exploring the sequence space for tetracycline-dependent transcriptional activators: novel mutations yield expanded range and sensitivity. *Proc. Natl Acad. Sci. USA* **97**, 7963–7968 (2000).
57. Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nat. Protoc.* **2**, 1467–1478 (2007).
58. Gu, H., Zou, Y.-R. & Rajewsky, K. Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting. *Cell* **73**, 1155–1164 (1993).
59. Sanulli, S. et al. Jarid2 methylation via the PRC2 complex regulates H3K27me3 deposition during cell differentiation. *Mol. Cell* **57**, 769–783 (2015).
60. Wang, Y. et al. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**, 2019–2026 (2011).
61. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
62. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteom.* **13**, 2513–2526 (2014).
63. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
64. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "Proteomic Ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteom.* **13**, 3497–3506 (2014).
65. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
66. Splinter, E., de Wit, E., van de Werken, H. J. G., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
67. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
68. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
69. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
70. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
71. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
72. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

nature structural & molecular biology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41594-019-0231-0

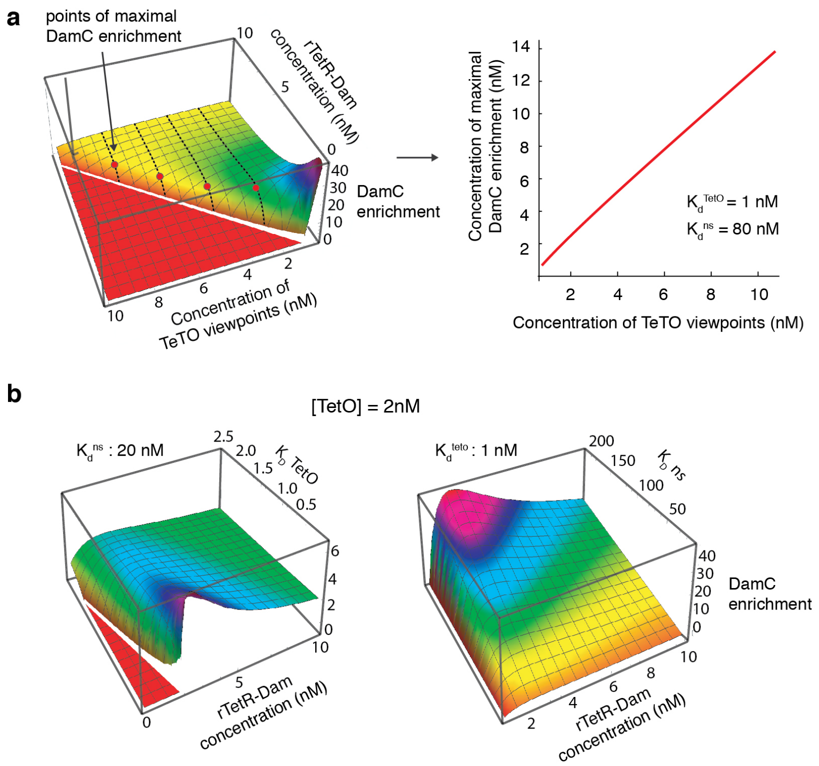
In the format provided by the authors and unedited.

DamC reveals principles of chromatin folding in vivo without crosslinking and ligation

Josef Redolfi[1,2,8], Yinxiu Zhan[1,2,8], Christian Valdes-Quezada[3,4,8], Mariya Kryzhanovska[1], Isabel Guerreiro[3,4], Vytautas Iesmantavicius[1], Tim Pollex[5], Ralph S. Grand[1], Eskeatnaf Mulugeta[6], Jop Kind[3,4], Guido Tiana[7], Sebastien A. Smallwood[1], Wouter de Laat[3,4] and Luca Giorgetti[1]*

---

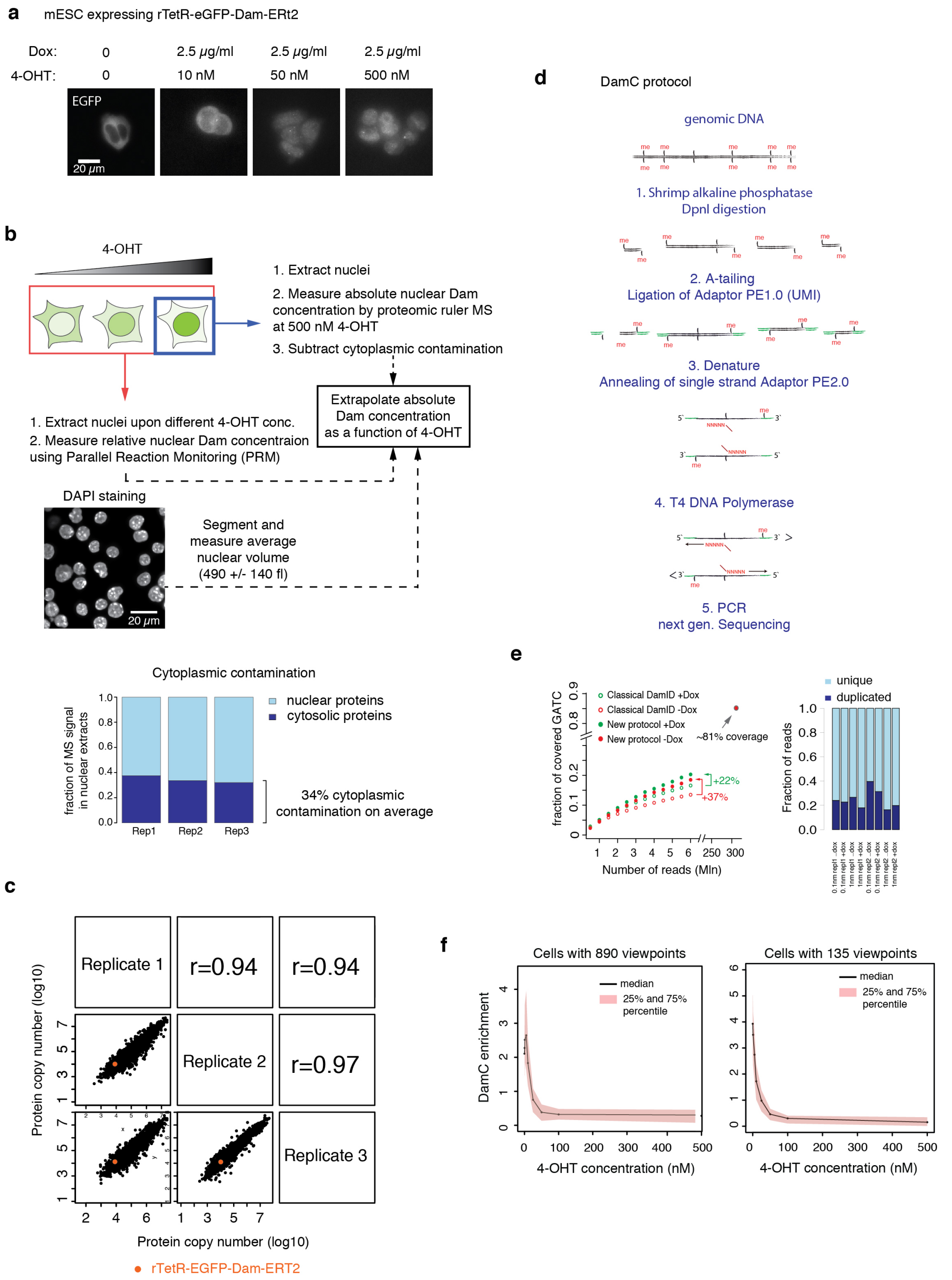
[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Oncode Institute, Hubrecht Institute–KNAW, Utrecht, the Netherlands. [4]University Medical Center Utrecht, Utrecht, the Netherlands. [5]EMBL, Heidelberg, Germany. [6]Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands. [7]Università degli Studi di Milano and INFN, Milan, Italy. [8]These authors contributed equally: J. Redolfi, Y. Zhan, C. Valdes-Quezada. *e-mail: luca.giorgetti@fmi.ch



Supplementary Figure 1

Parameter study of model predictions

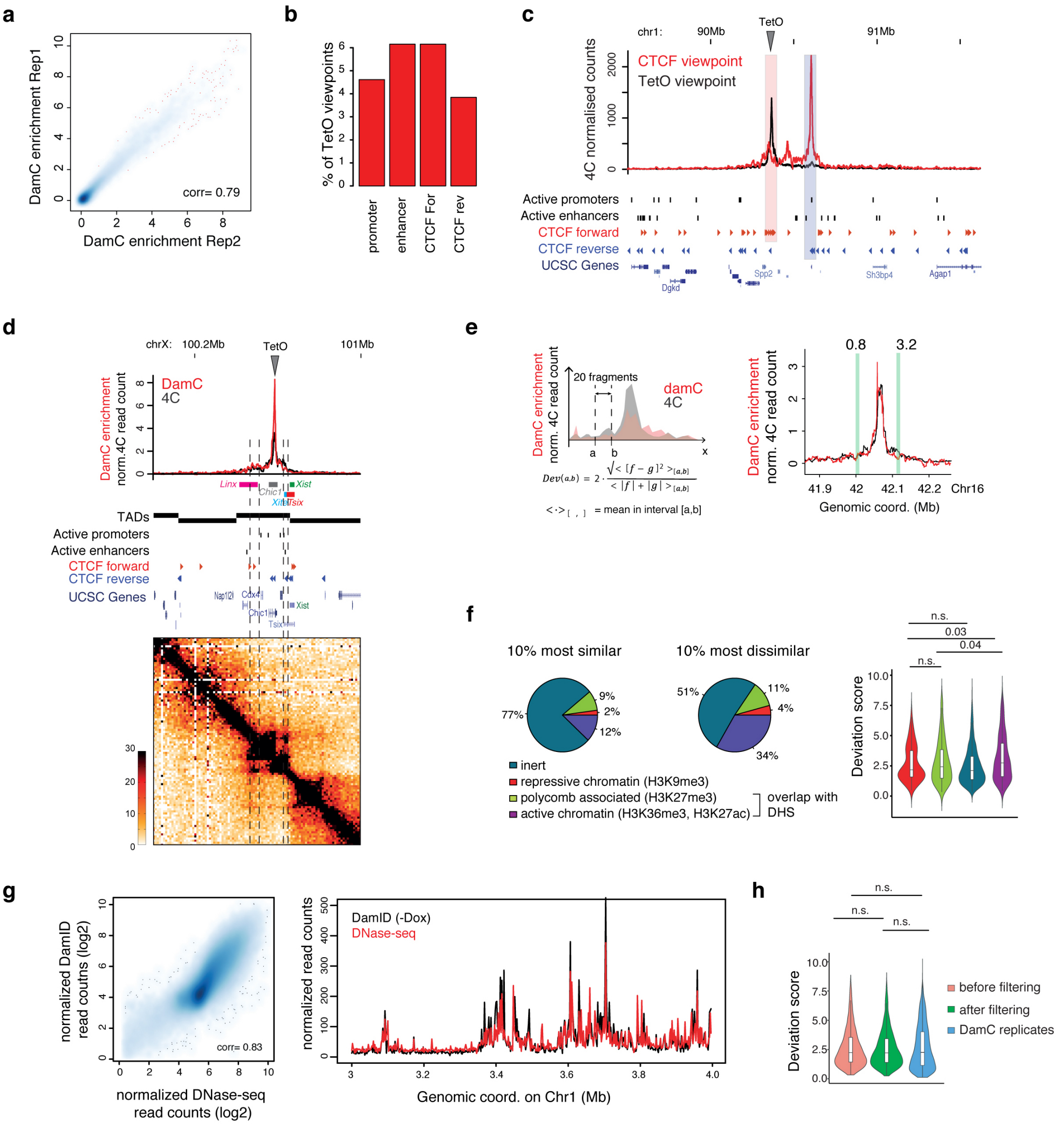
a) Left: DamC enrichment is plotted as a function of the concentrations of rTetR-Dam and TetO viewpoints, imposing specific and non-specific dissociation constants to 1 nM and 80 nM respectively. Right: the rTetR-Dam concentration where the DamC enrichment is maximal is linearly correlated with the concentration of TetO viewpoint. **b)** DamC enrichment shows a maximum irrespective of the choice of the numerical parameters. This is exemplified by plots of DamC enrichment as a function of rTetR-Dam concentration when varying the TetO specific affinity and keeping the nonspecific affinity fixed (left panel) and vice versa (right panel).



Supplementary Figure 2

Experimental system and optimized DamC protocol.

a) rTetR-Dam-EGFP-ERT2 becomes increasingly localized to the nucleus upon increasing 4-OHT concentration in the culture medium, as shown by the increasingly nuclear accumulation of EGFP. Maximum intensity projections of 10 wide-field Z planes are shown. Bright spots indicate binding of rTetR-Dam-EGFP-ERT2 to the 256x TetO array on chromosome X (see Figure 1c). **b)** Schematics of the strategy for measuring rTetR-Dam-EGFP-ERT2 nuclear concentrations as a function of 4-OHT concentration. After exposing the cells to different concentrations of 4-OHT, nuclei were extracted and prepared for mass spectrometry. The relative abundance of nuclear rTetR-EGFP-Dam-ERT2 was measured using parallel reaction monitoring (PRM) using two replicate samples from all 4-OHT concentrations. Absolute quantification was performed in triplicate uniquely in the 500 nM 4-OHT sample using proteomic-ruler based mass spectrometry measurements (Wiśniewski et al. Mol. Cell. Proteomics 13, 3497–3506, 2014). We then extrapolated absolute nuclear rTetR-Dam copy numbers at all concentrations of 4-OHT based on the absolute quantification at 500 nM 4-OHT and the relative PRM quantification. Finally, the nuclear concentration of Dam-fusion Protein was calculated based on the average nuclear volume determined based on DAPI staining. Contamination from cytoplasmic proteins was estimated by comparing protein copy numbers of nuclear and whole-cell extracts, and subtracted from nuclear copy numbers. **c)** Protein copy numbers determined in nuclear extracts at 500 nM 4-OHT using the proteomic ruler strategy (Wiśniewski et al. Mol. Cell. Proteomics 13, 3497–3506, 2014). Data from three biological replicates are plotted before correction for cytoplasmic contamination. **d)** Schematics of the DamC library preparation. Genomic DNA is extracted from cells expressing the Dam-fusion protein. To avoid nonspecific ligation events in step 2, DNA is treated with shrimp alkaline phosphatase prior to DpnI digestion. After digestion with DpnI, a non-templated adenine is added to the 3' blunt end of double-stranded DNA followed by ligation of the UMI-Adapter. Next, double-stranded DNA is denatured before random annealing of the second single stranded Adapter. In step 4, a T4-DNA-Polymerase is used for removal of 3’ overhangs and synthesis in the 5´→ 3´ direction. Finally, libraries are amplified by PCR and prepared for next generation sequencing. UMI: Unique Molecular Identifier. **e)** The DamC sequencing library preparation protocol includes UMIs allowing to filter ~40% of duplicated reads, and increases by roughly 30% the coverage of methylated GATC sites genome-wide compared to classical DamID (Peric-Hupkes et al. Mol. Cell 38, 603–613, 2010). at the same sequencing depth. **f)** Median DamC enrichment at the same viewpoints used for Figure 3d as a function of 4-OHT concentration. Significant amounts of DamC enrichment in our experimental system can be observed in a range of rTetR-Dam nuclear concentrations corresponding to 5-10 and 0.1-1 nM 4-OHT for the lines carrying 890 and 135 viewpoints, respectively.

