Measuring chromosomal interactions in living cells

Inauguraldissertation

Zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Josef Redolfi

aus Deutschland

Basel, 2020

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von:

Prof. Dr. Susan Gasser

Dr. Luca Giorgetti

Prof. Dr. Peter Meister

Basel, 19/11/2019

Prof. Dr. Martin Spiess

(Dekan)

to Martha

Table of Contents

Summary

3C based high-throughput sequencing methods such as Hi-C, 5C and 4C have substantially contributed to our current understanding of genome folding. These techniques have been instrumental in demonstrating that mammalian chromosomes possess a rich hierarchy of structural layers at the heart of which topologically associating domains (TADs) stand out as preferential functional units in the genome. TADs have been suggested to establish the correct interaction patterns between regulatory sequences, supported by genetic studies where the deletion of boundary elements resulted in ectopic gene expression in the neighboring domain. Within TADs, looping interactions occur between regulatory sequences and convergent binding sites of the architectural protein CTCF, the latter as a consequence of loop extrusion by cohesin that is blocked by CTCF bound to DNA in a defined orientation. The dominant role of CTCF in loop formation is further highlighted by induced depletion experiments and targeted deletions and inversion of CTCF sites manifesting in loss of these interactions.

Despite these fundamental discoveries and their implications for transcriptional control by *cis*-regulatory sequences, 3C and derivatives are based on formaldehyde crosslinking and ligation, which have been often criticized as a source of important experimental bias. This has actually raised the question if structures detected by 3C methods do really exist in living cells. Based on discrepancies between 5C and DNA-FISH data, it was suggested that 3C based methods might not always capture spatial

proximity or molecular-scale interactions, but rather detect DNA fragments which are hundreds of nanometers apart through crosslinking of macromolecular protein complexes between them. At the same time, it was debated whether capturing of ligation products might be variable depending on sequence context, therefore over- or underrepresenting some interactions detected in 3C based methods. Even though several other methods including native 4C/Hi-C, GAM and SPRITE have also detected chromatin compartmentalization, TADs and looping interactions, they still involve substantial biochemical manipulation of cells, notably either crosslinking or ligation. Importantly, many mechanistic models of chromosome folding rely on 3C based data, making the assumption that crosslinking frequency is proportional to absolute contact frequency. However, a formal proof of this is still missing.

In order to measure chromosomal contacts directly in living cells, without using chemical fixation nor ligation, I developed an alternative approach based on the DamID technique that exploits detection of ectopic adenine methylation by the bacterial methyltransferase Dam. In the original version of DamID, Dam is fused to a DNA binding protein of interest resulting in adenine methylation within GATC motifs in the neighborhood of the DNA binding sites. The methylation-sensitive restriction enzyme DpnI is then used to detect methylated GATCs followed by high throughput sequencing of the restriction sites. After mapping the reads and normalizing for non-specific methylation by freely diffusing Dam, the binding sites of the protein of interest can be detected genome wide.

I established a new modified version of this technique called DamC, where Dam is recruited in an inducible way to ectopically inserted Tet operators through fusion to the reverse tetracycline receptor. The detection of methylated DNA by high-throughput sequencing then allows to identify chromosomal contacts at high genomic resolution across hundreds of kilobases around viewpoints. Importantly, modeling of this process provides a theoretical framework showing that the experimental output of DamC is indeed proportional to chromosomal contact probabilities.

DamC provides the first crosslinking- and ligation-free validation of key structural features of mammalian chromosomes identified by 3C methods. It confirms the existence of TADs and CTCF loops as well as the scaling of contact probabilities measured in 4C and Hi-C, which supports the validity of physical models of chromosome folding based on 3C-based data. Finally, it demonstrates that ectopic insertion of CTCF sites can lead to the formation of new loops with endogenous CTCF-bound sequences. This shows that chromosome structure can be engineered by inserting short ectopic sequences that rewire interactions within TADs, opening interesting avenues for modifying gene expression by altering chromosomal interactions rather than regulatory DNA sequences directly.

List of figures

Abbreviations

ATP adenosine triphosphate

bps base pairs

BRE B-response element

CpG 5'-cytosine-phosphate-guanine-3'

dCas9 catalytically inactive CRISPR associated protein 9

CTCF CCCTC-binding factor

CTD carboxy terminal domain

Dam DNA adenine methyltransferase

DamID DNA-adenine-methyltransferase identification

DNA deoxyribonucleic acid

DNA-FISH DNA fluorescence in situ hybridization

DNMT1 DNA cytosine-5-methyltransferase 1

DPE promoter element

HP1 hetero-chromatin protein 1

HATs histone acetyltransferases

HDACs histone deacetylases

INR initiator element

KMTs lysine methyltransferases

KDMs lysine demethylases

TFs transcription factors

TSS transcriptional start site

Mb mega base

kbs kilobase

LCR locus control region

mESC mouse embryonic stem cells

mRNA messenger RNA

NTPs nucleoside triphosphates

PCR polymerase chain reaction

p-TEFb positive transcription elongation factor

PTMs post translational modifications

RNA ribonucleic acid

RNAPII RNA polymerase II

Shh sonic hedgehog

TADs topologically associating domains

TET ten-eleven translocation

TALEN transcription activator-like effector nucleases

3C chromosome conformation capture

3D 3-dimensional

1. General introduction

1.1 Gene expression in eukaryotes

Identifying the mechanisms which facilitate and maintain cellular diversity in complex organisms like humans has been a fascinating question for biologists for a long time. Remarkably, diversity originates from a single fertilized egg which transmits an identical genome to daughter cells during cell division. Despite sharing the same genomic information, specialized cells different in function and shape can arise from transcriptional programs which are tightly regulated in space and time. In higher eukaryotes, expression of protein coding genes by ribonucleic acid (RNA) polymerase II (RNAPII) is controlled by non-coding deoxyribonucleic acid (DNA) elements including core promoters, promoter proximal elements and enhancers which serve as regulatory platforms by recruiting transcription factors (TFs) through short DNA binding motifs typically 6-12 base pairs (bps) in size[1]. Among these elements enhancers stand out as they can map as far as one mega base (Mb) away from the transcriptional start site (TSS) of the target gene[2]. It is the genome wide occupancy of TFs at those sequences and their combinatorial physical interaction within the space of the nucleus that transmits the appropriate information to the RNAPII transcriptional machinery. Thus, it is critical to regulate TF binding to DNA in order to achieve correct expression patterns which ultimately determine cellular identity. The first layer of TF binding regulation happens at the level of DNA binding specificity. At the same time, more

indirect mechanisms of recruiting, like protein-protein interactions between TFs and transcriptional co-factors like mediator, without directly contacting DNA contribute to alternating TF occupancy[3,4]. In order to understand how these mechanisms of TF binding are regulated, it is fundamental to introduce the spatial organization of DNA within the nucleus.

1.1.1 Packaging of eukaryotic DNA into nucleosomes

The basic packaging of DNA occurs at the nucleosome where 146 bps are wrapped around four histone core proteins, H2A, H2B, H3, and H4 which are present in two copies each forming the histone octamer[5]. Together with a variety of proteins and RNA molecules which can undergo biochemical modification they are the backbone of chromatin. Chromatin does not only allow to compact the entire eukaryotic genome into the confined space of the nucleus but it is also key to regulate access of DNA to proteins involved in transcription. Histones and DNA itself can be modified with important implications for transcriptional regulation. Such modifications can alter chromatin structure via two possible mechanisms: First, by compacting and de-compacting DNA through changes in electrostatic charge providing or blocking access for DNA binding proteins such as TFs. Second, by introducing biochemical modifications at the nucleosome surface which attract chromatin binding proteins[6]. In fact, chromatin state and TF binding are strongly interconnected and distinct histone modifications have emerged as good predictors of the transcriptional state, with some

marks correlating with gene expression and others with gene repression. However, it is important to note that this classification may not always be black and white and some marks can be associated with either repressed or active genes. One example is the methylation of lysine 9 at histone H3 which normally has a negative effect on transcription at promoters but could also be found in coding regions of transcribed genes in mammalian genomes[7].

1.1.2 Chromatin marks and their role in transcription

One layer of transcriptional control on the chromatin level is mediated through a large variety of post translational modifications (PTMs) of histones, including methylation, phosphorylation, acetylation, ubiquitylation and SUMOylation. PTMs of histones can either change chromatin compaction and localization within the nucleus or serve as binding sites for effector proteins like chromatin remodelers. Most PTMs occur in a reversible manner at the histone tails, modulated by separate enzymes known as writers and erasers which add and remove the modifications respectively. Many writers including histone acetyltransferases (HATs) and lysine methyltransferases (KMTs), and erasers including histone deacetylases (HDACs) and lysine demethylases (KDMs) have been identified[8]. Together they maintain the homeostasis of histone modifications which are a central regulator of transcriptional programs. PTMs have been associated with two different categories of chromatin: Transcriptionally inactive heterochromatin which is tightly packed and localized at the periphery of nuclei in differentiated cells,

and transcriptionally active euchromatin which is less dense and localized towards the center of nuclei. However, it is important to further distinguish between facultative heterochromatin which is considered to be regulated and associated to differentiation and constitutive heterochromatin composed of high copy number repeats. In both cases the repressive marks restrict access for RNAPII and cofactors[9].

One other chromatin mark which has already been demonstrated to correlate with transcriptional repression in experiments from the early 80s, is the methylation of DNA itself. When transfected in *Xenopus* oocytes or cultured mammalian cells, it was shown that in vitro methylated DNA is transcriptionally silent[10,11]. Later, it became clear that DNA methylation in vertebrate genomes often occurs at cytosine residues within 5'-cytosine-phosphate-guanine-3' (CpG) dinucleotides which are concentrated at repeated regions and transposons[12]. In humans, DNA cytosine-5-methyltransferase 1 (DNMT1) is responsible for maintaining DNA methylation whereas de novo methylation is carried out by DNMT3A and DNMT3B. The methylation of DNA is reversible and ten-eleven translocation (TET) family members have been shown to catalyze active demethylation[13]. The patterns of methylation are very dynamic in developing embryos and tend to be more stable in differentiated somatic cells. Thereby, DNA methylation contributes in regulating gene expression during development but also in keeping chromatin transcriptionally silent in terminally differentiated cells. Interestingly, about 70% of all human promoters contain CpG islands which normally remain un-methylated but can contribute to gene silencing when methylated[14].

Thanks to next generation sequencing methods including DNase I hypersensitive sites sequencing (DNase-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) among others, it was also demonstrated that nucleosome positioning varies strongly across the chromatin landscape. At regulatory sequences such as promoters and enhancers, TFs have to compete for access to DNA with histones impacting on expression variation[15,16]. At those sites, pioneering factors and chromatin remodeling enzymes which can destabilize, restructure or eject nucleosomes, enable the dynamic access of TFs to packaged DNA. The general logic in activating transcription is by evicting nucleosomes at promoters and enhancers to expose the DNA to transcriptional activators, co-activators and the basal transcriptional machinery.

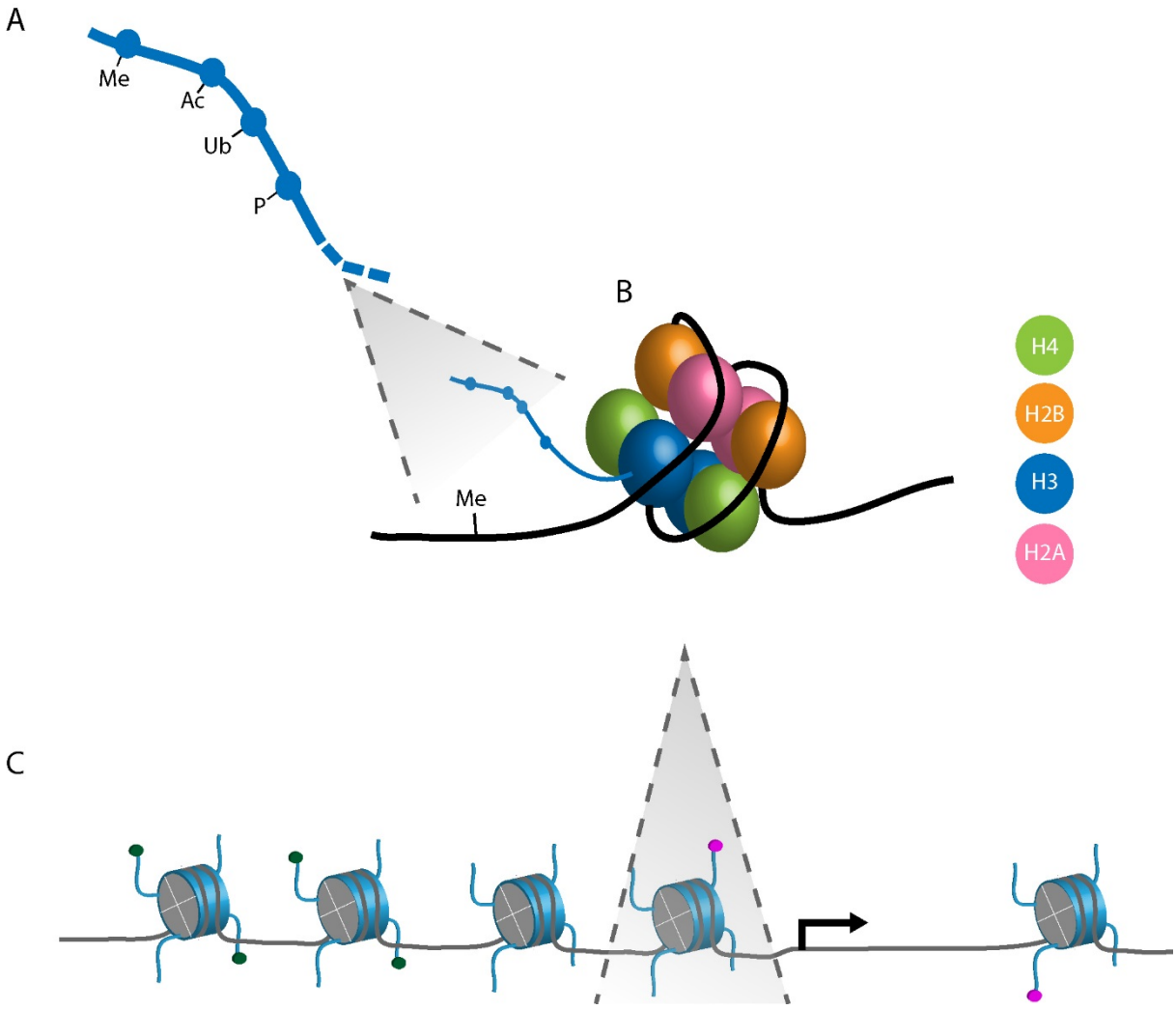


Figure 1 Regulatory features of chromatin: **A** Amino acid residues within histone tails can undergo biochemical modification including methylation (Me), acetylation (Ac), ubiquitination (Ub) and phosphorylation (P). The modifications impact on the interaction of nucleosomes with DNA and can recruit regulatory proteins. **B** 146 bps of DNA are wrapped around four histone core proteins, H2A, H2B, H3, and H4 which are present in two copies each forming the histone octamer. DNA itself can also be methylated at cytosine residues. **C** Nucleosome positioning on DNA regulates the accessibility for transcription factors at regulatory DNA sequences. Histone PTMS, which occur in clusters are a major determinant of DNA accessibility.