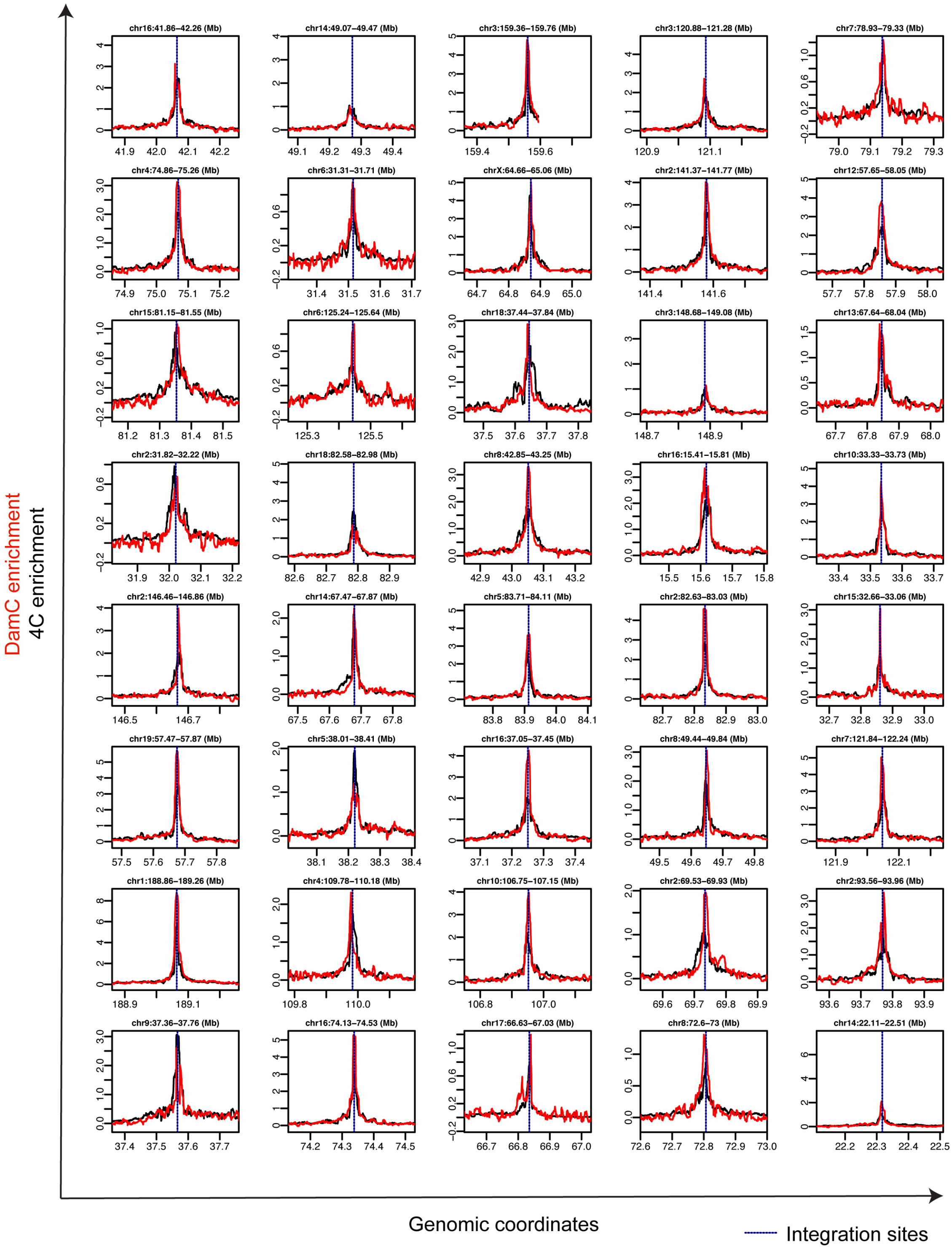


Supplementary Figure 3

Characterization of the TetO-piggyBac clonal cell line and saturation analysis.

a) DamC enrichment from single DpnI fragments within +/- 100 kb from individual TetO viewpoints is plotted for two biological replicates performed with 0.1-1 nM 4-OHT. The Spearman correlation coefficient between the two replicates is indicated. **b)** The percentage of TetO viewpoints inserted in close proximity (<1 kb) from an active promoter or enhancer, or from a CTCF site that is bound in ChIP-seq (Nora et al. Cell 169, 930-944.e22, 2017). **c)** 4C interaction profiles obtained using a TetO viewpoint within 2kb from an endogenous CTCF site and the partner CTCF locus as a reverse viewpoint. **d)** DamC and 4C interaction profiles measured from a TetO viewpoint inserted at the 3'UTR of the *Chic1* gene within the *Tsix* TAD in the X inactivation center. Dashed lines indicate the interactions of *Chic1* with the *Linx* and *Xite* loci. **e)** Definition of a deviation score measuring local differences between DamC and 4C. The deviation score is defined as the average quadratic difference between the DamC and the 4C signal in a 20-restriction fragment interval, normalized by the mean of the signal in the same interval. Two intervals are shown on the right to illustrate the differences between deviation scores of ~1 and ~3. **f)** Left: the 10% most dissimilar 20-fragment intervals are enriched in active chromatin, based on the dominant ChromHMM state (Ernst & Kellis. Nat. Methods 9, 215–216, 2012) in the interval using four chromatin states (ChromHMM emissions)

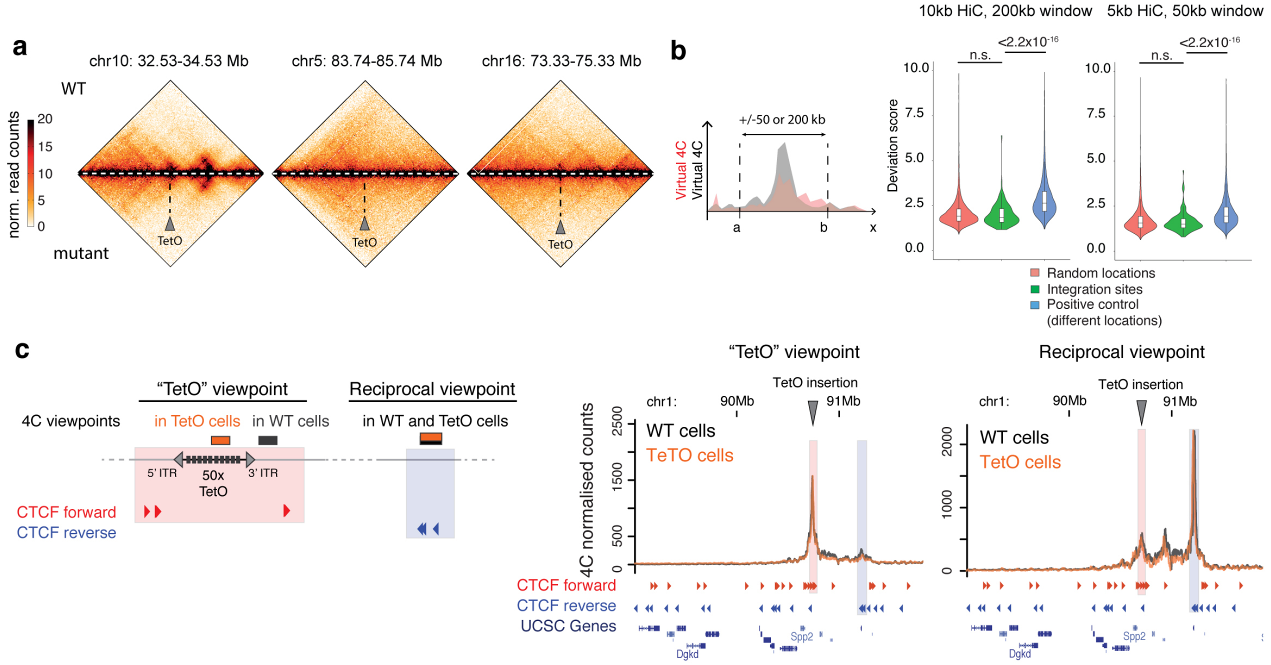
(Chi-Square Test: pvalue < 10^{-9}). ‘Inert’ corresponds to chromatin that is not enriched in H3K9me3, H3K27m3, H3K36me3, H3K9ac, nor H3K27ac. See the Methods section for more details. Right: The distributions of deviation scores in 20-fragment intervals where the dominant ChromHMM state is either inert, repressive, polycomb-associated or active, showing that active chromatin tends to show higher local dissimilarity between 4C and DamC (p-values from Wilcoxon test, two-sided). Cf. panel f for an example of a deviation score of ~3, corresponding to the average dissimilarity at active chromatin regions. **g)** Left: correlation between DamC signal in the -Dox sample and DNase-seq in mESC from ENCODE datasets. Each point in the scatter plot represents the aggregated signal in 20 kb; all 20kb intervals genome-wide are shown along with their Spearman correlation. Right: One representative megabase on Chr1 showing the high correlation between the two signals. DamC and DNase-seq data were normalized to have equal average signal over the genomic interval shown here. **h)** Left: Removing DNase hypersensitive GATCs (see Methods) does not lead to increased local similarity between DamC and 4C. Distributions of local deviation scores are calculated over all 130 valid profiles and deviation scores between two DamC biological replicates is shown for comparison (p-values from Wilcoxon test, one-sided).



Supplementary Figure 4

Additional DamC and 4C profiles from TetO viewpoints.

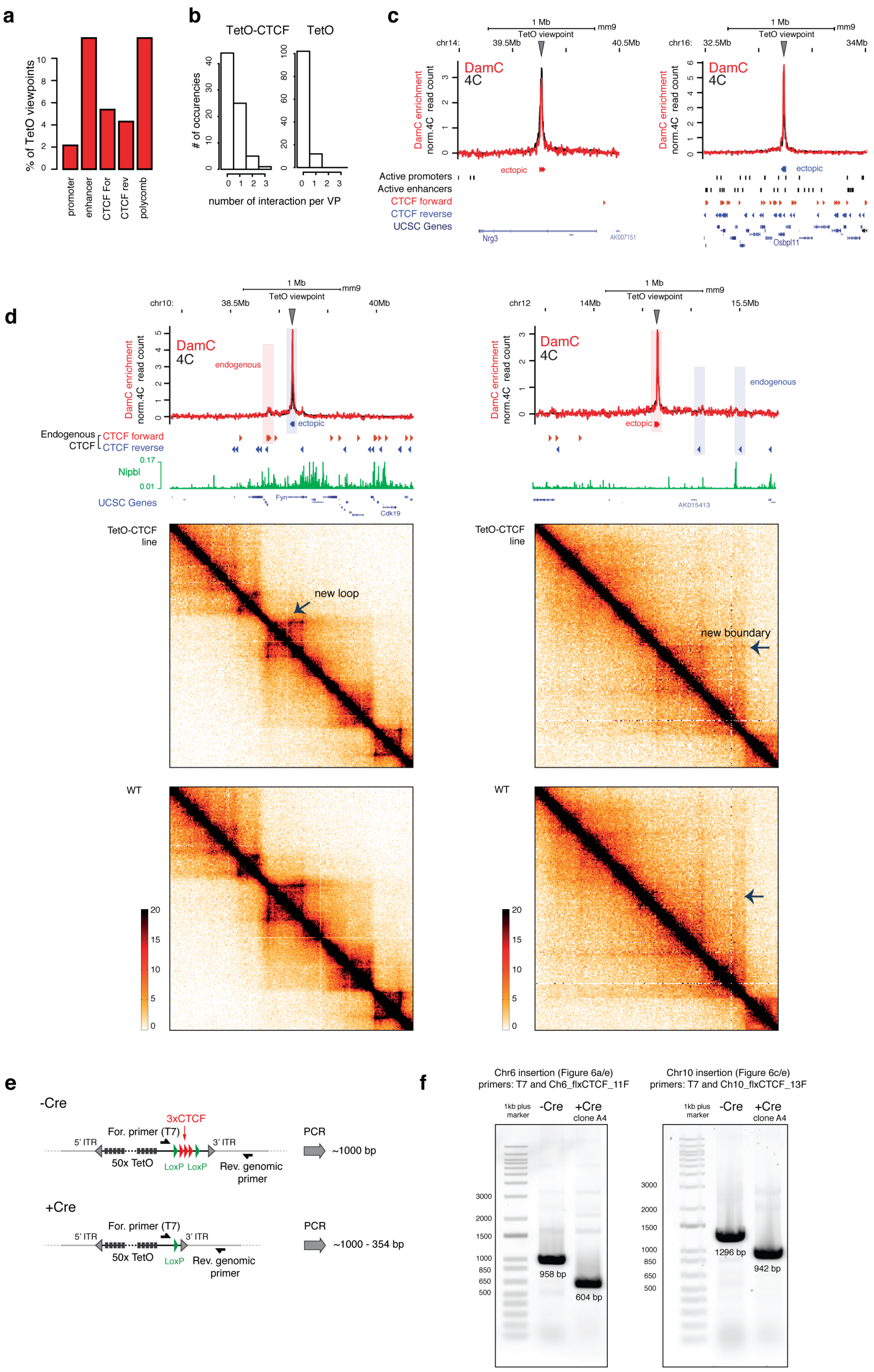
DamC (red) and 4C (black) profiles from forty TetO viewpoints in the pure clone with 135 TetO insertions.



Supplementary Figure 5

TetO-piggyBac insertions do not perturb chromosome structure.

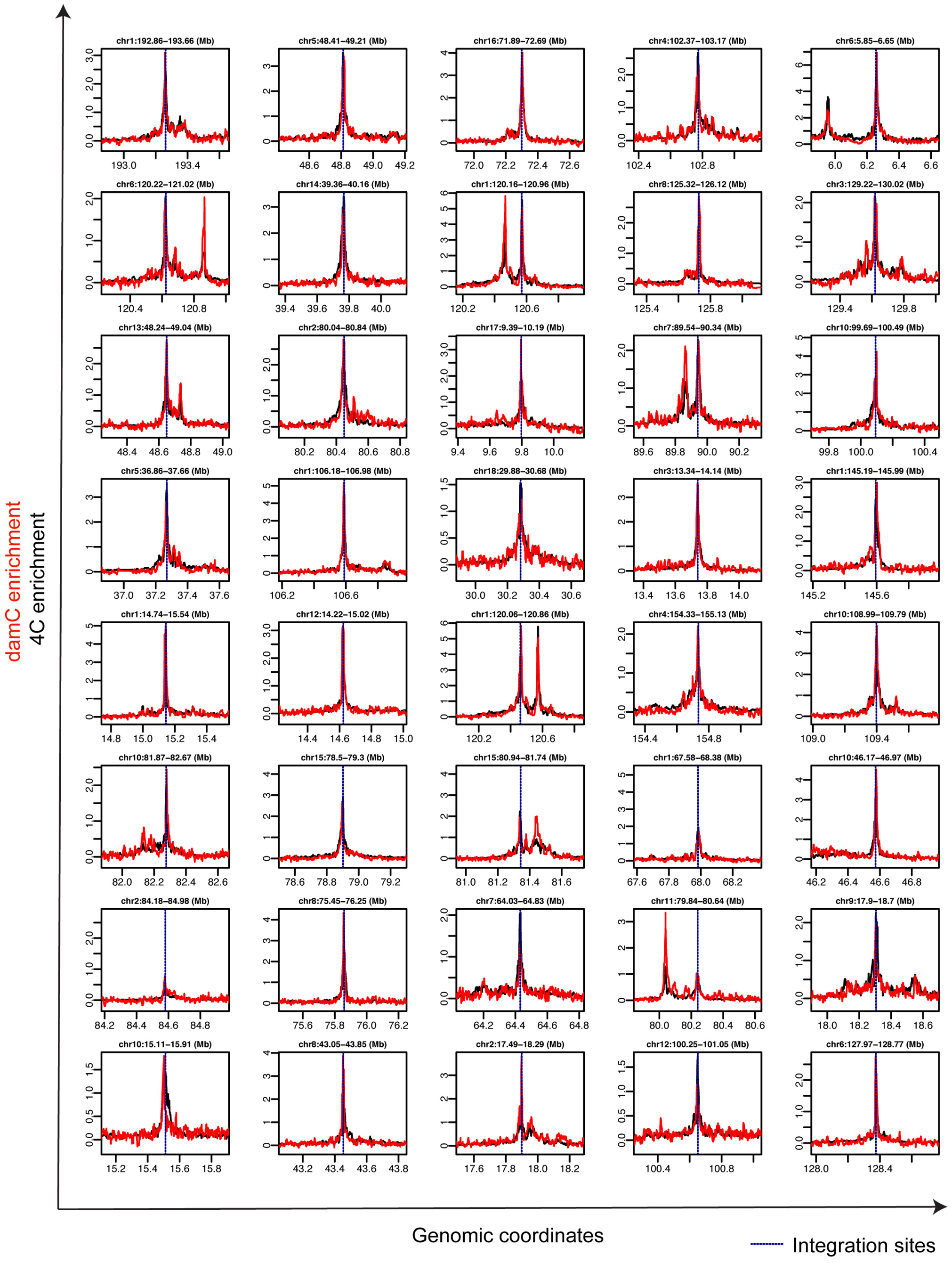
a) Insertion of TetO arrays does not perturb genome structure. Hi-C heatmaps of three different genomic locations harboring an array of 50xTetO sites and the corresponding wild-type locus are shown. Hi-C data are binned at 10 kb resolution. **b)** In windows of +/- 50 or +/- 200kb surrounding the TetO integration sites, no significant changes can be detected in Hi-C at 5 and 10 kb resolution, respectively. Indeed, deviation scores between wild-type and TetO cells obtained at TetO insertion sites (green violin plot) are similar to those obtained at random wild-type genomic viewpoints (pink violin plot), and significantly smaller than those obtained by comparing virtual 4C profiles from pairs of *different* random genomic viewpoints (blue) (p-values are from Wilcoxon test, one-sided). **c)** Left: scheme of viewpoints used for the 4C experiment shown on the right. In cells harboring the TetO insertions, the 'forward' 4C viewpoint is within the TetO array as in main Figure 3; in wild-type cells, the viewpoint is adjacent to the insertion genomic coordinate. The reciprocal viewpoint is the same in the two cases. Right: 4C profiles at the locus shown in panel c using the viewpoints shown on the left are indistinguishable.



Supplementary Figure 6

Analysis of TetO-CTCF insertions.

a) Percentage of TetO-CTCF viewpoints occurring in close proximity (<1 kb) from an active promoter or enhancer, or a CTCF site that is bound in ChIP-seq (Nora et al. Cell 169, 930-944.e22, 2017). **b)** Distribution of peaks detected by peakC per viewpoint in TetO-CTCF (left) and TetO line (right) **c)** Examples of interaction profiles from TetO-CTCF viewpoints occurring in regions that are either devoid of (left) or densely bound by CTCF (right). **d)** Two further examples of ectopic structures formed as a consequence of the insertion of TetO-CTCF viewpoints. Hi-C data are binned at 10 kb resolution. **e)** Scheme of Cre-mediated excision of the ectopic CTCF cassette and genotyping. **f)** Genotyping PCR showing Cre-mediated excision of the CTCF cassette from the two integration sites shown in Figure 6 in the same mESC clone (A4).



Supplementary Figure 7

Additional DamC and 4C profiles from TetO-CTCF viewpoints.

DamC (red) and 4C (black) profiles from forty TetO-CTCF viewpoints in the pure clone with 91 TetO insertions.