1.1.3 Initiation, Elongation and Termination in RNA-Polymerase II transcription

Although the architecture of promoters, including the methylation status at CpGs, histone PTMs and positioning of nucleosomes can differ substantially between them, the process of RNAPII transcription itself is equivalent. RNAPII transcription can be partitioned in a sequence of events including initiation, elongation and termination[17]. This multistep process starts with the formation of the preinitiation complex (PIC) at the core promoter which requires the binding of general TFs (GTFs) IIB, IID, IIE, IIF, IIH, additional cofactors including the mediator complex and RNAPII. At most promoters, PIC formation involves a set of characteristic DNA sequences including the TATA-box, a B-response element (BRE), the downstream promoter element (DPE) and a initiator element (INR) which all serve as binding hubs for various GTFs and TFs[18]. After GTFs and TFs have bound to promoters in a sequence specific way, they are involved in recruiting the mediator complex through diverse protein-protein surface interactions at virtually all RNAPII transcribed genes. First discovered by Kornberg and colleagues, mediator was shown to be essential for transcription in-vitro before it was purified from yeast providing evidence for a multi subunit complex which interacts with RNAPII[19,20]. Although the composition of the mediator complex can differ across species, mediator has a conserved function in transcription which is communicating signals between GTFs and RNAPII. Structural studies revealed that mediator is composed of four main modules including the head, middle, tail and CDK module, each

of which in turn is made of various polypeptides[21]. Importantly, the structural data indicates that polypeptides from the head and middle module interact with the carboxy terminal domain (CTD) of the Rbp1 subunit of RNAPII, highlighting its role in PIC formation.

After formation at the promoter, the PIC cooperates in an adenosine triphosphate (ATP) driven process to melt double stranded DNA, resulting in an open configuration referred to as transcriptional bubble. RNAPII can then start processing the DNA template strand in the presence of nucleoside triphosphates (NTPs) mediating RNA chain synthesis during elongation[22]. RNAPII promoter clearance coincides with the phosphorylation of the Pbp1 CTD (which contains multiple repeats of the heptapeptide YSPTSPS) involving the kinase activity of TFIIH or the positive transcription elongation factor (p-TEFb). Some elongation factors have also been shown to interact with RNAPII leading to poised polymerases at promoters counteracting p-TEFb[23]. Both, positive and negative regulators of elongation provide another level of rapid control of gene expression after the formation of the PIC.

After synthesis of the nascent transcript by RNAPII, the transcriptional cycle is terminated including cleavage and polyadenylation resulting in a mature messenger RNA (mRNA) which is exported to the cytoplasm where it eventually will be translated. When RNAPII approaches the end of a gene it slows down over a terminator sequence due to the recruitment of enzymes involved in mRNA processing. This includes 3'-end cleavage and polyadenylation of the nascent transcript when poly-adenylation signals appear in the sequence. The RNA can also invade the DNA and form an R-loop

structure that also slows down RNAPII[24]. One model of termination proposes a conformational change within RNAPII after associating with the 3'-end cleavage and polyadenylation complex resulting in pausing and finally release of RNAPII[25]. A second alternative model instead suggests RNA is still synthesized after 3'-end cleavage. In this case the 5'-3' exonuclease Xrn2 is recruited to the poly-adenylation site followed by degradation of the downstream transcript and release of the still elongating but slower RNAPII when taking over.

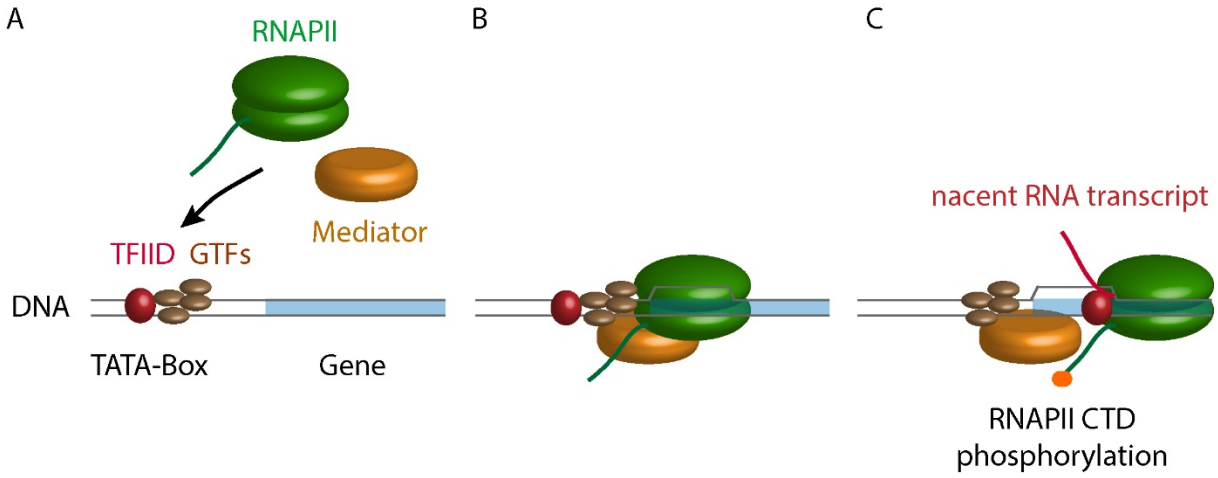


Figure 2 Scheme of RNA Polymerase II initiation and elongation at promoters: **A** The TATA binding protein which is a subunit of TFIID binds to the TATA box within the gene promoter. Next, other GTFs and the mediator complex are recruited promoting the correct binding of RNAPII. **B** Initiation: Mediator interacts with the PIC composed of RNAPII and the GTFs TFIID, TFIIB, TFIIE, TFIIE and TFIIH to initiate transcription. **C** Elongation: Promoter clearance coincides with the phosphorylation of the RNAPII CTD (which contains multiple repeats of the heptapeptide YSPTSPS) involving the kinase activity of TFIIH or the positive transcription elongation factor p-TEFb. RNAPII starts processing the DNA template strand mediating RNA synthesis in the presence of NTPs.

1.2 Enhancers and their role in gene expression

1.2.1 Enhancers: non-coding regulatory sequences in metazoan genomes

Enhancers are important regulatory sequences that have emerged as key distinctive genomic features of complex organisms and vertebrates in particular. They can be defined as portions of non-coding DNA, typically 100bp to 1kb in size, which contain binding motifs for transcription factors. As distal regulatory sequences which can map up to 1Mb away from the target promoter, they play a major role in gene regulation. According to the widely accepted looping model, they impact on transcriptional outputs of genes by engaging in physical contacts with promoters and looping out the intervening sequence. Different from other regulatory elements like promoters, enhancers can activate transcription over large genomic distances independent of their orientation relative to the cognate promoter, as reported in early studies where enhancer sequences were characterized by their activity when cloned in front of a reporter gene[26–28]. It is important to note that a gene can be regulated by multiple enhancers with different spatial and temporal activity, thus enabling complex combinatorial, time- and tissue-specific regulation of gene expression[29]. Deletions and mutations within enhancers also seem to be associated with human diseases including thalassemias and polydactyly, indicating their important role in regulating gene

expression[30,31]. Furthermore, mutations in enhancer sequences have been suggested to be the dominant mechanism driving morphological divergence within species[32].

1.2.3 Enhancers physically interact with promoters to regulate gene expression

Similar to promoters, activation of an enhancer generally requires multiple TFs to bind to overcome the energetic hurdle of nucleosome eviction. This is thought to be fostered by pioneer TFs (also named master regulators) which are able to bind to nucleosomal DNA and recruit chromatin remodelers that prime the enhancer for subsequent activation. A second class of activator TFs can then bind in a sequence-specific way to nucleosome-free regions, which are usually detected at active regulatory sequences as they are hypersensitive to nuclease digestion. At the enhancer, activators then recruit co-activators such as the mediator complex, p300 or the CREB binding protein (CBP) which can modify chromatin and/or directly interact with the transcriptional machinery. Although the mechanisms by which this happens are largely unknown and a matter of active research, it is possible that when active enhancers engage in physical contact with a target promoter, they can contribute to the recruitment of the PIC to, or the release of poised RNAPII from promoters[33]. Large multiprotein complexes including mediator are thought to function as a bridge for signal integration between enhancers and promoters through TFs and components of the PIC[34].

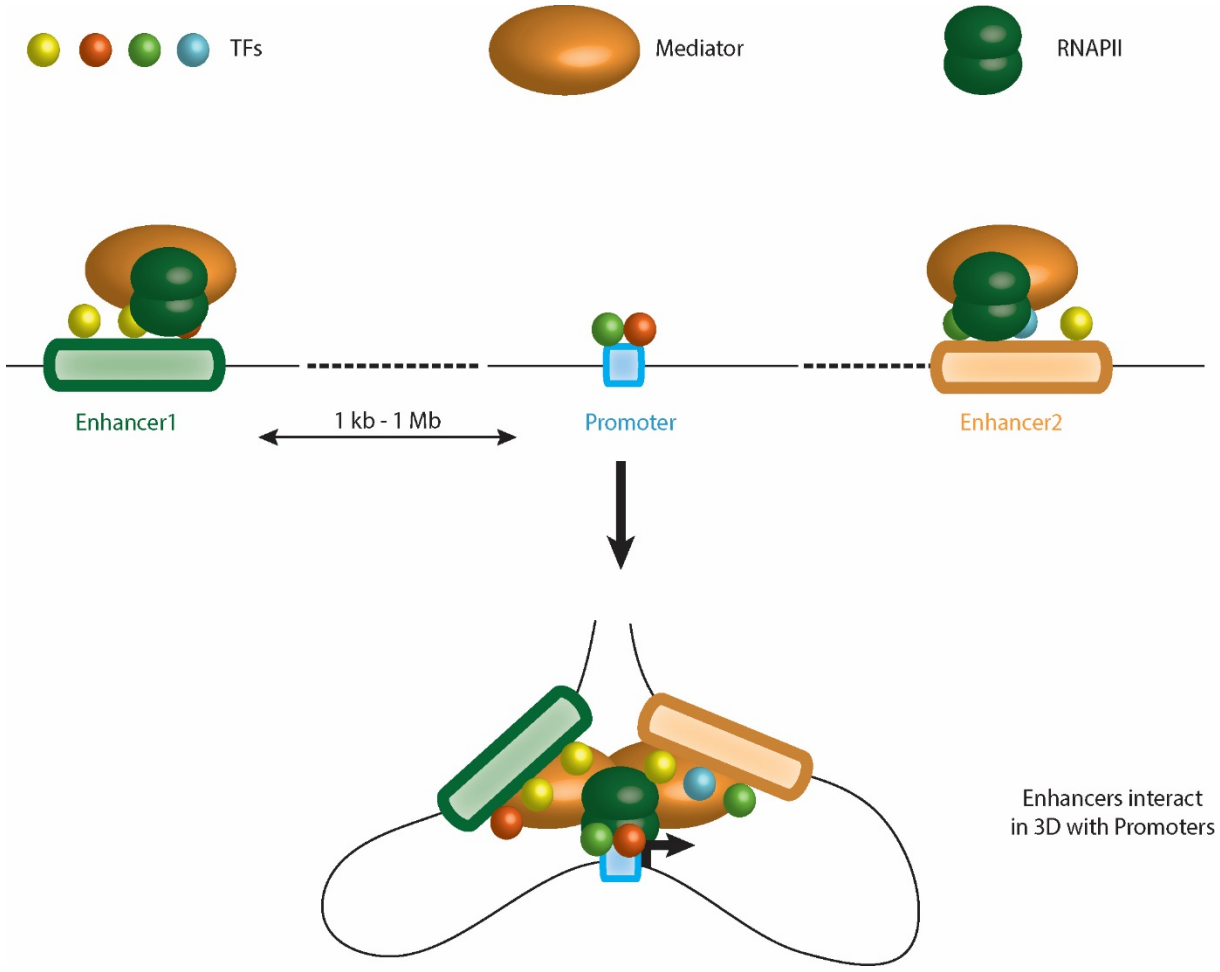


Figure 3 Enhancer promoter looping model: The looping model implies two or more regions of the genome to directly interact over large genomic distances by looping out the intervening DNA sequence. Enhancers impact on the transcriptional output of a gene by engaging in physical contact with the promoter. The mediator complex acts as a functional bridge between TFs and RNAPII. It has been shown to promote RNAPII binding and its transition to elongation.

The fact that the distance to target genes can exceed 1 Mb on the linear genome, as in the case of the *Shh* (sonic hedgehog) limb enhancer, raises the question of how enhancers specifically communicate with a given promoter without activating other

genes. Is it possible that factors binding simultaneously to the promoter and the enhancer can tether them together through specific protein-protein interactions? Studies of the paradigmatic β-globin gene cluster in mammals suggested the TFs GATA1 and FOG1 bridging the locus control region (LCR), a powerful enhancer characterized by DNase hypersensitivity and TF binding, to the beta major promoter[35]. Tethering experiments have also shown that Lbd1, which is recruited by GATA1, is necessary to bring the LCR and the promoter in close proximity[36]. Interestingly, when the LCR was ectopically inserted in other genomic locations, it was found to contact genes regulated by similar TFs. Based on these and other experiments, it was concluded that biochemical compatibility between some TFs and co-activators may underlie specific enhancer-promoter contacts in mammals by stabilizing a looping conformation after binding to enhancers and/or promoters. However, biochemical compatibility of TFs alone is not sufficient to promote specific interactions between regulatory elements. Thanks to the development of techniques that allow to study how chromosomes are spatially organized within the nucleus, it became evident that genome folding provides physical constrains for DNA interactions with important functional implications for enhancer-promoter communication.

1.3 Genome folding and enhancer biology

1.3.1 Microscopy based methods to study genome folding

The predominant method to study genome organization before the advent of next-generation sequencing techniques was DNA *in situ* hybridization (FISH), where cells are crosslinked using formaldehyde, followed by denaturation of DNA and hybridization with sequence specific fluorescent DNA probes. This method revealed that chromosomes occupy distinct territories in the nucleus and that nuclear positioning of a gene can impact on its expression[37–39]. More recently, electron microscope tomography has complemented this picture at a finer scale providing insights in the nuclear ultrastructure by visualizing 3-dimensional (3D) organization of Mb-scale DNA regions in fixed nuclei[40]. However, contrary to DNA-FISH, this method does not allow to study genome folding in a sequence specific way. The general advantage of light microscopy based methods utilizing fluorescent sequence specific probes is their ability to visualize directly how genomic sequences are organized within the space of the nucleus providing information about physical distances between loci of interest. Spectacular recent developments of this approach using super-resolution microcopy-based methods allow to extract spatial information on DNA conformation below the diffraction limit with increased sequence context[41,42]. However, at the current stage of development even these technical breakthroughs are still limited to consecutive DNA sequences of few hundred kb.

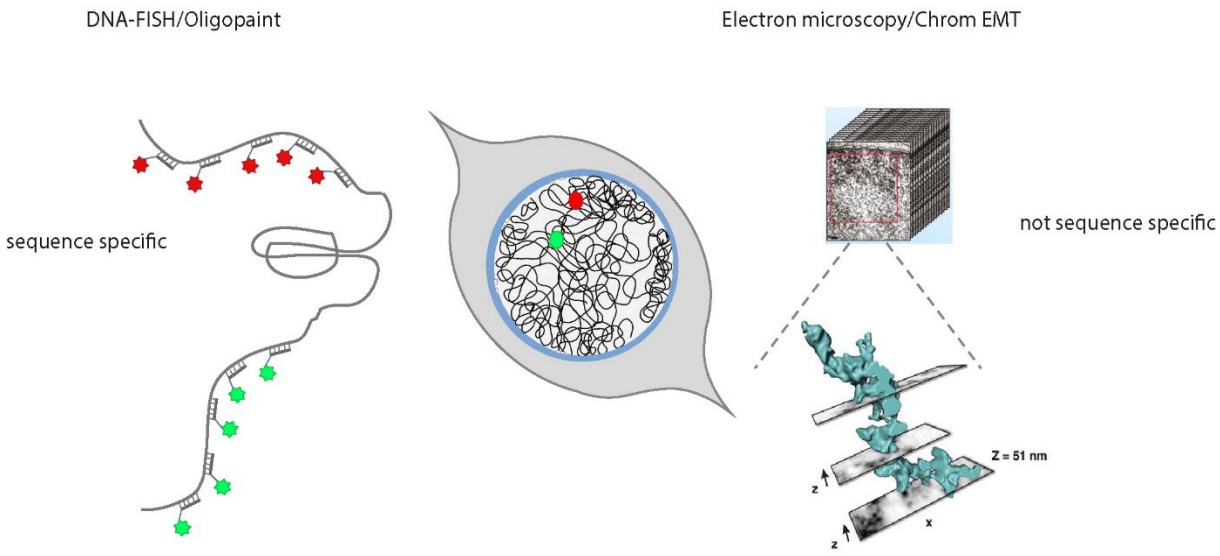


Figure 4 Microscopy based methods to study genome folding: In DNA-FISH, cells are crosslinked using formaldehyde, followed by denaturation of DNA and hybridization with sequence specific fluorescent DNA probes. In oligo paint, sequential incubation with fluorescent probes are used to label kilo-to-megabase regions of the genome. In combination with super-resolution microcopy-based methods this allowed to extract spatial information on DNA conformation below the diffraction limit with increased sequence context. Electron microscope tomography tiltseries (Chrom EMT) allows to visualize ultrastructure and 3D organization of individual chromatin polymers. However, Chrom EMT is not sequence specific. (This figure includes graphical objects from Ou et. al. Science 2015 and Parmar et. al. Annu. Rev. Biophys. 2019)

1.3.2 Sequencing based methods to study genome folding

Next-generation-sequencing based methods based on chromosome conformation capture (3C) have revolutionized our understanding of chromosome structure and its functional implications in enhancer biology. 3C-based methods rely on formaldehyde crosslinking of chromatin in cell nuclei, followed by restriction enzyme digestion and

subsequent re-ligation of free DNA ends. In the original 3C protocol, PCR primers where designed to amplify ligation junctions of two potentially interacting sequences (requiring to know the DNA sequence)[43]. This allowed a semi-quantitative one-vs.-one detection of crosslinking frequency by PCR when comparing two biological conditions[44]. Several next-generation sequencing variants of this method have been later developed including 4C, 5C and Hi-C[45,46]. Since 4C and Hi-C have been extensively used in this thesis, they will be discussed in more detail. Other techniques including 5C, ChIA-PET, ChIP-Hi-C have been extensively reviewed in de Wit and de Laat Genes Dev. 2012 and Denker and de Laat Genes Dev. 2016. The 4C technology allows to identify regions in the genome that contact a region of interest which is referred to as viewpoint. Different from 3C, the advantage is that all contacts made with the viewpoint can be appreciated in the experimental output. Technically 4C requires to start from a 3C template followed by a second round of digestion and one additional ligation step what generates circularized DNA of which a fraction contains the viewpoint and the contacting sequence. Using PCR with viewpoint specific primers then allows to amplify the interacting DNA for high-throughput sequencing. In contrast, Hi-C was developed to generate all-versus-all interaction maps of genomic sequences. As for 3C and 4C, the initial steps of the protocol involve crosslinking of chromatin with formaldehyde and digestion. The major difference in Hi-C being that 5-prime overhangs generated by the restriction enzyme digest are filled in with biotinylated nucleotides. After blunt end ligation and chromatin sharing, this allows to enrich the ligation junctions by biotin pulldown. Next, DNA-molecules are sequenced paired-end what generates a matrix of read counts of pair wise interactions. This matrix is then

transformed in a color coded heat map to visualize genome wide all-versus-all interactions. However, to which extent the experimental output of 3C based methods, generated through crosslinking over variable distances and proximity ligation, really reflects contact probabilities has remained so far unclear.

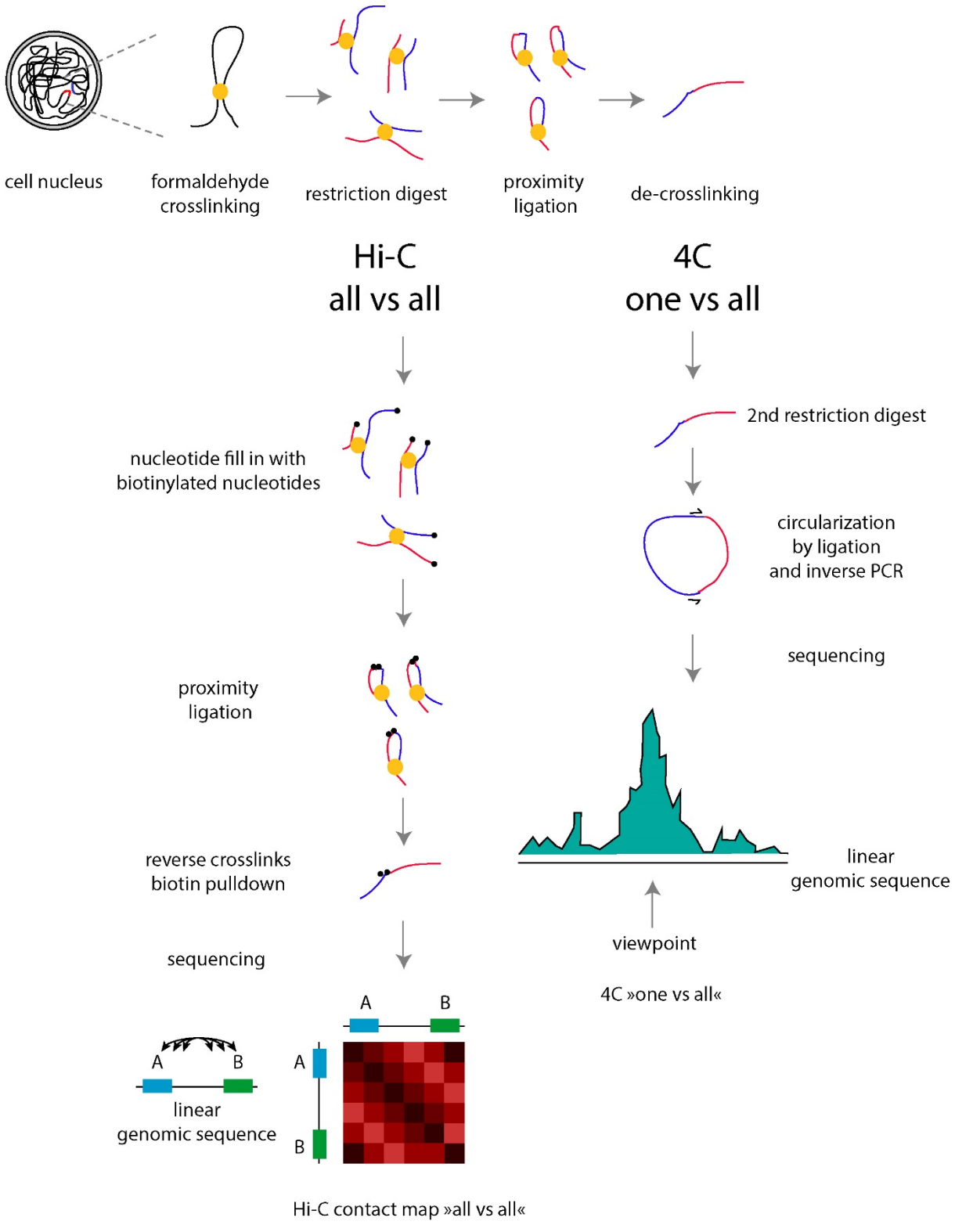

cell nucleus
formaldehyde crosslinking
restriction digest
proximity ligation
de-crosslinking
Hi-C
all vs all
4C
one vs all
nucleotide fill in with biotinylated nucleotides
2nd restriction digest
circularization by ligation and inverse PCR
sequencing
proximity ligation
reverse crosslinks biotin pulldown
sequencing
linear genomic sequence
viewpoint
4C »one vs all«
A
B
A
B
linear genomic sequence
A
B
Hi-C contact map »all vs all«

Figure 5 4C and Hi-C: The scheme illustrates formaldehyde crosslinking, restriction digest, proximity ligation and de-crosslinking which are common for 4C and Hi-C experiments. 4C involves a second round of digestion and one additional ligation step what generates circularized DNA. Using PCR with viewpoint specific primers then allows to amplify the interacting DNA for high-throughput sequencing. The experimental output of an 4C experiment allows to visualize crosslinking frequency of one genomic location (indicated as viewpoint) with all other sequences. In a Hi-C experiment, 5-prime overhangs generated by the restriction enzyme digest are filled in with biotinylated nucleotides. After blunt end ligation and chromatin sharing, this allows to enrich the ligation junctions by biotin pulldown. DNA-molecules are sequenced paired-end what generates a matrix of read counts of pair wise interactions. This matrix is then transformed in a color coded heat map to visualize genome wide all-versus-all interactions

As a consequence of increasing sequencing capacities and throughput, 4C, 5C and Hi-C showed that chromosomes are organized at different genomic scales following a hierarchical folding pattern. The first Hi-C studies revealed that chromatin interacts in an exclusive manner at the Mb scale separating active from inactive regions into preferential contact patterns that were called 'A' and 'B' compartments[47]. Subsequent studies using higher resolution Hi-C revealed that compartments are further segregated in sub-Mb domains often referred to as topologically associating domains (TADs)[48–50]. These self-interacting units in the genome manifest as squares along the diagonal of Hi-C contact maps indicating higher interaction frequency of genomic sequences within than across TAD boundaries. Accumulating genetic evidence shows that TADs function as regulatory neighborhoods which contribute to establish 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[51–55]. In agreement with this idea, the deletion of chromosomal regions that overlap with TAD boundaries, has also been shown to lead to ectopic gene expression in the neighboring domain which can be the cause of genetic disease or oncogene activation[51,54–56]. Their relevance is further highlighted by coordinated gene expression and DNA replication timing between TADs[57–59]. At an even shorter genomic scale, the dominant structural feature within TADs are ‘looping’ interactions which coincide with CTCF bound to DNA at the anchor sites[60]. This includes ‘stripe’-like structures which can be observed within TADs when clusters of CTCF sites map close to a cohesin loading site characterized by strong ChIP signal for the cohesin loading factor Nipbl[61].

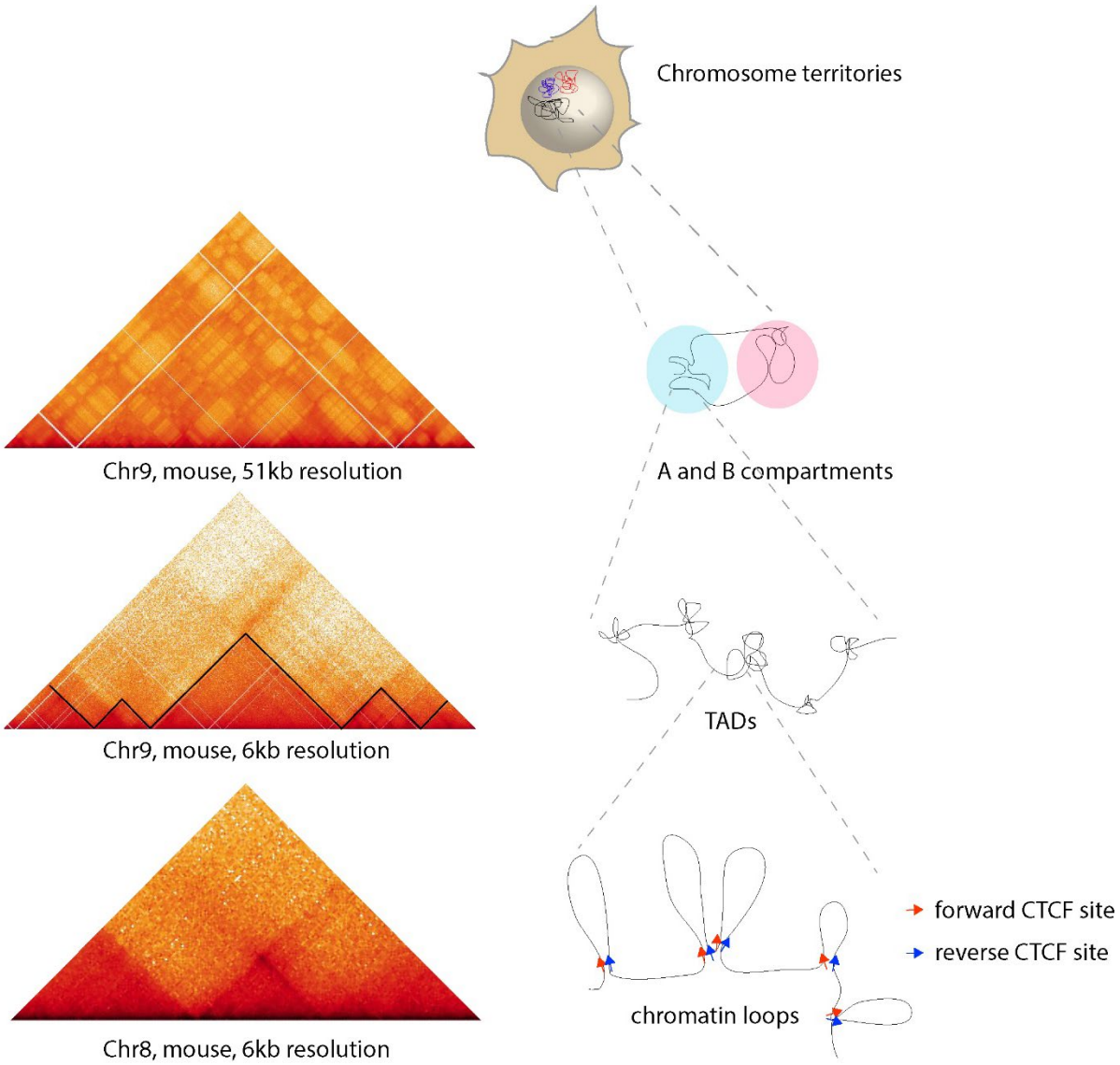


Figure 6 Genome folding is hierarchical: Chromosomes of mammalian genomes occupy distinct territories within the cell nucleus. Each chromosome further folds at different genomic length scales following a hierarchical folding pattern revealed by Hi-C. At the multi Mb scale, active regions cluster together in so called A compartments whereas inactive regions cluster in B compartments. Compartments are further portioned into TADs (denoted by solid black lines) where the chromatin fiber interacts more frequently with itself compared to any other region of the genome. Within TADs, chromatin loops, which form between convergent CTCF sites are the dominant structure.