Supplemental Note 1:
Model description

1. Biophysical modelling of methylation dynamics in DamC

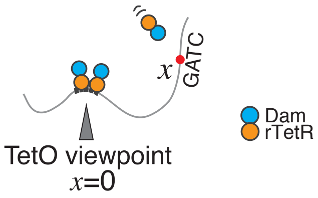


Figure M1.1. Scheme of the DamC physical model. The GATC located at x is methylated by freely diffusing and TetO-bound rTet-Dam.

The aim of the model is to describe rTetR-Dam mediated adenine methylation kinetics at an arbitrary GATC site located at genomic coordinate x. We suppose that the locus interacts in some cells with a TetO array located at the origin of genomic coordinates ($x = 0$), where rTetR-Dam is recruited in the presence of doxycycline (Dox) (**Figure M1.1**).

We consider that the methylation level at locus x at a certain time t depends on five factors:

1) The **concentration of freely diffusing rTetR-Dam**, which we indicate simply with $[Dam]$. This determines the **amount of rTetR-Dam that is bound non-specifically** at site x, which we indicate with F^{ns} and drives background methylation at locus x. Please note that 'non-specific' in this context means that binding does not occur through the rTetR-TetO affinity but rather via non-specific interactions of rTetR <u>plus</u> the intrinsic affinity of Dam for the GATC motif.
2) The **fraction of sites within the TetO array that is bound by rTetR-Dam** at a given concentration, F^{TetO}.
3) The **contact probability** between the TetO and x, $p(x)$. This is defined as the fraction of cells in the cell population where x and the TetO are in molecular proximity, i.e. at a distance small enough (≲10 nm) that Dam can bind and methylate the GATC site at x;
4) **Local biases** dependent on the genomic context (e.g. chromatin accessibility) that modulate the intrinsic methylation rate by a factor $a(x)$.
5) Adenine **methylation and demethylation rates**.

<u>Determination of the GATC demethylation rate</u>

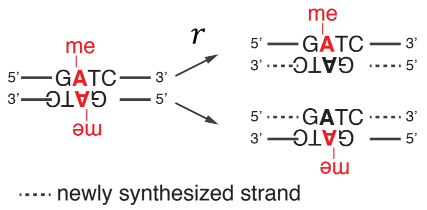


Figure M1.2. Effective demethylation rate in DamC. Methylated GATC sites are replicated into two hemi-methylated motifs, which are cut at very low frequency by DpnI compared to fully methylated GATCs. Hemi-methylated sites are thus essentially not detected in a DamC library and the effective demethylation rate equals the cell division rate r.

Note that due to the absence of an endogenous adenine demethylase, demethylation only occurs through dilution of GmATCs following DNA replication. All DamC experiments described in the manuscript are performed in cycling, unsynchronized cells. To determine the rate at which adenine methylation is diluted through DNA replication, let us assume that cells duplicate with rate r. Every GATC motif that has become methylated at some point during the cell cycle, either because of freely diffusing or TetO-bound Dam, will be replicated in the following S phase. After DNA replication, the adenine at the same genomic position will be hemi-methylated on the two resulting sister chromatids (**Figure M1.2**). Since DpnI cuts hemi-methylated GATC motifs at ~60x lower rate than fully methylated GATCs [1], hemi-methylated GATCs are virtually absent from DamC libraries. Thus the number of methylated adenines that are <u>detected</u> in DamC effectively decreases in time with rate r, irrespectively of the phase of the cell cycle when every adenine has been methylated.

<u>A system of ordinary differential equations describing GATC methylation dynamics</u>

Let us consider the case where cells are treated with doxycycline (+Dox), and the TetO array is bound by rTetR-Dam. We will indicate the methylation rate at locus x with $g_s(x)$ (where s stands for 'sample' as opposed to 'background' methylation in the absence of Dox). The form of $g_s(x)$ will be discussed later. For the moment, it is only important to keep in mind that it contains the contributions of both TetO-bound and freely diffusing Dam molecules. Let $N_0(x,t)$ and $N_1(x,t)$ be the number of GATC motifs at locus x, across the cell population, that are respectively unmethylated and fully methylated at time $t.$ Given that methylation in each cell is independent on the other cells, the dynamics of the population-averaged methylation states can be written in terms of ordinary differentially equations as follows:

$$\begin{cases} \frac{\partial N_1(x,t)}{\partial t} = -rN_1(x,t) + g_s(x)\, N_0(x,t) \\ \frac{\partial N_0(x,t)}{\partial t} = 2rN_1(x,t) + rN_0(x,t) - g_s(x)\, N_0(x,t) \end{cases} \tag{1}$$

where the factor $2rN_1(x,t)$ in the second equation accounts for the fact that every fully methylated GATC generates two hemi-methylated ones (see **Figure M1.2**), whereas each unmethylated GATC replicates into two unmethylated sites with an effective increment of one ($(rN_0(x,t))$). By diagonalizing the matrix of rates, and using the initial conditions $N_1(x,0) = 0$ and $N_0(x,0) = 1$ (i.e. no adenine is methylated before Dox is added to the culture medium), one finds

$$N_1(x,t) = g_s(x)\frac{\left(e^{rt} - e^{-(r+g_s)t}\right)}{2r+g_s(x)}. \tag{2}$$

Since the total number of cells $N(t) = N_0(t) + N_1(t)$ increases as $N(t) = exp[rt]$, the <u>fraction</u> of methylated cells $N_1(x,t)/N(x,t)$ can be written as

$$\frac{N_1(x,t)}{N(x,t)} = \frac{g_s(x)}{2r+g_s(x)}\left(1 - e^{-(2r+g_s)t}\right). \tag{3}$$

This quantity is proportional to the experimental output $S(x,t)$ in a DamC experiment in the presence of Dox through a multiplicative constant $W,$ which converts methylation probabilities into read counts:

$$S(x,t) = W\frac{g_s(x)}{2r+g_s(x)}\left(1 - e^{-(2r+g_s)t}\right) \tag{4}$$

Let us now consider the -Dox condition, where Dam is *not* bound to the TetO array. Methylation dynamics at x can be similarly written in terms of the DamC read counts as

$$B(x,t) = W\frac{g_b(x)}{2r+g_b(x)}\left(1 - e^{-(2r+g_b)t}\right) \tag{5}$$

where $g_b(x)$ stands for the "background" methylation rate at x in the absence of Dox, which only arises from freely diffusing rTetR-Dam.

<u>Expressions of the methylation rates in the presence and absence of Dox</u>

In order proceed it is now important to write the expressions of the methylation rates $g_s(x)$ and $g_b(x)$. In absence of Dox, the methylation rate is simply

$$g_b(x) = a(x) \cdot k \cdot F^{ns} \tag{6}$$

where k is the intrinsic Dam methylation rate, $a(x)$ is a multiplicative factor describing local biases such as chromatin accessibility, and F^{ns} is the fractional occupancy of rTetR-Dam on site $x,$ i.e. the fraction of cells in the population where rTetR-Dam is non-specifically bound at location $x.$ The latter depends on the genome-wide average of affinity constants of rTetR-Dam for non-TetO sites:

$$F^{ns} = \frac{[Dam]}{[Dam]+K_d^{ns}}. \tag{7}$$

In the presence of Dox, the methylation rate can instead be written as the sum of two terms: one accounting for the freely diffusing component, which only occurs if the GATC is *not* in contact with the TetO viewpoint, and the other describing methylation through rTetR-Dam that bound to the TetO array, which takes place only when x is in contact with the viewpoint:

$$g_s(x) = a(x) \cdot k \cdot \Big(F^{ns} \cdot \big(1 - p(x,t)\big) + F^{TetO} p(x,t) \Big)$$

where $p(x)$ is the contact probability between x and the TetO. This can be rewritten in the following form:

$$g_s(x) = g_b(x)\,[1 + Y\,p(x)], \tag{8}$$

where we have now defined $Y \equiv \frac{F^{TetO} - F^{ns}}{F^{ns}}$. In this expression $F^{TetO} = \frac{[Dam]}{[Dam] + K_d^{TetO}}$ measures the fractional occupancy of the TetO array, with K_d^{TetO} being the dissociation constant of rTetR-Dam for the TetO array.

Note that $K_d^{TetO} \ll K_d^{ns}$ since nuclear foci corresponding to TetO arrays where rTetR-Dam-EGFP is recruited in the presence of Dox can be clearly identified under the microscope, see Figure 3c in the main text.

The DamC enrichment is proportional to contact probabilities with the viewpoint

Equations (4), (5) and (8) can now be used to extract the contact probability $p(x)$ as a function of $S(x,t)$ and $B(x,t)$, both of which can be measured directly in DamC experiments. This result can be derived analytically in several asymptotic cases corresponding to well-defined kinetic regimes.

Case 1: $t \gg r^{-1}$, i.e. the duration of the experiment is much larger than the cell cycle duration. In this case, one easily obtains

$$p(x) = \frac{W}{Y} \frac{S(x) - B(x)}{B(x)(W - S(x))}\,. \tag{9}$$

Case 2: $r \gg g_b(x)^{-1}$ and $t \gg g_s(x)^{-1}$, i.e. methylation kinetics by TetO-bound Dam is much slower than the cell division rate, and the duration of the experiment is larger than the time it takes for freely diffusing Dam to methylate non-specific sites. Then

$$S(x,T) \approx W \frac{g_s(x)}{2r + g_s(x)}$$

$$B(x,T) \approx W \frac{g_b(x)}{2r} (1 - e^{-2rT}), \tag{10}$$

and thus $2rB(x)/\,W(1 - e^{-2rT}) \simeq g_b(x)$. One therefore obtains:

$$p(x) = \frac{W\alpha(T)}{YB(x)} \frac{S(x) - B(x)(W - S(x))/(W\alpha(T))}{(W - S(x))} \tag{11}$$

with $\alpha(T) = (1 - e^{-2rT})$.

Case 3: $r \ll g_b(x)\; and\; r \ll g_s(x)$, i.e. methylation rates are much faster than cell cycle duration, and $t \approx r^{-1}$ i.e. the experimental duration is comparable to the cell cycle duration. Equations 4 and 5 become

$$S(x,T) \approx W \frac{g_s(x)}{2r + g_s(x)} \text{ and } B(x,T) \approx W \frac{g_b(x)}{2r + g_b(x)} \tag{12}$$

which is equivalent to case 1:

$$p(x) = \frac{W}{Y} \frac{S(x) - B(x)}{B(x)(W - S(x))} \,. \tag{13}$$

Case 4: $r \gg g_b(x)\ and\ r \gg g_s(x)$, i.e. the inverse of the previous case. Equations 5 and 6 become:

$$\begin{aligned} S(x,T) &\approx W \frac{g_s(x)}{2r}\left(1 - e^{-2rT}\right) \\ B(x,T) &\approx W \frac{g_b(x)}{2r}\left(1 - e^{-2rT}\right) \end{aligned} \tag{14}$$

and thus

$$p(x) = \frac{1}{Y} \frac{S(x) - B(x)}{B(x)}. \tag{15}$$

Remarkably, if the GATC at site x is methylated in a small fraction of the cell population (in other words if $S(x) \ll W$), **the expressions for cases 1, 3 and 4 reduce to the same form**:

$$p(x) = \frac{1}{Y} \frac{S(x) - B(x)}{B(x)} \tag{16}$$

whereas for case 2,

$$p(x) = \frac{\alpha(T)}{Y} \frac{S(x) - B(x)/\alpha(T)}{B(x)}. \tag{17}$$

Note that our experimental setup satisfies the condition $S(x) \ll W$. W indeed represents the number of counts that we would observe if the adenine at location x was methylated in 100% of cells. We can safely assume $W \gg S(x,t)$ because at the vast majority of GATC sites surrounding the TetO viewpoints we observe up to ~3-fold increase in methylation in +Dox conditions compared to uninduced samples (see main Figure 3d), i.e. $S(x,t)$ is at least 3 times smaller than W.

Based on Ref. [1], it is known that Dam methylates genomic sequences over time scales of less than 5 hours, which is much faster than the duration of one cell cycle in mouse ESCs (~16 hours). Our overnight experiments last for ~18 hours, which is similar to the cell cycle duration; hence **our experiments are performed in the regime described in Case 3** and equation (16) can be rewritten in the form

$$\frac{S(x) - B(x)}{B(x)} = a \cdot p(x) \tag{18}$$

with $a = Y$. Thus, the relative difference between methylation in the presence and absence of Dox is directly proportional to the contact probability with the TetO array. Note that this quantity **does not depend on time**.

Generalization: depletion of freely diffusing rTetR-Dam in +Dox conditions

When doxycycline is added in the culture medium, rTetR-Dam binds with high specificity to the TetO arrays, resulting in a depletion of freely diffusing rTetR-Dam. In this case Equation 8 becomes

$$g_s(x) = g_b(x) \cdot \frac{F_+^{ns}}{F^{ns}} \cdot [1 + Y^+\, p(x)] = g_b(x) \cdot \sigma \cdot [1 + Y^+\, p(x)] \tag{19}$$

with now $Y^+ = \frac{F^{TetO} - F_+^{ns}}{F_+^{ns}}$, $\sigma = \frac{F_+^{ns}}{F^{ns}}$ and $F_+^{ns} = \frac{[Dam]^+}{[Dam]^+ + K_d^{ns}}$, where $[Dam]^+$ is the concentration of freely diffusing rTetR-Dam in the presence of Dox. Analytic solutions to cases 1-4 above can be derived also in this more general case. In particular for **case 3**, which is relevant to our experiments, one gets

$$p(x) = \frac{W}{\sigma Y^{+}} \frac{S(x) - B(x)}{B(x)(W - S(x))} - \frac{\sigma - 1}{\sigma Y^{+}} \tag{20}$$

which in the limit $W \gg S(x)$ becomes

$$p(x) = \frac{1}{\sigma Y^{+}} \frac{S(x) - B(x)}{B(x)} - \frac{\sigma - 1}{\sigma Y^{+}}. \tag{21}$$

This can be rewritten in the form

$$\boldsymbol{\frac{S(x) - B(x)}{B(x)} = a + b \cdot p(x)} \tag{22}$$

with $a = \sigma Y^{+}$ and $b = \frac{\sigma Y^{+}}{\sigma - 1}$. This is the equation reported in main Figure 2c in the main text and used to fit the experimental data in main Figure 3d (see Methods section in the manuscript for fitting details).

Expression of $[Dam]^{+}$

To estimate the concentration of freely diffusing rTetR-Dam in the presence of depletion due to TetO-bound molecules in _Dox conditions, we used a simple thermodynamic model describing the binding-unbinding of ligand to receptors. Let us assume to have n indistinguishable rTetR-Dam molecules and m independent TetO binding sites in a volume V (**Figure M1.3**). Let's call x the number of rTetR-Dam molecules bound to TetO sites. The probability of having x molecules bound is given by:

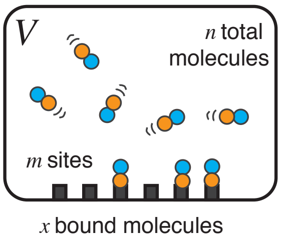


Figure M1.3. Scheme for the calculation of the concentration of freely diffusing rTetR-Dam in the presence of Dox.

$$p(x) = \frac{\Omega(x) \cdot e^{-\beta x \varepsilon}}{Z} \tag{23}$$

Where $\Omega(x)$ represents the number of ways x out of n molecules can bind to m sites (entropy) and it is given by:

$$\Omega(x) = \binom{m}{x} \cdot \frac{V^{n-x}}{(n-x)!},$$

ε represents the binding energy and Z is the partition function that is given by:

$$Z = \sum_y \Omega(y) \cdot e^{-\beta y \varepsilon} = \sum_y Exp[y \cdot \log([Dam]) - \beta y \varepsilon - y \cdot \log(y) - (m - y) \cdot \log(m - y) + m] \tag{24}$$

Where we used the following approximations:

$$(n - x)! \approx \frac{n!}{n^x}$$

and

$$\log(n!) \approx n \cdot \log(n) - n$$

The average number of bound rTetR-Dam is given by:

$$< x > = -\frac{\partial}{\partial(\beta\varepsilon)} \log(Z) = \begin{cases} \frac{[Dam]}{[Dam] + K_d} m & if\ 1 + \frac{[Dam]}{K_d} < \frac{m}{\min(n,m)} \\ \min(n, m) & if\ 1 + \frac{[Dam]}{K_d} > \frac{m}{\min(n,m)} \end{cases} \tag{25}$$

Where we used the saddle point approximation to estimate Z with respect to y. Thus,

$$[Dam]^{+} = [Dam] - < x > = [Dam] - \frac{[Dam]}{[Dam]+K_d} m. \tag{26}$$

2. Polymer model describing the scaling of contact probabilities in DamC

We reasoned that the leveling-off of contact probabilities at short genomic distances (≲10 kb) measured in DamC (Figure 4d in the main text) could have at least two alternative explanations: 1) TetO viewpoints have a finite genomic size (~2.7kb); and 2) chromosomes show local correlations (binding/rotational stiffness) below a certain genomic distance. We tested whether any of the two hypotheses (or both) would reproduce the observed scaling behavior, and found that only the second does as detailed below.