Since looping interactions often occur between convergent CTCF motifs and based on simulations exploiting Hi-C data, it was proposed that a one-dimensional tracking mechanism rather than physical encounters through the three-dimensional space of the nucleus is at the basis of the observed phenomenology. According to the highly influential loop extrusion hypothesis, the cohesin complex and CTCF are major drivers of genome architecture. In this model, cohesin would function as a loop extruder progressively forming bigger loops until it is blocked by CTCF bound to DNA in a specific orientation[62,63].

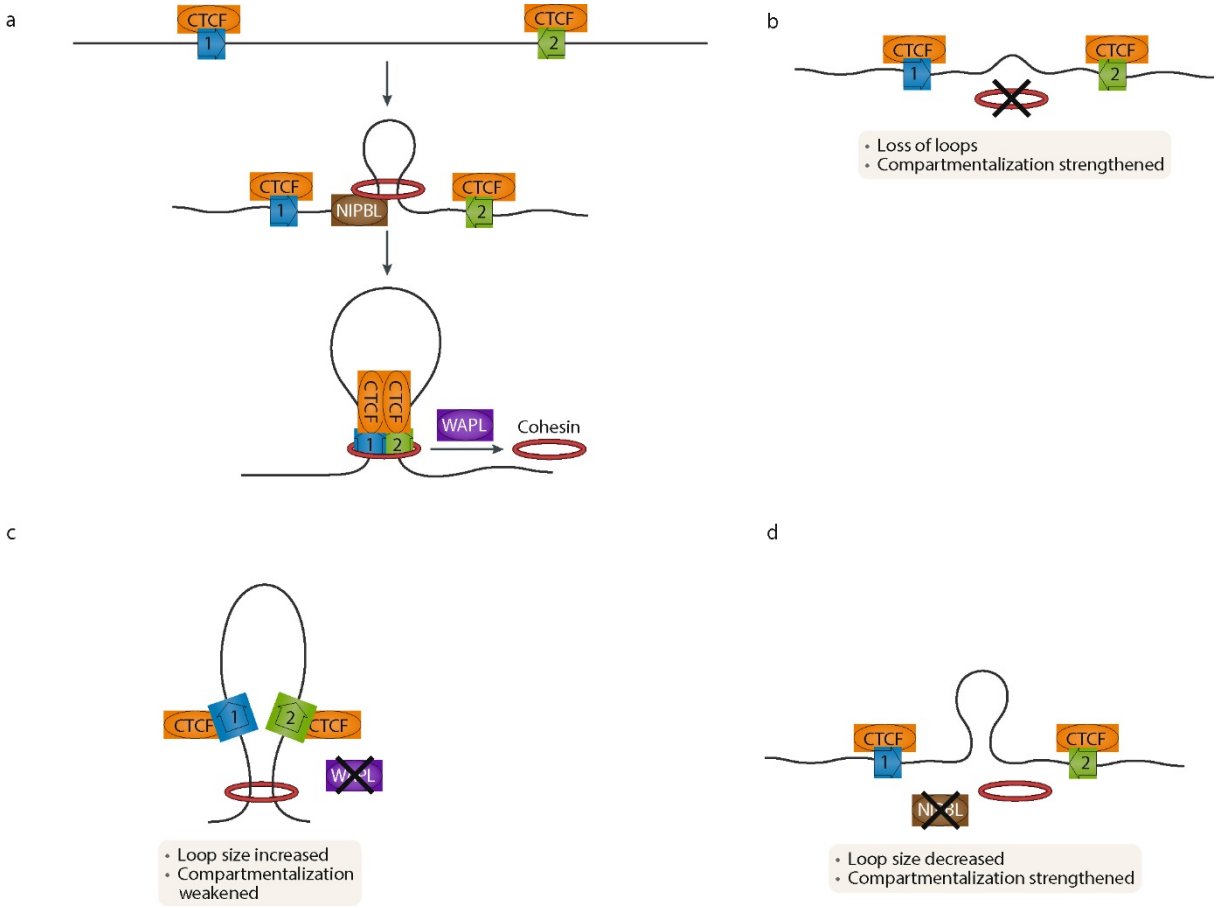


Figure 7: Loop extrusion model and TADs: **a** The loop extrusion model proposes cohesin to continuously extrude chromatin loops after being loaded to DNA by Nipbl. The extrusion of the growing chromatin loop stops when CTCF sites in a convergent orientation are encountered. **b** After degrading Rad21, a subunit of the cohesin complex using an auxin-inducible degron system, TADs and CTCF loops globally disappear in Hi-C contact maps whereas compartments are strengthened. **c** In cells where the cohesin unloading factor Wapl was deleted, chromatin loop size is increased whereas compartmentalization was weakened. This phenotype was suggested to be a consequence of the increased residence time of cohesin on DNA. **d** After depleting the cohesin loading factor Nipbl, loop size is decreased and TAD structure is substantially diminished whereas compartmentalization is strengthened. (This figure was adapted from Schoenfelder and Fraser Nat. Rev. Genet. 2019)

There are several lines of experimental evidence that support the idea of active loop extrusion by cohensin as the underlying mechanism of TAD formation. First, the inducible degradation of Rad21, which is a subunit of the cohesin complex, leads to disappearance of TADs and CTCF loops, a phenotype that is reversible when restoring cohesin[64]. A similar effect can be observed when the cohesin loading factor Nipl is deleted[65]. In line with this idea, it was shown that cells deleted for Wapl, a factor that unloads cohesin from DNA, have enlarged TADs and proliferation of looping interactions[66]. Additionally, the continuous extension of DNA loops driven by condensin, which like cohesin belongs to the structural maintenance of chromosomes (SMC) family, was recently imaged in-vitro[67]. Finally, the degradation of CTCF was shown to be essential for looping interactions between convergent CTCF binding sites but does not have an effect on compartmentalization of chromatin[68]. Interestingly, compartment size is increased when Nipl is deleted but decreased after Wapl deletion. This suggests that TAD formation and compartmentalization are driven by two independent mechanisms: loop extrusion and compartmental interactions which

possibly counteract each other. Given experimental evidence in favor of loop extrusion by cohesin, identifying CTCF as the determining factor that stalls cohesin activity, it should be possible to rewire contacts within TADs by introducing short ectopic sequences including CTCF binding motifs. This has not been demonstrated and would provide a genetic tool to study enhancer biology, either by tethering enhancers and promoters or insulating them from each other.

Despite the increasing amount and resolution of methodologies which helped to connect structural properties of chromosomes to function, many central questions regarding the causal links between the two remain unanswered. It is not clear how stable looping interactions are and what is their exact role in shaping enhancer-promoter communication. Even though CTCF and cohesin have been shown to be essential for the formation of site-specific chromatin loops, which seems to be key to partition chromosomal interactions into TADs, it is not yet understood if loops are directly involved in tethering regulatory elements together[60]. It is also not known how transcriptional variability is linked to interaction frequency of enhancers and promoters and if such interactions have a direct causal impact on transcription. Furthermore, it is possible that regulatory elements do not have to be in direct molecular contact but can communicate somehow through phase-separated compartments[69–71]. However appealing, this theory seems to lack any explanations for what could create specificity between interactions of regulatory elements, TFs and the transcriptional machinery.

1.3.3 Comparing 3C and DNA-FISH

Technical advances in both microscopy and sequencing-based 3C methodologies opened the way to explore the relationship between chromosome folding and biological function. Considering the far-reaching implications for our understanding of nuclear architecture and transcriptional control, it is important to discuss the synergies between different classes of techniques but also their limitations.

In DNA FISH experiments, fluorescent probes are used to visualize loci of interest in fixed cells exploiting complementary base pairing with the target regions. This allows to directly extract 3D distances between targeted loci in single cells as well as the associated distribution of distances in the population. In turn, useful statistical parameters like the average or median of 3D distances can be extracted. However, given the fact that physical proximity in DNA FISH experiments can only be defined using arbitrary distance thresholds and is biased as a consequence of the diffraction limit, it is not possible to infer if two loci are really in direct molecular contact or not.

In 3C and related methods like 4C, 5C and Hi-C, proximity ligation is used to join free DNA ends in crosslinked nuclei after digesting the genome with a restriction enzyme. When the DNA ends are close enough, hybrid molecules are generated followed by de-crosslinking and high throughput sequencing. Alignment of reads to the reference genome then results in a population averaged contact map which reflects relative crosslinking frequency between different genomic loci. Different from DNA-FISH, it is

not possible to measure physical distances between loci in 3C based assays since crosslinking of chromatin by formaldehyde occurs at distances that might range from a few to hundreds of nanometers[72,73]. In order to understand what is the relationship of crosslinking frequency and spatial proximity in 3C based methods, 3C-based data have been extensively benchmarked again DNA FISH experiments. Even though few studies reported discordant results, it was shown for a large number of loci pairs that physical distances measured by DNA-FISH decrease as an inverse power law of Hi-C reads from the same regions[74–77]. This provides evidence that 3C methods generally detect proximity events between genomic sequences. However, despite the increase of read counts in 3C based methods when the average distance between two loci decreases, it is important to bear in mind that DNA-FISH is not suitable to make conclusions on whether sequences are in direct molecular contact or not. This has important consequences on how 3C data should be interpreted. Especially since several studies postulated mechanistic models of chromosome folding based on the assumption that read counts in Hi-C are directly proportional to actual contact probabilities[62,63]. Surprisingly a formal proof of this assumption has never been provided. It is also important to note that proximity in 3C based methods is averaged across the cell population, not excluding the possibility of very different conformational states between individual cells. The recent improvement of super-resolution microscopic techniques using multiplexed DNA FISH probes to measure distances between sequences provide an excellent complement to this limitation of 3C based methods[42,75]. Using this methods will allow to dissect the relationship of transcriptional activity and single cell conformation at higher throughput. It will be interesting if

contacts of regulatory elements generally occur in the context of fluctuating conformations as it was shown in earlier experiments for a single locus in mESC combining polymer simulations with lower-throughput DNA- and RNA-FISH [78].

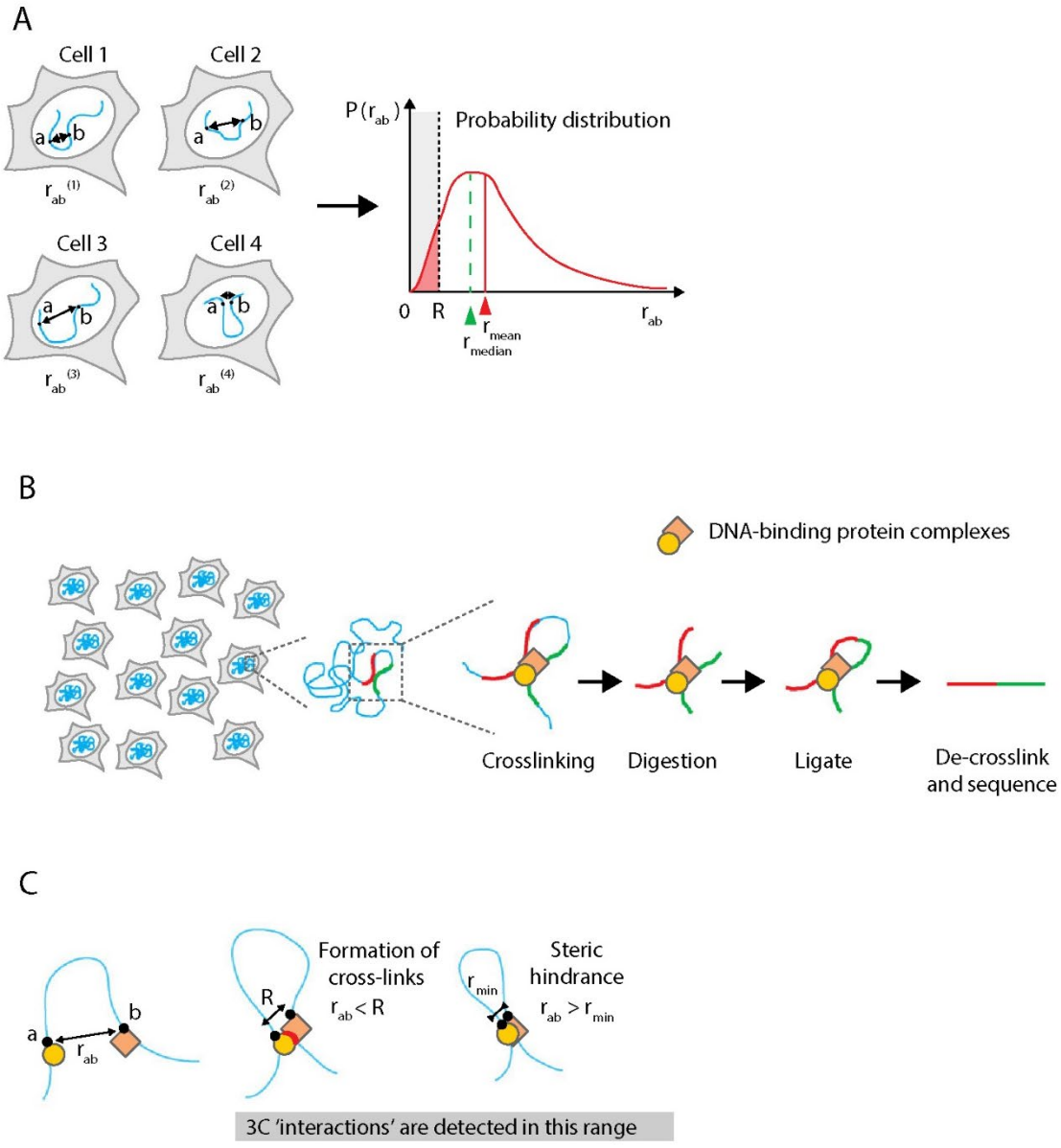


Figure 8 Comparing DNA-FISH and 3C: **A** DNA-FISH experiments allow to extract 3D distances between two loci a and b (r_{ab}) in single cells. This provides important information on the distribution of distances ($P_{(rab)}$) across the population. Additionally, the mean and median 3D distances between the loci a and b can be calculated. The red shaded area reflects a fraction of the population in which the distance between a and b is smaller than a certain threshold (R). **B** In 3C based experiments, only DNA fragments which are in sufficient physical proximity can be crosslinked via protein bridges. Following digestion and re-ligation, the hybrid DNA-molecules emerge from 3D proximity events of DNA sequences across the cell population. **C** A signal in 3C experiments can only be obtained if the distance r_{ab} is smaller than the locus specific crosslinking range R and bigger than a minimum distance r_{min} which arises from steric repulsion of a and b if they are too close to each other. It is important to note that the same signal in 3C experiments can arise from various distributions of distances across the cell population. (This figure was modified from Giorgetti and Heard Genome Biol. 2016)

1.3.4 Orthogonal methods to study genome folding

Recently several other sequencing based methods which are orthogonal to DNA-FISH and 3C techniques provided further evidence for structural features of chromosomes including compartments, TADs and CTCF loops. Native 4C and Hi-C, do not involve crosslinking in the experimental protocol but still detect CTCF mediated looping interactions and TADs[79]. Two other techniques, notably genome architecture mapping (GAM) and split-pool-recognition of interactions by tag extension (SPRITE), are based crosslinking but do not rely on ligation. In GAM, contact maps including compartments, TADs and inter-TAD interactions were generated from a large number of ultrathin nuclear cryosections combining laser microdissection and whole genome sequencing[80]. In SPRITE, cells are crosslinked followed by nuclei isolation, chromatin fragmentation and barcoding of individual crosslinking complexes using a split-pool strategy[81]. Interactions of the same crosslinking complex are then detected by

assigning the reads to the matching barcode. Besides confirming pairwise structures identified by Hi-C, SPRITE provided evidence for simultaneous interactions of multiple DNA regions in the same cell and identified clusters of long-range interactions that might be mediated through sub-nuclear compartments. However, key structural features of chromosomes including TADs and CTCF loops have not been validated in living cells, since these methods still involve chemical fixation of cells or proximity ligation to detect chromosomal interactions.

2. Aim of the thesis

As discussed in the introduction, chromosome conformation is thought to play an important role in regulating gene expression by instructing the physical interactions between promoters and enhancers. Existing methods like DNA-FISH and 3C techniques have been important tools in understanding how chromosomes are organized within the nucleus and what are the consequences of 3D genome organization on gene expression. Despite the importance of this findings and the implications on gene regulation by cis-regulatory elements, a major limitation of these methods is that they rely on chromatin crosslinking (and proximity ligation in the case of 3C-based techniques), two steps that were extensively discussed as sources of experimental biases in the detection of chromosomal interactions. Given the limitations of these methods, the aim of the thesis was to provide quantitative measurements of chromosomal interaction frequencies relying on direct enzymatic, molecular-range protein-DNA contacts in living cells. These measurements have important consequences on how 3C based data should be interpreted and used, notably to establish quantitative models of chromosome folding and its molecular determinants.

The main questions addressed in my PhD project are:

- Do structural feature of chromosomes detected by 3C-based methods such as TADs and CTCF loops exist in living cells?
- Does formaldehyde crosslinking distort the detection of chromosomal interactions?

- Are crosslinking frequencies in 3C based experiments quantitatively proportional to actual contact probabilities between DNA sequences?
- Is it possible to rewire contacts within TADs by adding ectopic CTCF sites, and how does this impact on actual contact probabilities?

3. Results

Contribution

I performed proof-of-principle experiments, cloned the majority of plasmids, generated cell lines involved in the study and performed DamC experiments. I wrote the paper on DamC with Luca Giorgetti and Yinxiu Zhan.

Yinxiu Zhan, who analyzed the data and wrote the model of methylation kinetics, and Christian Valdes, who performed 4C experiments, contributed equally to this work.

3.1 Published manuscript: DamC reveals principles of chromatin folding in vivo without crosslinking and ligation

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

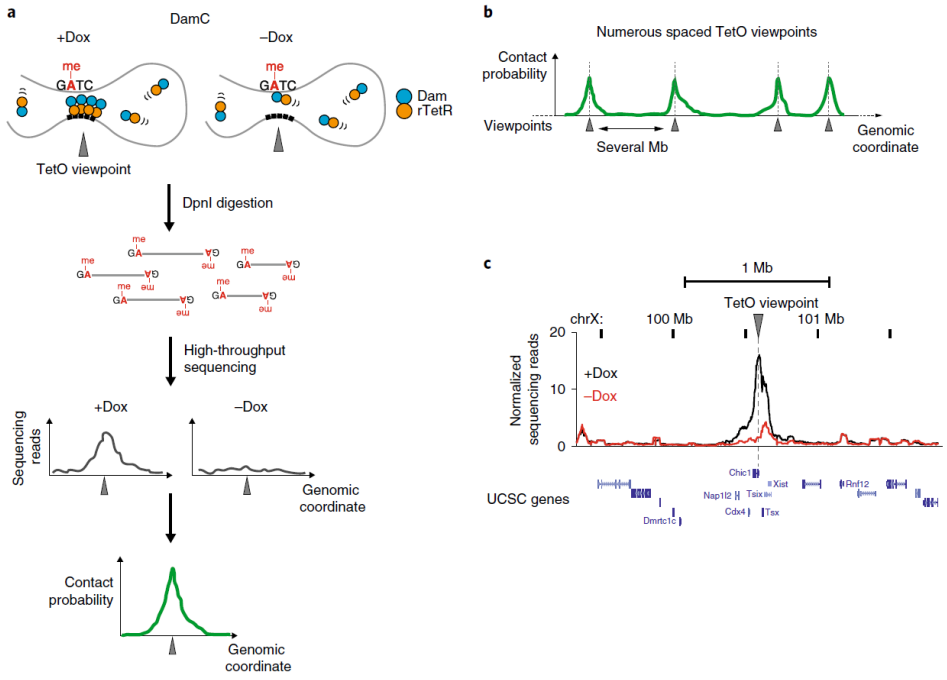


Fig. 1 | DamC: methylation-based measurement of chromosomal interactions. a, Scheme of DamC experiments. In the presence of doxycycline (+Dox), rTetR-Dam binds to a genomic viewpoint through a TetO array and methylates adenines in GATC sites that contact the viewpoint. In the absence of doxycycline (–Dox), 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.

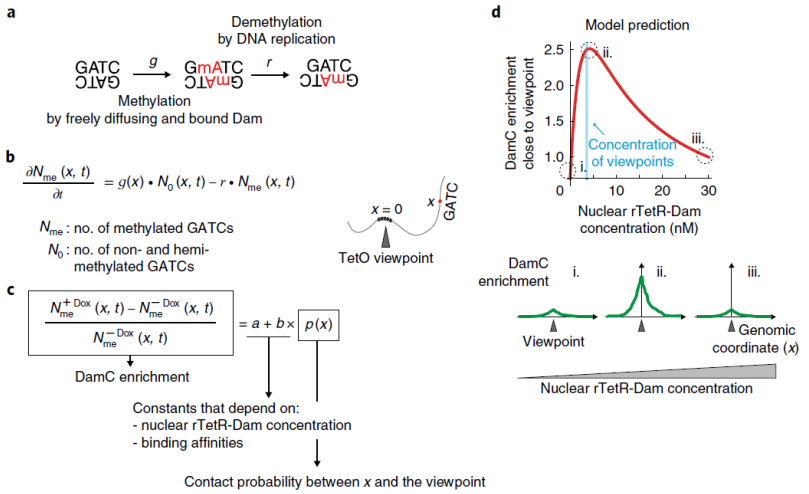


Fig. 2 | Physical model of methylation dynamics. a, Unmethylated GATCs can be methylated by either freely diffusing or TetO-bound rTetR-Dam at rate *g*, and partially demethylated during DNA replication at rate *r*. Partially demethylated GATCs are inefficiently cut by DpnI and do not contribute to the 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

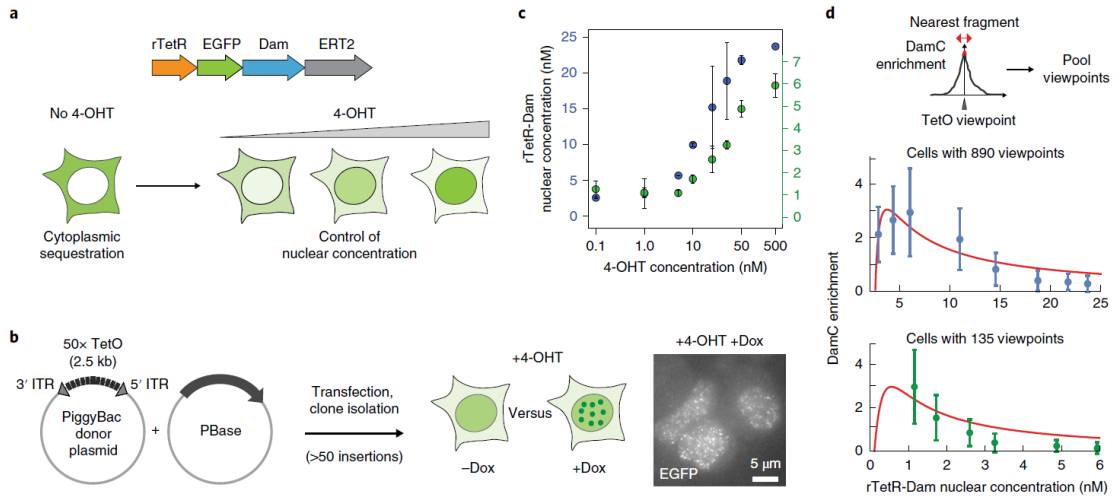


Fig. 3 | An inducible mESC line used to perform DamC and test the model predictions. a, mESCs expressing rTetR-Dam-EGFP-ERT2 allow control of the nuclear concentration of the fusion protein by changing the amount of 4-OHT in the culture medium. **b**, Nuclear concentration of the rTetR-Dam fusion protein 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

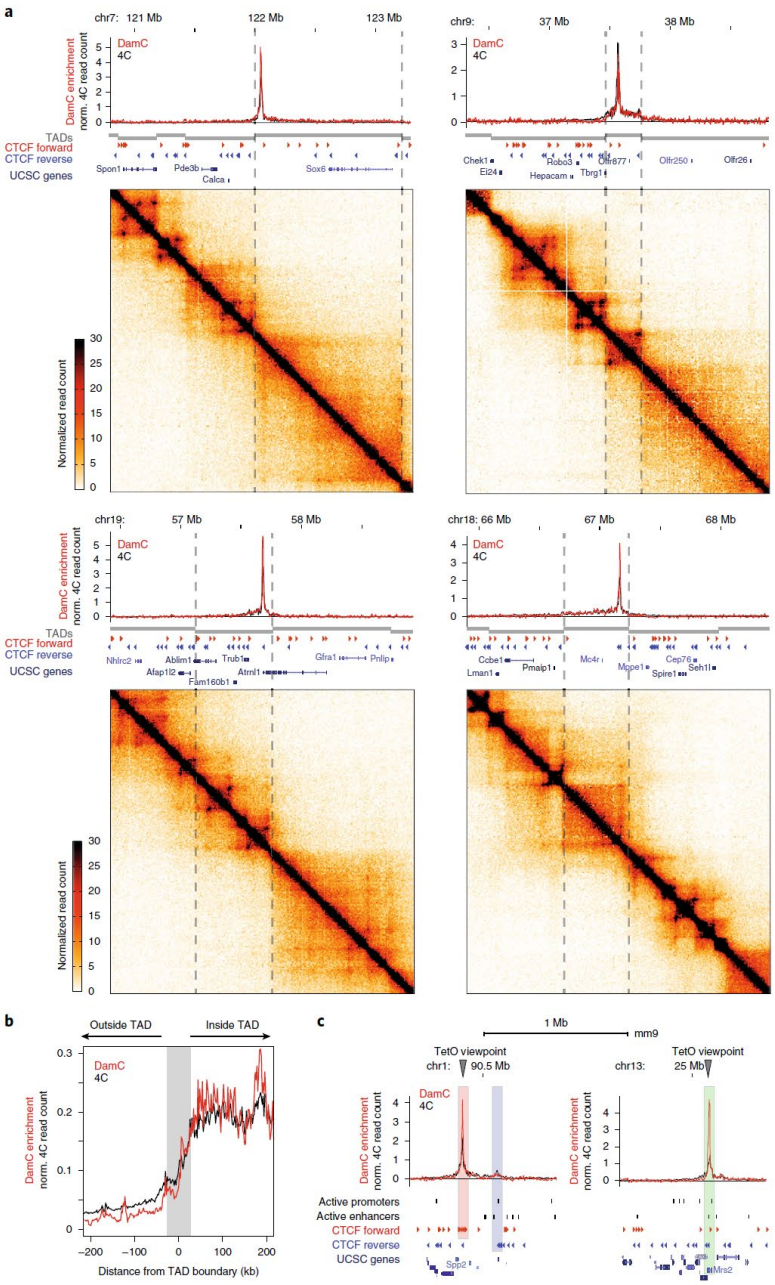


Fig. 4 | DamC confirms the existence of TAD boundaries and quantitatively correlates with 4C and Hi-C. a, Four representative DamC and 4C interaction profiles from the same piggyBac-TetO viewpoints, aligned with Hi-C experiments performed in the same cell line. Dashed lines mark TAD boundaries in mESC, 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.

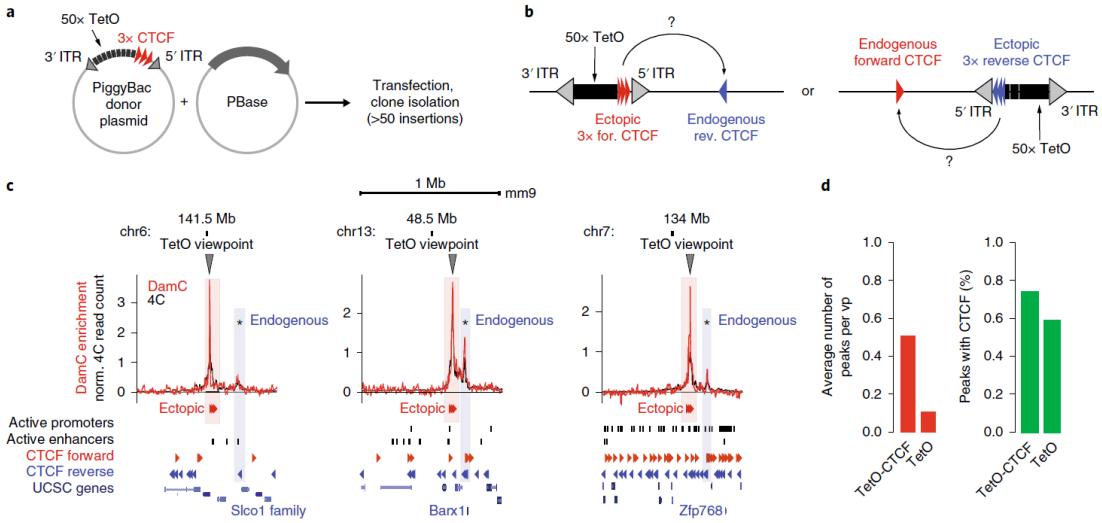


Fig. 5 | DamC-based detection of CTCF loops. a, Modified piggyBac strategy used to insert TetO viewpoints flanked by three CTCF sites oriented outwards. **b**, The TetO-CTCF cassette can insert in the genome in either direction and lead to the formation of interactions with either forward (for.) or reverse (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