Hypothesis 1: The finite size of TetO viewpoints determines the scaling behavior

Assuming that the TetO viewpoint extends over a distance L (from -L to 0), the contact probability between any genomic location *x* and the viewpoint can be written as

$$p(x) = \frac{1}{L} \int_{-L}^{0} p_0(x - y) dy, \tag{1}$$

where $p_0(l)$ is the interaction probability between two sites separate by a genomic distance $l = |x - y|$. At large genomic distances ($l \gg L$) the TetO viewpoint can be approximated as being infinitesimally small, so we expect $p(x) \approx p_0(x)$. The DamC and 4C data show that at large genomic distances (in the limit of large *x*), contact probabilities decay as

$$p(x) \sim \frac{1}{x^{\beta}} \tag{2}$$

with β =0.79. Assuming that the same functional form applies at all genomic distances, i.e.

$$p_0(x) = \frac{1}{x^{\beta}} \tag{3}$$

and plugging Eq. (3) into Eq. (1) we obtain

$$p(x) = \frac{(x)^{-\beta+1} - (L + x)^{-\beta+1}}{L(1-\beta)}. \tag{4}$$

Eq. (4) correctly scales as Eq. (2) in the limit of large *x* and is flatter at small *x* (see Fig. 1), decaying as $1/x^{1-\beta}$ in this regime. Eq. (4) however does not satisfy the experimental data as can be seen in **Figure M2.1**.

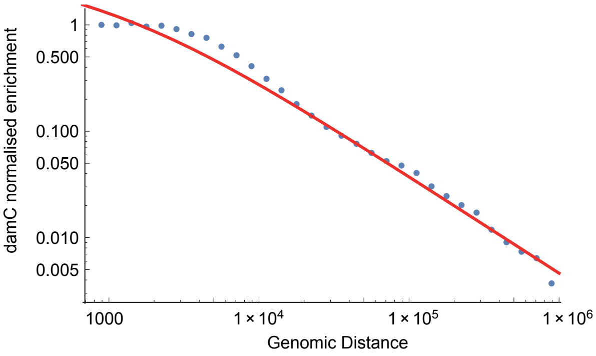


Figure M2.1 Black dots: Relative interaction frequency (normalized DamC enrichment) as in main Figure 3e-d. Red line: Fit to the data using Eq. (4).

Hypothesis 2: Short-range effects determine the scaling behavior

We then reasoned that the bending of the scaling behavior at short genomic scales could be due to local effects related to the stiffness of the polymer (e.g. bending stiffness, rotational stiffness, presence of obstacles bound to the fibre). A few models model have been derived to describe looping probabilities across stiff homopolymers [1,2]. However, they cannot be applied to our case because 1) none of them displays the experimentally observed power law at large *x*; 2) they present an analytical solution only in the small- and large-*x* limit; 3) they are based on specific hypotheses on bending energies that are difficult to relate to the actual biology of chromosomes and to test *in vivo*.

To describe proximity effects, we therefore decided to use the simplest, most general model we can derive on the basis of scaling arguments. The properties this model should have are:
1) The contribution of local effects should disappear in the limit of large *x*;
2) It should combine the probabilities of multiple points in a multiplicative way; i.e., if it affects the looping probability between points A and B separated by a genomic distance *n* by a factor *g(n)*, and points B and C separated by a distance *m* as *g(m),* then its overall effect should scale as $g(n)g(m)\sim g(n+m)$.

Requirement 2) suggests that the correction is an exponential function of the genomic distance; requirement 1) that it has the form $1-\exp\left[-x/x_0\right]$, where x_0 is a characteristic length that delimit all possible proximity effects. We will thus suppose that the contact probability at all scales is described by the function

$$p_0(x)=\frac{1-\exp\left[-x/x_0\right]}{x^{\beta}}. \tag{5}$$

Using Eq. (1) we obtain

$$p(x)=\frac{(x)^{-\beta+1}-\ (L+x)^{-\beta+1}}{L(1-b)}+\frac{x_0^{1-\beta}}{L}\left[\Gamma\left(1-\beta,\frac{x+L}{x_0}\right)-\Gamma\left(1-\beta,\frac{x}{x_0}\right)\right], \tag{6}$$

where Γ is the incomplete gamma function. Equation (6), using *L*=3 kb and β =0.78 and fitting to the experimental data using only x_0 as a fitting parameter reproduces very well the DamC-based experimental data as shown in **Figure M2.2,** returning a best estimate of ~2.5 kb for x_0.

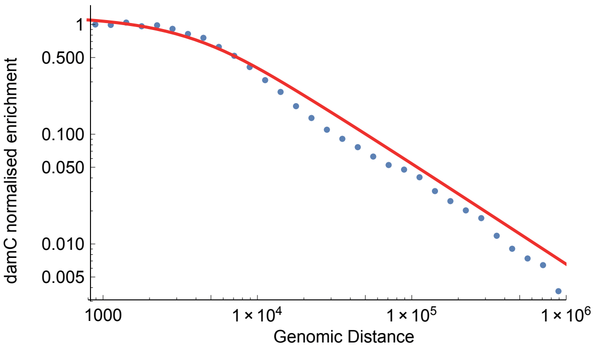


Figure M2.2 Black dots: Relative interaction frequency (normalized DamC enrichment) as in main Figure 3e-d. Red line: Fit to the data using Eq. (6). Fit done using NonlinearModelFit in Mathematica.

References

1. Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, et al. Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. Cell. 2013;153:178–92.

2. Shimada J, Yamakawa H. Ring-closure probabilities for twisted wormlike chains. Application to DNA. Macromolecules. 1984;17:689–98.

3. Rosa A, Becker NB, Everaers R. Looping Probabilities in Model Interphase Chromosomes. Biophys J. 2010;98:2410–9.

Chapter III: Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers

Y. Zhan, L. Giorgetti, G. Tiana

I performed and analyzed the simulations.

Summary

The development of 3C methods boosted our understanding of chromatin folding, making it possible to measure the frequency of pairwise interactions in a genome-wide manner. Building on this, several physical models have been developed to elucidate the mechanisms that drive chromatin folding. In these models, the scaling of contact probability has been used to discriminate alternative hypotheses. However, if equilibrium homopolymers can be univocally identified with the scaling, this is not the case for heteropolymers where interactions are heterogeneous, such as chromosomes. In this study, we showed that finite size, together with randomness in interactions between monomers, can reproduce the range of scaling values detected in Hi-C, suggesting that caution is needed in using the scaling to discriminate alternative physical models.

PHYSICAL REVIEW E **94**, 032402 (2016)

Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers

Y. Zhan and L. Giorgetti[*]
Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

G. Tiana[†]
Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN,
via Celoria 16, 20133 Milano, Italy

(Received 18 April 2016; published 2 September 2016)

Random heteropolymers are a minimal description of biopolymers and can provide a theoretical framework to the investigate the formation of loops in biophysical experiments. The looping probability as a function of polymer length was observed to display in some biopolymers, like chromosomes in cell nuclei or long RNA chains, anomalous scaling exponents. Combining a two-state model with self-adjusting simulated-tempering calculations, we calculate numerically the looping properties of several realizations of the random interactions within the chain. We find a continuous set of exponents upon varying the temperature, which arises from finite-size effects and is amplified by the disorder of the interactions. We suggest that this could provide a simple explanation for the anomalous scaling exponents found in experiments. In addition, our results have important implications notably for the study of chromosome folding as they show that scaling exponents cannot be the sole criteria for testing hypothesis-driven models of chromosome architecture.

DOI: 10.1103/PhysRevE.94.032402

I. INTRODUCTION

Most biological molecules are polymers, and the formation of contacts between monomers which are not close along the chain often plays an important biological role. For example, in the nucleus of mammalian cells, the encounter of an enhancer and a gene promoter that can be millions of base pairs away along the chromatin fiber is often necessary for the expression of the gene [1]. In the case of proteins, the formation of noncovalent interactions between distant amino acids is, in many cases, among the first steps in the folding process [2].

There are several experimental techniques to study, either directly or indirectly, the formation of contacts between pairs of monomers as a function of their distance N along the polymeric chain. Arguably, when N is large enough, the detailed chemistry of the system loses importance and one can highlight its more general physical properties. The looping probability of peptides with repeated AGQ sequence, measured by Förster resonance energy transfer, displays a power law with exponent 1.55 in water and 1.7 in urea and guanidine [3]. The folding rate of proteins, measured by stopped-flow experiments, was shown to correlate with the (rescaled) average value of N of pairs of amino acids which are in contact in the native state [4]. In long RNA chains the contact probability displays an exponent $\beta \approx 1$ [5]. In the case of chromosome folding, a class of biochemical techniques collectively known as chromosome conformation capture (3C) makes it possible to measure contact probabilities along the chromatin fiber following chemical cross linking of nuclei [6]. In human and mouse chromosomes, these techniques revealed that the looping probability between chromosomal loci depends on N as a power law $N^{-\beta}$ with exponent $\beta \approx 1$ above the 10^6-base-pairs scale [7] and even lower at a smaller scale [8]. Importantly, these scaling exponents have been used to derive and test models regarding the mechanisms that could give rise to the peculiar folding patterns observed in the genome (see Sec. VIII below). It is therefore important to understand if anomalous scaling exponents necessarily arise from specific model-specific mechanisms or can rather emerge as general properties of biopolymers.

The simplest theoretical framework to describe the contact formation in a biopolymer at equilibrium as a function of N is that of two interacting monomers linked by a homopolymer. One can employ a two-state description of the system, assuming that the formation of the contact between the two ends does not change the density of the polymer. In this case, if $\epsilon < 0$ is the energy gain of the system upon formation of the contact, the associated probability can be approximated as

$$c(N) = \frac{\exp(-\epsilon/T)}{g(N) + \exp(-\epsilon/T)}, \tag{1}$$

where $g(N)$ is the density of state of the system displaying the contact with respect to the unbound state. Its shape depends on the properties of the linking homopolymer. If this can be regarded as an ideal chain, then $g(N) = N^{3/2}$; if it is a random coil due to the repulsion between its elements, $g(N) = N^{9/5}$, while it is constant in a globule [9]. In the limit of large N one then expects a scaling law of the type $c \propto N^{-\beta}$, with $\beta = 0, 1/2$, or $9/5$, as discussed above. The scaling exponents found for repeat peptides [3] lie between those expected for an ideal chain and a random coil. In the case of chromatin, the anomalous exponent $\beta \leqslant 1$ found in experiments is not compatible with the above model and several mechanisms have been evoked to explain this finding: nonequilibrium effects similar to what observed in the "crumpled globule" state [10,11], looping interactions mediated by soluble DNA-binding molecules [12] or energy-driven mechanisms

---

[*] luca.giorgetti@fmi.ch

[†] guido.tiana@unimi.it

©2016 American Physical Society

such as loop extrusion by DNA-bound protein complexes [8,13].

However, in most cases, the monomers which build polymers of biological interest are chemically heterogeneous, and the homopolymeric assumption is questionable. The problem we would like to address in the present work is the role of heterogeneous interactions in determining the scaling properties of the contact probability between monomers. Specifically, we study the looping probability of random heteropolymers [14], regarding them as a minimal model for biomolecules.

To investigate this problem, we use a simple model, in which the polymer is described as a chain of beads connected by rigid links. Pairs of beads interact through a spherical-well potential with a hard core of radius r_H, a width r, and a depth B_{ij} which depends on the specific pair. For the sake of generality, we considered the energies B_{ij} as quenched stochastic variables, defined by a Gaussian distribution. In this way we are not focusing on a particular kind of biopolymer, but we are looking for the general properties which arise only because of the heterogeneity of the interactions.

Operatively, we investigated the equilibrium contact probability of heteropolymeric chains by means of numerical simulations. The stochasticity of the interaction energies was modeled by generating several realizations of the set of Gaussian variables, and for each of them carrying out a conformational sampling. This approach poses the problem of averaging the results of the samplings over the quenched energies. The contact probability itself does not result to be a self-averaging quantity, and consequently its average over the realizations of the quenched variables B_{ij} is poorly informative [15]. In Sec. IV we discuss under which conditions the average of quantities associated with the contact probability are informative.

Another problem one has to face is that the conformational sampling of disordered systems is computationally cumbersome, due to the roughness of the associated energy landscape. There are several computational techniques based on the multicanonical ensemble which, sampling the system simultaneously at different temperatures, facilitate conformational sampling [16,17]. However, they rely on the choice of a set of temperatures that are optimized to enhance diffusion in the temperature space. This set is not self-averaging, and consequently requires a manual fine tuning for each realization of the quenched variables. This is impractical if one wants to collect results from enough replicas to calculate reliable averages. To solve this problem in an automatic way, we made use of an adaptive simulated-tempering scheme developed in Ref. [18].

In Sec. II we describe a consistent theoretical framework which is necessary to study quantitatively the looping probability in heteropolymers. This framework is applied to a simple model of random heteropolymers, described in Sec. III. In Sec. IV we analyze the main obstacle one finds in a naive derivation of the scaling properties of the looping probability in polymers with a disordered interaction. We then analyze in a consistent way the looping properties (in Sec. V) and the related compactness (in Sec. VI) of a set of heteropolymers as a function of their length. Then, in Sec. VII, we perform a similar analysis on the scaling properties describing the formation of loops in the different segments of a fixed-length polymer, a case which is relevant for recent experiments on chromosome systems [6–10,12,13,19], as described in Sec. VIII. We then discuss the consequences of the model in Sec. IX and draw some conclusions in Sec. X.

II. THE THEORETICAL FRAMEWORK

In order to find the most appropriate way of calculating the scaling properties of the looping probability of a random heteropolymer, one can use a two-state model. One can assume that the bound and unbound states display, respectively, energies $E_1+\epsilon$ and E_2, where E_1 and E_2 are quenched random variables regarded as the sum of the internal contact energies of the chain, while ϵ is the interaction energy between the ends of the chain. Further assuming that E_1 and E_2 are uncorrelated and that the two states have the same density, the central-limit theorem suggests that

$$p(E_1)=p(E_2)=\frac{1}{\sqrt{2\pi N\sigma^2}}\exp\left[-\frac{(E_{1,2}-N\epsilon_0)^2}{2N\sigma^2}\right], \tag{2}$$

where N is the length of the chain, ϵ_0 the average interaction between the monomers, and σ their standard deviation. We define $\Delta E \equiv E_1-E_2$ and assume a density of states of the unbound state with respect to the looped state in the form of a power law of the kind N^β. Thus, the entropy difference is $\beta\ln N$ and the free-energy difference between the two states is given by

$$\Delta F=\Delta E+\epsilon+T\beta\ln N, \tag{3}$$

where ΔE is a stochastic variable with distribution

$$p(\Delta E)=\frac{1}{\sqrt{4\pi N\sigma^2}}\exp\left(-\frac{\Delta E^2}{4N\sigma^2}\right). \tag{4}$$

According to this model, the variability of the looping free energy, and consequently of the looping probability, at a given value of N is due to the variability of the internal energy difference ΔE. In other words, ΔE plays the role of the quenched disorder affecting the looping free energy defined as a function of N. The associated probability can be obtained by inverting Eq. (3) and substituting it into Eq. (4), that is,

$$p(\Delta F)=\frac{1}{\sqrt{4\pi N\sigma^2}}\exp\left[-\frac{(\Delta F-T\beta\ln N-\epsilon)^2}{4N\sigma^2}\right]. \tag{5}$$

This probability can be maximized with respect to β and ϵ according to a maximum-likelihood principle, obtaining

$$\beta=-\frac{1}{T}\frac{\sum_N\frac{1}{N}\sum_N\frac{\ln(N)\Delta F}{N}-\sum_N\frac{\ln(N)}{N}\sum_N\frac{\Delta F}{N}}{\sum_N\frac{1}{N}\sum_N\frac{\ln^2(N)}{N}-\left[\sum_N\frac{\ln(N)}{N}\right]^2}, \tag{6}$$

formally identical to the expression of a weighted linear regression.

From the simulations (or from a set of experiments) one can calculate the free-energy difference ΔF from the contact probability

$$\Delta F=-T\ln\left[\frac{c}{1-c}\right] \tag{7}$$

and use Eq. (6) to obtain β from a linear regression of F versus $\ln N$ with weights N^{-1}. This weighting is a consequence of the extensivity of the energy of the chain and has as consequence that larger-N points contribute less to the determination of β.

III. THE COMPUTATIONAL MODEL AND ALGORITHM

In the present work heteropolymers are described as chains of beads linked by rigid links of length $a = 1$ (which sets the length scale of the system). Beads interact with a two-body potential $U = \sum_{i<j} u_{ij}$, where the two-body terms are defined as

$$u_{ij} = \begin{cases} +\infty & \text{if } |r_i - r_j| < r_H, \\ B_{ij} & \text{if } r_H \leqslant |r_i - r_j| < r, \\ 0 & \text{if } |r_i - r_j| \geqslant r. \end{cases} \tag{8}$$

The B_{ij} are quenched stochastic energies, distributed according to a Gaussian function with zero mean and standard deviation $\sigma_B = 1$ (which sets the energy scale of the system). In the calculations, we chose [20] $r_H = 0.6$ and $r = 1.5$ (in units of a). The equilibrium properties of random heteropolymers are studied generating 500 realizations of the random interactions B_{ij}, sampling the conformational space of each of them, and performing averages over the realizations of the random interactions as described in Sec. IV below.

Conformational samplings are carried out with an iterative simulated-tempering algorithm [18]. It is based on a Metropolis scheme in which elementary moves are flips of single beads and pivot moves (see Ref. [21], where the code used for the simulations is described). A simulated tempering is then applied in which the temperatures $\{T_i\}$ and the free-energy factors $\{g_i\}$ which define the simulated tempering [17] are adjusted during the simulation to optimize the diffusion of the temperature, from scratch in each realization of the interaction matrix. Specifically, the simulation starts with a plain Metropolis at high temperature $T_0 = 2$ (in units of σ, setting Boltzmann constant to 1). From the distribution of energies calculated from this sampling, the ideal values of T_1 and g_1 to have a temperature-exchange rate of 0.1 are estimated and a simulated tempering over these two temperatures is carried out. A weighted-histogram algorithm is then applied to obtain the distribution of energies from the energy distributions obtained so far, and a further pair T_2 and g_2 is added to the tempering. This procedure is iterated until the target temperature T is reached. A set of rules is also applied in the case where actual exchange rates depart from the predicted ones, as described in detail in Ref. [18]. An example of this procedure results in a sampling of different temperatures as that displayed in Fig. 1, which makes it possible to calculate equilibrium averages of polymers up to $\sim 10^2$ monomers.

IV. THE SELF-AVERAGING ISSUE

The average $\overline{x}$ of a conformational property x of the random heteropolymer over the quenched stochastic energies provides valuable information only if the associated standard error σ_x is small, namely if the quantity is self-averaging [15]. In the thermodynamic limit, this corresponds to the condition

$$\xi_x \equiv \frac{\sigma_x}{|\overline{x}|} \to 0. \tag{9}$$

Usually extensive properties are self-averaging [22], while intensive properties, probability distributions, and partition functions are not. Thus, we do not expect $c(N)$ to be self-averaging, and in fact ξ_c is quite large, increasing above 1 quite fast as a function of N at low temperatures [cf. Fig. 2(a)]. This is the reason why in the context of disordered systems one focuses the attention on free energies. However, in the present case we are considering a free-energy difference between two states of the system, which is expected to scale as $\ln N$ according to Eq. (3). The associated self-averaging parameter thus scales as $\chi_{\Delta F} \sim N^{1/2}/\ln N$, which has a nonmonotonic behavior as a function of N, eventually diverging in the thermodynamic limit, although not very fast [cf. Fig. 2(b)].

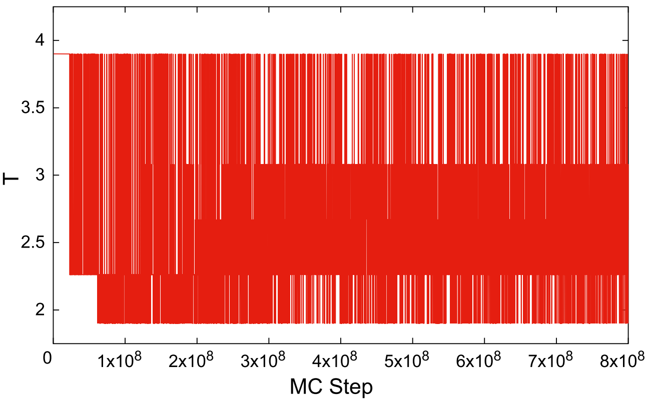


FIG. 1. An example of evolution of the temperatures in the self-adjusting simulated-tempering simulation.

Thus, strictly speaking, ΔF is not self-averaging. Nor it is any quantity which can be derived by the contact probability c. However, if one is interested in finite systems of the typical size of biopolymers, a sufficient request is that the variability of ΔF associated with the disorder is smaller than its average; that is, $\xi_{\Delta F} \ll 1$ in a specified interval of N.

Equation (5) suggests that the variability of ΔF over the quenched disorder should follow

$$\xi_{\Delta F} = \frac{2\sigma N^{1/2}}{|\epsilon + T\beta \ln N|} \tag{10}$$

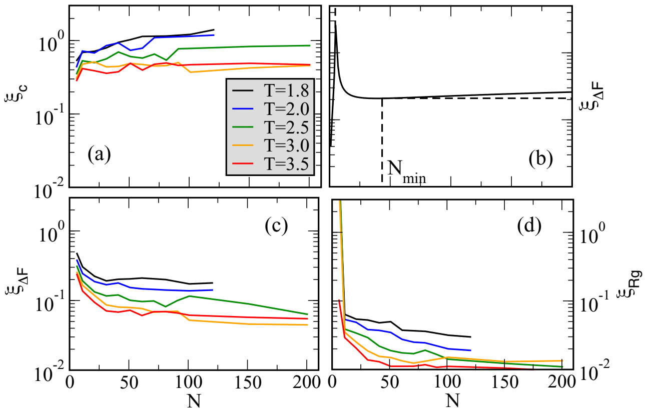


FIG. 2. (a) The relative error ξ_c associated with c; (b) a sketch of the theoretical behavior of $\xi_{\Delta F}$ according to Eq. (10); (c),(d) the relative error ξ calculated for ΔF and for the gyration radius R_g, respectively.

and consequently display a divergence at $N_{\rm div} = \exp[-\epsilon/T\beta]$ and a minimum at $N_{\rm min} = \exp[2 - \epsilon/T\beta]$, diverging at large N [cf. Fig. 2(b)]. Thus, we can expect ΔF to be representative of a typical realization of the disordered interactions if $N > N_{\rm div}$ and $N \sim N_{\rm min}$.

In Fig. 2(c) is plotted the value of $\xi_{\Delta F}$ at different temperatures as a function of the length N of the chain in semilog scale, calculated over 500 realizations of the random interactions. For each temperature we show the points up to the largest value of N for which we can guarantee the correct equilibration of the simulated-tempering algorithm. In the studied range of N, the calculated $\xi_{\Delta F}$ is decreasing, thus suggesting that $N_{\rm div} < N < N_{\rm min}$. Moreover, already for $N > 10$ the $\xi_{\Delta F}$ assumes small values, indicating that the standard error on ΔF is of the order of a few percent of the mean. That is, except for very short chains, the average of ΔF over the stochastic interactions are representative of their typical values. A similar behavior is observed for the gyration radius R_g of the polymer [see Fig. 2(d)].

V. SCALING OF THE FREE ENERGY ASSOCIATED WITH THE LOOPING PROBABILITY

From the same simulations used to estimate the degree of self-averageness, we calculated the values of $\overline{\Delta F}$ as a function of N, in order to estimate its scaling properties.

The linear fit of ΔF as a function of $\ln N$ is displayed in Fig. 3 for simulations carried out at different temperatures. The linear fit appears good at $T > 2.0$ and seems to worsen at lower temperatures. In particular, at $T \leqslant 2.0$ a power-law behavior applies up to $N \approx 60$, while $\overline{\Delta F}$ appears weakly dependent on N above ≈ 60, similarly to the behavior of a collapsed globule in a homopolymer.

Interpreting Eq. (5) as the likelihood of observing a value of ΔF in a chain of specified length, the quality of the linear fit can be expressed in terms of the average log-likelihood, that is nothing else but

$$\chi^2 = \frac{1}{Z_N}\sum_n^N \frac{(\overline{\Delta F}(n) - \epsilon - T\beta\ln n)^2}{n\sigma^2}, \tag{11}$$

where N is the length of the longest chain considered in the fit and $Z_N = \sum_n^N (n\sigma^2)^{-1}$. The values of χ^2 as a function of N are reported in the inset of Fig. 3. The fits of the points at $T > 2.0$ display a constant or decreasing χ^2 of the order of 10^{-2}, while at lower temperatures it increases with N. However, even at low temperatures the value of χ^2 remains lower than 1 for all the N studied, indicating that the fitting line matches the points within their error bars.

This is a result of the fact that both the estimation of β and the quantification χ^2 of the error of the fit emphasize smaller polymers because for them the variability of ΔF due to the disordered interactions is smaller. In the case of longer polymers, $\overline{\Delta F}$ seems to become independent on N, but at the same time it becomes less and less representative of a typical heteropolymer. In fact, even if $\overline{\Delta F}$ were constant at large N, the leading term of Eq. (11) would be $\chi^2 \sim N^{-1}\sum_n \ln^2 n/n$; approximating the sum with an integral gives $\chi^2 \sim \ln^3 N/N$, which vanishes at large N. In other words, it is the small-N slope that determines β, because at large N the free energy is dominated by the disorder. If the small-N scaling properties are due to finite-size effects, these will thus dominate the results even when considering longer chains.

The values of the parameter β obtained from the fits at different temperatures are reported as solid circles in Fig. 4. At high temperature ($T = 3.5$) the scaling exponent β converges to 2.06, which is comparable with the value 2.10 ± 0.15 obtained numerically for self-avoiding walks in three dimensions [23] and somewhat larger than the theoretical

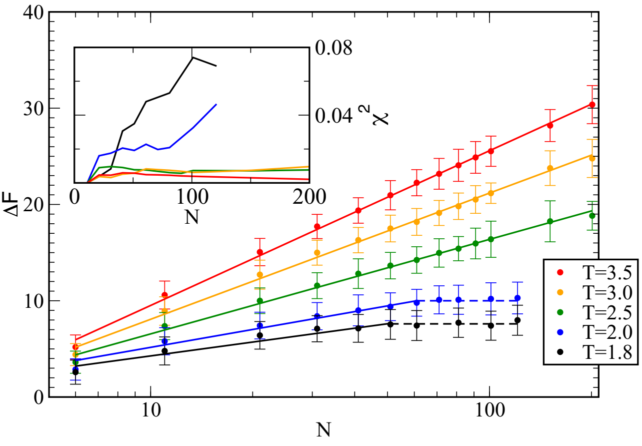


FIG. 3. The average value of ΔF as a function of N, the latter displayed in a logarithmic scale. For each value of N, 500 realizations of the disordered interaction are simulated. The points are fitted according to Eq. (6), and the corresponding line is drawn in the figure. (Inset) The χ^2 associated with the fits calculated up to length N.

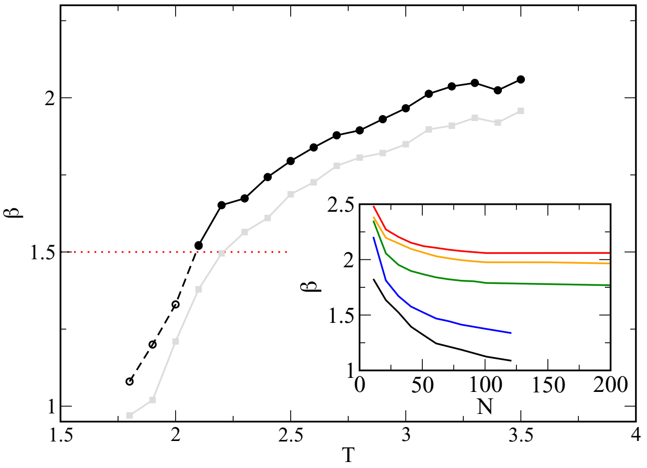


FIG. 4. The exponents β obtained using Eq. (6) at different temperatures from the fits of the simulated data up to the largest polymer we could equilibrate (circles). As a reference, the dotted curve indicates the exponent 3/2 expected for an ideal chain. Empty circles indicate the exponents below the θ point, strongly affected by finite-size effects. The gray squares indicate the exponents found in a fit of $\ln c$ versus $\ln N$. (Inset) The exponent calculated from fits up to length N.

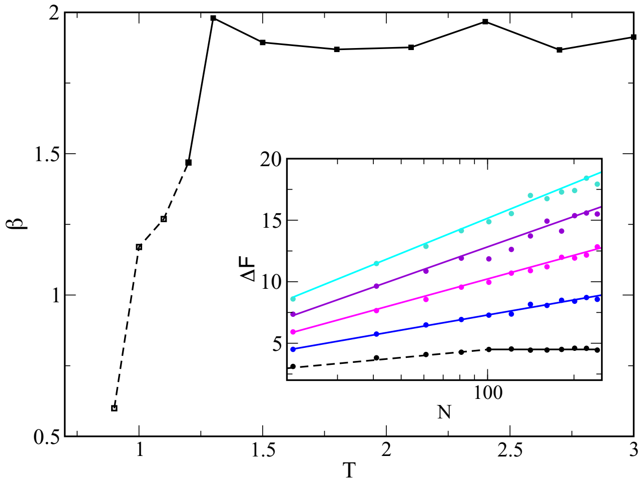


FIG. 5. The scaling exponent β calculated for a homopolymer (i.e., $\epsilon_0 = -0.1, \sigma = 0$) as a function of temperature T. Open symbols indicate the exponents associated with finite-size behavior (cf. dashed line in the inset). (Inset) The binding free energies whose fits were used to obtain the scaling exponents (the different sets correspond, starting from above, to $T = 2.1$, $T = 1.8$, $T = 1.5$, $T = 1.2$, and $T = 0.9$).

result 9/5 obtained by de Gennes solving a zero-dimensional Ising model [9].

As the temperature is decreased, β decreases continuously to the value $\beta = 3/2$ typical of the θ point at $T \approx 2.0$. This plot is markedly different from that of a homopolymer, in which case only two kinds of exponents are expected, associated with the coil state and the ideal behavior at the θ point. In fact, the exponents found from numerical simulations of homopolymers of comparable size are displayed in Fig. 5. Moreover, even a random heteropolymer in the coil or θ state in the limit of short interaction range is expected to display the same exponents of the homopolymer, superposed to an exponential cutoff [24].

Below the θ point the fit gives exponents $1 \lesssim \beta \lesssim 1.5$ (cf. empty circles in Fig. 4). Since the small-N contribution dominates due to the dependence on N of the denominator at the exponent of Eq. (5), the exponents β seem to converge to a N-independent value, different from zero, even below the θ point (cf. inset of Fig. 4).

The scaling of $\overline{\Delta F}$ below the θ point with exponents lower than 3/2 is a finite–size effect, also present in homopolymers (cf. Fig. 5). This is a consequence of the fact that if the polymer is too short, it is not able to define a bulk volume, necessary for the looping entropy to lose its dependence on N, but its volume essentially coincides with its surface. The order of magnitude of N below which this effect takes place is found by $4\pi R^2 2r_H = 4/3\pi R^3$, with $R = r_H N^{1/3}$ in a globule, that is, $N = 6^3 \approx 10^2$, in agreement with what shown in Fig. 5.

Often a simple regression of $\ln c$ versus $\ln N$ was applied to the analysis of the scaling properties of the contact probability [3] of biopolymers. This is more difficult to justify theoretically than the fit described in Sec. II. Anyway, the results of such a fit are displayed with gray squares in Fig. 4. The resulting exponents are slightly smaller than those obtained with the two-state model described above, but in this case the (unweighted) χ^2 of the fit ranges from 0.2 at high temperature to $\approx$1.8 at low temperature. At variance with the the weighted fit described above, in this case the χ^2 of the fit, as well as the value of the exponents, depend on the specific range of N employed in the simulations.

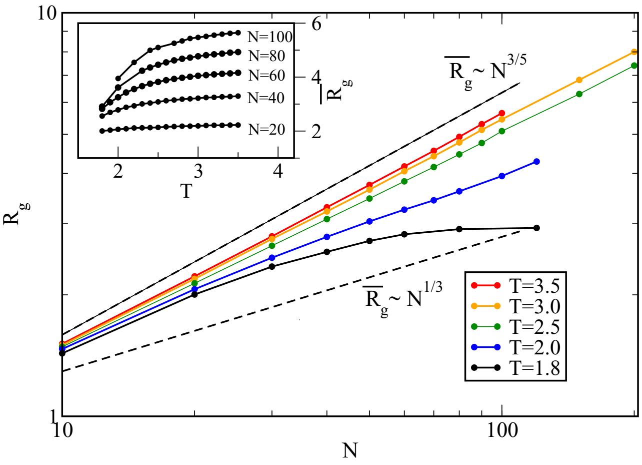


FIG. 6. The average gyration radius $\overline{R_g}$ at different temperatures as a function of the length of the chain plotted in log-log scale. As a reference, we indicate with dashed lines the $N^{3/5}$ curve expected for a random coil and the $N^{1/3}$ curve expected for a globule. (Inset) The value of R_g as a function of temperature for different lengths N.

VI. COMPACTNESS OF THE POLYMER

In order to compare the exponents β found for the random heteropolymer with those known from the theory of homopolymers, it is interesting to understand whether the polymer is, at the different temperatures studied above, in a globular or in a coil state. This problem is well-defined because the resulting thermal average R_g of the gyration radius is self-averaging (see Sec. IV), and consequently we can study its average $\overline{R_g}$ over the realizations of the disordered interaction. On the other hand, it is complicated by the small size of the system, while a globule-coil phase transition is defined, strictly speaking, only for an infinitely long polymer.

The average value of R_g as a function of N is displayed in log-log scale in Fig. 6 at different temperatures. For $T \geqslant 3.0$ the curves overlap almost perfectly to each other, with a slope of $\approx$3/5, that of a random coil in the case of a homopolymer. This is not unexpected, since at high temperature the heterogeneity in the interactions within the chain becomes negligible with respect to T, and the heteropolymer behaves effectively as a homopolymer.

For temperatures $T < 3.0$ the slope of $\ln \overline{R_g}$ versus $\ln N$ decreases and reaches 1/2, the value that homopolymers display at the θ point, at $T \approx 2.1$. If one decreases the temperature further, the curve is no longer linear in the range of N under consideration. This is likely to be a finite-size effect, since the gyration radius has to grow at least as $N^{1/3}$, corresponding to a fully compact structure.

The decrease of $\overline{R_g}$ as a function of T can also be visualized directly in the inset of Fig. 6 for each value of

N. A clear transition in $\overline{R_g}$ cannot be seen at any value of N. At large values of N, where transitions are expected to be sharper, we are not able to equilibrate the lowest temperatures, corresponding to the compact phase. Consequently, we are not able to highlight clearly a globule-coil transition, similar to that of homopolymers.