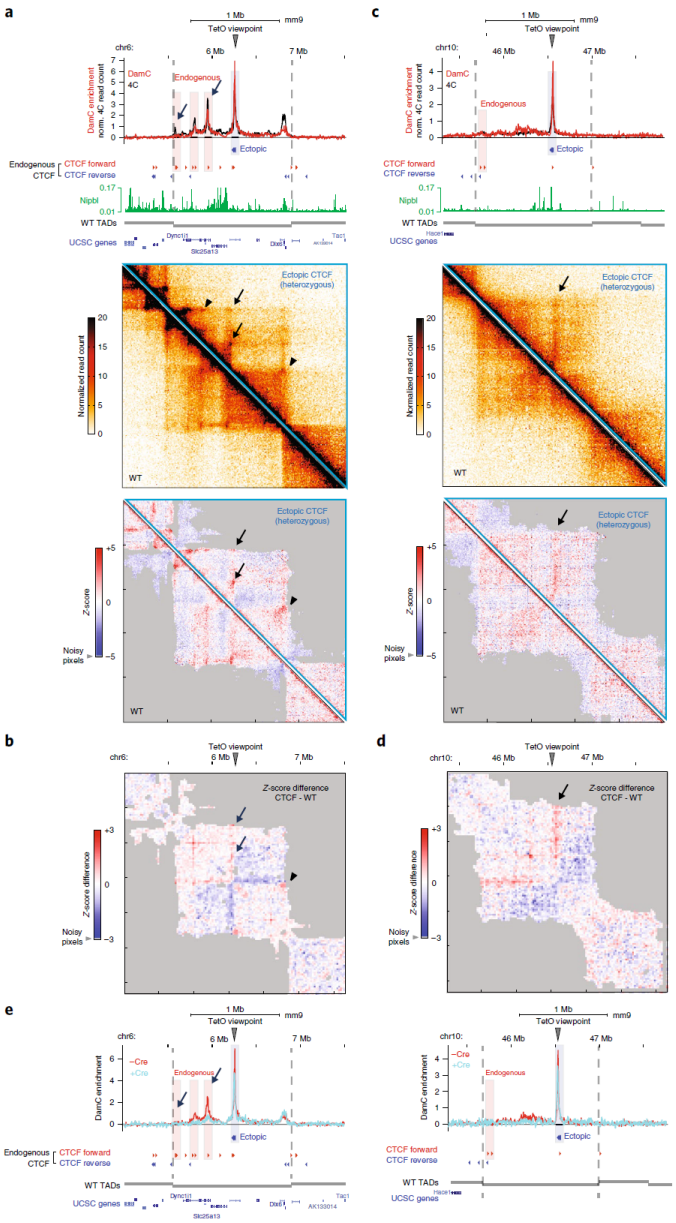


Fig. 6 | Ectopic CTCF insertion leads to the formation of new loops and stripes. a, TetO-CTCF insertion site giving rise to ectopic loops with convergently oriented endogenous CTCF sites. Top, interaction profiles measured with DamC and 4C are overlaid with the position of CTCF ChIP-seq sites from ref. [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).

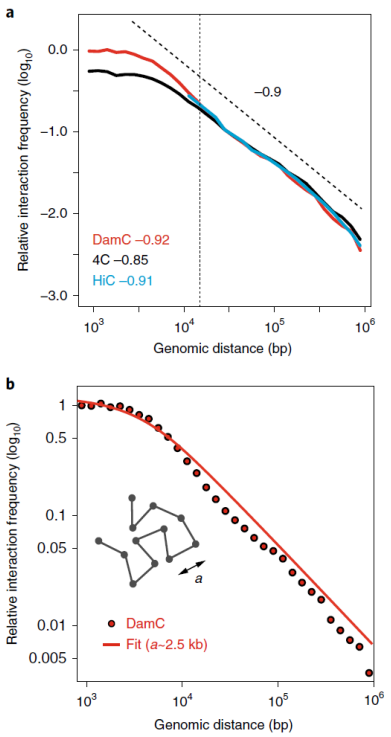


Fig. 7 | Scaling analysis of contact probabilities in vivo. a, Scaling of contact probabilities measured in DamC, 4C and Hi-C from all 130 TetO and 91 TetO-CTCF viewpoints. Power-law fitting was performed between 15 kb (dotted vertical line) and 1 Mb. **b**, Best fit of scaling behavior measured by DamC 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.

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[58] 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 analyis.

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]ATCTATAACAAGAAAATATATATAATAAGTTA
TCACGTAAGTAGAACATGAAATAACAATATAATTATCGTATGAG
TTAAATCTTAAAAGTCACGTAAAAGATAATCATGCGTCATTT,

ITR3-2 [Btn]TCCAAGCGGCGACTGAGATGTCCTAAATGCACAG
CGACGGATTCGCGCTATTTAGAAAGAGAGAGCAATATTTCAAG
AATGCATGCGTCAATTTTACGCAGACTATCTTTCTAGGGTTAA,

ITR5-1[Btn]TTAACCCTAGAAAGATAATCATATTGTGAC
GTACGTTAAAGATAATCATGCGTAAAATTGACGCATGTGT
TTTATCGGTCTGTATATCGAGGTTTATTTATTAATTTGAA,

ITR5-2 [Btn]ATTAAGTTTTATTATATTTACACTTACATACTAATAATAA
ATTCAACAAACAATTTATTTATGTTTATTTATTTATTAAAAAAAAACAAAA
ACTCAAAATTTCTTCTATAAAGTAACAAA.

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× 0.227× 0.73 µm^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-cm² 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 mm² 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′-AATGATACGGCGACCACCGAGATCTACACNNNNN NNNACACTCTTTCCCTACACGACGCTCTTCCGATC*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, chr8:96719438, 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 research

Corresponding author(s): Luca Giorgetti

Last updated by author(s): Apr 15, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo Xcalibur 4.1 was used to collect mass spectrometry data.
Data analysis	The following software was used to analyse mass spectrometry data: Maxquant version 1.5.3.8 (Cox et al, Genome Res. 27, 479–490 (2017)) Perseus version 1.5.0.0 (Tyanova et al, Nat. Methods 13, 731–740 (2016)) MaxLFQ (Cox et al, Mol. Cell. Proteomics 13, 2513–2526 (2014)) Skyline 4.135 (MacLean et al, Bioinformatics 26, 966–968 (2010)). The following software and code were used to analyse sequencing experiments: QuasR version 1.20.0 (Gaidatzis et al Bioinformatics 31, 1130–1132 (2015)) HiC-Pro version 2.7.10 (Servant et al,Genome Biology 16, 259 (2015)) CaTCH version 1.0 (Zhan et al, Genome Res. 27, 479–490 (2017)) ChromHMM version 1.14 (Ernst and Kellis, Nat. Methods 9, 215–216 (2012)) PeakC version 0.2 (Geeven et al, Nucleic Acids Res. doi:10.1093/nar/gky443) Mathematica version 11.3 Matlab version 9.6.0.1072779 csaw version 1.14.1 (Lun et al, Nucleic Acid Research (2016)) Additional custom scripts in R and bash are available upon request

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO GSE128017; http://www.ncbi.nlm. nih.gov/geo), including bedgraph files for the visualization of damC and 4C profiles from all the samples described in the manuscript. A UCSC session has been created containing all the damC and 4C tracks used in the experiment and can be found at https://genome.ucsc.edu/s/zhan/damC_publication_2019

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	damC, 4C and Hi-C experiments were performed at least in two biological replicates. Read sequencing depth in 4C was determined based on Splinter et al., Methods 58, 221–230 (2012). Hi-C samples were sequenced to a depth allowing to exclude that data at the genomic length scale of TADs (~1Mb), which is relevant to this study, are affected by large number of zeros. Based on scaling analysis, read statistics are zero-inflated starting from ~10Mb . In damC, read depth was chosen so that biological duplicates have 0.8 correlation between enrichments calculated in the first +/-100 kb from TetO viepoints (see Supplementary Figure 3a). Detailed read counts for all experiments are provided in Supplementary Figure S6.
Data exclusions	No data were excluded.
Replication	All experiments were performed at least in two biological replicates.
Randomization	Randomization of data was done only in Suppl Figure 5b, where 10'000 random genomic viewpoints were used to generate a control distribution of deviation scores.
Blinding	Blinding was not necessary. Analyses of high-throughput sequencing data was based on quantitative measures (e.g. deviation scores defined in Suppl. Figure 3) that are computed using bioinformatic algorithms.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All cell lines are based on feeder-independent PGK12.1 female mouse embryonic stem cells (mESC), kindly provided by Edith Heard's laboratory. The founder cell line in our study is an X0 sub-clone of the PGKT2 10 clone described in (Masui et. al. Cell 2011), carrying the insertion of a 256x TetO array within the 3'UTR of the Chic1 gene on chromosome X and the additional deletion of the Linx promoter (Nora et al Nature 2012)

Authentication	Heard et al have authentificated PGK cells by genomic sequencing
Mycoplasma contamination	Cells were tested for mycoplasma contamination once a month. No contamination was detected.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

nature structural & molecular biology

SUPPLEMENTARY INFORMATION

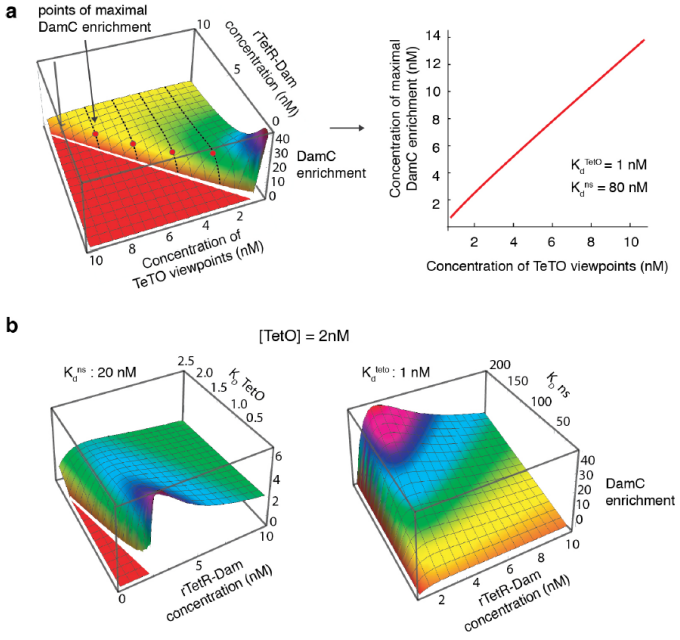
https://doi.org/10.1038/s41594-019-0231-0

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]*

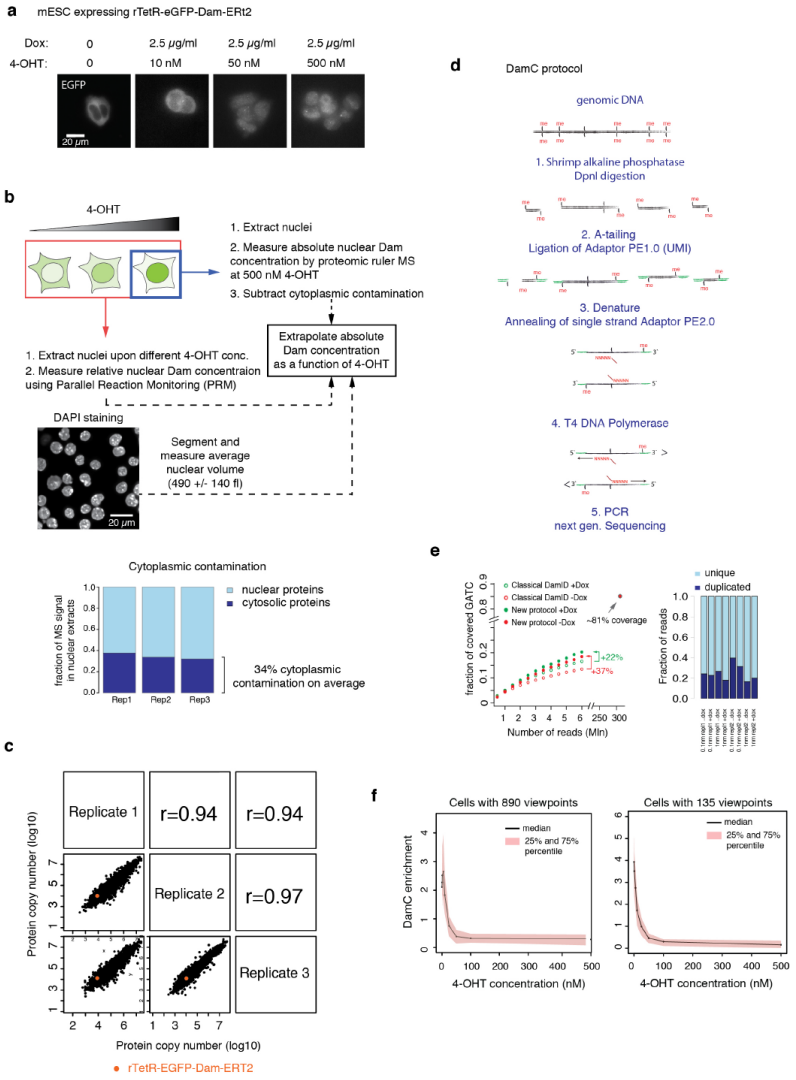
[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Oncode Institute, Hubrecht Institute–KNAW, Utrecht, the Netherlands. [4]University Medical Center Utrecht, Utrecht, the Netherlands. [5]EMBL, Heidelberg, Germany. [6]Department 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).

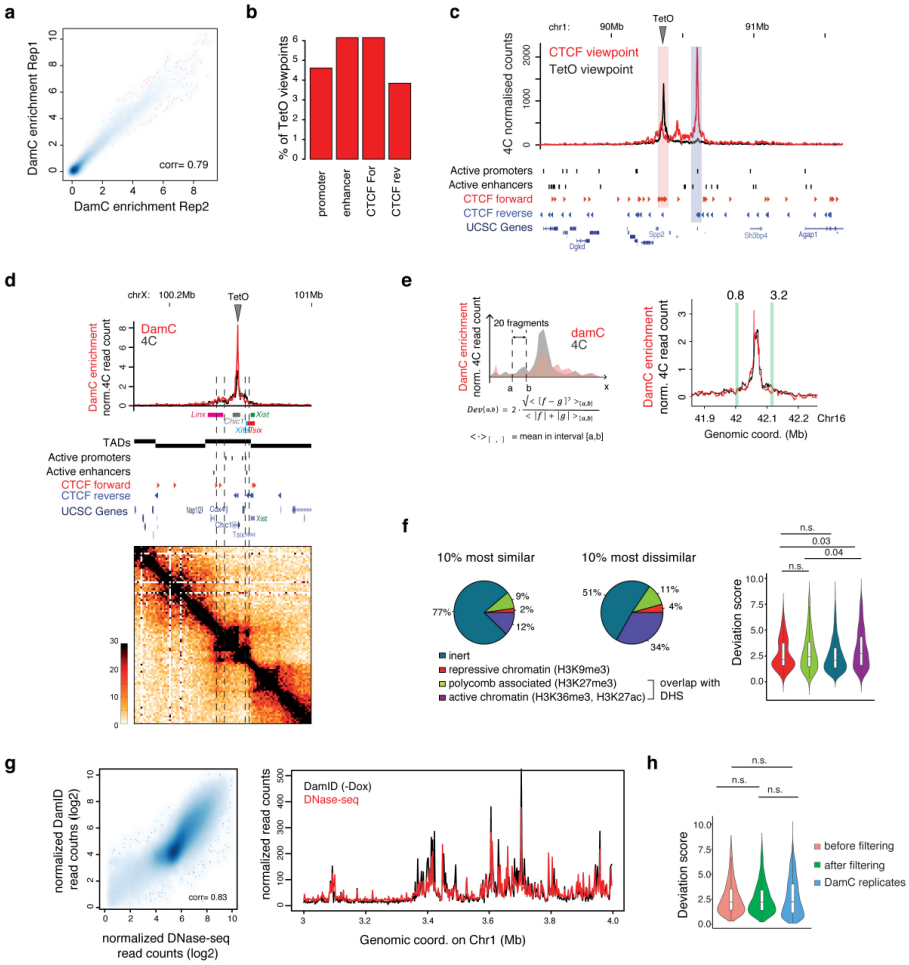


Supplementary Figure 2

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.