The clearest set of data is that calculated for $N = 60$. At $T = 1.8$ the mean gyration radius is 2.7, not far from that of a maximally compact globule, that is $N^{1/3} r_H = 2.4$. At $T = 2.0$ the value of $\overline{R_g}$ is 3.2, close to that associated with that of an ideal chain, that is, $0.41 N^{1/2} = 3.18$. Anyway, the curve increases smoothly from the more compact to the more elongated conformations.

Summing up, the random heteropolymer displays at high temperature properties of the radius of gyration similar to those of homopolymers, including a θ point at which the size of the heteropolymer scales as that of an ideal chain. At lower temperatures, in the range of lengths we could equilibrate, the size is dominated by finite-size effects.

VII. SCALING PROPERTIES WITHIN A FIXED-LENGTH CHAIN

Sometimes the experimental data to analyze are not the looping probability of polymers of different lengths, but the looping probabilities of the various segments, of different lengths, within a given polymer. This is, for example, the case of chromosome conformation capture experiments on the chromatin fiber [7]. The standard way of extracting the scaling exponent is a linear regression of $\ln c(i,j)$ versus $\ln |i-j|$ of the whole set of data, where $|i-j| \leqslant N$ is the length of the segment starting at monomer i and ending at monomer j of the N-bead polymer. It was also suggested that fitting c versus n is a better strategy [25]; this is, however, unwise in the case of heteropolymers, because of the lack of self-averaging of c (cf. Sec. IV).

In any case, if the heterogeneity in the looping probability at fixed intermonomer linear distance is due to the variability of the interactions, the correct way of extracting the scaling behavior is similar to that described in Sec. II. As in the case of heteropolymers of different lengths, one can define a looping free energy ΔF [cf. Eq. (7)] and develop calculations similar to those which lead to Eq. (6). However, now Eq. (3) depends on $|i-j|$ instead of N; that is,

$$\Delta F(i,j) = \Delta E + \epsilon + T\beta' \ln |i-j|, \tag{12}$$

where we define the scaling exponent as β' to distinguish it from that of varying-size polymers. Now Eq. (2) is still valid, but N is fixed. The result is that, according to this model, β' should be obtained by an unweighted linear regression of $\Delta F(i,j)$ versus $\ln |j-i|$. Here, the main difference with Eq. (6) is the lack of weights in the sum.

As one is usually interested in the scaling properties of any two monomers as a function of their distance n along the chain, and not of two specific monomers i and j (which is, anyway, hardly self-averaging), a more convenient quantity to study is $\Delta F(n) = (N-n+1)^{-1} \sum_j \Delta F(j,j+n)$. From the properties of convolutions of Gaussian distributions and Eq. (5) one obtains

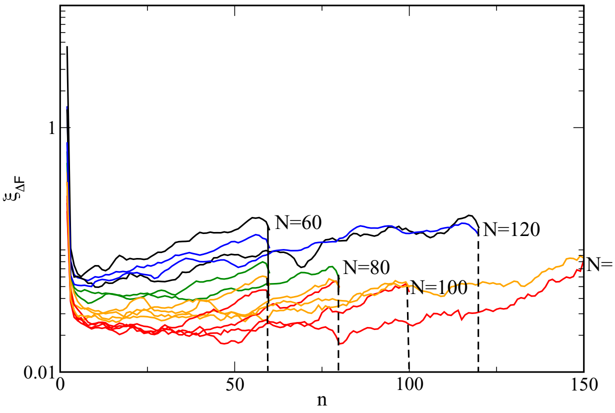


FIG. 7. The degree of self-averaging of $\Delta F(n)$ calculated at different values of N and of the temperature. The color code indicates the temperature and is the same as in Fig. 2.

$$p[\Delta F(n)] = \sqrt{\frac{(N-n+1)}{4\pi N\sigma^2}} \exp\left[-\frac{(\Delta F - T\beta \ln n + \epsilon)^2}{4N(N-n+1)^{-1}\sigma^2}\right]. \tag{13}$$

Consequently, β' can be found, in analogy with Eq. (5), from a linear fit of $\Delta F(n)$ versus $\ln n$, weighted by $(N-n+1)/N$. Operatively, this is not different from a linear regression of $\Delta F(i,j)$ versus $\ln |i-j|$, since $(N-n+1)$ is just the multiplicity of pairs of monomers at linear distance n.

The parameter $\xi^2_{\Delta F(n)}$ which describes the degree of self-averaging of $F(n)$ is displayed in Fig. 7. For each T and N it displays a nonmonotonic behavior as a function of n. At low n, $\xi^2_{\Delta F(n)}$ is large as in the case of fixed-length heteropolymer (cf. Fig. 2); then it drops because each value of $\Delta F(n)$ is the average not only on the realizations of the disorder, but also on the $N-n+1$ segments of length n, and each of them can be regarded as a realization of the disorder as well (see the discussion in Ref. [24]). As n increases, this effect diminishes, and $\xi^2_{\Delta F(n)}$ increases. For fixed n, $\xi^2_{\Delta F(n)}$ displays at each temperature in the region $n \sim N$ a decreasing behavior, which suggests the self-averaging character of this quantity.

The behavior of $\overline{\Delta F(n)}$ as a function of $\ln n$ is displayed in Fig. 8, obtained from polymers with $N = 60, 80, 100, 120$ at different temperatures. The χ^2, weighted according to Eq. (13), associated with the fit from $n = 6$ (below which self-averaging is absent, cf. Fig. 7) to varying n is displayed in the inset of Fig. 8. At $T > 2.0$, corresponding to the elongated phase of the polymer (cf. previous section), the linear fit is very good except when $n \approx N$. At lower temperatures, only the central region is linear ($6 \lesssim n \lesssim 60$), while for $n \sim N$ the curve bends down similarly to that expected for a homopolymeric globule. However, in all cases the associated χ^2 remains lower than 1, due to the larger weight of small n to the fit.

The values of β' obtained from the fits is displayed in Fig. 9. Overall, the values of β' are smaller than those of β corresponding to the same temperature. At the highest temperature it displays the value $\approx 9/5$ predicted for self-avoiding walks. At low temperatures, β' can reach values as low as 0.92. The reason is again that finite-size effects are

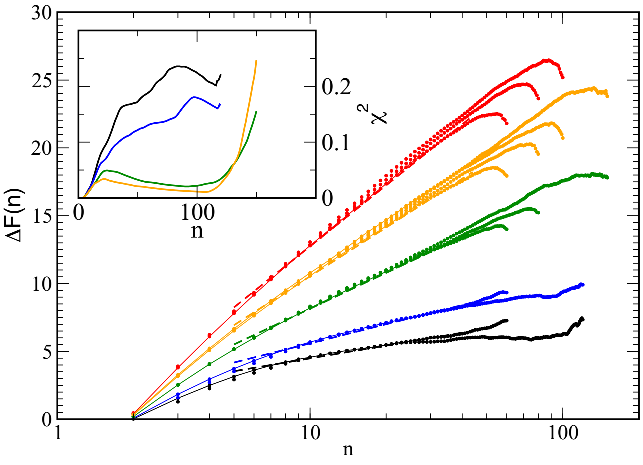


FIG. 8. The scaling of $\overline{\Delta F(n)}$ as a function of $\ln n$ at different temperatures (color code of Fig. 2) for different values of N. The fit, done between $N = 6$ and $n = 60$, is displayed with a dashed line. (Inset) The χ^2 associated with the fit up to length n.

amplified by the larger weight of small fragments of the chain, which is anyway unavoidable because fragments with $n \sim N$ are dominated by disorder.

VIII. IMPLICATIONS FOR CHROMOSOME CONFORMATION CAPTURE EXPERIMENTS

These results have important implications in the context of studies of chromosome conformation based on chromosome conformation capture (3C) experiments. In 3C-based methods, digestion and successive religation of formaldehyde-cross-linked chromatin in cell nuclei allows the detection of spatial proximity between DNA sequences (Fig. 10). In recent versions of 3C methods such as Hi-C, 4C, and 5C (reviewed in Ref. [6]), high-throughput sequencing is used to detect 3C DNA ligation products, making it possible to extract actual interaction frequencies. 3C-based experiments have allowed fundamental discoveries, notably that the folding of mammalian chromosomes is highly hierarchical. Each chromosome displays large-scale patterns of preferential associations into two so-called compartments, spanning several million base pairs of either active or inactive chromatin [7]. Compartments are further subdivided into smaller blocks of preferential interactions, referred to as topological associating domains (TADs) [26,27]. TADs are further characterized by the presence of smaller structures that occasionally define smaller domains dubbed loops domains [28].

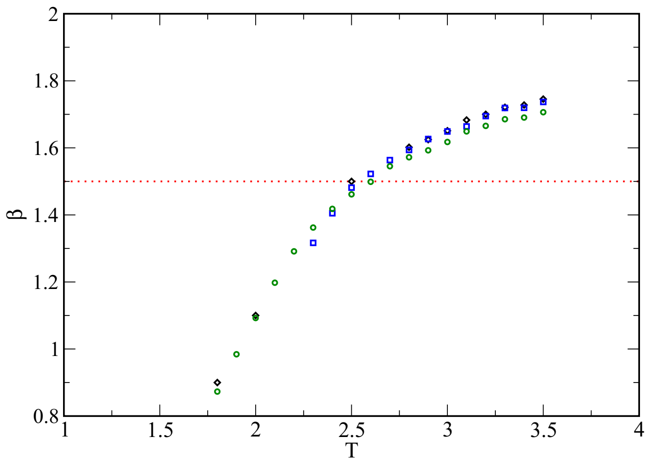


FIG. 9. The exponents β' associated with the fits of $\Delta F(n)$ versus n (solid black symbols), for the cases $N = 60$ (green circles), $N = 80$ (blue squares), and $N = 120$ (black diamonds).

In addition, 3C-based experiments make it possible to access the scaling behavior of chromosomes. Linear fitting of the logarithm of the experimentally determined contact probability versus the logarithm of the linear distance along the chain gives scaling exponents that are lower than those that are typical for homopolymers. Hi-C-based measurements led to scaling exponents of 1 over large genomic distances (between 10^6 and 10^7 base pairs) [7] and even smaller ($\sim$0.75) at shorter genomic distances [8]. Importantly, these scaling behaviors have been often used to test alternative models for how chromosomes are folded in the three-dimensional space, and what mechanisms give rise to the observed hierarchical structure, at various genomic length scales [7,8,12,29,30]. In the earliest application of this strategy [7] it was shown that the $\beta \sim 1$ behavior observed on human chromosomes in the megabase range can be explained in terms of fractal globule (or crumpled globule, according to the original nomenclature [11]). A fractal globule is the out-of-equilibrium structure obtained by a rapid collapsed of a swollen coil; not having the time to explore the associated conformational space, in this metastable globular state the polymer partially retains the correlations it displayed in the coil state, and in particular the fact that each monomer binds preferentially to those which are close along the chain. Successive investigations suggested that other models must be invoked to explain the deviations from the $\beta \sim 1$ behavior, which are observed either when studying shorter genomic ranges [8] or when considering single chromosomes instead of their average behavior [12]. In addition, scaling exponents were recently used to support the validity of models based on energy-driven mechanisms such as loop extrusion by DNA-associated protein complexes [8,13], which could explain how specific chromosome structures such as TADs and loop domains emerge. Finally, mitotic chromosomes have been shown to display a peculiar double-decay regime, which was used to infer a model where loop extrusion leads to chromosome condensation [29].

Importantly, our calculations suggest that finite-size effects, combined with the heterogeneity of the interactions in the chain, are sufficient to account for the observed range of scaling exponents. Of course the model we described does not provide a mechanistic interpretation of the observed exponents. Nevertheless, it suggests that scaling exponents cannot be the only quantitative observable used to construct and validate a model for chromosome folding. Other properties of the chain, in particular, distance distributions between pairs of loci, correlations between them, or even their dynamic properties, which can all be measured experimentally should be also used to distinguish between alternative models.

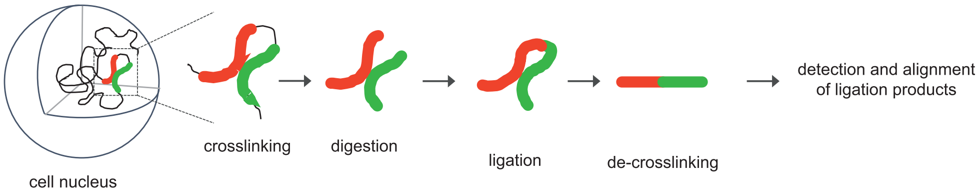


FIG. 10. Schematics of 3C-based techniques. Chromatin is cross linked with formaldehyde in the nuclei of a population of cells, digested with a restriction enzyme, and religated to favor the formation of hybrid DNA molecules that represent physical interaction events. After cremation decrosslinking, ligation products are purified, detected by DNA sequencing, and aligned to the reference genome.

IX. DISCUSSION

A. The polymer two-state model

The free-energy difference between looped and unlooped states within a two-state model provides a consistent way of studying the scaling properties associated with the looping mechanism with respect to the length of the random heteropolymer. From a theoretical argument and from numerical simulations, based on a self-adjusting simulated-tempering technique, the fluctuations about the average over the realizations of the random interaction within the heteropolymer are small, in the range of length of the order of 10^2 monomers but not in the thermodynamic limit.

Polymers of $\sim 10^2$ monomers are the longest systems for which we could guarantee equilibration, although with a consistent computational effort. Fortunately, this is the typical size of biological polymers. In fact, protein domains have an average length of 150 residues [31]. Topological associating domains in mammalian chromatin display a typical length of 10^6 bases, corresponding to 10^2 Kuhn lengths [32].

At high temperature, where the polymer is elongated, the looping probability of random heteropolymers displays a scaling exponent which varies continuously with respect to the temperature from ≈ 2.05 to 1.5. This is different from the behavior of homopolymers, for which only two possible exponents are expected.

At lower temperatures, corresponding to a compact phase of the heteropolymer, the determination of the scaling exponent is more cumbersome. Short chains display significant finite-size effects, resulting in a scaling of the looping probability with exponents smaller than 1.5. Longer chains display large disorder-dependent variability, which down-weights the determination of the exponent and the evaluation of the associated error. This amplifies the role of finite-size effects in the determination of the exponents even of large chains.

This phenomenon operates, for different reasons, both when considering chains of different lengths and segments of different lengths in a fixed-length heteropolymer. In the former case, the looping free energy is affected by the disorder provided by the internal energy of the chain, which is an extensive quantity. In the latter case, the free energy must be averaged over all the segments of the same length to be self-averaging, and the number of such segments decreases with the overall length of the chain. Anyway, fits of self-averaging free energies at low temperatures emphasize finite-size effects, resulting in exponents smaller than $3/2$.

B. Comparison with other models

Other investigations of the role of disorder in the looping of polymers were described in the literature, especially to describe the DNA double helix. In Ref. [33] it was shown that quenched randomness in the rest angles of a Kratky-Porod model result in a persistence length and response to external forces which are self-averaging (the latter under the hypothesis of small forces) and which are simply renormalized by the disorder.

A transfer-matrix formalism was used to study the effect of quenched (nonrandom) defects in the helasticity [34], resulting in a consistent increase in the looping probability of the polymer model. The same model was the extended [35] including random defects; a strong dependence of the looping probability was observed, suggesting a non-self-averageness of this property.

However, these models are controlled by the elasticity of the polymer and were designed to describe the properties of DNA strands of length comparable with their persistence length. The present model is thought to describe polymers, like chromatin and proteins, of length much larger than their persistence length (cf. Sec. IX A), and consequently no rigidity is modeled beyond the (inextensible) distance between consecutive monomers.

A perturbative calculation describing a flexible heteropolymer with random two-body interactions [24] showed that in the limit of small interaction volume the contact probability displays the standard homopolymeric exponents, affected by an exponential cutoff (cf. Sec. IX C below).

C. Role of excluded volume

The values of β found in the variable-length segments of a fixed-length chain are smaller than those of a set of chains of different lengths. There are two differences between the two cases. The former is that considering the variable-lengths segments of the same chain leaves correlations in the contact energies, which are absent when considering different realizations of varying-length chains. Moreover, when studying the variable-lengths segments of the same chain, the "tails" of the chain (i.e., the segments 1 to $i-1$ and $j+1$ to N, when studying the looping of i with j) may play a role. As a matter of fact, also for homopolymers it was shown [36] that the length of the tail can affect considerably the looping mechanism. The reason is that the excluded volume of the tail can shield the two monomers defining the loop, decreasing their binding probability.

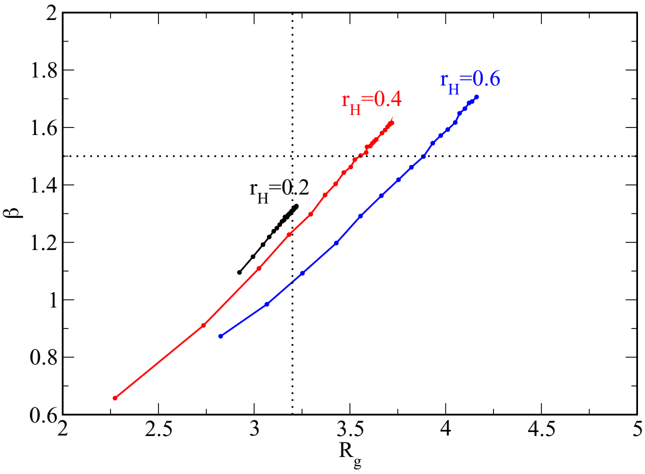


FIG. 11. The exponents β found at different temperatures, corresponding to different gyration radii R_g, using models with different length scale of the interaction potentials. The segments of a chain with $N = 60$ are used to calculate the values of β. The dotted lines indicate the expected values of R_g and of β at the θ point.

To investigate this point, we have repeated the simulations with different potentials, defined by different choices of the hardcore radius r_{HC} (and interaction radius proportional to r_{HC}), calculating the value of the exponent β for each of them. In Fig. 11 we show the result of these calculations. Since models with different r_{HC} display different temperature scales for the coil-globule transition, we use as an independent variable the gyration radius R_g. For each value of R_g, with decreasing r_{HC} the resulting β increases towards the values found with chains of different lengths, suggesting that the shielding effect plays a role in determining the difference between the two cases.

These results also suggests that the difference between the present numerical calculations and the analytical results found in Ref. [24], namely that for $T \geqslant \theta$ the exponent of a heteropolymer should not change with respect to the homopolymeric case, while only an exponential cutoff appears in the looping probability, can be associated with the hypothesis $r_{HC} \rightarrow 0$ used in the analytical calculations.

X. CONCLUSIONS

In random heteropolymers, scaling exponents relating the contact probability between monomers with their linear distance along the chain display strong finite-size effects which are amplified because of the large variability of the probability of long-range contacts, which are consequence of their lack of self-averageness. We suggest that this effect can strongly affect the interpretation of experimental data describing the scaling of contact probability in biopolymers. In the case of chromosome folding, our results suggest that one should be careful in selecting a physical model to describe the behavior of chromosome based on its scaling exponents, as a random heteropolymer can show exponents similar to those observed in experiments.

[1] F. Spitz, Semin. Cell Dev. Biol. **57**, 57 (2016).
[2] S. W. Bruun, V. Iesmantavicius, J. Danielsson, and F. M. Poulsen, Proc. Natl. Acad. Sci. USA **107**, 13306 (2010).
[3] M. Buscaglia, L. J. Lapidus, W. A. Eaton, and J. Hofrichter, Biophys. J. **91**, 276 (2006).
[4] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[5] L. Liu and C. Hyeon, Biophys J. **110**, 2320 (2016).
[6] A. Denker and W. de Laat, Genes Dev. **30**, 1357 (2016).
[7] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, Science **326**, 289 (2009).
[8] A. L. Sanborn, S. S. P. Rao, S.-C. Huanga, N. C. Duranda, M. H. Huntley, A. I. Jewett, I. D. Bochkova, D. Chinnappan, A. Cutkosky, J. Li, Kristopher P. Geeting, A. Gnirkee, A. Melnikove, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, Proc. Natl. Acad. Sci. USA **112**, E6456 (2015).
[9] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
[10] L. Mirny, Chromosome Res. **19**, 37 (2011).
[11] A. Yu. Grosberg, S. K. Nechaev and E. I. Shakhnovich, J. Phys. (France) **49**, 2095 (1988).
[12] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, Proc. Natl. Acad. Sci. USA **109**, 16173 (2011).
[13] Goloborodko, J. F. Marko, and L. A. Mirny, Cell Reports **15**, 2038 (2016).
[14] E. I. Shakhnovich and A. M. Gutin, J. Phys. (France) **50**, 1843 (1989).
[15] I. M. Lifshits, Zh. Eksp. Teor. Fiz. **12**, 117 (1942).
[16] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[17] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[18] G. Tiana and L. Sutto, Phys. Rev. E **84**, 061910 (2011).
[19] J. Johnson, C. A. Brackley, P. R. Cook, and D. Marenduzzo, J. Phys.: Condens. Matter **27**, 064119 (2015).
[20] L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard, Cell **157**, 950 (2014).
[21] G. Tiana, F. Villa, Y. Zhan, R. Capelli, C. Paissoni, P. Sormanni, E. Heard, L. Giorgetti, and R. Meloni, Comput. Phys. Commun. **186**, 93 (2014).
[22] R. Brout, Phys. Rev. **115**, 824 (1959).
[23] A. J. Guttman and M. F. Sykes, J. Phys. C **6**, 945 (1973).
[24] G. Tiana, Phys. Rev. E. **92**, 010702R (2015).
[25] A. Clauset, C. S. Shalizi and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[26] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, and J. D. E. Heard, Nature (London) **485**, 381 (2012).
[27] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, Nature (London) **485**, 376 (2012).

[28] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, Cell **159**, 1665 (2014).

[29] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker, Science **342**, 948 (2013).

[30] F. Benedetti, J. Dorier, Y. Burnier, and A. Stasiak, Nucl. Acid Res. **42**, 2848 (2014).

[31] D. Xu and R. Nussinov, Folding Des. **3**, 11 (1998).

[32] J. Dekker, J. Biol. Chem. **283**, 34532 (2008).

[33] D. Bensimon, D. Dohmi, and M. Mézard, Erophys. Lett. **42**, 97 (1998).

[34] J. Yan, R. Kawamura, and J. F. Marko, Phys. Rev. E **71**, 061905 (2005).

[35] P. Ranjith, P. B. Sunil Kumar, and G. I. Menon, Phys. Rev. Lett. **94**, 138102 (2005).

[36] H. S. Chan and K. A. Dill, J. Chem. Phys. **90**, 492 (1989).

Discussion

Proper development requires the activation of the right gene expression programs at the right time and place. Transcriptional control represents a fundamental mechanism that cells use to fine-tune their gene expression programs. In metazoans, transcriptional control critically depends on long-range *cis* regulatory elements such as enhancers that can be located hundreds of kilobases away from the target promoter[9,106]. How enhancers control specifically their target promoters and avoid aberrant interactions in a such crowded environment as the nucleus is still poorly understood. The current dominant model of enhancer function is through direct physical contact with the target promoter. The three-dimensional organization of chromatin, which accommodates promoter-enhancer interactions, therefore might play an important role in specifying the correct interactions. The development of chromosome conformation capture (3C) methods have enhanced our understanding of chromatin folding, especially at the scale where promoter-enhancer interactions occur. 3C-based techniques, and in particular the genome-wide version called Hi-C, revealed that folding of mammalian chromosomes folding is hierarchical.
Among the hierarchy, topologically associating domains (TADs) have been extensively characterized and many properties have been attributed to them, including importantly their role in instructing promoter-enhancer interactions[57,86,87,107].

The scale of TADs optimizes promoter-enhancer interactions

Despite the massive efforts in characterizing the properties of TADs, many fundamental open questions remain. First of all, TADs are defined as regions of enriched internal interactions, but it is not always obvious how to define a TAD given a Hi-C heatmap as illustrated in Figure 6. More importantly, it is not known whether the properties that have been attributed to TADs are specific to them or rather shared by all the folding layers across the hierarchy. To address these questions, I implemented an algorithm to detect all the folding layers in the hierarchy. To ensure the results to be the most general possible, I required the algorithm to be parameter free in order to avoid any parameter dependent biases. Moreover, the algorithm must be able to detect the previously characterized folding levels. Finally, the algorithm must stratify the genome into folding layers based on a biological meaningful measure. CaTCH, which stands for ***Ca****ller of* ***C****hromosomal* ***T****opological* ***H****ierarchies*, was designed to fulfil these criteria. Indeed, the only parameter in the algorithm, called reciprocal insulation, is used to define the hierarchy itself. In addition, the reciprocal insulation has a direct biological meaning: it describes how two domains are depleted in interactions (insulated) from each other compared to the local internal enrichment.

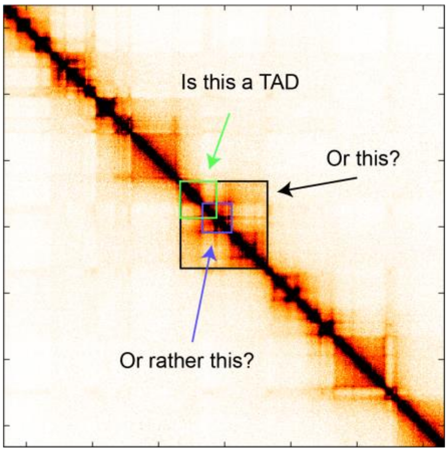


Figure 6: The problem of how defining a TAD

By applying CaTCH to published datasets, I found that a continuum of folding layers exists in the hierarchy and that previously characterized folding layers, namely TADs, compartments and sub-TAD structures, appear at different insulated scales. Here I want to stress the word **scale,** that indicates very similar but not identical sets of domains. CaTCH detects a continuum of domains where sets of domains close in the hierarchy (with comparable reciprocal insulation values) are very similar. The difficulties in defining TADs in some parts of the genome very often resulted in debates on what TADs really are. Given the population averaged nature of 3C methods, it might be more meaningful to define TADs as different sets of similar domains. Indeed, in a population of cells, different sub-populations may result in slightly different sets of domains giving rise to the nested pattern detected in Hi-C. Thus, using CaTCH to define a range of domains instead of a specific partitioning of the genome as TADs might be more biologically meaningful since it would in part account for the variability in chromatin folding in a population of cells.

The ability of CaTCH to detect the folding hierarchy in a parameter-free manner allowed me to perform a comparative analysis of structural, as well as functional properties, across the whole hierarchy in an unbiased way. By studying purely structural properties, I showed that none of the folding layers constitutes an intrinsically privileged scale. However, by looking at the functional properties, we found that although these functional properties are widespread across the hierarchy, they appear to be the most prominent at the scale of TADs. In fact, while CTCF clustering at domain boundaries is enriched across the whole hierarchy, it is maximized at the scale of TADs. Transcriptional coregulation during differentiation was also maximal at the scale of TADs, but only for down-regulated genes. It is interesting that transcriptional coregulation for up-regulated genes was maximized using TADs detected in differentiated cells suggesting that only active TADs are predictive for transcriptional coregulation. This might be explained by the fact that TADs functionally constrain only the action of active enhancers. Indeed, turning off the active enhancers would result in a coordinated lower expression of all genes within active TADs. In contrast, the activation of an enhancer in an inactive TAD might result in the formation of new structures (new TADs) that are then predictive for transcriptional coregulation. In line with this, I showed that the scale of TADs corresponds to domains where interactions between active promoters and enhancers are strongly enriched within the domains and start to be depleted across the boundaries. This is compatible with the idea that the scale of TADs corresponds to the domains where promoter-enhancer interactions are optimal, with the best trade-off between maximizing interactions within the interior of domains, and not enriching interactions across domain boundaries.

Studying chromatin structure in living cells without crosslinking and ligation

The advent of 3C-methods really paved the way for a more mechanistic understanding of chromatin folding and its relationship with transcription regulation. Despite fundamental discoveries enabled by 3C techniques, their existence is essentially based on a single technology that detects chromatin interactions as ligation products after crosslinking the DNA. Crosslinking and ligation have been often criticized as potential sources of experimental biases. In fact, crosslinking by nature cannot distinguish between direct and indirect molecular interactions since formaldehyde, the crosslinking compound, may cause the formation of networks of crosslinked chromatin leading to the detection of interactions that do not occur in living cells. Moreover, the ligation efficiency critically depends on the length of fragments. These experimental biases very often raised the question of whether fundamental structures such as TADs and chromatin loops exist in living cells.

To overcome these 3C techniques limitations, we developed a new method named DamC, where we recruit the *Escherichia coli* (*E. coli*) deoxyadenosine methylase (Dam) to arrays of TetO sites, with the idea that Dam methylates adenines in GATC motifs within regions that are in molecular contact with TetO array. By developing a biophysical model for methylation kinetics, we demonstrated that we can infer the molecular contact probability from methylation signal.

DamC has several advantages compared to 3C methods:

1) It detects chromatin interactions at the molecular scale since Dam can methylate only if it is directly bound to DNA.
2) It does not involve crosslinking and ligation
3) The data interpretation is based on rigorous physical modeling of methylation kinetics, providing a rational basis for the quantitative interpretation of the results
4) It can be used to study chromosomal interactions in a tissue-specific context by expressing the Dam protein under a tissue-specific promoter
5) It can be applied when the number of cells is limiting: indeed, if in 3C methods only two fragments per cell can be retrieved, in DamC many GATC sites can be methylated during the duration of the experiment resulting in many more fragments that can be retrieved.

At the current state, DamC has some limitations that could be improved in future implementations:

1) It requires genetic manipulation for the insertion of TetO arrays. In particular, studying chromatin conformation in a specific locus would require the targeted insertion of TetO arrays.
2) High resolution DamC requires a much higher sequencing depth than standard 4C
3) DamC requires tight control of nuclear Dam concentration: indeed, as my biophysical model of methylation kinetics has demonstrated, DamC experiments work optimally only in a narrow range of nuclear Dam concentrations.

To overcome the first point, one could use TAL effector proteins, similar to the semi-quantitative PCR-based approach (TALE-ID)[108], or catalytically inactivated Cas9 to perform targeted damC experiments; however, the Dam concentration might be tricky to control. To improve the resolution, oligonucleotides that recognize the regions of interest could be used to enrich for the fragments within these region as is done in Capture-C[109].

Despite the limitations, the findings from the DamC project represent an important contribution to the field of chromatin organization for several reasons. By using DamC in mouse embryonic stem cells carrying hundreds of TetO viewpoints, we provide the first orthogonal *in vivo* evidence for the existence of TADs and CTCF mediated chromatin loops. Moreover, DamC confirms the same modest drop in contact probability across TAD boundaries as detected in 4C and Hi-C suggesting that the mild drop might represent the best trade-off between enriching for interactions within the domain and depleting interactions across boundaries. In addition, insertion of ectopic CTCF sites demonstrated that chromatin structure can be manipulated in living cells. This is really exciting since it might open the way to new types of medicine for diseases which are primarily caused by deleterious rearrangements of chromatin structure[88–90]. Finally, DamC reports the same scaling of contact probability as in Hi-C and 4C. This is a fundamental result since it supports all the physical models where the scaling of contact probability was compared to the scaling of Hi-C crosslinking frequency to benchmark polymer simulations. In fact, if Hi-C crosslinking frequency has always been assumed, without any proof, to be proportional to absolute contact probability, we demonstrated through rigorous physical modelling that DamC enrichment is directly proportional the absolute contact probability.

Random heteropolymer as a model for studying chromatin

One of the reasons why I find the field of chromatin architecture so fascinating is that it lies at the interface between biology and physics. Chromatin constitutes a perfect example of a polymer for which physicists have developed models and formulated testable hypotheses to understand their arrangement and dynamic properties using principles of polymer physics. Building on the findings of 3C methods, several polymer models have been developed to better understand the mechanisms underlying chromatin folding[97,98], with the scaling exponent being used to test alternative models[78,110–112]. While for an equilibrium homopolymer the scaling can uniquely identify the corresponding polymer model, this is not the case for heteropolymers. Since chromatin is an heteropolymer, it is important to understand whether it would be enough to use only the scaling to benchmark the polymer simulations and test alternative models. To this aim, I have built finite size heteropolymers models with random interactions and simulated the corresponding scaling of contact probability. I showed that the heterogeneity, together with the finite size effect, can reproduce the whole range of observed scaling exponents. This suggests that one should be careful in discriminating polymer models to describe the folding of chromatin using only the scaling exponents since heterogeneity and finite size effect can reproduce the scaling values observed experimentally in Hi-C.

In summary, my contribution adds to the current debate on the role of chromosome structure in several respects. First, through an unbiased comparative analysis of functional and structural properties across the folding hierarchy, I have revealed why and how TADs represent a privileged folding scale in the hierarchy. Second, through the development of an orthogonal method to measure chromatin interactions in living cells and at the molecular level, we have provided an *in vivo* validation of folding structures where previously the evidence had come from a single technology. Finally, the ability of finite size random heteropolymers to reproduce the wild range of scaling exponent found in Hi-C suggests caution is needed in discriminating polymer models based solely on the scaling exponent.

Despite great progress, still a lot needs to be done to fully elucidate the mechanisms driving the establishment of the folding hierarchy. Indeed, if compelling evidence suggested that TADs and chromatin loops, in mammals, arise through a process called loop extrusion[78,82] with cohesin and CTCF being the main players, what drives the formation of compartments remains an open question. Compartments are mutually exclusive associations between active and inactive chromatin and have been observed for all mammalian cells. Only a few studies have reported loss of compartments, namely for the maternal genome in the zygote[54] and during mitosis[113]. Both in the zygote and during mitosis, extensive epigenetic reprogramming occurs with high histone PTMs dynamics[114–116]. Whether PTMs are the drivers of compartments formation remains an interesting hypothesis. Knocking out the readers/ and or writers that recognize/add PTMs, either fully or a conditional KO using the degron system when the full KO is lethal, will be essential to elucidate the role of histone modifications in driving compartmentalization.

Genetic evidences have shown that the higher order chromatin folding plays an important role in establishing the correct pattern of promoter enhancer interactions[86,88,89], which is essential for the spatial and time control of gene expression. However, how physical interactions between promoters and enhancers are translated into transcriptional output remains completely obscure. Does a physical contact between enhancer and promoter lead to transcriptional bursting? Or rather many contacts are needed to turn on transcription? The rapid progresses that we are currently experiencing in improving imaging techniques, fluorophores stability and genetic engineering will allow to study simultaneously the dynamics of chromatin folding and transcription in living cells for long time. This will pave the way for a quantitative understanding chromatin folding and its relationship with transcription.

Acknowledgement

First and foremost, I would like to thank Luca for giving me the opportunity to pursue my PhD in his lab. He has been always supportive and helpful along the whole way. He gave me always the freedom to do my research as well as essential advices when needed. He thought me so much in these years that I would need a book to list everything.

I would like to thank the members of my thesis committee Nils Blüthgen, Michael Stadler and Attila Becskei for all the fruitful discussions and suggestions along my PhD. I would like to thank Antoine Peters for chairing my defense. Furthermore, I would like to thank Nils for being my co-referee and Attila for being my faculty representative.