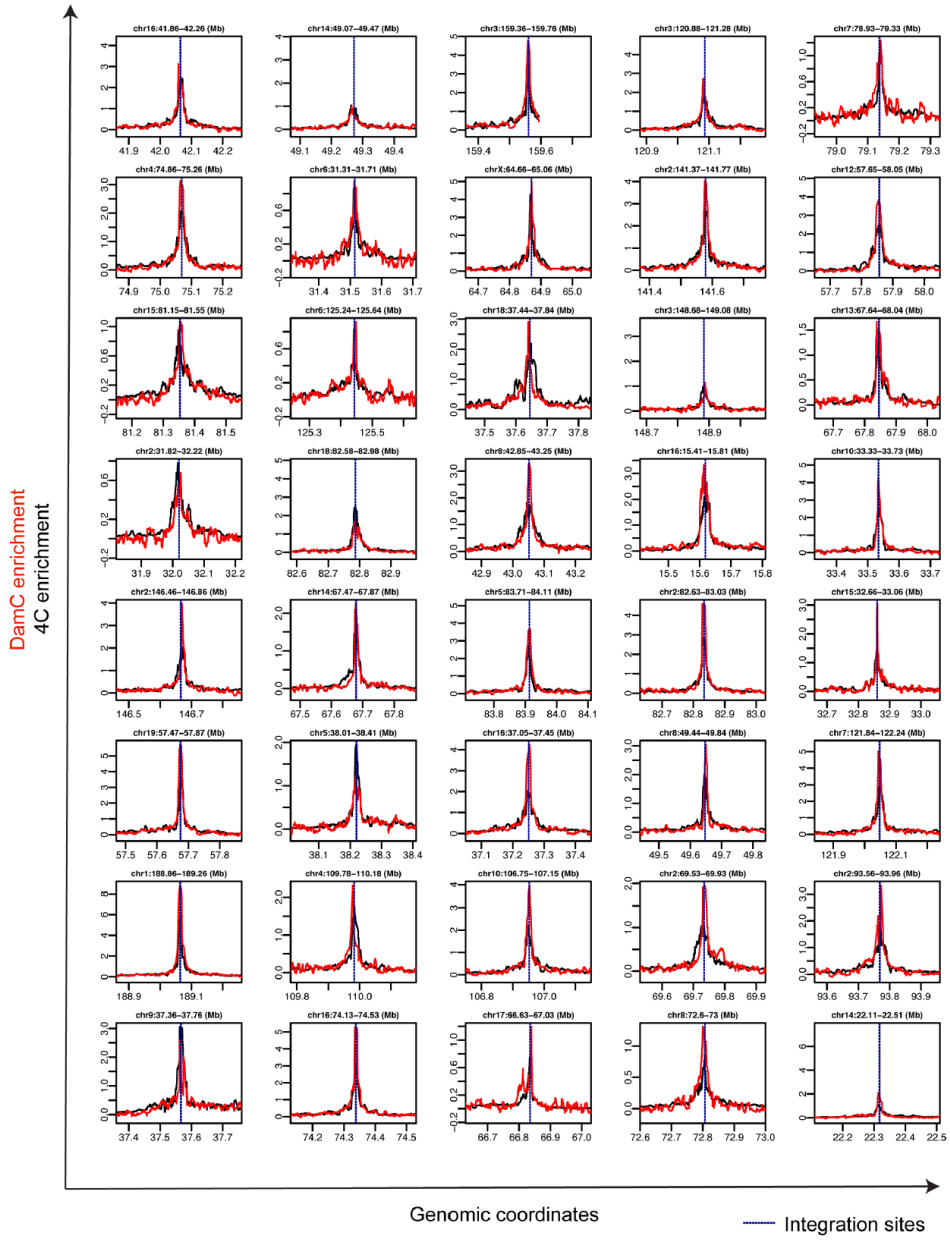


Supplementary Figure 3

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)

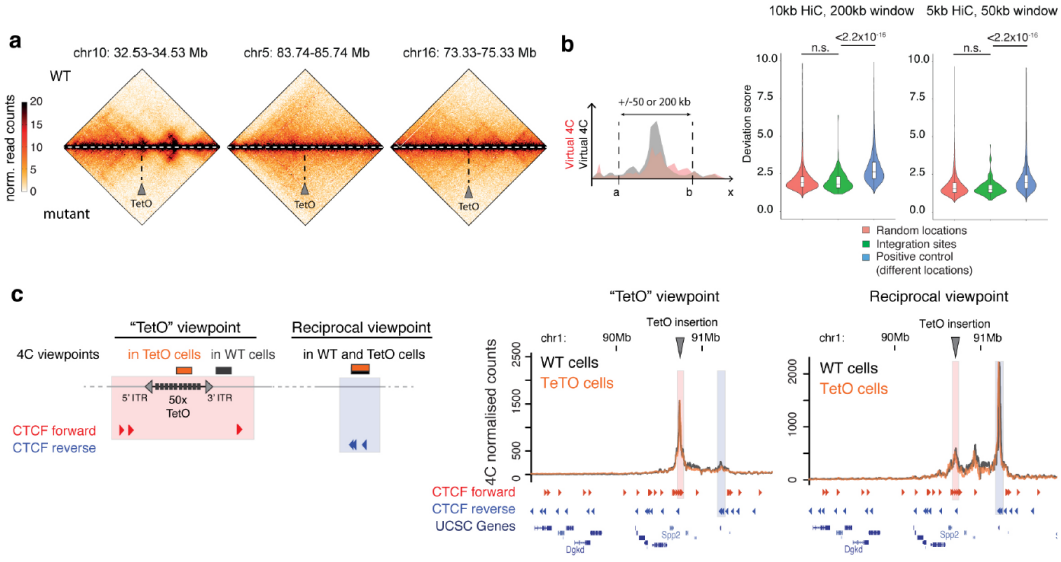
(Chi-Square Test: pvalue < 10^{-9}). 'Inert' corresponds to chromatin that is not enriched in H3K9me3, H3K27m3, H3K36me3, H3K9ac, nor H3K27ac. See the Methods section for more details. Right: The distributions of deviation scores in 20-fragment intervals where the dominant ChromHMM state is either inert, 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.

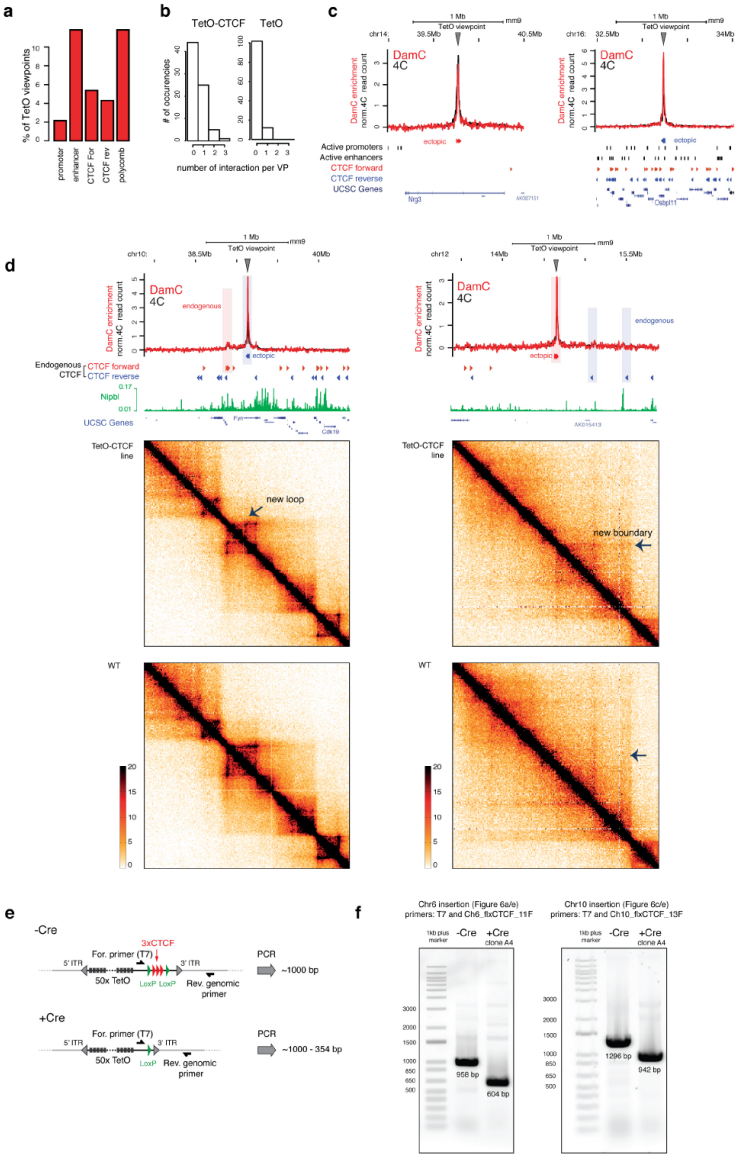
DamC (red) and 4C (black) profiles from forty TetO viewpoints in the pure clone with 135 TetO insertions.



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.

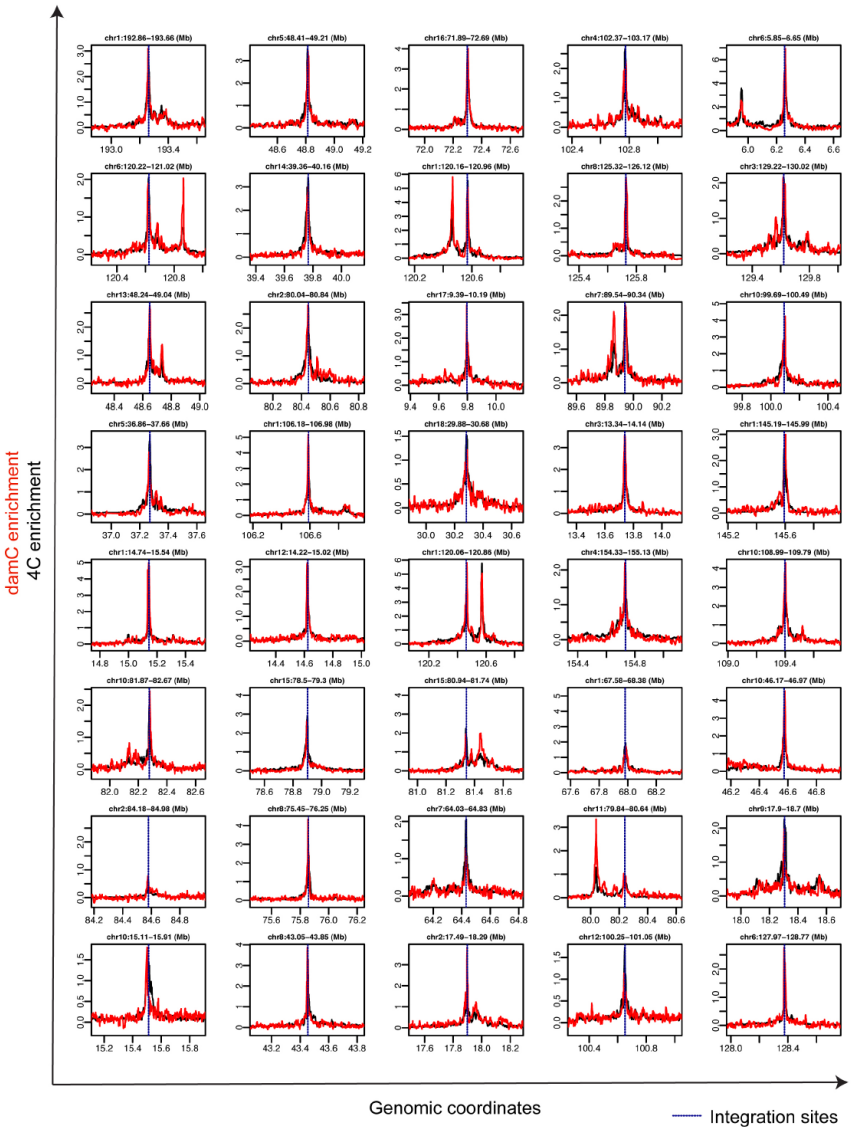


Supplementary Figure 6

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

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**).

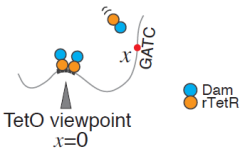


Figure M1.1. Scheme of the DamC physical model. The GATC located at x is methylated by freely diffusing and TetO-bound rTet-Dam.

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 plus 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 ($\lesssim$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**.

Determination of the GATC demethylation rate

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 detected in DamC effectively decreases in time with rate r, irrespectively of the phase of the cell cycle when every adenine has been methylated.

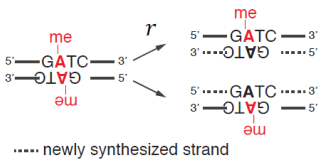


Figure M1.2. Effective demethylation rate in DamC. Methylated GATC sites are replicated into two hemi-methylated motifs, which are cut at very low frequency by DpnI compared to fully methylated GATCs. Hemi-methylated sites are thus essentially not detected in a DamC library and the effective demethylation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

A system of ordinary differential equations describing GATC methylation dynamics

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:

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

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

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

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

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

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:

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

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

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

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

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

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:

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

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

$$\begin{aligned} S(x,T) &\approx W \frac{g_s(x)}{2r + g_s(x)} \\ B(x,T) &\approx W \frac{g_b(x)}{2r}(1 - e^{-2rT}), \end{aligned} \tag{10}$$

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:

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

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:

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

and thus

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

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**:

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

whereas for case 2,

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

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 <u>Case 3</u>** and equation (16) can be rewritten in the form

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

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**.

<u>Generalization: depletion of freely diffusing rTetR-Dam in +Dox conditions</u>
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

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

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

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

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

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

This can be rewritten in the form

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

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:

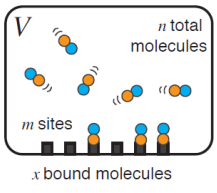


Figure M1.3. Scheme for the calculation of the concentration of freely diffusing rTetR-Dam in the presence of Dox.

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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**.