Special thanks to Michael Stadler, Lukas Burger and Iros Barozzi for introducing me to the world of bioinformatics and being always there when I needed advices and help.

A huge thank to the whole Giorgetti Lab. I feel truly privileged to have worked with all of you! The atmosphere in the lab is just awesome, I am sure there is no better working environment. Big thanks to Jessica that guided me when I had the crazy idea to experience the wet lab! I still remember the first gel without a ladder! Big thanks also to Josef, with whom I shared the PhD roller coaster!

Big thanks to FMI facilities. Thanks to Stefan and Enrico for always being there when I messed up the server. Thanks to Sebastien for sequencing everything we asked for.

Big thanks to the officemates! You made my daily life in the office a pleasure.
Big thanks to Joanna Mitchelmore and Marco Michalski for proof-reading my thesis.

I would like to express my deepest gratitude to all the friends, inside and outside FMI, who made the PhD journey much easier. Special thanks to the “Italian mafia”, for all the dinners, food, parties, grilling. You made my life Basel like at home!

Big thanks to Elida, who has been always there to help!

Last but not least, I would like to thank my family for their endless support. This thesis would not have been possible without Leilei, who supported me along the whole way, and made always every of my days better!

References

1. Carroll, S. B. Evolution at two levels: On genes and form. *PLoS Biology* (2005). doi:10.1371/journal.pbio.0030245
2. Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* **22**, R898–R899 (2012).
3. Cavalier-Smith, T. The evolution of genome size. *Evol. Genome Size . John Wiley Sons, New York.* (1985). doi:10.1016/0092-8674(86)90278-3
4. Hahn, M. W. & Wray, G. A. The g-value paradox. *Evol. Dev.* **4**, 73–75 (2002).
5. Craig Venter, J. *et al.* The sequence of the human genome. *Science (80-.).* (2001). doi:10.1126/science.1058040
6. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).
7. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* (1961). doi:10.1016/S0022-2836(61)80072-7
8. Walters, M. C. *et al.* Enhancers increase the probability but not the level of gene expression. *Proc. Natl. Acad. Sci.* (1995). doi:10.1073/pnas.92.15.7125
9. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
10. Maurano, M. T. *et al.* Systematic Localization of Common. *Science (80-.).* **337**, 1190–1195 (2012).
11. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics* (2013). doi:10.1038/nrg3373
12. Stanojevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science (80-.).* (1991). doi:10.1126/science.1683715
13. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* (2010). doi:10.1038/nature09326
14. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).
15. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics* (2011). doi:10.1038/ng.759
16. Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.* (2008). doi:10.1101/gad.1693008
17. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* (2012). doi:10.1038/nature11243
18. The ENCODE Project Consortium *et al.* 06 An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012). doi:10.1038/nature11247
19. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* (2012). doi:10.1038/nature11279
20. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* (2015). doi:10.1101/gr.185272.114
21. Bulger, M. & Groudine, M. Looping versus linking: Toward a model for long-distance gene activation. *Genes Dev.* **13**, 2465–2477 (1999).
22. Heitz, E. Das Heterochromatin der Moose. *Jahrbücher für wissenschaftliche Bot.* (1928). doi:10.5244/C.2.23
23. Dillon, N. & Festenstein, R. Unravelling heterochromatin: Competition between positive and negative factors regulates accessibility. *Trends in Genetics* (2002). doi:10.1016/S0168-9525(02)02648-3
24. Olins, A. L. & Olins, D. E. Spheroid chromatin units (v bodies). *Science (80-.).* (1974). doi:10.1126/science.183.4122.330
25. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* (1997). doi:10.1038/38444
26. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.* **79**, 347–372 (2015).
27. Han, M. & Grunstein, M. Nucleosome loss activates yeast downstream promoters in vivo. *Cell* (1988). doi:10.1016/0092-8674(88)90258-9
28. Allfrey, V. G. & Mirsky, A. E. Structural modifications of histones and their possible role in the regulation of RNA synthesis. *Science (80-.).* (1964). doi:10.1126/science.144.3618.559
29. Zhao, Y. & Garcia, B. A. Comprehensive catalog of currently documented histone modifications. *Cold Spring Harb. Perspect. Biol.* (2015). doi:10.1101/cshperspect.a025064
30. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Experimental and Molecular Medicine* (2017). doi:10.1038/emm.2017.11
31. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2016.59
32. Vermeulen, M. *et al.* Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* (2007). doi:10.1016/j.cell.2007.08.016
33. Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.* (2003). doi:10.1016/S0960-9822(03)00432-9
34. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412 (2007).
35. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics* (2008). doi:10.1038/nrg2206
36. Vernimmen, D. & Bickmore, W. A. The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. *Trends Genet.* **31**, 696–708 (2015).
37. Yun, M., Wu, J., Workman, J. L. & Li, B. Readers of histone modifications. *Cell Research* (2011). doi:10.1038/cr.2011.42
38. Cremer, T., Lichter, P., Borden, J., Ward, D. C. & Manuelidis, L. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Hum. Genet.* (1988). doi:10.1007/BF01790091
39. Lichter, P., Cremer, T., Borden, J., Manuelidis, L. & Ward, D. C. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum. Genet.* (1988). doi:10.1007/BF01790090

40. Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and tanslocations of chromosomes 4. *Proc. Natl. Acad. Sci.* (1988).
41. Boyle, S. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* (2002). doi:10.1093/hmg/10.3.211
42. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science (80-.).* (2002). doi:10.1126/science.1067799
43. Cullen, K. E., Kladde, M. P. & Seyfred, M. A. Interaction between transcription regulatory regions of prolactin chromatin. *Science (80-.).* (1993). doi:10.1126/science.8327891
44. Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & De Laat, W. Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol. Cell* (2002). doi:10.1016/S1097-2765(02)00781-5
45. Wit, E. & Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, (2012).
46. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* (2006). doi:10.1038/ng1896
47. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* (2006). doi:10.1101/gr.5571506
48. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
49. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* **485**, (2012).
50. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
51. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* (2009). doi:10.1126/science.1181369
52. Gibcus, J. H. & Dekker, J. The Hierarchy of the 3D Genome. *Molecular Cell* **49**, 773–782 (2013).
53. van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* (2017). doi:10.1016/j.cell.2017.04.022
54. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).
55. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
56. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
57. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–5 (2012).
58. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* (2015). doi:10.1038/nature14222
59. Van Bortle, K. *et al.* Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* (2014). doi:10.1186/gb-2014-15-5-r82
60. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* (2010). doi:10.1101/gr.099655.109
61. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.* (2009). doi:10.1073/pnas.0912402107
62. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
63. Berlivet, S. *et al.* Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs. *PLoS Genet.* (2013). doi:10.1371/journal.pgen.1004018
64. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* (2013). doi:10.1016/j.cell.2013.04.053
65. Rao, S. S. P., Huntley, M. H., Durand, N. C. & Stamenova, E. K. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
66. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **13**, 1 (2018).
67. Allen, B. L. & Taatjes, D. J. The Mediator complex: A central integrator of transcription. *Nature Reviews Molecular Cell Biology* (2015). doi:10.1038/nrm3951
68. Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* (2013). doi:10.1038/nature11884
69. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
70. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* (2017). doi:10.1016/j.cell.2017.08.002
71. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).
72. Nasmyth, K. & Haering, C. H. Cohesin: Its Roles and Mechanisms. *Annu. Rev. Genet.* (2009). doi:10.1146/annurev-genet-102108-134233
73. Ong, C. T. & Corces, V. G. CTCF: An architectural protein bridging genome topology and function. *Nature Reviews Genetics* (2014). doi:10.1038/nrg3663
74. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* (2017). doi:10.1016/j.cell.2017.05.004
75. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
76. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* (2017). doi:10.15252/embj.201798004
77. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* (2017). doi:10.1016/j.cell.2017.09.026
78. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* (2015). doi:10.1073/pnas.1518552112
79. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).
80. de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* (2015). doi:10.1016/j.molcel.2015.09.023
81. Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science (80-.).* **7831**, eaar7831 (2018).
82. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

83. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* (2016). doi:10.1016/j.cell.2016.09.037
84. Bartman, C. R., Hsu, S. C., Hsiung, C. C. S., Raj, A. & Blobel, G. A. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Mol. Cell* (2016). doi:10.1016/j.molcel.2016.03.007
85. Chen, H. *et al.* Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0175-z
86. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).
87. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* (2016). doi:10.1016/j.devcel.2016.10.015
88. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
89. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (80-.).* **351**, 1454–1458 (2016).
90. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* (2016). doi:10.1038/nature19800
91. Le Dily, François; Bau, D. Supplemental material. (2014).
92. Sazer, S. & Schiessel, H. The biology and polymer physics underlying large-scale chromosome organization. *Traffic* **19**, 87–104 (2018).
93. Parmar, J. J., Woringer, M. & Zimmer, C. How the Genome Folds: The Biophysics of Four-Dimensional Chromatin Organization. *Https://Doi.Org/10.1146/Annurev-Biophys-052118-115638* **48**, annurev-biophys-052118-115638 (2019).
94. Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosom. Res.* (2011). doi:10.1007/s10577-010-9177-0
95. de Gennes, P.-G. *Scaling Concepts in Polymer Physics*. (1979).
96. Van Den Engh, G., Sachs, R. & Trask, B. J. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science (80-.).* (1992). doi:10.1126/science.1388286
97. Fudenberg, G. & Mirny, L. a. Higher-order chromatin structure: Bridging physics and biology. *Curr. Opin. Genet. Dev.* **22**, 115–124 (2012).
98. Tiana, G. & Giorgetti, L. Integrating experiment, theory and simulation to determine the structure and dynamics of mammalian chromosomes. *Curr. Opin. Struct. Biol.* **49**, 11–17 (2018).
99. Baú, D. *et al.* The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* (2011). doi:10.1038/nsmb.1936
100. Di Stefano, M., Paulsen, J., Lien, T. G., Hovig, E. & Micheletti, C. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci. Rep.* (2016). doi:10.1038/srep35985
101. Giorgetti, L. *et al.* Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–63 (2014).
102. Gavrilov, A., Razin, S. V. & Cavalli, G. In vivo formaldehyde cross-linking: It is time for black box analysis. *Brief. Funct. Genomics* (2015). doi:10.1093/bfgp/elu037
103. Gavrilov, A. A. *et al.* Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt067
104. Williamson, I. *et al.* Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* (2014). doi:10.1101/gad.251694.114
105. Belmont, A. S. Large-scale chromatin organization: The good, the surprising, and the still perplexing. *Current Opinion in Cell Biology* (2014). doi:10.1016/j.ceb.2013.10.002
106. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0128-0
107. Le Dily, F. *et al.* Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* **28**, 2151–62 (2014).
108. Brant, L. *et al.* Exploiting native forces to capture chromosome conformation in mammalian cell nuclei. *Mol. Syst. Biol.* **12**, 891 (2016).
109. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* (2014). doi:10.1038/ng.2871
110. Barbieri, M. *et al.* Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci.* (2012). doi:10.1073/pnas.1204799109
111. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342**, 948–53 (2013).
112. Benedetti, F., Dorier, J., Burnier, Y. & Stasiak, A. Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1353
113. Gibcus, J. H. *et al.* A pathway for mitotic chromosome formation. *Science (80-.).* (2018). doi:10.1126/science.aao6135
114. Fraser, R. & Lin, C.-J. Epigenetic reprogramming of the zygote in mice and men: on your marks, get set, go! *Reproduction* (2016). doi:10.1530/rep-16-0376
115. Schulz, K. N. & Harrison, M. M. Mechanisms regulating zygotic genome activation. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-018-0087-x
116. Wang, F. & Higgins, J. M. G. Histone modications and mitosis: Countermarks, landmarks, and bookmarks. *Trends in Cell Biology* (2013). doi:10.1016/j.tcb.2012.11.005

Yinxiu Zhan

Personal Data

Date of Birth	30th April 1989
Phone	+41 7 66 73 45 58
Email	yinxiu.zhan@fmi.ch
Citizenship	Italian

Scientific Interests

The extraordinary progress in DNA sequencing technologies provides the scientific community with the tools to quantitatively tackle cutting-edge problems in molecular biology. Taking advantage of the enormous amount of quantitative data available, predictive theoretical models can ensure the optimal experimental design which is required to efficiently answer leading questions. My main research interest is to build and adapt theoretical mathematical and physical models to biological systems in order to guide experimental design, perform experiments and rigorously interprete experimental data.

Education

2015–2019 **PhD in Biophysics**, *Friedrich Miescher Institute for Biomedical Research*, Basel, Switzerland.
Supervisor : Dr. Luca Giorgetti
Research interests : My primary research goal is to understand whether and how the three-dimensional conformation of chromatin is involved in the control of gene expression. I use a combination of biophysical modelling, high-throughput sequencing and imaging techniques to gain insight into how chromosome conformation affects transcription.

2012–2014 **Master Degree in Physics**, *Università degli Studi di Milano*, Milan.
Supervisor : Prof. Guido Tiana
Thesis : ”Computational study of conformational fluctuactions of chromatin based on Hi-C data“
Final Grade : 110/110 *cum laude*

2009–2012 **Bachelor Degree in Physics**, *Università degli Studi di Milano*, Milan.
Supervisor : Prof. Matteo Paris
Thesis : ”The problem of discriminating quantum ensembles“
Final Grade : 110/110 *cum laude*

2004–2009 **High School Degree**, *Liceo Scientifico ”Elio Vittorini“*, Milan.
Final Grade: 85/100

Programming languages

Advanced	R, Awk, Bash, C
Intermediate	C++, Matlab, Mathematica, LaTeX, MS Office, ImageJ, Python
Basic	Adobe Illustrator, Foltran 90

Experimental and computational skills

◇ Bioconductor, Polymer Physics, Stochastic Simulation

◇ Molecular Cloning, RNA and DNA FISH, Fluorescence Microscopy, Culture of Mammalian Cells

Languages

Italian	Native
English	Highly proficient in spoken and written
Chinese	Basic

Publications

◇ J. Redolfi*, **Y. Zhan***, C. Valdes*, M. Kryzhanovska, I. Guerreiro, V. Iesmantavicius, T. Pollex, R. Grand, E. Mulugeta, J. Kind, G. Tiana, S. Smallwood, W. de Laat, L. Giorgetti “ DamC reveals principles of chromatin folding in vivo without crosslinking and ligation” **Nat. Struct. Mol. Biol.** doi: 10.1038/s41594-019-0231-0 (2019) *equal contribution

◇ J. G. van Bemmel, R. Galupa, C. Gard, ... , **Y. Zhan**, ... ,E. Heard “ The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist ” **Nat. Genet.** doi: 10.1038/s41588-019-0412-0 (2019)

◇ J. H. Wilbertz, F. Voigt, I. Horvathova, G. Roth, **Y. Zhan**, J. A.Chao“ Single-Molecule Imaging of mRNA Localization and Regulation during the Integrated Stress Response ” **Mol. Cell.** doi: 10.1016/j.molcel.2018.12.006 (2019)

◇ I. Horvathova, F. Voigt, A.V. Kotrys, **Y. Zhan**, C.G. Artus-Revel, J. Eglinger, M. Stadler L. Giorgetti, J.A. Chao “ The Dynamics of mRNA Turnover Revealed by Single-Molecule Imaging in Single Cells ” **Mol. Cell.** doi: 10.1016/j.molcel.2017.09.030 (2017)

◇ **Y. Zhan**, L. Giorgetti, G. Tiana “ Modelling genome-wide topological associating domains in mouse embryonic stem cells” **Chromosome Research**, Vol. **25**(1), 5-14, (2017)

◇ **Y. Zhan**, L. Mariani, I. Barozzi, E.G. Schulz, N. Bluthgen, M. Stadler, G. Tiana, L. Giorgetti “ Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes” **Genome Res.** doi: 10.1101/gr.212803.116 (2017)

◇ **Y. Zhan**, L. Giorgetti, G. Tiana “ Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers.” **Physical Review E**, Vol. **94**, 032402, (2016)

- G. Tiana, F. Villa, **Y. Zhan**, et al. " MonteGrappa: An iterative Monte Carlo program to optimize biomolecular potentials in simplified models." **Computer Physics Communications**, Vol. **186**, 93-104, (2015)
- **Y Zhan**, M. G. A. Paris. "Quantum ensembles and the statistical operator: a tutorial." **International Journal of Software and Informatics**, Vol. **8**, 241-253, (2014)

Conferences and Summer Schools

- Summer School, *Statistical Data Analysis for Genome-Scale Biology*, Bressanone (2019).
- Conference, *TriRhena Transcription and Chromatin Club*, IGBMC, Strasbourg (2019). **Talk**
- Conference, *Evolution, Structure and Function of Chromosomes High Order Structure, Institut Pasteur, Paris (2019).* ***Poster***
- Conference, *TriRhena Transcription and Chromatin Club*, IGBMC, Strasbourg (2019). **Talk**
- Keystone Symposia, *Chromatin Architecture and Chromosome Organization*, Whistler, Canada (2018). **Poster**
- Conference, *TriRhena Transcription and Chromatin Club*, FMI, Basel (2016). **Talk**
- Conference, *Genome architecture in Space and Time*, ICTP, Trieste (2016). **Poster**
- Conference, *EpiGeneSwiss kick-off meeting*, Weggis, Switzerland (2016). **Poster**
- Summer School, *Molecular and Atomistic Computational Techniques*, SISSA, Trieste (2013).