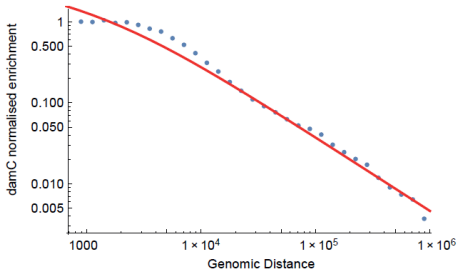


Figure M2.1 Black dots: Relative interaction frequency (normalized DamC enrichment) as in main Figure 3e-d. Red line: Fit to the data using Eq. (4).

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 $1-\exp[-x/x_0]$, 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

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

Using Eq. (1) we obtain

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

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.

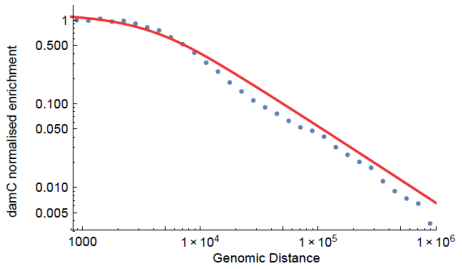


Figure M2.2 Black dots: Relative interaction frequency (normalized DamC enrichment) as in main Figure 3e-d. Red line: Fit to the data using Eq. (6). Fit done using NonlinearModelFit in Mathematica.

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.

4. Discussion

Binding of TFs to DNA, which is both regulated by and determines chromatin accessibility, coordinates into cell type specific gene expression patterns. However, beside *trans* acting factors, temporal and spatial control of transcription also requires the coordinate action of *cis*-regulatory sequences such as enhancers that communicate with promoters over large genomic distances and integrate complex regulatory information[29]. Thanks to DNA-FISH and 3C methods, we know that these interactions occur in context of hierarchically folded chromosomes which occupy distinct territories that further organize into A and B compartments, TADs and chromatin loops[47–50,60]. These findings had major impact on our current understanding of genome folding and how it relates to biological function. TADs, which are self-interacting sub-megabase regions in the genome, have been suggested to limit the genomic search space of enhancers by preventing interactions across their boundaries thereby contributing in establishing the correct interaction patterns between *cis*-regulatory sequences. This concept is supported by genetic experiments, were deletions of DNA-elements containing boundary regions resulted in ectopic gene expression in the neighboring domain[51,52,54,55]. Furthermore, TADs are evolutionary conserved across species and invariant between cell types highlighting their functional relevance[82]. Based on Hi-C experiments and polymer simulations, it was proposed that DNA loop extrusion by cohesin is the underling mechanism of TAD formation. The model suggested that cohesin continuously creates chromatin loops until convergent

CTCF sites are encountered, thereby stalling its activity and stabilizing the looping conformation[62,63]. This is in line with two independent studies showing that inducible degradation of the cohesin subunit Rad21 or CTCF respectively, results in disappearance of TADs in Hi-C contact maps[64,68]. Concordant with this, deletion of the cohesin loading factor Nibpl leads to a similar phenotype whereas deletion of the cohesin unloading factor Wapl manifests in enlarged TADs[65,66].

These findings have far-reaching implications for our understanding of genome folding and its functional connection to enhancer-promotor communication. It should be noted however that discoveries in this field have been fueled by 3C methods and DNA FISH, which rely on chromatin crosslinking (and proximity ligation in the case of 3C-based techniques), two steps that were discussed as sources of experimental biases in the detection of chromosomal interactions[74,83,84]. By nature, 3C techniques cannot distinguish between direct molecular contacts or indirect contacts of sequences through macromolecular protein complexes what raised the question if crosslinking frequency is proportional to contact probability. Additionally, TADs and CTCF loops have never been validated in living cells.

My PhD work has provided systematic measurements of chromosomal interactions at high genomic resolution using DamC, a new method relying on an enzymatic readout of molecular-range protein-DNA contacts in living cells. Importantly, one essential feature of the method is that results are directly proportional to contact probabilities. This is supported by rigorous modeling of Dam methylation kinetics setting a theoretical basis to interpret sequencing results in a quantitative way. A

systematic comparison of DamC with 4C and Hi-C experiments from the same cells, shows that the relative average contact probability drops by a factor of 2 across TAD boundaries in both methods. This raises the interesting question of how such a small differences in relative interaction frequency can sufficiently insulate enhancers from off-target promoters. In order to understand the relationship of enhancer function and interaction frequency, static population averaged contact profiles will not be sufficient. Answering this question requires methods which can resolve single cell conformations over time and allow to understand if contacts are functional or not. In preliminary experiments performed in collaboration with Jop Kind, we could obtain single cell DamC contact profiles. However, technical noise did not allow to extract single cell conformations at a genomic scale which could resolve looping interactions, albeit pooling reads from 1000 cells recapitulated chromosomal interactions detected in population based DamC experiments. This shows that DamC can measure chromosomal interactions within TADs starting from a low number of cells, what is currently not possible using 3C based methods.

Since DamC relies on the methylation of adenines within GATC motifs, requiring Dam to directly bind to DNA, it detects chromosomal interactions at short spatial distances. More specifically, the detection range of contacts by DamC should be less than 10 nm provided a physical size of the rTetR-EGFP-Dam-ERT2 fusion protein of approximately 3 nm[85]. Given the similarity to 4C profiles and Hi-C data suggests that formaldehyde crosslinking in 3C based methods does not substantially distort the detection of chromosomal interactions. Additionally, DamC also confirms the scaling of crosslinking

probabilities which were previously measured using 3C based methods. This was reassuring, since the scaling exponent detected by Hi-C was extensively used to build models of chromosome folding from which mechanistic determinants, including active loop extrusion by cohesin, have been inferred. Since the enrichment in DamC, which was formally shown to be proportional to actual short range contact probabilities, confirms the scaling exponent measured in 4C and Hi-C, it provides strong evidence in favor of such models. This is an important validation of 3C based findings given that the experimental readout of 3C based methods is a convolution of crosslinking over variable distances and ligation efficiency that was not shown to be proportional to contact probabilities before.

DamC could also confirm the existence of TADs, reflected by a sharp drop of contact probabilities across TAD boundaries, and CTCF mediated loops *in-vivo*. The finding that new loops can be established by inserting short sequences including ectopic CTCF binding motifs further shows that chromosome structure can be manipulated by the addition of new structures. It was interesting to note that these structures, notably loops and stripes, only formed within TADs, not violating existing boundaries. This suggests that TAD boundaries, which are enriched in clusters of CTCF binding sites, provide efficient physical barriers for loop extrusion. It will be interesting to study the functional consequences of either tethering enhancers to promoters or insulating them from each other by introducing ectopic binding sites for architectural proteins like CTCF in the genome.

A disadvantage of DamC preventing its adoption as an alternative for 3C based methods in routine experiments, is the need for genetic engineering of cells before experiments can be carried out. On the one hand, genomic viewpoints have to be integrated at the regions of interest, and on the other hand rTetR-Dam-ERT2 has to be stably expressed to ensure a tight control of nuclear Dam concentration. However, future developments could make the integration of viewpoints redundant. By either fusing Dam to catalytically inactive CRISPR associated protein 9 (dCas9) or transcription activator-like effector nucleases (TALEN) proteins, it should be possible to recruit Dam to specific genomic locations without targeted insertion of TetO arrays.

Different from 3C methods where only one ligation event per allele can be detected at most, several GATCs can be contacted by a Dam bound TetO viewpoint in the course of a typical DamC experiment (18 hours). Based on previous measurements from the *Chic1* locus in mESC, it seems to be likely that several contacts with neighboring sequences occur within 18 hours[86]. Under the assumption that *n* GATCs are methylated in this time window, DamC would require *n* times less cells than 4C in order to obtain similar contact profiles what is supported by our proof-of-principal single cell DamC data.

DamC can also be used to study tissue-specific chromosomal interactions by controlling the expression of the Dam fusion protein with a tissue specific promoter. This is particularly interesting in cases where the amount of starting material is limited given the low number of cells needed for DamC experiments.

Overall, DamC enables crosslinking- and ligation free measurements of chromosomal interactions in living cells by combining a sequencing based readout with physical modelling of methylation kinetics. Besides providing an important validation of 3C based findings it allows for interesting future developments.

5. References

1. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
2. Amano, T. *et al.* Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16**, 47–57 (2009).
3. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1–8 (2017).
4. Quevedo, M. *et al.* Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. *Nat. Commun.* **10**, 2669 (2019).
5. Luger, K. Crystal structure of the nucleosome core particle at 2.8 A˚ resolution. **389**, 10 (1997).
6. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412 (2007).
7. Vakoc, C. R., Mandat, S. A., Olenchock, B. A. & Blobel, G. A. Histone H3 Lysine 9 Methylation and HP1γ Are Associated with Transcription Elongation through Mammalian Chromatin. *Mol. Cell* **19**, 381–391 (2005).
8. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* **15**, 93–106 (2014).
9. Harr, J. C., Gonzalez-Sandoval, A. & Gasser, S. M. Histones and histone modifications in perinuclear chromatin anchoring: from yeast to man. *EMBO Rep.* **17**, 139–155 (2016).
10. Vardimon, L., Kressmann, A., Cedars, H., Maechler, M. & Doerfler, W. Expression of a cloned adenovirus gene is inhibited by in vitro methylationt. *Proc Natl Acad Sci USA* 5 (1982).

11. Stein, R., Razin, A. & Cedar, H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl. Acad. Sci.* **79**, 3418–3422 (1982).

12. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).

13. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

14. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).

15. Li, X.-Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol.* **12**, R34 (2011).

16. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

17. Sims, R. J. Elongation by RNA polymerase II: the short and long of it. *Genes Dev.* **18**, 2437–2468 (2004).

18. Vo ngoc, L., Wang, Y.-L., Kassavetis, G. A. & Kadonaga, J. T. The punctilious RNA polymerase II core promoter. *Genes Dev.* **31**, 1289–1301 (2017).

19. Flanagan, P. M., Kelleher, R. J., Sayre, M. H., Tschochner, H. & Kornberg, R. D. A mediator required for activation of RNA polymerase II transcription in vitro. *Nature* **350**, 436–438 (1991).

20. Kim, T.-K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).

21. Harper, T. M. & Taatjes, D. J. The complex structure and function of Mediator. *J. Biol. Chem.* **293**, 13778–13785 (2018).

22. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–143 (2015).

23. Nechaev, S. & Adelman, K. Promoter-proximal Pol II: When stalling speeds things up. *Cell Cycle* **7**, 1539–1544 (2008).

24. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926–aad9926 (2016).

25. Zhang, H., Rigo, F. & Martinson, H. G. Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Mol. Cell* **59**, 437–448 (2015).

26. Banerji, julian, Rusconi, sandro & Schaffner, walter. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

27. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J.* **11**, 4047–4057 (1992).

28. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. A Tissue-specific Transcription Enhancer Element Is Located in the Major lntron of a Rearranged lmmunoglobulin Heavy Chain Gene. 12

29. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

30. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).

31. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

32. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).

33. Vernimmen, D. & Bickmore, W. A. The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. *Trends Genet.* **31**, 696–708 (2015).

34. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).

35. Vakoc, C. R. *et al.* Proximity among Distant Regulatory Elements at the β-Globin Locus Requires GATA-1 and FOG-1. *Mol. Cell* **17**, 453–462 (2005).

36. Deng, W. *et al.* Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell* **149**, 1233–1244 (2012).

37. Bolzer, A. *et al.* Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biol.* **3**, e157 (2005).

38. Finlan, L. E. *et al.* Recruitment to the Nuclear Periphery Can Alter Expression of Genes in Human Cells. *PLoS Genet.* **4**, e1000039 (2008).

39. Pirrotta, V. & Li, H.-B. A view of nuclear Polycomb bodies. *Curr. Opin. Genet. Dev.* **22**, 101–109 (2012).

40. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, eaag0025 (2017).

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

42. Mateo, L. J. *et al.* Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49–54 (2019).

43. Dekker, J. Capturing Chromosome Conformation. *Science* **295**, 1306–1311 (2002).

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* **10**, 1453–1465 (2002).

45. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).

46. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–1382 (2016).

47. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).

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

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

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

51. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

52. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* (2016). doi:10.1038/nature19800

53. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* (2016). doi:10.1016/j.tig.2016.01.003

54. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* (2016). doi:10.1126/science.aad9024

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

56. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0128-0

57. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).

58. 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).

59. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).

60. 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).

61. Vian, L. *et al.* The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* (2018). doi:10.1016/j.cell.2018.03.072

62. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

63. 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.* **112**, E6456–E6465 (2015).

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

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

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

67. Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).

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

69. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).

70. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* (2018). doi:10.1016/j.cell.2018.10.042

71. Chong, S. *et al.* Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, eaar2555 (2018).

72. Giorgetti, L. & Heard, E. Closing the loop: 3C versus DNA FISH. *Genome Biol.* **17**, (2016).

73. Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat. Methods* **14**, 673–678 (2017).

74. Williamson, I. *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* **28**, 2778–2791 (2014).

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

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

77. Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* **176**, 1502-1515.e10 (2019).

78. Giorgetti, L. *et al.* Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell* **157**, 950–963 (2014).

79. Brant, L. *et al.* Exploiting native forces to capture chromosome conformation in mammalian cell nuclei. *Mol. Syst. Biol.* **12**, 891 (2016).

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

81. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).

82. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62**, 668–680 (2016).

83. 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).

84. Belmont, A. S. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.* **26**, 69–78 (2014).

85. 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**, 32–51 (2009).

86. Tiana, G. *et al.* Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains. *Biophys. J.* **110**, 1234–1245 (2016).

6. Acknowledgements

If I think retrospectively about the time as a PhD student, I realize it's the people I met which made this period a very good and valuable time for me. I guess not only because they were critical for my wellbeing in the last four years but also in helping me to develop my personal and professional identity.

First, I want to thank Luca: Doing a PhD in your lab was a fantastic experience which will retain a very good memory. I am grateful for all the inspiring discussions we had about science and the many many things I could learn from you. I had a lot of freedom and still always found your office door open when needed. Thank you for listening to my ideas and your continuous support in the last couple of years. I could grow intellectually and personally and learned to believe in my scientific value.

I want to thank my committee members, Susan Gasser, Peter Meister and Marc Bühler for spending their precious time in coming to my yearly meetings to listen to a young scientist. Thank you for sharing your experience and providing valuable input to my PhD project!

On this occasion I want to highlight the great facilities (in particular genomics and FAIM) for the outstanding technical support and their patience when dealing with PhD students.

I want to thank all lab members of the Giorgetti group: It was a privilege to work with you! A special thanks to Zhan who spend a lot of time and nerves in analyzing data and preparing plots. A big thanks to Mariya for the great support in the lab and pipetting side by side for a long time. Thanks to Gregory for proof reading my thesis.

Thanks to all of you, we also had a lot of fun outside the lab. I remember the Canada trip, dancing at conference parties, mouth-watering barbecues, beers at happy hours and cheerful dinners. Not to forget the nice Apero you organized when Martha was born: A truly thrilling experience (special thanks to the knitting club)!

Also a big thanks to Piera who was always there when I asked for help!

I also highly appreciated the help from Elida when dealing with administrative things!

Finally, I want to express my deepest gratitude to my family. Without the endless support of my parents and brothers I would not be where I am today. Thank you Luise for listening and dealing with me when I am stressed, but especially thank you for your love. You make me a very happy person – every day.